



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Research Commons

<https://researchcommons.waikato.ac.nz/>

Research Commons at the University of Waikato

Copyright Statement:

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

The thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of the thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from the thesis.

**Enhancing Consumer Health Question Answering Systems
through Rhetorical Structure Theory-Guided Large Language
Model**

A thesis
submitted in fulfilment
of the requirements for the degree
of
Doctor of Philosophy in Management Information System
at
The University of Waikato
by
CHEN LIU



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

Abstract

Consumer Health Question Answering (CHQA) systems have traditionally been designed around isolated capabilities—either factual medical correctness or empathetic support—rather than delivering integrated, multi-dimensional assistance. This separation fails to reflect the reality of consumer health consultations, where individuals typically express intertwined medical concerns, personal circumstances, and emotional distress within a single query. Existing CHQA architectures, often optimised for short, single-intent inputs, prioritise answer generation over question understanding and lack explicit mechanisms to jointly interpret informational and emotional support needs. Although Large Language Models (LLMs) have advanced the state of medical QA, most LLM-based systems still struggle to reconcile clinical reliability with context-sensitive empathy in complex consumer health questions.

To address this gap, this study adopts a Design Science Research (DSR) paradigm and introduces Joint Medical–Emotional Question Answering (JMEQA) as a new task that jointly understand and response medical informational needs and emotional support needs in Consumer Health Questions (CHQs). Grounded in Rhetorical Structure Theory (RST) and Appraisal Framework for Clinical Empathy (AFCE), the research proposes the Med-Emo CHQA architecture, a neural-symbolic unified, LLM-based system that augments answer generation with explicit question understanding. The architecture integrates hierarchical modules for question understanding and answer generation, unifying symbolic knowledge (an RST-based discourse structure tree that encodes intents, contexts, and rhetorical relations during the question-understanding stage) with neural network (LLM) via supervised fine-tuning and structured prompting.

To support system training and evaluation, this study constructs CHQA-MedEmo, the first large-scale, multi-layer corpus of Chinese health consultation records. The corpus is annotated for both medical informational needs and emotional support needs, with explicit contextual spans and discourse relations. System performance is assessed through controlled experiments and an expert reader study using a human-centred evaluation framework that jointly measures accuracy, personalisation, and empathy, alongside latency. Experimental and reader-study results show that the proposed architecture outperforms baseline and ablated models in full

medical-need coverage and empathetic alignment, while maintaining lower latency.

Overall, this study advances the design of CHQA systems by demonstrating how symbolic, discourse-aware question understanding can be embedded into LLM workflows to produce responses that are simultaneously medically reliable, personally tailored, and emotionally supportive. The findings contribute design knowledge for Neural-Symbolic Integration (NSI), LLM-based CHQA in online health consultation settings and offer a scalable architectural paradigm for building complex, human-centred AI systems that balance informational rigour with affective care.

Acknowledgement

My deepest gratitude goes to my Chief Supervisor, Associate Professor William Wang, for granting me the invaluable opportunity to pursue my doctoral studies in the beautiful land of New Zealand. His guidance has been both a lighthouse and a compass, illuminating the path of academic exploration while allowing me the freedom to chart my own course. I am profoundly thankful for his respect to my ideas, even when they were still unformed, for his patience in every discussion, and for his constant encouragement to think deeper, question further, and dream bolder. I also wish to extend my heartfelt thanks to my second supervisor, Associate Professor Gohar Khan, whose insight and kindness have been a steady source of support and inspiration throughout this research journey.

My heartfelt appreciation goes to my husband, Xi Zuo, and my daughter, Dulin Zuo, whose love has been my unshakable foundation. Their unwavering support in financial, emotional, and spiritual has sustained me through long nights of writing and moments of uncertainty. This achievement belongs as much to them as it does to me.

I am deeply grateful to my parents, Gaoyu Xiao and Shunchang Liu, who have devoted their lives to nurturing my dreams, offering their strength and love across oceans and years. Their faith in me has been the quiet power behind every step I have taken.

I would also like to express my sincere appreciation to the administrators of the Waikato Management School for their unfailing assistance, and to the Student Learning Support team for their invaluable guidance and resources. My gratitude extends to my fellow doctoral students within the Management School, whose warmth, sincerity, and perseverance have created a community that has inspired and uplifted me in countless ways.

Lastly, I wish to thank myself for holding on when things were difficult, for believing when outcomes were uncertain, and for continuing to move forward with courage and curiosity. Through more than four years of exploration, challenges and growth, I have learned that persistence is its own form of grace.

Research Activities and Publications

The main publications and completed manuscripts associated with this thesis are as follows.

Journal Articles:

1.Liu, C., & Wang, W. Y. C. (2025). *Understanding medical information and emotional support needs in mental health questions with large language models. Industrial Management & Data Systems. (Accepted; SCI; SSCI; JCR-Q1; ABDC-A).*

This article introduces Multi-Needs and Context Recognition (MNCR) and MHQ-MedEmo, the first framework and benchmark for simultaneously understanding medical informational needs and emotional support needs in consumer health questions. It provides a robust empirical and methodological foundation for this thesis by underpinning the construction of the CHQA-MedEmo dataset and Med-Emo CHQA system, as well as the formal definition of the Joint Medical–Emotional Question Answering (JMEQA) task.

2.Liu, C., Wang, W. Y. C., & Khan, G. (2023). *Digital transformation in measuring social determinants of health: A systematic review. Journal of Organizational and End User Computing, 35(3), 1–21. (SSCI; JCR-Q1; ABDC-B).*

This systematic review synthesises how digital technologies are used to measure SDOH, highlighting opportunities and challenges for leveraging digital traces (including OHC data) in health information system research.

Conference Papers:

3.Liu, C., Wang, W. Y. C., & Khan, G. (2023). *Online health consultation: A systematic review for future research directions. Paper presented at the 7th International Conference on Medical and Health Informatics, Kyoto, Japan. (EI; Best Presentation Award).*

This paper reviews the OHC literature, identifies key research gaps, and motivates the need for advanced CHQA systems capable of addressing complex consumer health needs, therefore directly feeding into the problem awareness stage of this thesis.

4.Liu, C., Wang, W. Y. C., & Khan, G. (2022). *Digital public health surveillance using online health consultation data: A case study on HIV monitoring. Paper presented at the*

6th International Conference on Medical and Health Informatics, Kyoto, Japan. (EI; Best Presentation Award).

This paper demonstrates how OHC data can be used for digital public health surveillance, providing methodological experience and domain knowledge that inform the data selection and contextual analysis in the present thesis.

5.Liu, C., Wang, W. Y. C., & Khan, G. (2021). Matching medical demand in the context of online consultation data. Paper presented at the 5th International Conference on Medical and Health Informatics, Kyoto, Japan. (EI).

This work investigates how to use OHC data in medical demand forecasting, offering early quantitative insights into how consumer health informational needs are articulated and responded to in online consultations.

Completed Manuscripts:

6.Liu, C., & Wang, W. Y. C. (2025). Enhancing consumer health question answering systems through Rhetorical Structure Theory-guided large language models. Manuscript completed; intended submission to the Information Systems Research Special Issue on Compassionate AI.

This paper consolidates the core contributions of the thesis by presenting an RST-guided LLM architecture for CHQA. It formalises how discourse structures and dual medical-emotional intents can be encoded as symbolic knowledge and integrated into LLM workflows, thereby positioning the Med-Emo CHQA system as a generalisable design blueprint for compassionate, neural-symbolic CHQA in OHC settings.

7.Liu, C., & Wang, W. Y. C. (2025). Discourse analysis of consumer health questions to differentiate conscious and formalised information needs. Manuscript completed; intended submission to Information Processing & Management.

This paper conducts a fine-grained discourse analysis of consumer health questions and differentiates conscious and formalised information needs across temporal stages of help-seeking. It deepens the conceptualisation of consumer health-related needs and provides a richer theoretical foundation for modelling temporal context in CHQA, extending the situation-awareness and need-modelling components of this thesis.

Table of Content

Abstract	I
Acknowledgement	III
Research Activities and Publications	IV
Table of Content	VI
List of Figures	IX
List of Tables	X
List of Abbreviations	IX
Chapter 1 Introduction	1
1.1 Background and Context.....	1
1.2 Research Motivation.....	3
1.3 Research Objectives and Questions.....	5
1.4 Scope and Boundaries.....	6
1.5 Research Methodology Overview.....	8
1.6 Expected Contributions.....	9
1.7 Thesis Structure.....	11
Chapter 2 Literature Review	14
2.1 Chapter Introduction.....	14
2.2 Consumer Health Question Answering System.....	15
2.2.1 Definition of CHQA System.....	15
2.2.2 Types of CHQA Systems.....	17
2.2.3 Question Understanding in CHQA.....	24
2.3 Consumer Health-Related Needs.....	31
2.3.1 Traditional Clinical Communication.....	32
2.3.2 Consumer Health Information Needs.....	34
2.3.3 Multi-Dimensional CHRN.....	36
2.4 Large Language Models.....	38
2.4.1 Definition of LLMs.....	38
2.4.2 Types of LLMs.....	44
2.4.3 Alignment Strategies of LLMs.....	48
2.5 Current State of LLM-based CHQA Systems.....	58
2.5.1 System Architecture.....	58
2.5.2 Datasets and Benchmarks.....	66
2.5.3 Evaluation Methodology.....	70
2.5.4 Research Gap.....	73
2.6 Theoretical Framework.....	77
2.6.1 Neural-Symbolic Integration.....	77
2.6.2 Rhetorical Structure Theory.....	84
2.6.3 Appraisal Framework for Clinical Empathy.....	87
2.6.4 Research Framework.....	89
2.7 Chapter Summary.....	91
Chapter 3 Research Methodology	92
3.1 Chapter Introduction.....	92
3.2 Research Paradigm.....	93
3.3 Methodological Framework.....	95
3.3.1 Design Science Research Framework.....	95

3.3.2 Design Science Research Process	97
3.4 Cycle 1: Problem Awareness and Solution Suggestion	101
3.4.1 Literature Review	101
3.4.2 Content Analysis	102
3.4.3 Observation	106
3.5 Cycle 2: Design, Development and Evaluation	108
3.5.1 Hierarchical System Design	108
3.5.2 Neural-Symbolic Integration	110
3.5.3 Experiment	113
3.5.4 Reader Study	115
3.6 Ethical Consideration	117
3.7 Chapter Summary	118
Chapter 4 Situation Awareness	119
4.1 Chapter Introduction	119
4.2 Positioning the Research Environment	120
4.3 CHRN in Real-World OHC	122
4.3.1 Data Collection and Preprocessing	124
4.3.2 Annotation Framework	125
4.3.3 Analyzing Results	130
4.4 Workflow of Real-World OHC	135
4.5 Synthesized Implications	138
4.5.1 Implications for Benchmark Construction	138
4.5.2 Implications for System Design	139
4.6 Benchmark Dataset Construction	140
4.6.1 Dataset Composition	140
4.6.2 Question Annotation	141
4.6.3 Answer Refinement	142
4.6.4 Dataset Statistics and Analysis	144
4.7 Chapter Summary	149
Chapter 5 System Design and Development	151
5.1 Chapter Introduction	151
5.2 Task Definition and Formulation	152
5.2.1 Task Definition	152
5.2.2 Task Formulation	154
5.3 Overall System Architecture	157
5.4 Fine-Tune on Question Understanding	159
5.4.1 Data Preparation	160
5.4.2 Models Fine-tuning	161
5.4.3 Models Evaluation	166
5.4.4 Evaluation Results	170
5.5 Inference on Answer Generation	174
5.6 Chapter Summary	177
Chapter 6 System Evaluation	178
6.1 Chapter Introduction	178
6.2 Experimental Setup	179
6.2.1 Models and Configurations	179
6.2.2 Baselines and Ablations	179
6.2.3 Implementation Details	181

6.3 Expert-Based Evaluation	182
6.3.1 Evaluation Criteria and Metrics	183
6.3.2 Reader Study Protocol	189
6.3.3 Inter-Rater Reliability Analysis	192
6.4 Evaluation Results	194
6.4.1 Accuracy and Personalisation Performance	194
6.4.2 Empathy Performance	197
6.4.3 Latency Performance	199
6.4.4 Trade-Off and Synergies Across Metrics	201
6.4.5 Case-Based Comparison	203
6.5 Chapter Summary	208
Chapter 7 Discussion	210
7.1 Chapter Introduction	210
7.2 Research Objects and Research Questions	211
7.3 Discussion of Sub-RQ 1	212
7.3.1 Practical Relevance	213
7.3.2 Theoretical Rigour	214
7.3.3 RST-Based DST for CHQs	215
7.3.4 Evaluation Results	215
7.4 Discussion of Sub-RQ 2	217
7.4.1 Practical Relevance	217
7.4.2 Theoretical Rigour	218
7.4.3 Neural-Symbolic Unified System Architecture	219
7.4.4 Reader Study Results	220
7.5 Compare with Prior Studies	221
7.5.1 Consumer Health Question Answering Task Definition	221
7.5.2 Discourse-Level Question Understanding	222
7.5.3 Neural-Symbolic Integration in the Era of LLM	224
7.5.4 Evaluation of LLM-Based CHQA System	225
7.6 Chapter Summary	227
Chapter 8 Conclusion	228
8.1 Research Implications	228
8.1.1 Theoretical Implications	228
8.1.2 Practical Implications	230
8.2 Limitations and Future Directions	231
8.2.1 Dataset Scope	231
8.2.2 Research Boundaries	232
8.2.3 System Architecture	233
8.2.4 System Evaluation	234
8.3 Overall Conclusion	235
References	237
Appendix	258

List of Figures

Figure 1-1. Overall Structure of the Thesis.....	13
Figure 2-1. Structure of Chapter 2.....	14
Figure 2-2. CHQA Position within QA. Adapted from Welivita & Pu (2023).....	16
Figure 2-3. Typical Architecture of a Symbolic-Based CHQA system.....	19
Figure 2-4. Typical Architecture of a Statistical-based CHQA system.....	20
Figure 2-5. Typical Architecture of a Neural-Based CHQA System.....	23
Figure 2-6. Dual Streams of Traditional Clinical Communication.....	32
Figure 2-7. Transformer Architecture. Left: encoder; centre: decoder; right: internal structure of the multi-head self-attention module.....	39
Figure 2-8. Comparing RNN and Self-Attention Architectures.....	40
Figure 2-9. Compare Classical Task-Specific Machine Learning Models with Foundation Models.....	41
Figure 2-10. End-to-End Architecture of LLMs-Based CHQA System.....	59
Figure 2-11. A Typical RAG Architecture for LLM-based CHQA System.....	63
Figure 2-12. A Typical Agentic Architecture for LLM-based CHQA System.....	65
Figure 2-13. Neural-Symbolic Learning Cycle.....	79
Figure 2-14. Hybrid (left) versus Unified (right) Neural-Symbolic Architectures.....	79
Figure 3- 1. Structure of Chapter 3.....	92
Figure 3-2. The Reasoning Approaches in DSRP.....	99
Figure 3-3. Research Process with Knowledge Flows, Activities and Intended Outputs.....	101
Figure 3-4. Neuro: Symbolic→Neuro Pathway for System Architecture Design in this Study.....	111
Figure 4-1. Structure of Chapter 4.....	119
Figure 4-2. Content Analyzing Process in This Section.....	124
Figure 4-4. Cross-Dimensional Bubble Chart of Emotional Needs (X), Medical Needs (Y), and Query Length (Size & Color).....	135
Figure 4-5. Typical Workflow of Text-Based Online Health Consultation.....	136
Figure 4-6. An Example of the Benchmark Generate and Structure.....	143
Figure 4-8. Thematic Distribution of Subjective Perspectives in E-N-VIEW Entries.....	147
Figure 5-1. Structure of Chapter 5.....	151
Figure 5-2. Illustration of Relationship among JMEQA Task and Other Three Tasks.....	153
Figure 5-3. Six Sub-tasks of JMEQA.....	154
Figure 5-4. Conceptual Architecture of the Proposed Med-Emo CHQA System.....	158
Figure 5-5. Fine-tuning Process on Question Understanding.....	160
Figure 5-6. An Example of JSONL Training Data in ChatML Schema.....	161
Figure 5-7. Illustration of (a) PEFT (b) Limited-Scope Adaptation.....	163
Figure 5-8. Training Loss Curve of GPT-4o Fine-Tuning.....	164
Figure 5-9. Training Accuracy Curve of GPT-4o Fine-Tuning.....	164
Figure 5-10. Training Loss Curve of Qwen2.5-72B Fine-Tuning with QLoRA.....	165
Figure 5-11. Training Loss Curve of Qwen3-32B Fine-Tuning with QLoRA.....	165
Figure 5-12. The Illustration of the Prompt Design with CoT and Few-Shot Prompt Strategies for Joint Question Understanding.....	168
Figure 5-13. Overall F1 Score and Processing Time Per Test Query Across 8 LLMs.....	172
Figure 5-14. The Illustration of the Prompt Design for Joint Answer Generation.....	176
Figure 6-1. Structure of Chapter 6.....	178
Figure 6-2. Overview of the Reader Study Evaluation Framework.....	183
Figure 6-3. Illustration of a Single Evaluation Group Presented to Readers under the Blinded Assessment Protocol.....	191
Figure 6-4. An Example of the Blinded Assessment under the Reader Study Protocol.....	192
Figure 6-5. Performance Comparison on Full Comprehensiveness Rate (FCR) and High Personalisation Rate (HPR) Across Models.....	196
Figure 6-6. Empathy Level Rating Distribution Across Models.....	199
Figure 6-7. Latency vs Empathy, Comprehensiveness, and Personalisation Trade-Off.....	202
Figure 7-1. Structure of Chapter 7.....	210
Figure 8-1. the DSR Knowledge Contribution Framework.....	228

List of Tables

Table 2-1. Types of CHQA Systems Based on Question-Knowledge Representation and Matching Paradigm.....	18
Table 2-2. Comparison between LLMs as Knowledge Bases and Traditional Knowledge Organisation Approaches.....	44
Table 2-3. Comparative Overview of LLM Alignment Strategies.....	48
Table 2-4. Representative LLM-based CHQA Models and Their Evaluation Strategies.....	71
Table 3-1. Comparison of MedDialog-CN Datasets with Other Chinese OHC Datasets. ...	104
Table 4-1. Comparison between Online Health Communities (OHCm) and Online Health Consultation (OHC) Contexts.....	121
Table 4-2. Overview of Collected Consumer Health Questions.....	125
Table 4-3. Overall agreement in 4 dimensions, 11 label units.....	130
Table 4-5. Number of Medical Needs per Query.....	131
Table 4-6. Number of Emotional Needs per Query.....	134
Table 4-7. Overall agreement in 4 dimensions, 11 label units.....	142
Table 4-8. Statistics of CHQA-MedEmo.....	145
Table 4-9. Comparison Between CHQA-MedEmo and Other CHQA Dataset.....	149
Table 5-1. Base LLMs Adopted for Fine-tuning and Their Specifications.....	162
Table 5-2. Other Base LLMs Adopted for Evaluation.....	167
Table 5-3. Performance Results of Models (Qwen2.5-72B, Qwen2.5-72B-finetuning, Qwen2.5-max & Qwen3-235B).....	173
Table 5-4. Performance Results of Models (GPT-4o, GPT-4o-finetuning, DeepSeek-V3 & DeepSeek-R1).....	173
Table 6-1. Configurations of Baseline, Ablation, and Proposed Models.....	181
Table 6-2. Description and Ordinal Scale of Evaluation Criteria and Metrics.....	188
Table 6-3. Intraclass Correlation for the 5 Metrics Among the 3 Readers.....	193
Table 6-4. Comprehensiveness-, Hallucination-, Risk-, and Personalisation-Level Counts for All Models.....	194
Table 6-5. Comprehensiveness, Hallucination, Risk and Personalisation Performance Results of All Models.....	195
Table 6-6. Empathy Level Rating Count and Average Empathy Score.....	198
Table 6-7. Latency and Output Length among Six Models.....	200
Table 6-8. An Example User Query in this Study.....	204
Table 6-9. Baseline Model Outputs to the Example User Query.....	205
Table 6-10. Abl-AG Model Output to the Example User Query.....	205
Table 6-11. Abl-QU Model Output to the Example User Query.....	206
Table 6-12. Abl-Arch Model Output to the Example User Query.....	207
Table 6-13. Abl-SFT Model Output to the Example User Query.....	207
Table 6-14. Med-Emo CHQA (Proposed) Model Output to the Example User Query.....	208

List of Abbreviations

AI	Artificial Intelligence
AFCE	Appraisal Framework for Clinical Empathy
AG	Answer Generation
ANN	Artificial Neural Network
API	Application Programming Interface
APO	Automatic Prompt Optimisation
BERT	Bidirectional Encoder Representations from Transformers
CHQ	Consumer Health Question
CHQA	Consumer Health Question Answering
CHI	Consumer Health Informatics
CHIN	Consumer Health Information Needs
CHRN	Consumer Health-Related Needs
CHV	Consumer Health Vocabulary
CNN	Convolutional Neural Network
CoT	Chain-of-Thought
CPT	Continued Pre-Training
CRF	Conditional Random Field
DAPT	Domain-Adaptive Pre-Training
DNN	Deep Neural Network
DPO	Direct Preference Optimisation
DSR	Design Science Research
DSRP	Design Science Research Process
DST	Rhetorical Structure Theory
E2E	End-to-End
EDU	Elementary Discourse Unit
ENA	Emotional Need Answering
ENCX	Emotional Need Context Extraction
ENR	Emotional Need Recognition
ESC	Emotional Support Conversation
FFN	Feed-Forward Network
FM	Foundation Model
GPT	Generative Pre-trained Transformer
HIS	Health Information System
HSD	Hierarchical System Design
IAA	Inter-Annotator Agreement
ICC	Intraclass Correlation Coefficient
IR	Information Retrieval

IS	Information System
ISDT	Information Systems Design Theory
JMEQA	Joint Medical-Emotional Question Answering
JSON	JavaScript Object Notation
KB	Knowledge Base
KGE-MD	Knowledge-Grounded Empathy-Enhanced Medical Dialogue
LLM	Large Language Model
LM	Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MCDTB	Macro-level Chinese Discourse Treebank
MLM	Masked Language Modelling
MNA	Medical Need Answering
MNCE	Medical Need Context Extraction
MNR	Medical Need Recognition
MoE	Mixture-of-Expert
MQA	Medical Question Answering
NLI	Natural Language Inference
NLG	Natural Language Generation
NER	Named Entity Recognition
NLM	National Library of Medicine
NLP	Natural Language Processing
NSI	Neural-Symbolic Integration
PaLM	Pathways Language Model
P2D	Patient-to-Doctor
P2P	Patient-to-Peer
PDP	Parallel Distributed Processing
PDTB	Penn Discourse TreeBank
PEFT	Parameter-Efficient Fine-Tuning
PPO	Proximal Policy Optimisation
QA	Question Answering
QU	Question Understanding
OHC	Online Health Consultation
OHCm	Online Health Community
RAG	Retrieval-Augmented Generation
RIAS	Roter Interaction Analysis System
RLHF	Reinforcement Learning from Human Feedback
RM	Reward Model
RNN	Recurrent Neural Network

RST	Rhetorical Structure Theory
RST-DT	RST Discourse Treebank
SDK	Software Development Kit
SDM	Shared Decision-Making
SFT	Supervised Fine-Tuning
T5	Text-to-Text Transfer Transformer
TAPT	Task-Adapted Pre-Training
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference
UMLS	Unified Medical Language System
USMLE	United States Medical Licensing Examination
XLNet	Cross-lingual Language Model - RoBERTa

Chapter 1 Introduction

1.1 Background and Context

As digitisation continues to permeate healthcare and everyday life, a diverse array of online channels, including patient portals, search engines, online health communities, social media, and Online Health Consultation (OHC) platforms, has become the primary means through which individuals seek information and support for their health needs (Demner-Fushman et al., 2020; Welivita & Pu, 2023). Within this landscape, OHC platforms have emerged as one of the most transformative forms of digital health service delivery, revolutionising medical consultations by enabling patients to access licensed healthcare professionals remotely (Liu et al., 2023). Through such platforms, patients can seek professional advice and preliminary assessments for various conditions without physically visiting healthcare facilities, thereby improving accessibility and convenience, especially for those living in rural or under-resourced areas.

Globally, the scale of OHC continues to expand. Recent industry analyses estimate that the global digital health market will approach or exceed USD 900 billion by 2030, driven by the widespread adoption of telehealth, mHealth, and virtual care technologies (Grand View Research, 2025; Mordor Intelligence, 2025). At the same time, large-scale surveys report that a majority of consumers have used telehealth or virtual consultation services at least once since the COVID-19 pandemic, and many intend to continue using them post-pandemic (Bestsenny et al., 2021; McKinsey & Company, 2022). The World Health Organisation (WHO) similarly notes that digital consultation platforms now complement or even replace in-person visits for millions of patients worldwide, and emphasises their potential for improving equity and continuity of care, particularly in low-resource settings (WHO, 2022).

In the Chinese context, this transformation has reached an unprecedented scale. According to the China Internet Network Information Center (CNNIC), by the end of 2024 more than 470 million users in China accessed online medical services, and over 60% of Internet users searched for health-related information (CNNIC, 2025). Leading platforms such as Haodf Online, Chunyu Doctor, and Ping An Good Doctor have grown into large-scale digital health ecosystems, collectively facilitating hundreds of millions of consultations annually (Ping An Healthcare and Technology Co., Ltd., 2023). Studies based on Haodf Online data further document the rapid expansion and evolving

usage patterns of online consultations across specialties and regions in China (Tran et al., 2025).

However, the demand for professional medical advice far exceeds the available physician capacity. This imbalance becomes particularly acute during major public-health events. During the COVID-19 pandemic, for example, multiple studies reported sharp spikes in online consultation volumes on Chinese OHC platforms (Liu et al., 2023; Jin et al., 2025; Tran et al., 2025). The rapid growth in user demand for OHC strained physicians' response capacity, highlighting the need for intelligent information systems capable of providing medical accurate yet emotional support.

Within this landscape, Consumer Health Question Answering (CHQA) systems have emerged as a promising class of intelligent information systems designed to automatically interpret and respond to health-related questions posed by lay users (Demner-Fushman et al., 2020; Welivita & Pu, 2023). Unlike traditional information-retrieval systems that merely return documents or keyword matches, CHQA systems aim to deliver concise, contextually relevant, and comprehensible answers tailored to users with varying levels of health literacy (Ben Abacha & Zweigenbaum, 2015). By bridging the gap between biomedical expertise and consumer comprehension, such systems hold the potential to alleviate physicians' workload and enhance the responsiveness and perceived empathy of online health communication.

Recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) have further accelerated the development of CHQA systems. Earlier rule-based or retrieval-driven models are increasingly supplanted by neural or neural-symbolic architectures capable of richer semantic representation, contextual reasoning, and discourse-level understanding (Demner-Fushman et al., 2020; He et al., 2020). In particular, Large Language Models (LLMs) demonstrate impressive proficiency in generating human-like responses that can integrate informational precision with more natural, conversational styles of interaction.

Nevertheless, a persistent limitation of existing CHQA systems, both traditional and LLM-based, is their predominant focus on informational support needs while neglecting users' emotional support needs. Most current models conceptualise consumer questions purely as factual inquiries, focusing on syntactic or semantic interpretation while overlooking affective

dimensions of user expression (Welivita & Pu, 2023; Yadav et al., 2021). Yet extensive clinical communication research has established that patients' questions often contain both informational and emotional components (Suchman et al., 1997; Roter & Hall, 2006). Emotional cues, such as anxiety, uncertainty, and reassurance-seeking, are integral to patient engagement, trust, and adherence (Roter & Hall, 2006; Topol, 2019). Recent analyses of consumer health questions further confirm that CHQs are inherently multi-intent, containing intertwined informational and emotional demands (Alasmari et al., 2023; He et al., 2020). The inability of current systems to recognise and respond to this dual nature results in emotionally detached interactions that may diminish user trust, satisfaction, and adherence (Topol, 2019).

In principle, LLMs offer new possibilities for overcoming these limitations by jointly modelling informational content and affective tone. However, most current implementations remain limited to the informational dimension, lacking a unified architecture that systematically models and balances the informational and emotional intents embedded in complex, discourse-rich CHQs (Welivita & Pu, 2023; Sorin et al., 2024). In particular, there is a lack of symbolic, interpretable representations that can anchor LLMs' understanding of dual intents and guide the generation of responses that are both medically reliable and emotionally supportive. This persistent gap between technological capability and empathetic communication provides the foundation for the present research, which seeks to design a LLM-based CHQA architecture capable of jointly understanding and responding to informational and emotional support needs in Chinese OHC settings.

1.2 Research Motivation

The motivation for this research arises from the intersection of practical necessity, theoretical insufficiency, and technological opportunity in the evolving domain of online health communication. As outlined in Sections 1.1, the Chinese online consultation environment illustrates both the social urgency and the research potential for developing intelligent systems capable of delivering accurate, empathetic, and contextually grounded medical communication.

Practical Motivation. China's digital healthcare ecosystem faces a structural imbalance between massive patient demand and limited physician availability, with over 400 million users seeking online medical services (CNNIC, 2025). During peak public-health events such as the COVID-19 pandemic,

consultation volumes on OHC platforms surged dramatically, leaving many patient questions partially answered or emotionally neglected (Liu et al., 2023; Jin et al., 2025; Tran et al., 2025). This not only undermines patients' sense of reassurance but also reduces trust in digital medical services (Zhu et al., 2025). Empirical studies of platforms such as Haodf Online show that most consumer health questions involve both informational support needs (e.g., diagnosis, treatment options, prognosis) and emotional support needs (e.g., anxiety relief, reassurance, validation of prior care) (Liu et al., 2022; Liu et al., 2023). Current systems, whether search engines, FAQ-style chatbots, or early medical LLMs, tend to address these two dimensions incoherently, resulting in fragmented interactions that weaken patient satisfaction, adherence, and long-term engagement (Topol, 2019; de Lima & Osório, 2021). Hence, there is an urgent practical need for an intelligent CHQA agent capable of providing both informational and emotional support to supplement overburdened physicians and improve the quality and humaneness of online consultations.

Theoretical Motivation. From an academic standpoint, CHQA research remains theoretically fragmented. Existing studies either focus on biomedical information retrieval and question answering, or on emotion-aware dialogue and empathic conversational agents, but seldom integrate the two within a unified theoretical framework (Demner-Fushman et al., 2020; de Lima & Osório, 2021; Welivita & Pu, 2023). This dichotomy mirrors the long-standing divide between cognitive reasoning and affective computing in AI more broadly. In clinical communication literature, however, effective consultation depends on the interplay of both instrumental (informational, problem-solving) and affective (emotional, relational) dimensions of interaction (Suchman et al., 1997; Roter & Hall, 2006). The absence of an integrative model for representing and operationalising this interplay in CHQA creates a significant theoretical void, particularly in linguistically and culturally specific contexts such as Chinese OHC settings.

Recent breakthroughs in LLMs have made it feasible to simulate human-like responses that are both informative and empathetic in form. Yet, these models still operate largely as black boxes, lacking transparency, reasoning consistency, and domain-specific reliability, issues that are especially critical in healthcare applications (Topol, 2019; Rajkomar et al., 2019). Purely data-driven architectures often fail to capture the logical structure and emotional subtleties of patient discourse, such as how different clauses in a long CHQ jointly express medical concerns and underlying worries. Neural-Symbolic

Integration (NSI) offers a promising paradigm to overcome these weaknesses by embedding structured, interpretable knowledge (e.g., rhetorical relations, medical intents, and affective categories) directly into neural learning pipelines (Bader & Hitzler, 2005; d'Avila Garcez & Lamb, 2020). In the context of CHQA, NSI can help make explicit how informational and emotional support needs co-occur and interact across the discourse structure of a CHQ, and can constrain LLM outputs to remain aligned with clinical relevance and culturally appropriate empathy.

Technological Motivation. Technologically, the emergence of powerful general-purpose LLMs and domain-adapted medical LLMs (e.g., BioGPT, ChatDoctor, DoctorGPT, Zhongjing, BianQue) creates an unprecedented opportunity to re-design CHQA architectures around joint informational-emotional understanding rather than purely factual answering (Luo et al., 2022; Li et al., 2023; Chen et al., 2023; Yang et al., 2023). At the same time, advances in fine-tuning, prompt engineering, retrieval-augmented generation, and structured output control make it possible to inject symbolic knowledge into LLM workflows without fully retraining models. This suggests a feasible pathway to implementing and evaluating neural-symbolic CHQA architectures driven by LLM in realistic OHC scenarios.

1.3 Research Objectives and Questions

To bridge the long-standing divide between informational and emotional support in CHQA system, this study aims to design a CHQA system capable of jointly understanding and responding to both informational and emotional needs within complex Chinese online health consultations through the application of LLM. Building on the promise of LLMs in medical question answering and on the emerging paradigm of neural-symbolic integration, the study proposes to embed a symbolic representation of dual-dimensional support needs into the LLM pipeline so as to enhance both response accuracy and empathetic quality under rigorous, multi-dimensional evaluation.

The central Research Question (RQ) guiding this investigation is formulated as follows:

RQ: *How can a large language model-based consumer health question answering architecture be designed to jointly understand and respond to*

informational and emotional support needs embedded in consumer health questions within online health consultation settings?

This overarching question is further decomposed into two subordinate sub-investigations, each addressing a distinct design challenge within the overall research aim:

Sub-RQ1: *How can a symbolic knowledge representation be designed to effectively support the joint understanding of informational and emotional support needs in large language model-based consumer health question answering systems?*

This sub-question addresses the representation design challenge underpinning the architecture. The symbolic knowledge representation provides interpretable anchors that guide the LLM in disambiguating intents and aligning emotional tone with clinical relevance. The effectiveness of the representation will be assessed through its impact on intent comprehension, dual-intent alignment, and knowledge interpretability, compared with purely neural baselines.

Sub-RQ2: *How can symbolic knowledge be integrated with large language models to enhance a consumer health question answering system's capability to jointly understand and respond to informational and emotional support needs?*

This sub-question examines how symbolic knowledge designed under Sub-RQ1 can be operationalised within LLM-based CHQA systems through a neural-symbolic integration approach. These integration strategies will be evaluated on their ability to improve response accuracy, personalisation, and empathy, as well as latency and practical deployability, within a multi-dimensional, human-centred evaluation framework

1.4 Scope and Boundaries

This study situates its research environment within the domain of Online Health Consultation (OHC), defined as asynchronous, text-based clinical interactions in which consumers seek medical advice from licensed physicians or certified health professionals through regulated, closed platforms (Al-Mahdi et al., 2015). Unlike open health forums, OHC platforms operate as formalised clinical micro-environments, characterised by

structured communication protocols, regulated participation, and clear accountability for the accuracy and safety of medical advice.

This environment represents a boundary object between medical and information systems: it preserves the professional norms of healthcare practice while adopting the interactional affordances of digital communication platforms. Such characteristics make OHC an ideal context for investigating how intelligent systems can mediate both informational reliability and emotional empathy, two dimensions that are often separated in traditional AI-driven communication systems.

Within this boundary, the proposed LLM-based CHQA system is conceptualised as an automated OHC agent capable of augmenting, or partially substituting, the interpretive and empathetic functions of human physicians under carefully specified constraints. Modelling the OHC problem space is therefore essential for ensuring the relevance, ethical safety, and evaluability of the proposed system within real-world conditions.

To maintain methodological focus and ensure analytical tractability, this study adopts three key delimitations:

- **Modality constraint.** The analysis is limited to text-based interactions. Audio or multimodal consultations are excluded in order to avoid the confounding influence of paralinguistic cues that would require different modelling paradigms and data resources.
- **Platform and language scope.** The empirical corpus is drawn from Chinese OHC platforms, representing a culturally and linguistically coherent environment with well-defined clinical communication norms. The findings are therefore most directly transferable to similar Chinese-language OHC settings.
- **Interaction granularity.** The study focuses on single-turn question–answer pairs, which constitute a fundamental unit of consumer health communication and enable a clear mapping between informational and emotional intents. Multi-turn dialogues are reserved for future research once the core JMEQA framework has been validated.

Through these boundaries, the research maintains alignment with DSR’s rigour–relevance balance, ensuring that artefact design is grounded in a realistic problem domain while remaining methodologically manageable for evaluation, replication, and future extension.

1.5 Research Methodology Overview

This research adopts a pluralistic methodological stance grounded in the DSR paradigm. DSR provides a rigorous and practical framework for systematically identifying a real-world problem, designing an artefact as a solution, and evaluating its effectiveness, thereby contributing to both theoretical understanding and practical innovation. The study avoids methodological monism, that is, rigid adherence to a single philosophical or methodological tradition, and instead integrates positivist and interpretivist elements in a complementary manner. This integration ensures methodological rigour while allowing for rich contextual interpretation of human communication in digital health settings.

Within this overarching DSR framework, the research process progresses through five interconnected stages: problem awareness, solution suggestion, design and development, evaluation, and reflection and conclusion. Each stage employs specific research methods tailored to its objectives, ensuring continuous interaction between empirical evidence, theoretical insight and artefact design.

The initial stage, problem awareness, establishes a comprehensive understanding of the research domain and its existing limitations. A literature review was conducted to synthesise prior work in CHQA, Consumer Health Related Need (CHRN), and LLM. This was complemented by a content analysis of real-world CHQs to empirically characterise users' informational and emotional support needs. Additionally, observation of authentic OHC records was undertaken to capture the dynamics of patient–clinician communication. Together, these methods enabled the identification of key research gaps and informed the development of subsequent research questions.

Based on the problem diagnosis, the study formulated a conceptual solution integrating discourse-level linguistic analysis with computational reasoning. The same three methods, literature review, content analysis and observation, were used to derive and refine the theoretical underpinnings of the proposed approach. This process led to the specification of a design framework for CHQA grounded in Rhetorical Structure Theory (RST) and clinical empathy models, establishing the conceptual basis for artefact design.

To operationalise the proposed framework, the study employed Hierarchical System Design (HSD) and Neural-Symbolic Integration (NSI) as its principal design methods. HSD was used to decompose the cognitively complex CHQA process into distinct yet interdependent layers—question understanding and answer generation—reflecting the staged structure of clinical reasoning. NSI was then applied to unify neural and symbolic reasoning within these layers, combining the contextual adaptability of LLMs with the interpretability of symbolic knowledge representations. This integration produced an end-to-end CHQA artefact capable of addressing both informational and emotional aspects of user queries.

The designed system was empirically evaluated using experiments and a reader study. The experiments quantitatively assessed the system's performance in accuracy, empathy and interpretability through controlled testing against benchmark datasets. The reader study provided a complementary qualitative evaluation by engaging human assessors to judge the naturalness, emotional appropriateness and trustworthiness of generated responses. These methods together ensured a comprehensive assessment of both functional and human-centred dimensions.

Finally, the research engaged in systematic reflection on the artefact, the design process and the theoretical implications. The reflection process synthesised insights from the evaluation phase, identifying areas for improvement and refining the underlying design principles. The conclusions derived from this stage articulate the study's theoretical, methodological and practical contributions, and propose directions for future research in empathetic, explainable and human-centred AI for healthcare communication.

1.6 Expected Contributions

This study is expected to make five major contributions spanning theoretical and practical dimensions, consistent with the dual build-evaluate logic of DSR.

First, the study introduces Joint Medical–Emotional Question Answering (JMEQA), a novel research task that jointly models medical informational needs and emotional support needs in complex consumer health queries. This task moves beyond the single-dimensional focus of traditional medical QA systems, which prioritise factual accuracy (Singhal et al., 2025), and of emotional support conversation tasks, which emphasise affective empathy

but lack medical reasoning capacity (Liu et al., 2021). By conceptualising these two dimensions as complementary rather than independent, JMEQA expands the theoretical boundary of CHQA research from pure information provision to interrelated informational and emotional support.

Second, the research advances the theoretical understanding of CHQs through the development of a multi-dimensional analytical framework that captures both informational and emotional support needs within long-form, complex queries. By integrating Rhetorical Structure Theory (RST), and the Appraisal Framework for Clinical Empathy (AFCE), this framework bridges cognitive reasoning and affective understanding. It extends IS theory by demonstrating that empathy and reasoning constitute mutually reinforcing processes in human–AI health communication, rather than separable functional components.

Third, the study constructs the CHQA-MedEmo Corpus, which represents the first large-scale, multi-layer annotated dataset of 703 real-world consumer health consultation threads sourced from online health platforms. Each instance is annotated for medical informational needs and emotional support needs with explicit contextual spans and discourse relations. This corpus fills a crucial methodological and resource gap by providing a robust empirical foundation for dual-mode need recognition, for training and fine-tuning models, and for reproducible system benchmarking.

Fourth, building upon this foundation, the study designs the Med–Emo CHQA system architecture. The architecture integrates hierarchical modules for question understanding and answer generation, unifying symbolic knowledge (an RST-based Discourse Structure Tree that encodes intents, contexts, and rhetorical relations during the question-understanding stage) with neural network (LLM) via supervised fine-tuning and structured prompting. This instantiation illustrates how theory-informed design principles can yield systems that are both functionally intelligent and emotionally attuned, thereby contributing a concrete design artefact to the IS discipline.

Finally, the study proposes a human-centred evaluation framework that measures Med–Emo CHQA system performance along three interrelated dimensions: (a) Accuracy, assessing the correctness and reliability of medical information; (b) Personalisation, evaluating how well responses adapt to users’ specific contexts and literacy levels; and (c) Empathy, measuring the contextual appropriateness, specificity, and emotional resonance of the

support provided. This framework operationalises the dual criteria of rigour and relevance fundamental to DSR by jointly considering informational reliability and affective alignment. It provides a replicable approach for assessing AI-driven health systems in ways that reflect the multidimensional nature of real-world human–AI interactions.

1.7 Thesis Structure

This thesis is organised into eight interrelated chapters, following the DSR framework. As illustrated in Figure 1-1, the overall structure embodies two iterative and mutually reinforcing DSR cycles: (a) Cycle 1 establishes problem awareness and solution suggestion through the development of theoretical and contextual understanding; (b) Cycle 2 focuses on system design, development, and evaluation through iterative refinement and empirical validation. Rather than progressing in a linear manner, these cycles interact dynamically. Insights from each stage continuously inform, refine, and reshape subsequent design decisions, ensuring both theoretical grounding and practical validation of the proposed system.

Chapter 1 (Introduction). This chapter introduces the research background, problem statement, motivation, and objectives. It highlights the significance of addressing both informational and emotional dimensions within CHQA systems and outlines the research scope, questions, and methodological orientation. The chapter also positions the study within the broader DSR paradigm, setting the stage for iterative knowledge building.

Chapter 2 (Literature Review). This chapter provides a comprehensive review of existing research related to research questions. By uncovering gaps in current research, this chapter contributes to the problem awareness phase of Cycle 1 and informs the methodological design of subsequent stages. It also identifies theoretical foundations, therefore provides potential solutions for system design.

Chapter 3 (Research Methodology). This chapter presents the overarching research paradigm and methodological strategy. It elaborates the DSR framework, explaining its principles, reasoning modes, and iterative process. It further outlines the multi-method approach employed across both cycles. The methodological blueprint provides the connective logic linking problem diagnosis, artefact construction, and evaluation.

Chapter 4 (Situation Awareness). This chapter operationalises the relevance cycle by analysing consumer health related needs and typical workflow in real-world OHC records. These observations not only contextualise the problem but also generate empirical insights that feed back into the solution suggestion and artefact conceptualisation stages, thereby bridging Cycle 1 and Cycle 2.

Chapter 5 (System Design and Development). Building upon the theoretical insights and contextual findings of earlier chapters, this chapter describes the iterative process of artefact creation. It details system objectives, architectural design, and component integration, illustrating how LLMs are unified with symbolic knowledge to support dual medical-emotional reasoning. Feedback from preliminary testing loops back to refine both the system architecture and its underlying design principles, exemplifying the build-evaluate-refine iteration central to Cycle 2.

Chapter 6 (System Evaluation). This chapter evaluates the artefact using a multi-method empirical design encompassing controlled experiments, and reader studies. The results not only validate the artefact but also generate design knowledge that informs iterative improvements, thus closing the loop between evaluation and redesign within Cycle 2 and feeding forward into the discussion chapter.

Chapter 7 (Discussion). This chapter interprets the evaluation findings in relation to the research objectives, theoretical constructs, and prior work. It synthesises insights from both cycles to articulate the study's design knowledge contributions, theoretical advancements, and methodological reflections.

Chapter 8 (Conclusion). The final chapter summarises the research outcomes, theoretical contributions, practical implications and limitations. It also suggests directions for future research on empathetic and interpretable AI systems in healthcare communication.

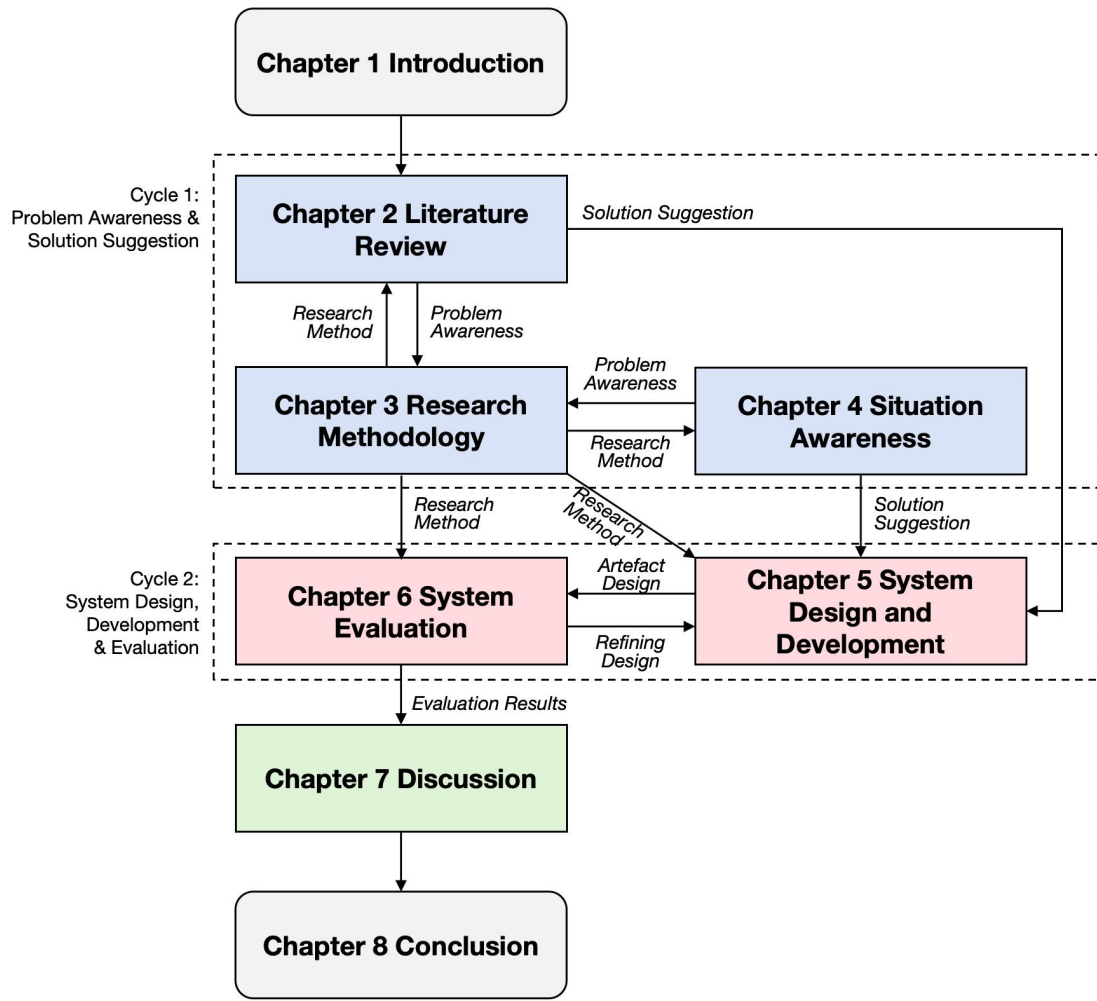


Figure 1-1. Overall Structure of the Thesis.

Chapter 2 Literature Review

2.1 Chapter Introduction

Chapter 2 develops the theoretical groundwork for this study by reviewing the evolution and interconnections among CHQA system, CHRN, and LLM within the broader domain of intelligent health communication. The chapter synthesises prior work in these areas to surface key research gaps and to derive the conceptual and technological foundations for the research framework adopted in this thesis. This framework, in turn, guides the subsequent DSR activities reported in later chapters.

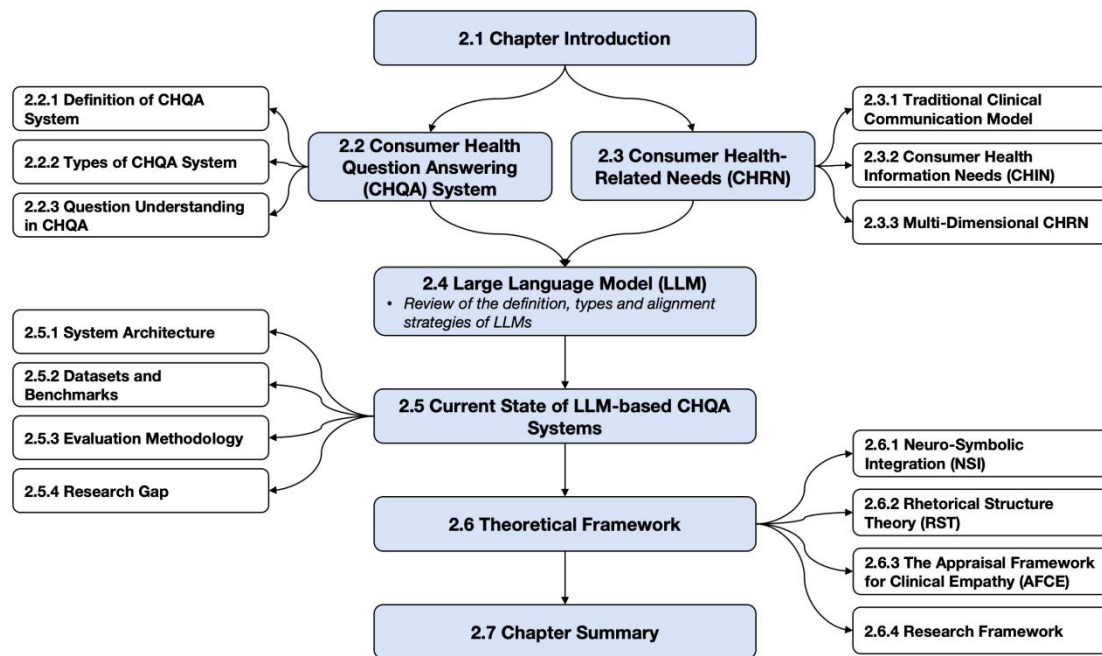


Figure 2-1. Structure of Chapter 2.

The structure of this chapter is summarised in Figure 2-1. Section 2.2 introduces CHQA systems, outlining their definitions, system types, and approaches to question understanding, and tracing the progression from early rule-based medical question answering to contemporary neural network-based systems. Section 2.3 then examines the CHRN that CHQA systems are intended to address. Drawing on clinical communication and consumer informatics literature, it differentiates informational and emotional support needs and contrasts traditional clinician-centred communication models with an emerging multi-dimensional, consumer-centred paradigm.

Building on these foundations, Section 2.4 reviews LLM, including their definitions, model families, and alignment strategies. This section establishes the technological basis for the proposed system and explains why LLMs are particularly well-suited to interpreting heterogeneous clinical and affective cues in health-related discourse. Section 2.5 integrates insights from the preceding sections to assess the current state of LLM-based CHQA research, comparing system architectures, datasets and benchmarks, and evaluation methodologies, and highlighting key research gaps. Section 2.6 then presents the theoretical framework underpinning this study. It draws on the NSI, RST, and the AFCE to construct an integrated lens for modelling how informational and emotional reasoning can be jointly operationalised within CHQA systems.

2.2 Consumer Health Question Answering (CHQA) System

2.2.1 Definition of CHQA System

CHQA systems are medical domain-specific Question Answering (QA) systems designed to help laypersons obtain accurate and comprehensible answers to health-related queries. Unlike general-purpose search engines or document retrieval systems, CHQA systems aim to provide concise, trustworthy, and contextually relevant responses directly, rather than requiring users to sift through entire documents (Deardorff et al., 2017; Welivita & Pu, 2023). They typically retrieve or generate answers from authoritative medical resources, such as MedlinePlus, PubMed and clinical practice guidelines, while ensuring that the language remains accessible to non-expert users (Ben Abacha & Demner-Fushman, 2019a).

As illustrated in Figure 2-2, CHQA systems occupy a specific branch within the broader hierarchy of QA research. QA systems represent a central paradigm in Artificial Intelligence (AI) designed to automatically generate precise and contextually appropriate answers to user-posed natural language queries (Hirschman & Gaizauskas, 2001). Unlike traditional Information Retrieval (IR) systems that merely return ranked document lists, QA systems directly extract or generate concise and semantically grounded answers (Voorhees & Tice, 2000; Hirschman & Gaizauskas, 2001). According to their scope and domain constraints, QA systems are generally divided into open-domain and restricted-domain categories. While open-domain QA systems, such as IBM Watson or recent LLMs, operate across heterogeneous corpora to answer general knowledge questions (Ferrucci et al., 2010; Kwiatkowski et

al., 2019), restricted-domain systems are tailored to specialized fields (e.g., medicine, law, or finance) where factual reliability and interpretability are essential (Ben Abacha & Demner-Fushman, 2019b; Moor et al., 2023).

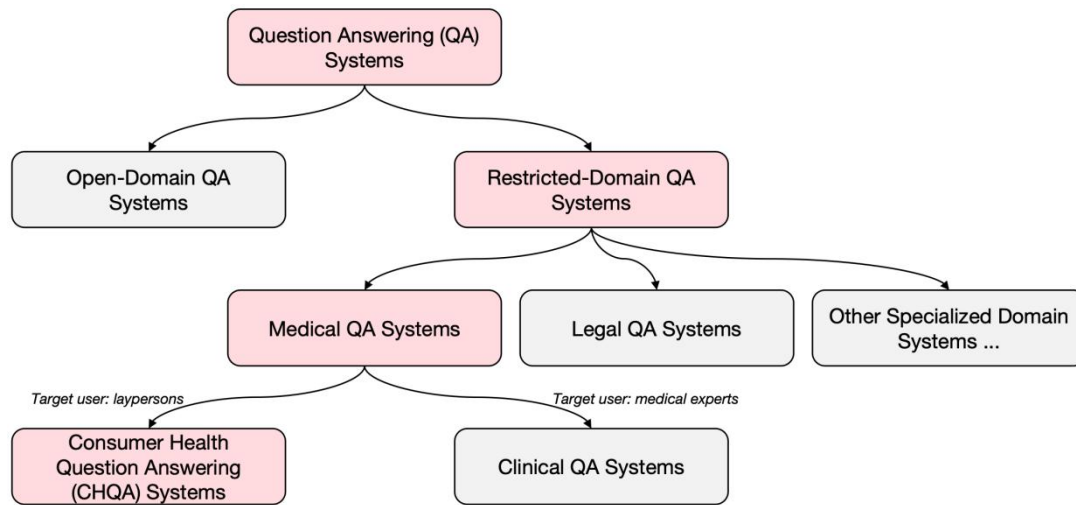


Figure 2-2. CHQA Position within QA. Adapted from Welivita & Pu (2023).

Within this restricted category, medical QA systems have emerged as a major research frontier in response to growing demand for accessible and trustworthy digital health information. Medical QA systems can be further subdivided according to their intended end users. Clinical QA systems target medical professionals and are optimised to interpret structured semi-structured medical data, including clinical trial results and electronic health records, to support of evidence-based decision-making (Jin et al., 2022). CHQA systems, by contrast, are oriented towards laypersons seeking to understand symptoms, diagnoses, treatments or prognoses, and therefore prioritise the generation of concise, reliable and easy-to-comprehend responses drawn from vetted medical sources (Deardorff et al., 2017; Ben Abacha & Demner-Fushman, 2019a; Welivita & Pu, 2023).

Compared with clinical QA, CHQA faces distinct challenges stemming from the heterogeneity and informality of consumer expressions. Consumers frequently formulate multi-intent queries that conflate diagnosis, treatment, prognosis, and preventive aspects in colloquial language. Consequently, CHQA systems must robustly process non-standard linguistic patterns, resolve multi-focal intents, and provide medically sound yet comprehensible explanations (Savery et al., 2020). This dual requirement, accuracy and accessibility positions CHQA at the intersection of medical informatics, natural language processing, and human-centred AI design.

Historically, the United States National Library of Medicine (NLM) consolidated early consumer-oriented practice through the Consumer Health Information and Question Answering (CHiQA) initiative, which formalised protocol-driven pipelines and released resources and methods for producing authoritative yet readable answers for lay users (Demner-Fushman et al., 2019). This line of work culminated in large-scale datasets and evaluation resources such as the MEDIQA Answer Summarisation (MEDIQA-AnS) dataset and related benchmarks for question-driven summarisation of answers to consumer health questions (Savery et al., 2020). Momentum further increased with TREC 2017 LiveQA Medical, which introduced a widely used benchmark for end-to-end answering of real CHQs together with standardised evaluation settings tailored to CHQA systems (Ben Abacha et al., 2017). These efforts established common tasks, datasets, and quality criteria that subsequently shaped system design and reporting practices in the field.

2.2.2 Types of CHQA Systems

CHQA systems typically comprise two tightly coupled functional modules: (a) a question understanding module that interprets the user's input to extract salient medical concepts, intents and contextual cues, and (b) an answer generation module that selects or produces a concise, accurate and context-appropriate response (Deardorff et al., 2017; Demner-Fushman et al., 2020). Underpinning these modules are mechanisms for representing questions and data sources, and for matching or reasoning over these representations. These mechanisms largely determine the overall architecture and performance profile of a CHQA system (Bouziane et al., 2015; Welivita & Pu, 2023).

Following prior work on QA architectures, this study categorises CHQA systems into three broad types based on their dominant question-knowledge representation and matching paradigm: symbolic-based, statistical-based and neural-based systems (Bouziane et al., 2015; Welivita & Pu, 2023). Symbolic-based systems rely on hand-crafted templates and symbolic knowledge bases, using pattern matching, rule-based inference or ontology mapping to derive answers. They offer high interpretability but tend to be brittle when confronted with noisy or unstructured input. Statistical-based systems represent questions using features such as keywords, Term Frequency-Inverse Document Frequency (TF-IDF) vectors, medical entities or question types and operate over unstructured information resources using

statistical matching or semantic similarity measures, achieving a moderate balance between interpretability and robustness. Neural-based systems encode questions and knowledge as distributed representations within neural network parameters and depend on implicit retrieval via attention mechanisms or parametric memory activation, trading off lower transparency for greater robustness to linguistic variability and domain shift (Chakraborty et al., 2021; Liang et al., 2025). As summarised in Table 2-1, these three categories differ systematically in their approaches to question representation, knowledge representation and matching strategy, as well as in their typical trade-offs between interpretability and robustness.

Table 2-1. Types of CHQA Systems Based on Question-Knowledge Representation and Matching Paradigm.

	Question Representation	Knowledge Representation	Matching Strategy	Interpretability	Robustness
Symbolic-Based	Hand-crafted Template	Symbolic knowledge base	Pattern matching / Rule inference / Ontology mapping	High	Low
Statistical-Based	Keyword / TF-IDF / Medical entities	Unstructured information resources	Statistical matching / Semantic similarity	Medium	Medium
Neural-Based	Contextual Embedding / Vector Representation	Distributed parameters encoded within model weights	Implicit retrieval via attention / Parametric memory activation	Low	High

(1) Symbolic-Based CHQA Systems

Early medical QA systems were primarily symbolic-based, relying on expert-defined rules, domain ontologies, and structured databases to provide medically accurate yet readable answers. Unlike statistical models that generate text based on probability distributions, symbolic systems operate on the principle of determinism and explicit knowledge representation. Early explorations in CHQA inherited the logical paradigms of classic expert systems, such as MYCIN (Shortliffe, 1975), which utilized hundreds of “IF-THEN” rules for diagnosis, and adapted them for consumer interaction. To bridge the gap between layperson expressions and professional medical terminologies, researchers developed the Consumer Health Vocabulary (CHV) to facilitate precise query interpretation and retrieval. As shown in Figure 2-3, the architecture of a typical symbolic-based CHQA system follows two main stages: question understanding and answer generation via inference.

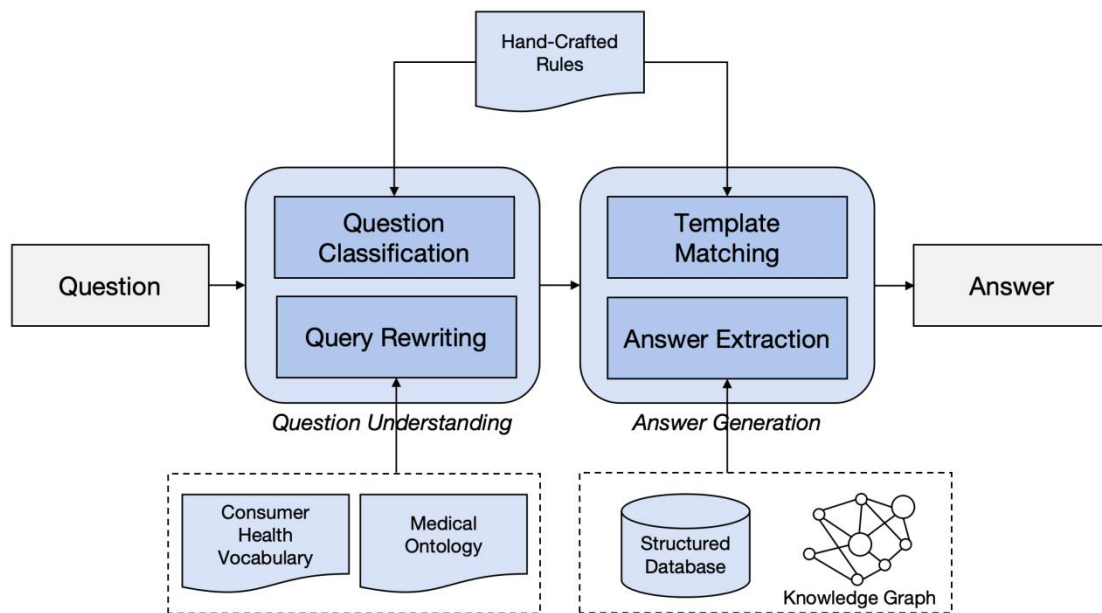


Figure 2-3. Typical Architecture of a Symbolic-Based CHQA system. Synthesised from Yu et al. (2007), Athenikos & Brooks (2009), Ben Abacha et al. (2019a), and Huang et al. (2021).

In the question understanding stage, the system interprets a consumer’s natural-language query through coordinated modules of question classification and query rewriting. The question classification module assigns each question to a predefined intent category based on lexical and syntactic cues. The query generation module then reformulates the question into a machine-understandable representation by leveraging a CHV or a medical ontology such as the Unified Medical Language System (UMLS) or SNOMED-CT (Bodenreider, 2004). This process normalises colloquial consumer expressions and ensures terminological consistency across the system’s rule base and knowledge sources (Zeng et al., 2005).

The answer generation stage integrates rule-based reasoning with template-driven retrieval. The template matching module applies handcrafted inference rules that map classified question types to domain-specific answer templates. The answer extraction module fills these templates with factual content retrieved from the structured database or performs traversing inference over a Knowledge Graph, then delivers a concise, human-readable response. Rules are typically encoded as pattern–action pairs or logical templates that specify how question patterns correspond to query forms and answer fields (Yu et al., 2007; Athenikos & Brooks, 2009; Ben Abacha et al., 2019a; Huang et al., 2021).

This architecture ensures interpretability, factual precision, and medical reliability, since each output can be traced back to explicit rules and verified knowledge sources. However, the rule-based CHQA architecture exhibits notable limitations in scalability and adaptability. Its heavy reliance on manually defined templates restricts domain coverage and leads to poor performance when faced with informal, ambiguous, or multi-intent consumer questions (Pugachev et al., 2023). Despite these drawbacks, rule-based systems laid the conceptual foundation for subsequent retrieval-based and neural-based CHQA architectures by formalising the pipeline of question interpretation, query formulation, and answer synthesis.

(2) Statistical-Based CHQA Systems

Statistical-based CHQA systems emerged as a data-driven transition between symbolic and neural paradigms. Instead of hand-crafted rules, these systems employ probabilistic and feature-based learning to model query–document similarity (Athenikos & Han, 2010; Welivita & Pu, 2023). Figure 2-4 summarises a typical architecture of a statistical-based CHQA system: each module is formulated as a separate learning problem trained on annotated corpora or probabilistic signals, such as term co-occurrence, lexical similarity, or distributional semantics, rather than on explicit logical rules (Voorhees & Tice, 2000; Roberts et al., 2014b).

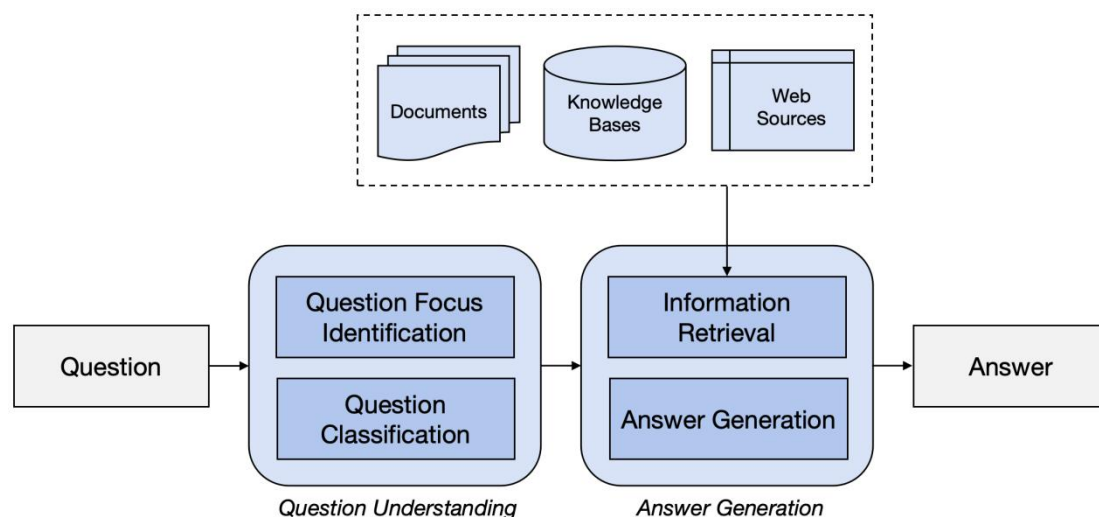


Figure 2-4. Typical Architecture of a Statistical-based CHQA system.
 Synthesised from Roberts et al. (2014b), Kilicoglu et al. (2018) and Savery et al. (2020).

Statistical-based CHQA systems represent user questions primarily through keyword or term-frequency vectors, such as term frequency–inverse document frequency (TF-IDF) or BM25. Instead of predefined semantic

templates, the question text is tokenised and normalised (e.g., lowercased, stemmed), and then mapped to a weighted vector in which each term's weight reflects its relative importance in the underlying corpus (Zhai & Lafferty, 2004). This bag-of-words representation obviates the need for explicit ontology alignment while offering broad coverage of naturally expressed queries—an essential property for consumer-generated health questions, which frequently exhibit informal and colloquial phrasing (Kilicoglu et al., 2018; Welivita & Pu, 2023).

The question-understanding stage aims to translate verbose, ambiguous, or colloquial consumer questions into structured representations suitable for retrieval (Roberts et al., 2014c). Typical subtasks include focus identification and question-type classification, which jointly specify what the consumer is asking and how it should be interpreted. Supervised classifiers such as Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) have been trained on corpora including the MeQSum question-summarisation dataset (Ben Abacha & Demner-Fushman, 2019b) and the semantically annotated CHQs dataset (Kilicoglu et al., 2018). In parallel, distributional semantic models help normalise lay terms using resources such as the CHV, thereby mapping non-expert expressions to medically standardised terminology and reducing lexical gaps between lay queries and clinical evidence sources (Zeng et al., 2006).

In the information retrieval module, candidate documents, QA pairs, or knowledge snippets are ranked as probabilistic evidence based on their estimated relevance to the structured query (Voorhees & Tice, 2000). Classical retrieval models such as BM25, TF-IDF, and language modelling approaches (Zhai & Lafferty, 2004) have been widely adopted, and have subsequently been complemented by embedding-based retrieval that exploits distributed word representations, for example word2vec and BioWordVec embeddings (Zhang et al., 2019).

The answer generation module extracts and synthesises content from the top-ranked items into a coherent and readable response. Early CHQA systems used template-based or statistical Natural Language Generation (NLG) methods to organise retrieved snippets into fluent sentences (Athenikos & Han, 2010). More recent approaches incorporate question-driven summarisation and readability-enhancement techniques that simplify medical expressions for non-expert audiences (Ben Abacha & Demner-Fushman, 2019b; Savery et al., 2020). For instance, lexical substitution,

syntactic simplification, or abstractive summarisation can transform complex clinical discourse into layperson-friendly language while preserving medical meaning.

Overall, statistical-based CHQA systems strike a middle ground between the interpretability of symbolic approaches and the adaptability of neural models. Their modular design enables targeted optimisation of individual components and facilitates transparent error analysis, which is crucial in safety-critical medical contexts (Welivita & Pu, 2023). However, because each component is trained and tuned separately, errors can propagate through the pipeline, motivating later neural and hybrid architectures that learn shared representations for question understanding, retrieval, and answer generation (Jin et al., 2022).

(3) Neural-Based CHQA System

The emergence of deep neural architectures in the mid-2010s reshaped the design paradigm of CHQA, shifting the field from rule-centred pipelines towards end-to-end semantic reasoning. Unlike symbolic systems that rely on rigid templates and explicit feature engineering, neural-based systems learn distributed semantic representations from large-scale data, enabling them to bridge the so-called vocabulary gap between layperson expressions and professional medical terminology by inferring latent semantic relationships rather than depending on exact keyword overlap.

The transition to neural CHQA was catalysed by two influential lines of research. In professional biomedical QA, Wiese and colleagues (2017) showed that a neural QA model trained on the large open-domain SQuAD dataset could be successfully adapted to the much smaller BioASQ benchmark through transfer learning and biomedical word embeddings, achieving state-of-the-art performance without relying on domain-specific parsers or ontologies. In the consumer health domain, Zhang et al. (2017) proposed an end-to-end character-level multi-scale convolutional neural network to perform question-answer matching on the cMedQA dataset, effectively handling informal user queries by jointly modelling character- and word-level features. Together, these studies demonstrated that deep learning could cope with both the precision required for biomedical facts and the noise inherent in consumer-generated language.

Subsequent work focused on enhancing the reasoning capacity of neural models. Memory Network (MemNN) architectures introduced an explicit, addressable memory component that can be read from and written to during inference, enabling multi-hop reasoning over stored representations (Weston et al., 2014; Sukhbaatar et al., 2015). Related attentional Long Short-Term Memory (LSTM) and gated recurrent models similarly learned to focus on salient parts of the input and external context (Bahdanau et al., 2015). Within a typical neural CHQA workflow, the architecture comprises three major components (see Figure 2-5). First, a semantic encoding module maps the input question into a dense vector representation. While early systems relied on static word embeddings (Mikolov et al., 2013), more recent approaches employ contextualised biomedical encoders such as BioBERT to capture fine-grained clinical nuances (Lee et al., 2020). Second, a neural retrieval and attention module acts as a differentiable evidence-matching mechanism, assigning attention weights over memory slots or retrieved passages and thereby approximating the retriever–reader paradigm that underpins many contemporary CHQA systems (Bahdanau et al., 2015; Lewis et al., 2020; Zakka et al., 2023). Third, an answer generation module aggregates the attended representations to produce the final output, ranging from answer selection in factoid settings to abstractive generation in systems built on top of LLMs such as Med-PaLM (Singhal et al., 2023).

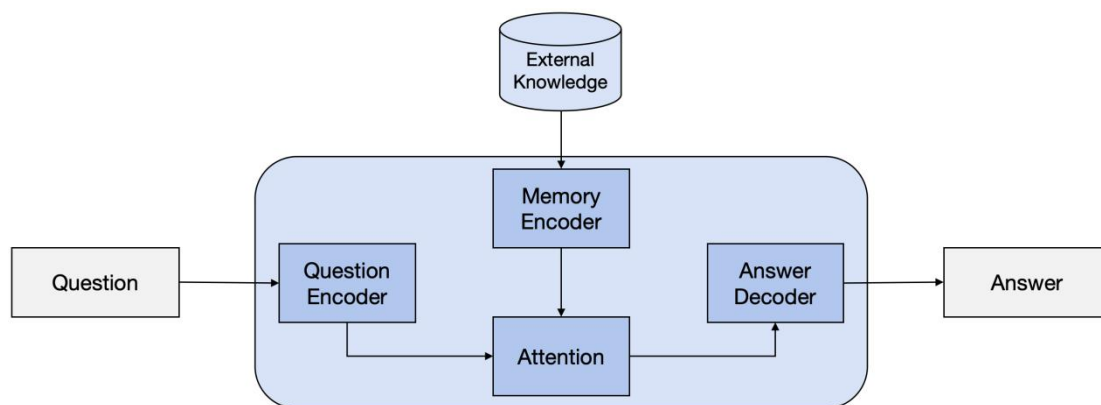


Figure 2-5. Typical Architecture of a Neural-Based CHQA System. Synthesised from Wiese et al. (2017) and Zhang et al. (2017).

A key advantage of neural architectures over their symbolic predecessors is their capacity to model informational content and emotional cues within a unified framework. This capability is particularly important for CHQA, where empathetic communication is integral to perceived usefulness and safety. Building on advances in open-domain dialogue generation, for example, Affect-LM, which conditions an LSTM language model on affect categories to

control emotional content (Ghosh et al., 2017), and the Emotional Chatting Machine (ECM), which integrates internal and external memory to generate emotionally consistent responses (Zhou et al., 2018), health-focused systems have begun to incorporate explicit affective objectives. Emotion-Aware mHealth Agent (EMMA), for instance, combines mood detection with an emotionally intelligent wellness chatbot that delivers brief, mood-congruent micro-interventions (Ghandeharioun et al., 2019). Likewise, the Topic-aware Emotional Response Generation (TERG) model adopts a joint framework to control both topic relevance and emotional tone in dialogue responses, demonstrating strong performance in producing emotionally appropriate and contextually relevant utterances (Huo et al., 2020). These developments illustrate how neural CHQA has evolved from pure fact retrieval towards more holistic, patient-centred conversational agents that can, at least partially, align medical accuracy with affective support.

Overall, neural-based CHQA systems represent a major paradigm shift, offering greater adaptability to heterogeneous user expressions and enabling the integration of empathetic behaviours into answer generation. At the same time, their largely opaque, probabilistic nature introduces new risks, including limited interpretability and the potential for hallucinations—plausible-sounding but medically incorrect statements (Lewis et al., 2020; Zakka et al., 2023). These limitations have directly motivated recent research on hybrid neural-symbolic architectures and Retrieval-Augmented Generation (RAG), which aim to combine the expressive power of neural models with the factual reliability and traceability of structured medical knowledge, as discussed in the following section.

2.2.3 Question Understanding in CHQA

A central challenge in CHQA lies in the linguistic and cognitive complexity of consumer-generated queries. Unlike clinicians' concise, information-seeking questions, CHQs are often long, narrative, and ambiguous, blending medical descriptions with personal emotions and social contexts. As Roberts et al. (2014a) and Alasmari et al. (2023) note, such questions are typically: (a) expressed in colloquial and non-standard language, (b) fairly long and often contain multiple sub-questions, and (c) shaped by informational, emotional, and social support needs.

Because of these characteristics, question understanding has become a foundational stage in CHQA pipelines, serving as the bridge between raw

user input and downstream retrieval or reasoning modules. Early studies thus focused on how to represent, normalise, and decompose consumer queries so that computational systems could interpret them in a medically meaningful way.

The methodological evolution of question understanding in CHQA reveals a clear trajectory across three levels:

- Syntactic level: lexical normalisation and query rewriting for improved retrieval precision;
- Semantic level: entailment recognition and abstractive summarisation for intent alignment and answer reuse;
- Discourse level: pragmatic and contextual modelling for multi-turn coherence.

In response to these challenges, prior studies have proposed several representative strategies:

- Terminology mapping: converting ambiguous or colloquial expressions into standardised medical concepts, enabling systems to retrieve accurate information from medical texts;
- Question focus and type identification: classifying questions by focus and type, allowing efficient knowledge graph retrieval for accurate answers;
- Question decomposition: distinguishing between main questions and contextual or subordinate information, enabling systems to generate more targeted and comprehensive responses.

However, a common limitation across these approaches is the insufficient understanding of multi-dimensional needs in consumer health questions. Most research continues to focus on a single need type, such as informational or emotional support. To date, the study by Alasmari et al. (2023) remains the only one explicitly examining multiple needs within CHQs. Yet, its scope is limited to annotating emotional and social support dimensions, without concurrently modelling informational support needs, which are equally crucial for holistic understanding in CHQA.

(1) Syntactic-Level Question Understanding

A fundamental challenge in CHQA lies in the linguistic divergence between health consumers and healthcare professionals, who often employ distinct lexical forms to describe the same medical concepts (Vydiswaran et al.,

2014). Early CHQA research therefore concentrated on lexical normalisation and query rewriting, aiming to translate consumer expressions into professional or controlled-vocabulary forms to improve retrieval precision and recall. Foundational resources such as the UMLS and the CHV played crucial roles in bridging this gap by systematically mapping lay expressions to expert terminology. Empirical studies have shown that vocabulary-level normalisation can enhance retrieval performance and support more effective access to high-quality health information (Patrick et al., 2001; Zielstorff, 2003; Zeng et al., 2007; Vydiswaran et al., 2014).

From the mid-2000s onwards, the systematic construction of CHVs—through initiatives such as the Open-Access Collaborative Consumer Health Vocabulary (OAC CHV) and related projects—formalised this approach via iterative processes of term mining, familiarity scoring, and semi-automatic expansion (Zeng & Tse, 2006; Zeng et al., 2007; Keselman et al., 2008). These efforts provided robust lexical resources that underpinned subsequent work on query rewriting and semantic standardisation. Complementary tools such as MetaMap (Aronson, 2001) and MedLEE (Friedman et al., 1994) offered automated concept extraction and mapping to UMLS, further enabling machine-level alignment between consumer and clinical terminologies.

Beyond lexical mapping, early studies also explored query reformulation strategies to reduce the consumer–professional language gap. Plovnick and Zeng (2004) showed that reformulating consumer queries with UMLS-preferred professional terms could, in many cases, improve retrieval coverage and precision, while other work highlighted that such reformulation may sometimes degrade retrieval when professional terminology is obscure or poorly aligned with user intent (Abidi et al., 2006; Patrick et al., 2001). Parallel research demonstrated that many consumer health concepts do not map cleanly to existing professional vocabularies, reflecting mismatches in granularity and users’ mental models (Keselman et al., 2008). Collectively, these findings underscored that purely syntactic rewriting and controlled-vocabulary translation, although beneficial, cannot fully capture the conceptual intentions behind consumer expressions, thereby motivating a gradual shift towards semantic-level understanding.

Evaluation campaigns such as the TREC Medical Records Track (2011–2012) and the CLEF eHealth evaluation labs (2013–2016) further validated the importance of syntactic-level preprocessing by introducing patient-centred

retrieval tasks and systematically assessing the impact of vocabulary normalisation, spelling correction, and synonym unification on retrieval accuracy (Voorhees & Hersh, 2012; Goeuriot et al., 2015). During this period, CHVs together with UMLS-based expansion techniques became a de facto standard for question preprocessing and reformulation in CHQA pipelines.

From the mid-2010s, data-driven lexical alignment increasingly complemented and, in some settings, replaced manually curated mappings. Gu et al. (2019), for example, leveraged word2vec-based distributional similarity to identify and add new consumer terms to existing CHVs, while more recent work has employed transformer-based contextual embeddings (e.g., BERT) to perform implicit query rewriting and question summarisation without explicit dictionaries (Yadav et al., 2022). These studies mark a transitional stage in which syntactic-level normalisation evolved towards more semantic representations of consumer queries, laying the foundation for subsequent transformer-based and entailment-aware CHQA models.

In summary, the syntactic-level phase of CHQA evolution primarily focused on word-level alignment and vocabulary harmonisation, improving retrieval precision through controlled-vocabulary translation, spelling correction, and query reformulation. Yet these methods remained largely surface-oriented—addressing lexical discrepancies without fully engaging with the deeper semantic and pragmatic dimensions of consumer questions. The emergence of embedding-based representations thus set the stage for a new research frontier centred on semantic-level understanding.

(2) Semantic-Level Question Understanding

Semantic-level approaches extend beyond lexical normalisation to model the meaning, intent, and latent reasoning embedded in CHQs. CHQs frequently contain implicit inferences, narrative background, and multiple intertwined intents that cannot be adequately captured by surface-level term matching alone (Roberts et al., 2014a). As a result, the focus of CHQA has progressively shifted from vocabulary alignment towards semantic modelling that captures entailment, paraphrase relations, and discourse structure within the query.

A pivotal development in this shift was the integration of textual entailment and Natural Language Inference (NLI) into medical and consumer QA pipelines. The MEDIQA 2019 shared task explicitly introduced question

entailment as a core subtask, requiring systems to determine whether candidate answers or related questions were entailed by, contradicted, or neutral with respect to a given consumer query (Ben Abacha et al., 2019). This was complemented by earlier work on recognising question entailment for medical QA, which constructed inference-oriented datasets to model semantic equivalence and subsumption between question pairs and demonstrated that entailment-based filtering improves both retrieval relevance and answer grounding (Ben Abacha & Demner-Fushman, 2017). Building on these foundations, subsequent studies adopted neural inference architectures such as Enhanced Sequential Inference Models (ESIM), RoBERTa, and DeBERTa to capture fine-grained semantic relations in medical and consumer question pairs (Yadav et al., 2022; Yadav et al., 2023). Collectively, these efforts reframed CHQA as an inference-driven reasoning task rather than a purely retrieval-based problem.

A pivotal development in this shift was the integration of textual entailment and Natural Language Inference (NLI) into medical and consumer QA pipelines. The MEDIQA 2019 shared task explicitly introduced question entailment as a core subtask, requiring systems to determine whether candidate answers or related questions were entailed by, contradicted, or neutral with respect to a given consumer query (Ben Abacha, Shivade, & Demner-Fushman, 2019). This was complemented by earlier work on recognising question entailment for medical QA, which constructed inference-oriented datasets to model semantic equivalence and subsumption between question pairs and demonstrated that entailment-based filtering improves both retrieval relevance and answer grounding (Ben Abacha & Demner-Fushman, 2017). Building on these foundations, subsequent studies adopted neural inference architectures such as Enhanced Sequential Inference Models (ESIM), RoBERTa, and DeBERTa to capture fine-grained semantic relations in medical and consumer question pairs (Yadav et al., 2022; Yadav et al., 2023). Collectively, these efforts reframed CHQA as an inference-driven reasoning task rather than a purely retrieval-based problem.

In parallel, advances in contextualised language models fundamentally transformed semantic representation for CHQA. Traditional static embeddings (e.g., word2vec, GloVe) were superseded by transformer-based models pre-trained on large biomedical and clinical corpora, such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and PubMedBERT (Gu et al., 2021). These models encode rich syntactic and semantic dependencies and have been shown to substantially improve

performance on question classification, entity linking, and intent recognition in health-related QA tasks (Lee et al., 2020; Gu et al., 2021). More recent work adapts these contextual embeddings to consumer-facing settings by incorporating lay expressions and web-scale CHQs, yielding representations that can better capture latent concerns such as treatment efficacy, prognosis expectations, and emotional reassurance (Yadav et al., 2022; Yadav et al., 2023).

Another major strand of semantic-level research focuses on the summarisation and decomposition of long, narrative-style CHQs. The MeQSum dataset introduced 1,000 pairs of verbose consumer questions and concise, expert-written summaries, demonstrating that abstractive summarisation can distil peripheral details into medically focused core questions and thereby improve downstream retrieval and answer quality (Ben Abacha & Demner-Fushman, 2019). Subsequent work proposed question-aware transformer models that jointly leverage medical entities, question types, and semantic cues to generate high-quality summaries on MeQSum and CHQ-Summ, achieving substantial gains over earlier pointer-generator baselines (Yadav et al., 2022; Yadav et al., 2022b). Complementary approaches further decompose complex CHQs into sub-questions aligned with distinct medical intents (e.g., symptom–cause, treatment–risk, prognosis–prevention), thereby enabling more targeted retrieval and response generation (Roberts et al., 2014c; Zhang et al., 2022). Recent frameworks combine summarisation, intent detection, and entailment recognition into unified question-understanding modules, delivering state-of-the-art performance across multiple CHQ benchmarks (Yadav et al., 2023).

In summary, semantic-level question understanding marks a transition from lexical equivalence to meaning-centred and inference-aware modelling. Through textual entailment, contextualised embeddings, and question summarisation/decomposition, CHQA systems gain a more robust grasp of user intent and underlying medical concerns. Nonetheless, semantic modelling alone remains insufficient to fully capture the affective and situational nuances that shape consumer communication, motivating further developments at the discourse level where informational, emotional, and contextual needs are jointly represented.

(3) Discourse-Level Question Understanding

CHQs are rarely isolated or syntactically simple; they are typically narrative, multi-intent, and context-dependent, often embedding a chief complaint together with background details, constraints, personal experiences, and affective cues (Roberts et al., 2014c). Such complexity makes sentence-level or purely semantic processing insufficient. A discourse-level perspective—concerned with how clauses and sentences cohere to express a complete communicative act—is therefore essential for accurate consumer-query understanding. However, explicit discourse-based studies within CHQA remain scarce. Consequently, this section draws on representative work from the broader QA literature in which discourse analysis has been used to enhance question interpretation and answer generation, and then distils methodological implications for CHQA.

Early attempts to incorporate discourse theory into QA emerged in the TREC question-series tasks. Ahn et al. (2005) integrated Discourse Representation Structures (DRS) into the QED system to track entities and quantifier scope across related factoid and list questions, enabling cross-question reference resolution in multi-question series. Similarly, Sun and Chai (2007) applied Centring Theory to model local coherence among context questions, using forward- and backward-looking centres to resolve pronouns and elliptical forms in sequences of related queries. Their results showed that such question sequences behave like short discourses exhibiting anaphora and topic shifts—phenomena that closely mirror multi-turn consumer health interactions in online forums or chat-based consultations.

Beyond entity tracking and pronominal reference, several studies have leveraged rhetorical or discourse-relation frameworks to improve question interpretation and answer organisation. Bosma (2005) used RST to build discourse trees over documents and showed that selecting evidence spans along nucleus–satellite relations yields more coherent, contextually appropriate answers than sentence-level retrieval alone. Verberne (2006, 2007) focused on “why-questions”, demonstrating that identifying causal and explanatory discourse relations in RST-annotated corpora significantly improves both recall and answer quality for complex explanatory questions. Complementary work based on the Penn Discourse Treebank (PDTB) has shown that recognising discourse connectives such as contrast, concession, and condition, and modelling their argument structures, can enhance evidence ranking and text coherence for answer presentation (Prasad et al.,

2008; Feng & Hirst, 2014). Together, these findings suggest that discourse cues—connectives, argument spans, and nucleus–satellite structure—guide both retrieval and generation beyond surface-level semantics.

More recently, research on conversational and multi-turn QA has revisited discourse modelling in dialogue form. Datasets such as QuAC (Question Answering in Context) and CoQA (Conversational Question Answering) were designed specifically to capture information-seeking dialogues in which each follow-up question depends on previous turns and on shared discourse context (Choi et al., 2018; Reddy et al., 2019). Models developed for these benchmarks maintain discourse continuity across turns through mechanisms such as coreference resolution, question rewriting, and hierarchical context encoders (Clark & Manning, 2016; Lee et al., 2017). Although these approaches are predominantly evaluated on open-domain content, they address the same challenges observed in consumer health forums, where users iteratively refine their questions or introduce new details over multiple posts.

Although the above investigations were conducted mostly in open-domain or explanatory QA rather than CHQA, they collectively illustrate how discourse-level modelling enables (a) identification of main versus supporting intents, (b) resolution of contextual dependencies across question turns, and (c) construction of coherent, contextually appropriate answers. For consumer health QA, these insights underline the importance of representing how background information, constraints, and affective expressions relate to the user’s central medical concern. Such discourse-aware understanding is essential for producing responses that are not only complete and medically safe, but also aligned with the user’s situation and emotional state.

2.3 Consumer Health-Related Needs (CHRN)

Consumer Health-Related Needs (CHRN) are the multidimensional requirements that individuals articulate or imply when navigating health-related uncertainty across clinical and digital environments. In this study, CHRN are conceptualised as encompassing at least two interrelated domains: (a) cognitive and informational clarity about symptoms, diagnoses, risks, and treatments; (b) affective and emotional validation and regulation in the face of distress, fear, or stigma. Rather than treating information and emotion as separate concerns, the CHRN construct emphasises that these needs are typically co-present and dynamically negotiated across episodes

of care, whether in face-to-face encounters or digitally mediated consultations (Engel, 1977; Emanuel & Emanuel, 1992; Street et al., 2009). This section reviews the theoretical evolution of CHRN in three stages: first, by examining its origins in Traditional Clinical Communication; second, by analysing the specific literature on Consumer Health Information Need (CHIN); and finally, by synthesizing these perspectives into a Multi-Dimensional CHRN framework.

2.3.1 Traditional Clinical Communication

Before the widespread digitisation of health services, in-person clinical encounters were the primary—and for many patients, the only—context in which health-related needs could be expressed and addressed. Early clinical communication research therefore provided a crucial foundation for understanding CHRN. This body of work situated the medical visit within a predominantly biomedical and paternalistic paradigm, in which clinicians controlled access to information and decision-making authority, while patients were positioned largely as recipients of expert judgement and guidance (Engel, 1977; Emanuel & Emanuel, 1992).

Within this paradigm, empirical studies converged on the idea that medical encounters perform two interdependent functions (Figure 2-6): an instrumental (task-focused) stream oriented towards diagnosis and problem solving, and an affective (socio-emotional) stream oriented towards relationship building and emotional support (Hall et al., 1987; Roter & Hall, 1989; Ong et al., 1995). This dual-function perspective offered one of the earliest theoretical footholds for conceptualising consumer needs as inherently multidimensional, rather than reducible to information transfer alone.

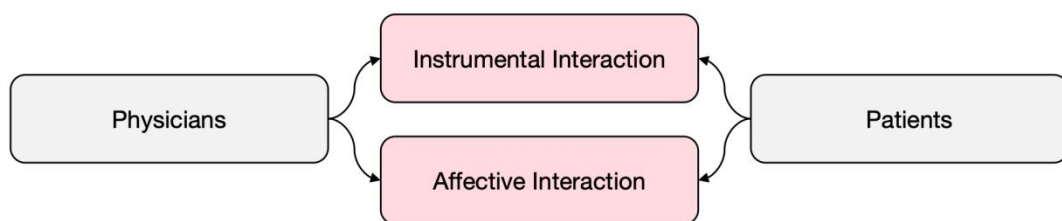


Figure 2-6. Dual Streams of Traditional Clinical Communication. Adapted From Ong et al. (1995).

Instrumental behaviours include history taking, physical examination, test ordering and explanation, treatment recommendations, and patient

education/counselling—activities closely tied to medical expertise and technical problem solving (Hall et al., 1987; Roter & Hall, 1989; Roter & Larson, 2002). Affective behaviours encompass empathy, expressions of concern, reassurance, small talk, and other rapport-building moves that acknowledge the patient’s illness experience and psychosocial context (Roter & Larson, 2002). Over time, patient-centred extensions of this work emphasised responsiveness to patients’ perspectives, contexts, and values, arguing that effective encounters integrate both instrumental and affective streams to support better understanding, adherence, satisfaction, and, in some cases, clinical outcomes (Street et al., 2009; Haskard Zolnieriek & DiMatteo, 2009).

To move beyond narrative description, researchers developed coding systems that render these two streams observable and measurable. The Roter Interaction Analysis System (RIAS) categorises discrete utterances into biomedical, psychosocial, and socio-emotional functions, enabling fine-grained quantification of information-giving, question-asking, partnership building, and affective talk (Roter & Larson, 2002). For the affective stream specifically, the Verona Coding Definitions of Emotional Sequences (VR-CoDES) framework distinguishes patient emotional cues (implicit signals of negative affect) from concerns (explicit expressions of emotion), and characterises clinician responses in terms of exploring versus reducing emotional content (Zimmermann et al., 2011). Complementing this micro-analytic lens, Suchman et al. (1997) modelled empathic communication as a process of recognising, eliciting, and responding to patients’ emotions. Together, these frameworks bridge theory and measurement, providing actionable targets for both clinical training and the design of digital systems that aim to emulate empathic communication.

Subsequent work reframed decision making itself as a communicative activity that should integrate clinical evidence with patient preferences and values. Shared decision-making (SDM) models describe practices such as presenting options, discussing risk–benefit trade-offs, eliciting preferences, and supporting deliberation (Charles et al., 1997; Elwyn et al., 2012, 2017). This line of work effectively stretches the instrumental stream to include risk and uncertainty communication, and stretches the affective stream to include validation of values-laden concerns, thereby tying decisional support directly to both informational quality and emotional needs.

Although classic clinical communication studies did not explicitly theorise “consumer health-related needs” as a formal construct, their dual-function account of medical encounters, along with SDM frameworks, implies that patients’ needs extend beyond information acquisition to encompass empathy, validation, emotional regulation, and value-congruent decision support. This implication motivates the multidimensional framing of CHRN developed in subsequent sections of this chapter.

2.3.2 Consumer Health Information Needs (CHIN)

As digitisation has diffused across health services and everyday life, a wide range of online channels—including patient portals, search engines, Q&A sites, online health communities, social media, and online health consultation platforms—has become central to how patients and their surrogates seek health information (Eysenbach, 2000).

The modern genealogy of consumer-oriented “health needs” is commonly traced to Consumer Health Informatics (CHI), where Eysenbach (2000) reframed the public as active information seekers operating outside traditional clinical hierarchies. Although CHI did not fully theorise “needs” as a formal construct, it decisively shifted emphasis from physician-driven information delivery to consumer-driven information seeking, foregrounding agency, access, and participatory decision-making (Eysenbach, 2000). In this early framing, informational needs—understood as cognitive demands for clarity, reliability, and accessibility of medical knowledge—were treated as the primary dimension of CHRN.

Building on this foundation, CHIN have been further clarified as a gap state: a health consumer’s recognition that their current knowledge is inadequate to achieve a specific health-related goal, within a particular context and at a particular point in time (Ormandy, 2011). This perceived gap initiates and orients information-seeking and evaluation behaviours, linking subjective experiences of uncertainty to observable search and appraisal practices (Ormandy, 2011).

Empirical work on CHIN has largely concentrated on two platform ecologies: Patient-to-Doctor (P2D) and Patient-to-Peer (P2P). In P2D contexts, studies examine patient portals and online consultation services with attention to the structure of user queries, content characteristics, interaction or service quality, and the extent to which these interactions furnish decision support.

Research on patient portals, for example, links portal use to information flow across episodes of care and shows how acute-care contexts trigger, transmit, and respond to information needs (Schnock et al., 2019). Work on formal online consultations and moderated Q&A communities underscores the roles of comprehensibility, provenance, and credibility in uncertainty reduction, and evaluates how content structuring and quality assessment can enhance usability (Huh et al., 2016). Together, these findings indicate that CHIN in P2D settings encompass not only demands for accuracy and completeness but also needs for actionable recommendations and explicit communication of risk and uncertainty (Schnock et al., 2019; Huh et al., 2016).

In P2P contexts, studies draw on data from large Online Health Communities (OHCm) to characterise user demographics, network structure and diffusion dynamics, and the ways in which information and support propagate among peers. Evidence from PatientsLikeMe shows that sharing structured outcomes and illness experiences facilitates experiential learning and self-management, highlighting patient-centred information generation and reuse (Wicks et al., 2010). A longitudinal study of the HealthUnlocked platform associates community participation with subsequent gains in patient activation, suggesting that P2P engagement may meet CHIN by strengthening information literacy and self-efficacy (Costello et al., 2019). Work on WebMD communities uses a natural experiment in moderator withdrawal to demonstrate how professional and peer moderation contribute to maintaining information quality, conversational order, and a supportive climate, while also identifying risks of misinformation and affective drift (Huh et al., 2016). Studies that mine open social platforms such as Reddit extract biomedical entities and topic clusters, map information-seeking trajectories across subcommunities, and quantify the influence of opinion leaders on content reach and persistence (Foufi et al., 2019; Garg et al., 2020). Collectively, this literature suggests that CHIN in P2P environments are shaped not only by the availability and credibility of information, but also by community-level empowerment processes and relational support (Atanasova & Petrič, 2019).

Disease-focused analyses provide additional granularity regarding the structure of CHIN. Large-sample studies of social Q&A services organise cancer-related questions into relatively stable thematic sets spanning diagnosis, treatment, prognosis, and everyday management (Oh et al., 2016). Diabetes-focused research maps consumer language to Unified Medical Language System (UMLS) concepts in blogs and Q&A sites, thereby

quantifying terminology gaps that can impede retrieval and comprehension (Park et al., 2016). More recent investigations examine multimorbidity and disease complexity, showing that the type, specificity, and temporal distribution of information needs vary across co-existing conditions. This evidence motivates a shift from single-disease framing to multi-intent and context-sensitive modelling of CHIN (Alasmari et al., 2021). In summary, this evidence base positions CHIN as the information-clarity axis within the broader CHRN framework, while underscoring that an information-only perspective is necessary but not sufficient for real-world support.

2.3.3 Multi-Dimensional CHRN

Recent scholarship increasingly shows that CHRN extend beyond the pursuit of factual correctness to encompass emotional validation, identity affirmation, and broader psychosocial support within digital health environments. Rather than merely seeking clinical information, many consumers engage with online health communities and peer-support platforms to cope with uncertainty, share lived experiences, and negotiate illness identities, illustrating that experiential knowledge and relational connection are central components of contemporary health engagement (Ziebland & Wyke, 2012; Rowlands et al., 2023; Chen, 2023).

A key milestone in this trajectory is the CHQ-SocioEmo corpus, the first publicly available dataset that explicitly annotates non-informational support needs in consumer health questions (Alasmari et al., 2023). Curated from community-based Q&A forums, CHQ-SocioEmo provides question-level labels for demographics, focus, basic emotions, emotion triggers, and social support types, alongside answer-level annotations of emotional responses. Analyses of this corpus demonstrate that a substantial proportion of consumer questions contain explicit or implicit appeals for emotional reassurance and social support, challenging the assumption that health queries are purely informational and underscoring the need to treat affective and social dimensions as first-order targets in CHQA.

Complementary evidence from digital health behaviour research further confirms that consumers use interactive platforms not only to obtain biomedical explanations but also to access experiential knowledge, empathy, and identity-safe spaces. Qualitative and mixed-methods studies of online support groups and long-term-condition peer platforms show that narratives of lived experience, affective reciprocity, and shared social identity contribute

to coping efficacy, empowerment, and perceived wellbeing (Ziebland & Wyke, 2012; Atanasova & Petrič, 2019; Rowlands et al., 2023; Mills et al., 2024). Parallel work on OHCm documents how participation changes health behaviours and supports self-management, with informational and emotional support often intertwined in discussion threads (Chen, 2023). These findings collectively position emotional and social support as core health-related needs rather than incidental by-products of information delivery.

Concurrently, methodological advancements have refined the conceptualisation and measurement of these multidimensional needs. Petrič et al. (2023) distinguish the quality of informational social support from other forms of support in cancer-related online discussions, showing that message features associated with clarity, accuracy, and responsiveness can be systematically evaluated. Scale-development work on collective empowerment in OHCm similarly differentiates informational quality from empowerment and emotional climate, providing validated instruments for assessing community-level processes (Atanasova & Petrič, 2019). Beyond disease-specific settings, emerging assessment frameworks for online support groups incorporate indicators of empathy, relational warmth, and peer validation, linking these to health and wellbeing outcomes (Mills et al., 2024; Mills et al., 2025; Knowles & Danzi, 2025). Together, these approaches enable CHRN to be operationalised across informational, emotional, and social dimensions in a manner that is amenable to both quantitative evaluation and system design.

In summary, although the empirical base remains uneven, converging evidence from corpus analysis, digital behaviour studies, and support-quality frameworks substantiates a multidimensional view of CHRN in digital health. Informational clarity, emotional reassurance, and social belonging emerge as coexisting and interdependent needs that shape how consumers formulate queries, interpret responses, and evaluate the usefulness of CHQA systems. Any CHQA architecture that focuses solely on informational accuracy while neglecting these affective and social dimensions is therefore likely to fall short of real-world expectations for support.

2.4 Large Language Models (LLMs)

2.4.1 Definition of LLMs

LLMs are a category of deep learning-based foundation models trained on immense volumes of text (and increasingly multimodal) data to understand, generate, and transform natural language across a wide range of tasks (Bommasani et al., 2021; IBM, 2023). Compared with classical machine learning models, LLMs embody three paradigm shifts that redefine how artificial systems represent, generalise, and reason over information: (a) Architecture-level shift: They are founded on the Transformer framework, allowing efficient modelling of long-range dependencies and dynamic contextual weighting, thereby overcoming the limitations of recurrent and convolutional architectures in handling long-form text (Vaswani et al., 2017); (b) Functional-level shift: They exhibit strong cross-task generalisation, performing diverse downstream tasks through shared latent representations instead of being confined to task-specific parameterisation (Wei et al., 2022). This enables a single model to adapt flexibly to summarisation, dialogue, reasoning, and translation tasks via prompting rather than retraining. (c) Epistemic-level shift: LLMs act as approximate implicit knowledge bases, encoding large-scale semantic regularities within their parameters (Petroni et al., 2019; Mialon et al., 2023). Such internalised representations allow inference and reasoning without explicit access to external databases, though the knowledge thus “stored” remains statistical and probabilistic rather than factual or symbolic.

(1) LLMs as Transformer-Based ANNs

The Transformer architecture constitutes the computational backbone of LLMs (Vaswani et al., 2017). It replaces recurrent computation with parallelised self-attention, allowing the model to weigh relationships among all tokens in a sequence simultaneously. The Transformer architecture comprises two main components: the encoder and the decoder (see Figure 2-7). At a high level, the Transformer encoder is a stack of multiple identical layers, where each layer has two sublayers. The first is a multi-head self-attention pooling and the second is a position-wise Feed-Forward Network (FFN). Specifically, in the encoder self-attention, queries (Q), keys (K), and values (V) are all from the outputs of the previous encoder layer.

The Transformer decoder is also a stack of multiple identical layers with residual connections and layer normalizations. As well as the two sublayers described in the encoder, the decoder inserts a third sublayer, known as the encoder–decoder attention, between these two. In the encoder–decoder attention, queries are from the outputs of the decoder’s self-attention sublayer, and the keys and values are from the Transformer encoder outputs. In the decoder self-attention, queries, keys, and values are all from the outputs of the previous decoder layer. However, each position in the decoder is allowed only to attend to all positions in the decoder up to that position. This masked attention preserves the autoregressive property, ensuring that the prediction only depends on those output tokens that have been generated.

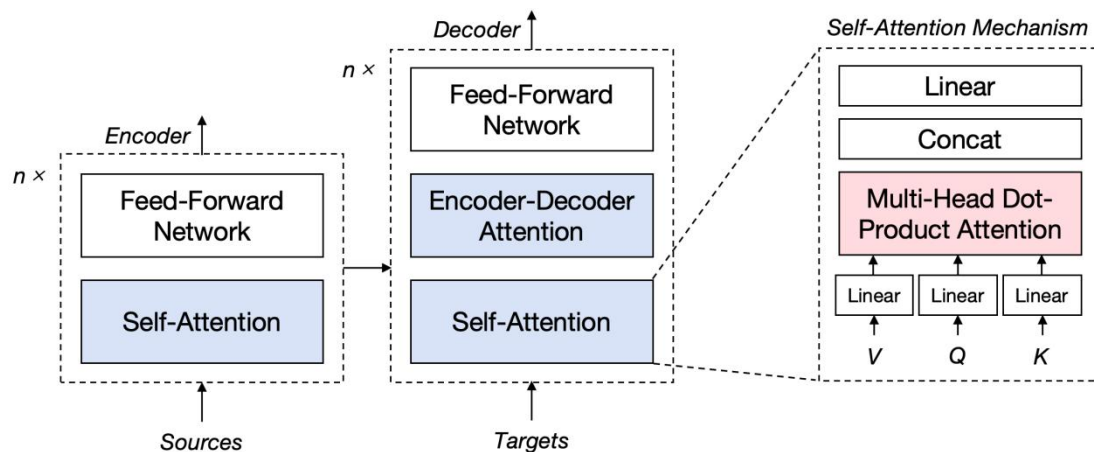


Figure 2-7. Transformer Architecture. Left: encoder; centre: decoder; right: internal structure of the multi-head self-attention module. Adapted from Vaswani et al. (2017).

At the core of the Transformer architecture lies the self-attention mechanism, a computational innovation that enables the model to capture contextual relationships across the entire input sequence simultaneously. Unlike traditional sequential models such as Recurrent Neural Networks (RNNs), which process tokens of a sequence one-by-one and struggle with long-range dependencies, self-attention allows each element in the sequence to directly “attend to” all others. This parallel processing dramatically improves both training efficiency and contextual understanding, forming the foundation of all modern LLMs (Vaswani et al., 2017).

Unlike RNNs, which recurrently process tokens of a sequence one-by-one, self-attention ditches sequential operations in favor of parallel computation. As shown in Figure 2-8, each token is directly connected to any other token via self-attention.

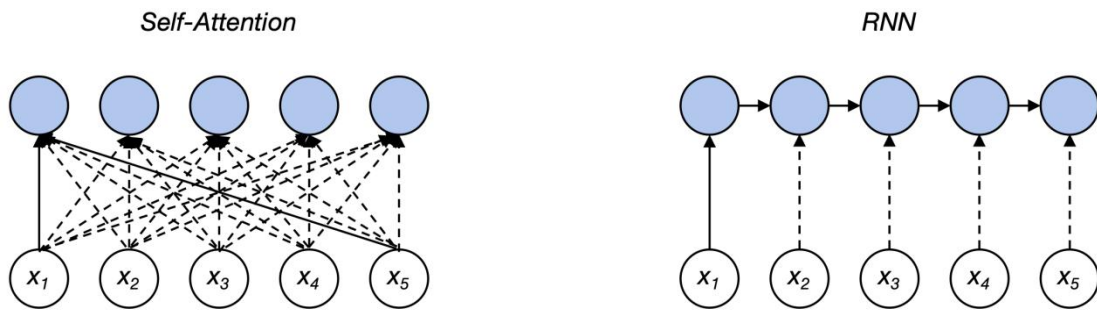


Figure 2-8. Comparing RNN and Self-Attention Architectures. Adapted from Zhang et al. (2023).

Formally, given a sequence of input representations H , the self-attention module computes a new sequence in which each output vector is a weighted combination of all input vectors. The weights reflect the relative importance of each token to the current position and are determined by an attention function that measures pairwise similarity between tokens. The most widely adopted variant is the scaled dot-product attention, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K , and V denote the query, key, and value matrices obtained by applying learned linear projections to the input sequence H . Here, d_k represents the dimensionality of the key vectors. The softmax function normalises the dot-product scores into a probability distribution, ensuring that attention weights sum to one.

By decoupling dependency modelling from sequential order, self-attention provides the Transformer with both global receptive field and parallelisable computation. Consequently, it not only alleviates the vanishing-gradient problem that affects recurrent networks but also scales effectively to extremely large datasets and parameter counts.

(2) LLMs as Foundation Models

LLMs are now widely recognised as the most mature realisation of the foundation model paradigm, which has reshaped how artificial intelligence systems acquire, transfer, and apply knowledge across tasks and domains. Foundation models are large-scale, pre-trained architectures that learn

general-purpose representations from immense and heterogeneous datasets, and from which a broad range of downstream applications can be derived via fine-tuning, prompting, or few-shot adaptation (Bommasani et al., 2021). In this sense, LLMs operationalise the theoretical vision of foundation models by demonstrating how a single unified model can support text generation, reasoning, summarisation, translation, and dialogue without task-specific re-training (Brown et al., 2020; Wei et al., 2022).

As shown in Figure 2-9, LLM-based foundation models are pre-trained on massive, heterogeneous corpora spanning multiple domains, genres, registers, and increasingly modalities (Raffel et al., 2020; Radford et al., 2019), in contrast to task-specific machine-learning models that are trained on narrow, carefully curated datasets aligned to a single objective (e.g., SQuAD for extractive QA, Stanford NLI Corpus for NLI) (Bowman et al., 2015; Rajpurkar et al., 2016). Canonical pre-training sources include large web crawls such as Common Crawl and the Colossal Clean Crawled Corpus (C4), community-assembled corpora such as The Pile, and, for multimodal variants, large-scale image-text collections such as LAION-5B (Dodge et al., 2021; Gao et al., 2020; Schuhmann et al., 2022). This breadth and diversity of data are not incidental. They are central to enabling the model to learn robust lexical and semantic priors that transfer across tasks and domains (Bommasani et al., 2021).

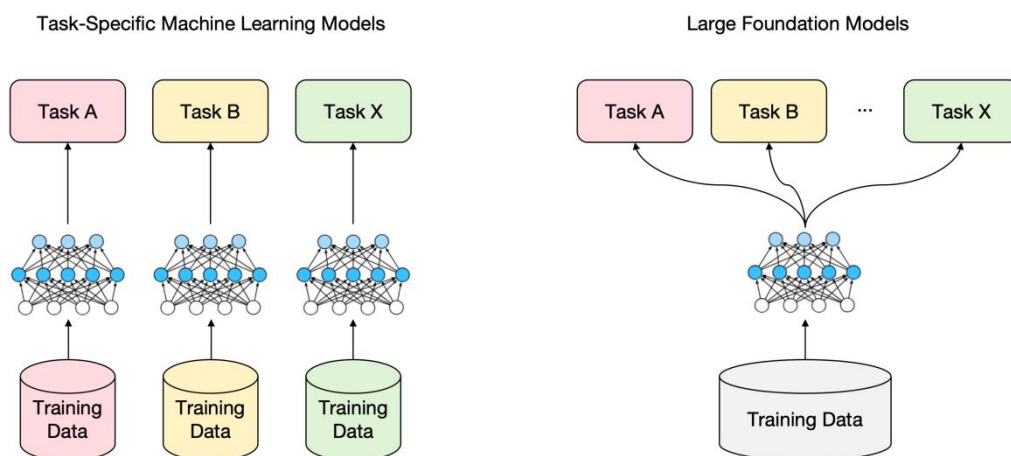


Figure 2-9. Compare Classical Task-Specific Machine Learning Models with Foundation Models. Adapted from Alto (2024).

From a scaling perspective, LLM pre-training departs sharply from conventional practice in statistical machine learning. Traditional models are typically trained on millions of examples matched closely to a specific label

space, whereas contemporary LLMs are trained on hundreds of billions to trillions of tokens, with parameter counts ranging from billions to hundreds of billions (Brown et al., 2020; Touvron et al., 2023). Empirical scaling laws show that capability improves in a predictable manner with data, parameters, and compute, and that, given a fixed compute budget, there exists a near-optimal trade-off between model size and dataset size (Kaplan et al., 2020; Hoffmann et al., 2022). As a result, LLM pre-training emphasises not only volume and heterogeneity, but also deduplication, quality filtering, and contamination control to avoid overfitting and evaluation leakage, supported by emerging dataset-governance practices such as corpus documentation and datasheets (Dodge et al., 2021; Gebru et al., 2021).

Methodologically, the training objective also differs. Task-specific models typically optimise supervised losses tied to a fixed label space, whereas foundation-model pre-training primarily relies on self-supervised next-token or masked-token prediction (Devlin et al., 2019; Radford et al., 2019). This objective yields a task-agnostic representation space in which linguistic and world-knowledge regularities are implicitly encoded within model parameters (Petroni et al., 2019; Mialon et al., 2023). Such representations support in-context learning, few-shot adaptation, and instruction-following without architectural modification - capabilities that are fundamentally unavailable to narrowly trained models and that underpin LLMs' versatility as general-purpose foundations for downstream systems (Brown et al., 2020; Wei et al., 2022).

(3) LLMs as Knowledge Bases

Beyond their versatility as generative systems, LLMs can also be conceptualised as parameterised implicit knowledge bases-systems that internalise factual, conceptual, and procedural knowledge within their parameters rather than via explicit symbolic structures such as triples or ontologies (Petroni et al., 2019; Raffel et al., 2020; Roberts et al., 2020; Bommasani et al., 2021). In this view, an LLM instantiates what Hinton (1986) termed distributed representations: knowledge is not localised in discrete entries, but diffused across a high-dimensional parameter space. Tokens, phrases, and concepts are embedded as vectors in a contextual manifold, where semantic similarity emerges from geometric proximity, and multi-head self-attention dynamically recomposes these representations during inference to retrieve, integrate, and generate knowledge implicitly.

From a knowledge storage perspective, an LLM is primarily trained to learn statistical patterns rather than to maintain explicit symbolic records. During pre-training, massive text corpora are processed under an autoregressive or masked language modelling objective, which predicts the conditional probability of each subsequent (or masked) token (Raffel et al., 2020). Over iterative optimisation, the model’s parameters capture latent regularities in semantics, syntax, and factual co-occurrence, effectively compressing aspects of world knowledge into neural weights (Roberts et al., 2020). Empirical studies confirm that sizeable portions of factual associations, such as entity-attribute or entity-relation pairs, can be recovered directly from model parameters without recourse to external knowledge bases (Petroni et al., 2019; Roberts et al., 2020).

From a knowledge representation perspective, the internal organisation of knowledge within an LLM can be described across at least three interacting representational layers. First, token embeddings provide base-level vector representations of words or subwords, capturing local lexical semantics and grounding the model’s vocabulary. Secondly, contextual representations, produced by stacked Transformer layers, yield dynamic, context-dependent meanings by integrating attention signals across tokens in the input sequence. Thirdly, the model parameters themselves, particularly within feed-forward and attention sublayers, encode high-dimensional statistical associations that approximate linguistic regularities, factual relationships, and reasoning heuristics (Geva et al., 2023; Hernandez et al., 2023; Meng et al., 2022). Recent work has shown, for example, that mid-layer multi-layer perceptron (MLP) modules can be interpreted as key-value memories storing factual associations that are selectively activated during prediction (Geva et al., 2023; Meng et al., 2022).

Taken together, these layers form a distributed representation system in which “knowledge” is not stored as explicit symbolic propositions but emerges from patterns of activation over many parameters. This perspective parallels connectionist theories of human memory, which posit that cognitive functions arise from the coordinated activation of numerous simple processing units rather than from explicit rule storage (Hinton, 1986; Rumelhart et al., 1988). Accordingly, an LLM’s knowledge is best understood as the statistical embedding of relational structure, learned from large-scale corpora during pre-training and encoded within a vast parameter manifold (Bommasani et al., 2021).

From a knowledge retrieval perspective, LLMs access this embedded knowledge via attention-based and activation-mediated mechanisms rather than through symbolic lookup. In base LLMs, that is, models without explicit retrieval augmentation, self-attention layers assign probabilistic weights to inter-token dependencies, dynamically activating those regions of the parameter space most relevant to the input prompt. This process has been interpreted as a form of parametric retrieval, in which factual associations are reconstructed on the fly by propagating information through specific attention heads and MLP sublayers (Geva et al., 2023; Hernandez et al., 2023; Meng et al., 2022). Unlike traditional knowledge bases grounded in explicit ontologies or logical triples, LLMs therefore operate as distributed memory systems in which knowledge is implicitly embedded, contextually reconstructed, and dynamically activated through probabilistic inference. This shift blurs the boundary between data storage and reasoning, positioning LLMs as hybrid cognitive systems that both contain and construct knowledge in response to prompts (see Table 2-2).

Table 2-2. Comparison between LLMs as Knowledge Bases and Traditional Knowledge Organisation Approaches

Feature	Database	Knowledge Graph	LLM
Storage Form	Tabular storage (rows × columns)	Graph structure (nodes + edges + relations)	Parametric storage (neural weights / matrices)
Knowledge Type	Explicit factual data	Explicit semantic relations	Implicit linguistic and world knowledge
Retrieval Mechanism	Deterministic querying (SQL)	Semantic traversal (graph search / SPARQL)	Probabilistic generation (contextual activation)
Interpretability	High (fully transparent schema and logic)	High (explicit relations and ontologies)	Low (black-box parameter space)
Adaptivity	Low (requires manual updates)	Medium (ontology- guided extensions)	High (context-adaptive, generalisable reasoning)

2.4.2 Types of LLMs

As large-scale language modelling has matured, diverse typologies of LLMs have emerged. These typologies reflect differences not only in scale or architecture, but also in cognitive strategy, openness, linguistic coverage, and modality. Clarifying these distinctions is important for informed model

selection, benchmarking, and system design in information-systems contexts (Bommasani et al., 2021).

Decoder-only vs. Encoder-only vs. Encoder-decoder. The fundamental typology of LLMs is defined by their Transformer backbone and pre-training objective: (a) Decoder-only, autoregressive models (e.g., the GPT family) predict the next token given the preceding context and excel at text generation, dialogue, and stepwise reasoning (Brown et al., 2020). (b) Encoder-only, masked-language models (e.g., BERT, mBERT) learn bidirectional contextual embeddings by predicting masked tokens, making them effective for comprehension-oriented tasks such as retrieval, classification, and natural language inference (Devlin et al., 2019). (c) Encoder-decoder, sequence-to-sequence models (e.g., T5) unify comprehension and generation under a text-to-text objective, supporting tasks such as summarisation, translation, and generative question answering within a single framework (Raffel et al., 2020).

These three architectures form the canonical Transformer triad: generation-oriented (decoder-only), understanding-oriented (encoder-only), and integrated sequence-to-sequence (encoder-decoder). Contemporary frontier LLMs predominantly adopt a decoder-only architecture. This design aligns naturally with generative next-token prediction, simplifies training and deployment, and supports a unified interface in which diverse tasks, from question answering to dialogue and code generation, are cast as conditional text generation. This architectural simplicity and generative generality have positioned decoder-only models such as GPT-4, LLaMA 2 and Qwen as the dominant paradigm in modern large language modelling (Brown et al., 2020; Touvron et al., 2023; Bai et al., 2023).

Fast Thinking vs. Slow Thinking. Inspired by Kahneman’s dual-process theory, a second typology differentiates LLMs by inference strategy rather than architecture. “Fast-thinking” LLMs prioritise low latency and high throughput. They generate answers in a single forward pass with minimal explicit deliberation, making them suitable for interactive dialogue systems, real-time agents and high-volume applications. Examples include production-optimised models such as GPT-3.5 and Gemini 1.5 Flash, which are explicitly engineered for fast response times at scale (Brown et al., 2020; Gemini Team, 2024). “Slow-thinking” LLMs incorporate multi-step reasoning, chain-of-thought prompting or self-reflection loops to emulate more deliberative cognition. Models such as GPT-4o and DeepSeek-R1 are

frequently used with explicit reasoning prompts, multi-round planning or tool use, trading latency for improved reliability on complex reasoning tasks (Open AI, 2024; DeepSeek-AI, 2025). This distinction parallels the trade-off between cognitive speed and epistemic reliability in human decision-making: fast-thinking configurations support responsiveness, whereas slow-thinking configurations seek deeper reasoning and error correction. In practice, many systems orchestrate both modes. For example, using a fast model for routine requests and deferring difficult cases to a slow, higher-capability model.

Large Models vs. Small Models. A third axis concerns parameter scale. Large models (tens to hundreds of billions of parameters) typically achieve superior generalisation, emergent reasoning, and cross-domain robustness, but require substantial computational resources for training and inference (Brown et al., 2020; Touvron et al., 2023). Small or compact models (often <10B parameters) are designed for efficiency, edge deployment, and on-device privacy. Recent work on models such as Phi-3 Mini demonstrates that carefully curated training data and alignment can yield performance competitive with much larger models while remaining deployable on resource-constrained devices (Abdin et al., 2024). Empirical scaling laws show that LLM capability typically improves predictably with data, parameters and compute, and that, given a fixed compute budget, there exists an approximate optimum trade-off between dataset size and model size (Kaplan et al., 2020; Hoffmann et al., 2022). In practice, organisations often adopt hybrid strategies, combining a high-capability large model for complex tasks with smaller specialist or distilled variants for routine or latency-sensitive workloads.

Open-Source vs. Closed-Source. LLMs also differ fundamentally in governance and accessibility. Open-source (or open-weight) models such as LLaMA 2, Qwen and many Mistral variants release model weights (and sometimes training recipes), enabling community scrutiny, fine-tuning, and integration into bespoke pipelines (Touvron et al., 2023; Bai et al., 2023). Closed-source models such as GPT-4, GPT-4o and the Gemini family retain proprietary weights and training data but provide managed API access with vendor-supported safety, monitoring and service-level guarantees (OpenAI, 2024; Gemini Team, 2024). In IS design, this openness dimension affects auditability, compliance, data-governance options and the ability to perform domain-specific alignment. Increasingly, hybrid ecosystems combine closed-source base models (for reliability and support) with open-source adapters,

retrieval modules or locally deployed models to satisfy organisational privacy and customisation requirements.

Monolingual vs. Multilingual. Language coverage is another key typology. Monolingual LLMs, typically English-centric, have dominated early development due to the abundance of English training data and benchmarks. Multilingual LLMs (e.g., mBERT, XLM-R, GPT-4o) are pre-trained on corpora spanning dozens or hundreds of languages, enabling cross-lingual transfer and more inclusive access to knowledge (Devlin et al., 2019; Conneau et al., 2020; Open AI, 2024). Multilingual capability is typically achieved through strategies such as shared subword vocabularies, cross-lingual alignment objectives, and inclusion of parallel or comparable corpora. For global information systems, multilingual LLMs are essential to mitigate linguistic inequity by supporting diverse user populations and enabling cross-lingual retrieval, summarisation and dialogue.

Dense vs. Sparse. At the parameter level, LLMs can be categorised into dense and sparse (Mixture-of-Experts, MoE) architectures: Dense models activate all parameters for every token during inference. This design simplifies optimisation and tends to yield stable training dynamics, as in GPT-3, GPT-4 and T5 (Brown et al., 2020; Raffel et al., 2020). Sparse or MoE models (e.g., Switch Transformer, some Gemini and DeepSeek variants) route each token to a small subset of “expert” subnetworks, increasing effective parameter capacity while keeping per-token compute nearly constant (Fedus et al., 2022; Gemini Team, 2024; Bai et al., 2022). MoE architectures introduce new challenges, such as routing stability, expert specialisation and load balancing, but offer a promising path for scaling capacity under compute constraints. For enterprise deployment, they make it feasible to field very large-capacity models while controlling inference cost.

Single-modality vs. Multi-modal. Finally, LLMs differ in modal scope: Single-modality LLMs are trained exclusively on text and are optimised for symbolic reasoning, language understanding and generation. Multi-modal LLMs extend training to images, audio and other modalities, allowing them to align representations across text and perceptual signals. Recent omni-models such as GPT-4o and Gemini models accept and generate combinations of text, images, audio and (in some cases) video, enabling cross-modal reasoning and grounded interaction with the physical and digital world (Open AI, 2024; Gemini Team, 2024). The transition from single- to multi-modal architectures reflects a broader shift from purely linguistic

intelligence toward perceptual-cognitive intelligence, expanding the applicability of LLMs to domains such as clinical image interpretation, document understanding, human-computer interaction and embodied agents.

2.4.3 Alignment Strategies of LLMs

Alignment ensures LLM outputs adhere to human preferences, ethical guidelines, or domain-specific rules (Gabriel, 2020). Effective alignment is essential for the safe and reliable deployment of LLMs in real-world applications. As summarised in Table 2-3, the principal alignment strategies of LLMs can be conceptualised as a progressive continuum of mechanisms operating at different stages of the model lifecycle. Prompt engineering functions at the input level, conditioning model behaviour through carefully designed contextual cues without modifying model parameters. Continued pre-training and fine-tuning operate at the parameter level, either unsupervised or supervised, to embed domain knowledge or task-specific expertise within the model’s internal representations. Finally, self-training and reinforcement learning from human feedback (RLHF) perform output-level alignment, iteratively refining model responses based on automated or human feedback signals.

Table 2-3. Comparative Overview of LLM Alignment Strategies

Strategy	Alignment Stage	Required Resources
Prompt Engineering	Input-based alignment	In-context data or task descriptions
Continued Pre-Training	Unsupervised parameter-based alignment	Unlabeled domain corpus
Fine-Tuning	Supervised parameter-based alignment	Labeled task-specific data
Self-Training	Output-based alignment	Model-generated feedback on outputs
RLHF	Output-based alignment	Human-evaluated feedback on outputs

These complementary approaches collectively advance model controllability and reliability, bridging the gap between general pre-trained knowledge and human-aligned reasoning.

(1) Prompt Engineering

Prompt engineering represents the most direct and lightweight way to align LLMs with user intent without modifying model parameters. It operates entirely at the inference stage, where the model's behaviour is guided through carefully crafted textual or structured inputs that frame the task, define roles, or impose constraints on the model's output (Liu et al., 2023; White et al., 2023). In essence, prompt engineering converts implicit user goals into explicit linguistic instructions that condition the model's probabilistic generation process.

At a theoretical level, prompt engineering embodies a contextual alignment mechanism. Instead of fine-tuning internal weights, the model's latent representations are steered via the conditioning effect of input tokens. When an LLM receives a prompt, the attention mechanism dynamically reweights contextual dependencies according to the textual cues in the input, allowing the model to simulate task-specific reasoning pathways (Brown et al., 2020). This means that alignment occurs through contextual modulation rather than parameter adaptation, enabling the same underlying model to perform multiple, distinct tasks simply by altering the input prompt.

From a functional perspective, prompt engineering can be broadly categorised into two complementary forms: hard prompting and soft prompting. Hard prompts are explicit natural-language strings manually designed by humans. The main forms of hard prompting are:

Instruction-Based Prompting. The most fundamental strategy is instruction prompting, in which the model is given explicit natural language commands describing the desired task (Brown et al., 2020). Clarity, specificity, and task framing critically affect output quality. Advanced variants include contextual prompting, where task-relevant background information or definitions are added, output-constrained prompting, where the output format (e.g., JSON, table) is explicitly specified to improve consistency (Zhou et al., 2022), and role prompting, which assigns the model a persona or professional identity to align responses with a specific tone or expertise (Ouyang et al., 2022).

Example-Driven Prompting. Example-based prompting improves model reliability by illustrating target behaviours through examples. Few-shot prompting (Brown et al., 2020) presents several input–output pairs to demonstrate desired reasoning patterns. One-shot and zero-shot prompting

test the model’s ability to generalize with minimal examples (Sanh et al., 2022). Research further shows that demonstration ordering and semantic similarity between examples significantly influence performance (Lu et al., 2022).

Reasoning-oriented Prompting. To enhance reasoning transparency, Chain-of-Thought (CoT) prompting encourages the model to “think step by step” (Wei et al., 2022), leading to substantial improvements in multi-step reasoning tasks. Self-consistency prompting refines this by sampling multiple reasoning paths and selecting the most consistent answer (Wang et al., 2023). Extensions such as Tree-of-Thought (Yao et al., 2023) and Graph-of-Thought (Besta et al., 2024) expand this idea into structured multi-path reasoning graphs.

Meta-Prompting and Self-Adaptive Prompting. Advanced frameworks employ self-reflection and meta-prompting, allowing the model to critique and refine its own output (Madaan et al., 2023). In ReAct (Reasoning and Acting) prompting (Yao et al., 2022), reasoning steps are interleaved with external actions, such as tool use or API calls, bridging symbolic and neural reasoning. Automated methods like Automatic Prompt Optimization (APO) use reinforcement learning or LLM-as-a-judge frameworks to iteratively improve prompt efficiency (Shin et al., 2020; Zhou et al., 2023).

Hard prompt engineering offers extremely high data efficiency—requiring no labelled data and no training overhead, but it also exhibits several limitations. Because alignment is achieved only at the contextual level, behaviour control remains soft rather than deterministic, and the model may still generate inconsistent or unsafe outputs under distributional shifts (Perez et al., 2022). Moreover, prompts can inadvertently encode biases or ambiguous intent, leading to unstable performance across tasks and domains.

Soft prompting, also known as continuous prompting, represents a parameter-efficient adaptation paradigm in which trainable continuous embeddings are prepended to model inputs while keeping the backbone parameters frozen. The earliest form, Prompt Tuning (Lester et al., 2021), demonstrated that a small number of learned embeddings can elicit task-specific behaviours without labeled data or full fine-tuning. Subsequent methods such as Prefix Tuning (Li & Liang, 2021) and P-Tuning v2 (Liu et al., 2021) extended this idea by inserting continuous prefix vectors into each Transformer layer, enabling deeper control and stronger generalization.

These techniques have inspired a broader family of parameter-efficient learning (PEFT) methods—such as Adapter Tuning (Houlsby et al., 2019) and LoRA (Hu et al., 2022)—that share the same design principle: controlling large models through small, learnable components. More recent frameworks, including SPoT (Vu et al., 2022) and CoOp (Zhou et al., 2022), demonstrate the transferability of soft prompts across tasks and modalities, suggesting that continuous prompting has become a general mechanism for efficient model steering.

(2) Fine-Tuning

Fine-tuning represents the parameter-level alignment strategy applied during the post-training phase of LLM development. Unlike prompt engineering, which modulates model behaviour solely through contextual inputs, fine-tuning directly adjusts a model’s internal parameters using curated task-specific or instruction-formatted datasets, thereby transforming a general-purpose model into a domain-adapted or behaviour-aligned system (Devlin et al., 2019; Ouyang et al., 2022). Through this process, the LLM learns to internalise linguistic, stylistic, and pragmatic cues that reflect domain conventions or user objectives.

At its core, fine-tuning is an optimisation procedure wherein a pre-trained model updates its parameters to better satisfy the requirements of a downstream task. Practically, the model reduces the discrepancy between predicted and reference (ground-truth) outputs through continued gradient-based optimisation from an existing checkpoint. This facilitates domain adaptation, instruction following, and behavioural specialisation without training from scratch—a process often described as transfer learning or parameter reuse (Howard & Ruder, 2018).

Fine-tuning practices have evolved into several canonical forms:

Supervised Fine-Tuning (SFT). This is the standard form of fine-tuning, where the model is trained on explicitly labelled input–output pairs to reinforce desired behaviours or domain-specific responses. In LLMs such as GPT-3.5 and PaLM, SFT serves as the foundation of alignment pipelines, teaching the model to respond helpfully and coherently in structured conversational formats (Ouyang et al., 2022).

Instruction Fine-Tuning. A specialised variant of SFT, instruction fine-tuning uses instruction–response datasets (e.g., “Explain X in simple terms.”) that explicitly encode task intent in natural language. This strategy improves the model’s task generalisation and zero-shot capability, allowing it to follow unseen instructions more faithfully (Wei et al., 2022).

Preference-Based Fine-Tuning (e.g., DPO). Recent extensions such as Direct Preference Optimisation (DPO) (Rafailov et al., 2023) refine the fine-tuning framework by optimising model parameters directly on human preference pairs, without requiring a separate reward model as in RLHF. Conceptually, DPO remains a fine-tuning procedure but functions as an efficient post-training alignment technique that better captures human-evaluated quality.

Parameter-Efficient Fine-Tuning (PEFT). Techniques such as LoRA, Prefix-Tuning, and Adapter-Tuning modify only a small subset of parameters while keeping the original backbone frozen (Hu et al., 2022). This drastically reduces computational cost and mitigates catastrophic forgetting, enabling flexible deployment across multiple domains.

Fine-tuning offers stable and deterministic behavioural control, ensuring consistent alignment across inference sessions. It also facilitates domain grounding. For example, adapting an LLM to biomedical, legal, or customer-service contexts. However, its effectiveness critically depends on the quality, representativeness, and diversity of the training data. Over-fitting to narrow data distributions may limit generalisation or exacerbate dataset-specific biases (Zhao et al., 2023). Moreover, parameter-level updates are resource-intensive, requiring substantial GPU capacity and carefully tuned hyperparameters to maintain training stability.

The general process of fine-tuning a LLM resembles the process of educating a human specialist: it begins with well-defined learning objectives, followed by iterative instruction, practice, and evaluation until optimal performance is achieved. In practical terms, fine-tuning involves several key stages (OpenAI, 2025):

- **Data Collection:** A domain-specific dataset is compiled to serve as the training corpus. This dataset should reflect the linguistic, contextual, and task-specific characteristics required for the target application.

- **Data Preparation:** The dataset is pre-processed and formatted according to the model's input schema, typically as JSONL (JSON Lines) files that pair input prompts with corresponding desired outputs.
- **Model Configuration:** A fine-tuning job is initialized using the selected training method (e.g., supervised fine-tuning, reinforcement learning, or reinforcement fine-tuning). The configuration defines hyperparameters, optimization settings, and target checkpoints.
- **Optional Evaluation Integration:** In reinforcement-based fine-tuning (RFT), a grader or reward model is additionally specified to evaluate and score model responses, providing feedback signals for optimization.
- **Training and Validation:** The model is iteratively trained on the fine-tuning dataset, with periodic validation to assess convergence, generalization, and alignment with intended behaviours.
- **Performance Evaluation:** Finally, the fine-tuned model is evaluated using held-out datasets or human assessments to determine its accuracy, safety, and reliability before deployment.

Through these steps, fine-tuning enables a pre-trained LLM to adapt its general knowledge to specialized domains, tasks, or organisational requirements, while maintaining the efficiency of transfer learning.

In practice, fine-tuning is often integrated within multi-stage alignment pipelines: models are first instruction-tuned (SFT), subsequently refined through preference-based optimisation (DPO or RLHF), and optionally complemented by reinforcement or constitutional alignment (Bai et al., 2022). Together, these methods transform LLMs from generic text generators into cooperative, value-aligned agents.

(3) Continued Pre-Training (CPT)

CPT, also known as domain-adaptive pre-training (DAPT) or task-adaptive pre-training (TAPT), denotes an unsupervised alignment strategy that bridges the gap between large-scale general pre-training and domain-specific language usage (Gururangan et al., 2020). Rather than introducing new supervision signals or modifying the model's behavioural objectives, CPT extends the original pre-training phase on unlabelled corpora that are thematically or stylistically aligned with a target domain. In so doing, the model internalises domain-relevant vocabulary, syntactic patterns and contextual semantics, thereby enhancing its performance on downstream tasks and its factual grounding.

In CPT, a pre-trained model is further trained on an unlabelled corpus that reflects the target domain, such as biomedical, legal or financial text. Rather than introducing a novel supervisory signal, CPT retains the model's original self-supervised learning objective (for example masked-language modelling in encoder-based architectures or next-token prediction in decoder-based ones). This means that the model continues to predict missing or subsequent tokens in the input sequence, but now within the context of domain-specific data.

Because the learning signal remains unsupervised, CPT does not depend on labelled data or human feedback. Instead, the critical design factor lies in corpus selection and curation. By exposing the model to domain-specific text (for example biomedical literature, legal contracts or financial reports), CPT reinforces domain-consistent statistical regularities while preserving the general linguistic capabilities acquired during broad pre-training.

This approach effectively acts as a knowledge-adaptation mechanism, allowing the model's internal representations to shift towards domain-relevant distributions in embedding space. For example, BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019) applied CPT to adapt general-purpose BERT models to biomedical and scientific domains respectively, demonstrating significant improvements in domain benchmarks such as Named Entity Recognition (NER) and QA.

CPT encompasses several implementation variants:

Domain-Adaptive Pre-training (DAPT). The model is exposed to unlabelled domain corpora (e.g., clinical notes, legal documents) to internalise domain-specific semantics. This improves factual precision and contextual relevance for specialised tasks.

Task-Adaptive Pre-training (TAPT). A lighter variant of CPT that uses task-related but unlabelled data, refining the model's contextual understanding of a target task before supervised fine-tuning. TAPT often yields substantial performance gains even with limited labelled examples (Gururangan et al., 2020).

Continual Learning and Lifelong Adaptation. Recent research extends CPT into continual pre-training, by updating LLMs periodically on streaming

or time-sensitive data to maintain factual recency and mitigate knowledge obsolescence (Ke et al., 2023).

CPT offers a cost-effective and data-efficient pathway to domain alignment. It avoids the intensive computational demands of full-scale pre-training while improving knowledge coverage and domain fidelity without labelled supervision. Moreover, CPT mitigates the catastrophic-forgetting often observed in fine-tuning by preserving the model’s general linguistic structure. However, it has notable limitations: (a) It does not directly address behavioural or value alignment, insofar as its learning signal remains purely statistical rather than preference-based. (b) Improper corpus selection can introduce domain bias or concept drift, thereby compromising the model’s generalisation ability. (c) Over-reliance on domain-specific text may reduce linguistic diversity, potentially leading to semantic over-fitting.

(4) Self-Training

Self-training represents a semi-supervised alignment strategy in which an LLM leverages its own high-confidence predictions to generate pseudo-labelled data for further training. Originating from classical semi-supervised learning (Yarowsky, 1995; Scudder, 1965), self-training has been adapted to the LLM paradigm as an autonomous self-improvement mechanism—allowing models to iteratively refine their performance without requiring extensive human annotation (Zhang et al., 2023). In essence, the model becomes both the teacher and the student, generating provisional labels for unlabelled inputs and using them as additional supervision signals in a bootstrapping loop.

The self-training process typically involves three stages: (a) Pseudo-labelling. The pre-trained model is applied to an unlabelled dataset to generate predictions that act as provisional labels. To maintain quality, only predictions made with high confidence, often determined by a predefined threshold, are retained for further use. (b) Augmented training. The pseudo-labelled data are combined with the original labelled dataset to form an expanded training corpus. The model is then retrained on this combined dataset, allowing it to refine its internal representations and improve task performance. (c) Iterative refinement. This procedure can be repeated iteratively, with each generation of the model producing improved pseudo-labels for subsequent training rounds.

Through this self-reinforcing cycle, the model progressively captures unlabelled knowledge embedded in large corpora, thereby improving task accuracy, domain adaptation, and contextual understanding even in the absence of additional human-labelled data.

In the context of LLMs, self-training is being increasingly explored as a low-cost alignment method, especially when human preference data or domain-specific annotations are scarce. Examples include: (a) Self-Instruct (Wang et al., 2023): GPT-based models generate synthetic instruction–response pairs, which are then used to fine-tune themselves, producing notable improvements in instruction-following ability. (b) STaR (Zelikman et al., 2022): The model generates reasoning chains (explanations) for tasks like question answering and then retrains on those self-generated rationales to improve reasoning accuracy. (c) Reflexion and Self-Refine (Shinn et al., 2023): Recent frameworks extend self-training to reflective alignment, where the model evaluates, critiques, and revises its own outputs—forming a self-corrective feedback loop.

In these approaches, self-training enables alignment through self-generated supervision, turning the LLM into an autonomous data generator that refines its internal representations and behaviour iteratively. This paradigm represents a shift from human-centric alignment (e.g., RLHF, DPO) to model-centric alignment, where improvement emerges from internal reflection and confidence estimation rather than external reward signals.

Self-training offers multiple advantages: (a) Data efficiency. It leverages vast unlabelled data sources without manual annotation. (b) Scalability. It can be deployed continuously, enabling models to self-update as new data becomes available. (c) Adaptability. It facilitates incremental learning across evolving domains or tasks. However, it also introduces significant risks: (a) Error reinforcement. Incorrect pseudo-labels can be propagated and amplified in successive iterations (confirmation bias). (b) Quality instability. Confidence thresholds are difficult to calibrate for probabilistic LLMs, especially in open-ended generation. (c) Value misalignment. Since self-training lacks explicit human preference signals, it cannot guarantee alignment with social or ethical norms.

Consequently, self-training is often integrated with human-in-the-loop mechanisms or critique-based reflection frameworks to ensure that self-generated data aligns with the intended behavioural and factual objectives.

(5) Reinforce Learning from Human Feedback (RLHF)

RLHF constitutes the most influential and mature framework for behavioural alignment in LLMs. Unlike fine-tuning or CPT, which primarily align models with task or domain distributions, RLHF explicitly aligns model outputs with human preferences, social values, and communicative norms (Christiano et al., 2017; Ouyang et al., 2022). Through iterative optimisation driven by human feedback, the model learns to produce responses that are not only factually accurate but also helpful, harmless, and honest, commonly referred to as the 3H objectives in alignment research (Bai et al., 2022).

At its core, RLHF reframes alignment as a reinforcement learning problem in which human evaluators serve as the reward signal. The model's parameters are optimised to maximise the expected human approval of its responses rather than traditional likelihood-based objectives. This approach allows LLMs to internalise normative constraints and social expectations that cannot be directly captured through text-only corpora.

The RLHF pipeline generally consists of three stages (Ouyang et al., 2022): (a) SFT. A pre-trained language model is first fine-tuned on high-quality instruction–response pairs, establishing an initial behaviour that roughly aligns with user expectations. (b) Reward Model (RM) Training. Human annotators compare multiple candidate responses generated by the SFT model for the same prompt. These preference judgements are used to train a reward model that quantitatively estimates the relative desirability of different outputs—essentially teaching the system what humans prefer. (c) RL, where it is further optimised using the feedback signal provided by the reward model. The training objective is to maximise the expected reward while maintaining behavioural stability and avoiding excessive deviation from the fine-tuned baseline. This is often achieved using policy-optimisation algorithms, such as Proximal Policy Optimisation (PPO), which balance exploration and alignment.

Through this iterative pipeline, RLHF gradually transforms a pre-trained model from a pattern-matching text generator into an adaptive, human-aligned conversational agent. By integrating human preference signals directly into the training process, RLHF enhances the model's ability to produce responses that are not only linguistically accurate but also contextually appropriate and ethically aligned.

RLHF remains the most robust approach for behavioural and value alignment, as it directly integrates human judgement into the optimisation loop. It has proven effective in mitigating unsafe or unhelpful behaviours and improving subjective user satisfaction, as demonstrated in OpenAI's InstructGPT and ChatGPT series. However, several limitations persist: (a) Scalability. Human feedback collection is costly, time-consuming, and prone to inconsistency across annotators. (b) Reward hacking. The reward model may capture superficial correlations rather than genuine human values, leading to undesirable optimisation shortcuts. (c) Instability. PPO-based optimisation is sensitive to hyperparameters and can cause catastrophic divergence or loss of linguistic fluency if not carefully regularised. (d) Ethical bias. The system inherits the cultural and demographic biases of its human raters, posing ongoing challenges for fairness and inclusivity (Gabriel, 2020).

For these reasons, contemporary alignment research increasingly combines RLHF with other complementary strategies, such as self-training, constitutional prompting, and AI-assisted feedback, to achieve a more sustainable balance between human oversight and model autonomy.

2.5 Current State of LLM-based CHQA Systems

2.5.1 System Architecture

(1) End-to-End Architecture

End-to-end neural architectures represent the earliest and most direct application of LLMs to CHQA (see Figure 2-10). In this configuration, the model itself encapsulates the entire reasoning and response-generation process within its parameters, transforming an input query directly into a textual answer without explicit intermediate stages. Typical variants include (a) pretraining-from-scratch models, which acquire domain-specific knowledge from large-scale biomedical corpora; (b) prompt-based adaptation on general-purpose LLMs, where domain alignment is achieved through carefully engineered instructions; and (c) fine-tuning general-purpose LLMs on curated medical QA datasets.

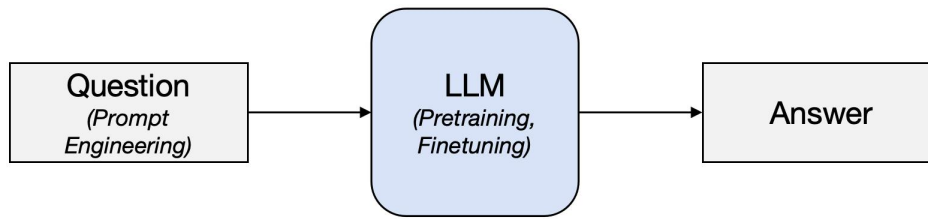


Figure 2-10. End-to-End Architecture of LLMs-Based CHQA System.

Pretraining From Scratch. A distinct line of end-to-end architecture research trains language models from scratch on in-domain corpora so that domain terminology, discourse patterns and reasoning shortcuts are embedded directly into model parameters. Two representative generative LMs are BioGPT and GatorTronGPT. BioGPT is a domain-specific GPT-style transformer pretrained on large-scale biomedical literature. It established strong results across multiple biomedical NLP tasks and reported 78.2% accuracy on PubMedQA, demonstrating that generative pretraining purely on biomedical text can yield competitive research-oriented QA performance (Luo et al., 2022).

Building on the “pretrain-on-biomedicine” premise at much larger scale, GatorTronGPT was trained from scratch on ~277 billion words (\approx 82 B de-identified clinical words from UF Health + 195 B general English from The Pile) using a GPT-3 style decoder-only architecture up to 20 B parameters. It achieved state-of-the-art performance on several biomedical NLP benchmarks and, notably, generated 20 B words of synthetic clinical text that enabled training “synthetic” models matching or surpassing models trained on real clinical notes—highlighting both the utility of domain-native generative pretraining and a practical path to alleviating data-access constraints in healthcare (Peng et al., 2023).

While most early domain models in biomedicine were encoder-only BERT variants, their evidence is instructive for CHQA: PubMedBERT showed that pretraining entirely from scratch on PubMed abstracts outperforms mixed-domain pretraining across many biomedical tasks, motivating the “all-in-domain” strategy (Gu et al., 2020). Likewise, BioBERT, pretrained on large biomedical corpora, consistently improved NER, relation extraction and biomedical QA versus general BERT, reinforcing that domain-specific pretraining provides better coverage of biomedical terminology and discourse structures relevant to question answering (Lee et al., 2020). In medical-exam style benchmarks that are widely used to gauge QA capability—MedQA (USMLE), MedMCQA, and PubMedQA—such domain-

pretrained models (either generative or encoder-only with task-specific heads) generally require fine-tuning or instruction-tuning but benefit markedly from in-domain pretraining during answer selection or generation (Jin et al., 2020).

Overall, from-scratch domain pretraining remains a cornerstone end-to-end approach for CHQA: it maximizes on-policy exposure to biomedical/clinical language, improves faithfulness to domain facts, and—at sufficient scale—can even substitute scarce clinical text with synthetic corpora. Yet it comes with high compute and data costs and does not, by itself, eliminate hallucination or calibration issues, which motivates complementary designs reviewed in later sections.

Prompting on General LLMs. A second end-to-end configuration relies on prompting and instruction-based adaptation of general-purpose foundation models such as GPT-4, PaLM, Claude, or Gemini. Instead of embedding medical knowledge within model parameters, these systems leverage in-context learning and instruction tuning to steer general LLMs toward biomedical or consumer-health reasoning. The approach capitalizes on the massive linguistic prior of general LLMs while achieving domain alignment through task-specific prompts, exemplars, or lightweight fine-tuning.

Early explorations demonstrated that instruction-tuned GPT-3 models could answer short biomedical questions competitively when supplied with few-shot exemplars or chain-of-thought reasoning cues (Brown et al., 2020; Wei et al., 2022). With the public release of ChatGPT and GPT-4, health-related prompting rapidly became feasible for lay users, spurring studies on prompt engineering strategies such as role prompting or multi-turn contextualization (Zhu et al., 2023). These methods improve response factuality and empathy without retraining, thus reducing computational cost compared with domain-specific pretraining.

PaLM 2 (Anil et al., 2023) introduced retrieval-chain prompting, a form of prompt chaining in which the model dynamically composes successive prompts that integrate retrieved evidence during reasoning. Although retrieval is invoked, the process remains prompt-driven rather than a distinct retrieval-augmented architecture. Such chaining improves factual grounding and interpretability while retaining the adaptability characteristic of prompt-based systems. Subsequent research on prompt-chaining and self-consistency decoding further boosted the reliability of medical advice generation (Wang et al., 2023). Complementary work examined few-shot prompt design for

biomedical question summarization (Li et al., 2023) and persona-based prompting to enhance empathetic tone in consumer-health counseling (Kim & Lee, 2024).

Recent advances have underscored that prompt engineering itself can rival or even surpass fine-tuning. OpenMedLM (Yang et al., 2024) systematically compared open-source LLMs (e.g., Llama 2, Falcon, Mistral) and showed that optimized multi-step prompts—with self-consistency voting and expert-role conditioning—can outperform fine-tuned biomedical models on MedQA and PubMedQA. This finding highlights the efficiency and accessibility of in-context learning for medical reasoning without additional training cost.

Prompt-based adaptation thus forms a pragmatic middle ground between static domain pretraining and retrieval-augmented paradigms. Its chief advantages are flexibility, accessibility, and rapid updatability, allowing new medical knowledge to be incorporated through prompt revision rather than costly retraining. However, prompt-only CHQA systems remain sensitive to prompt phrasing, context window limits, and factual drift. Recent studies highlight persistent hallucination, over-generalization, and ethical-safety concerns when general LLMs are applied to patient-facing tasks (Thirunavukarasu et al., 2023). Consequently, prompting on general LLMs is often combined with retrieval or multi-agent orchestration.

Fine-Tuning on General LLMs. Another major paradigm in end-to-end CHQA systems is fine-tuning general LLMs on medical or consumer-health datasets. Unlike pure prompting, fine-tuning modifies model parameters to internalize domain patterns, terminology, and reasoning styles. This approach bridges the gap between domain-agnostic general LLMs and fully domain-specific pretraining, typically through instruction tuning or SFT on curated medical QA corpora.

Early work such as BioGPT-Fine-Tuned and PubMedGPT-SFT demonstrated that adapting general architectures on biomedical question-answer pairs markedly improves factual accuracy and reasoning coherence (Luo et al., 2022; Gu et al., 2020). With the emergence of instruction-tuned models such as FLAN-T5 (Chung et al., 2022) and GPT-3.5-Turbo, researchers began applying instruction-style templates drawn from medical benchmarks (e.g., MedQA, MedMCQA, PubMedQA), yielding improvements in both correctness and explanatory quality (Singhal et al., 2023).

A landmark development was the introduction of Med-PaLM and Med-PaLM 2 (Singhal et al., 2023), which fine-tuned PaLM and PaLM 2 models via multi-stage instruction tuning. The first stage used broad medical instruction corpora constructed from UMLS and health-forum questions, while later stages employed clinician-validated QA pairs for safety alignment. Med-PaLM achieved 67 % accuracy on MedQA (USMLE) and Med-PaLM 2 further reached 86.5 %, approaching expert-level performance with significant gains in factuality, reasoning traceability, and helpfulness ratings from medical professionals. These studies established fine-tuning as a core mechanism for embedding clinical knowledge and ethical constraints within general LLMs.

Fine-tuning has also been used for adjusting general LLMs to the tone and safety requirements of health communication. Studies such as ChatDoctor and ClinicalGPT applied low-rank adaptation (LoRA) on medical dialogues or clinical notes to enhance contextual appropriateness and reduce hallucination (Li et al., 2023; Zhang et al., 2024). MedAlpaca, a lightweight model fine-tuned from LLaMA-7B on open-source clinical data, achieved competitive performance on MedQA and PubMedQA with only 0.5 % of the pretraining compute (Han et al., 2023). Similarly, MedLLama-2 extended the paradigm by introducing multi-round dialogue fine-tuning to support interactive consumer consultations while maintaining medical accuracy (Chen et al., 2024).

Recent studies further refine the fine-tuning process through alignment tuning with human or model-based feedback. For example, GatorTronSFT incorporated RLHF using clinician evaluations to balance helpfulness, safety, and factual consistency (Peng et al., 2023). Moreover, self-instruct pipelines and synthetic QA generation, where a base LLM creates additional medical QA pairs for its own tuning, have emerged as efficient methods for low-resource domains (Wang et al., 2023).

Fine-tuning thus provides stronger domain embedding and response consistency than prompt-only systems, particularly when reliable labeled data are available. However, it remains resource-intensive and may cause catastrophic forgetting of general reasoning abilities if domain datasets are narrow. As a result, hybrid frameworks now increasingly combine fine-tuned backbones with retrieval or multi-agent modules to retain factual grounding and adaptability.

(2) Retrieval-Augmented Generation (RAG) Architecture

While end-to-end approaches rely primarily on implicit knowledge encoded within model parameters, RAG architectures explicitly integrate external knowledge retrieval into the generation pipeline. In CHQA systems, this paradigm addresses a persistent limitation of purely parametric models—knowledge limitations, hallucinations, and outdated information—by coupling a retriever module with a generator (Gao et al., 2023).

When a consumer submits a health query, the retriever fetches relevant passages from external sources such as PubMed, MedlinePlus, or specialized consumer-health portals, and the generator (an LLM) conditions its response on these retrieved texts (see Figure 2-11). The resulting architecture combines the precision of information retrieval with the fluency and reasoning capacity of LLMs.

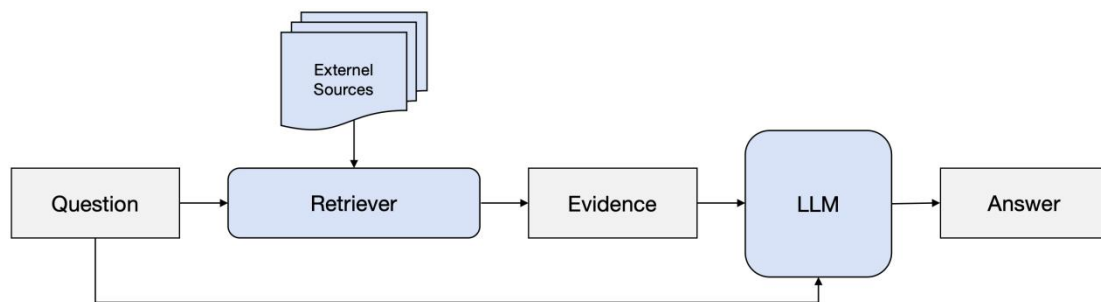


Figure 2-11. A Typical RAG Architecture for LLM-based CHQA System.

Synthesised from Lewis et al. (2020) and Lee et al. (2022).

The RAG framework was first formalized by Lewis et al. (2020) in open-domain QA and soon adapted to biomedical contexts (Lee et al., 2022). In medical and consumer-health QA, retrieval typically operates on hybrid indexes that merge dense embeddings (e.g., Sentence-BERT or BioE5) with symbolic keyword search to ensure both semantic and lexical coverage. This design improves factual grounding while mitigating hallucination, which is a key safety requirement in health communication.

Biomedical-RAG implementations such as PubMedGPT-RAG and BioMedRAG demonstrated that domain-specific retrievers significantly enhance factual accuracy on MedQA and PubMedQA benchmarks (Gupta et al., 2023). These systems retrieve top-k passages from PubMed abstracts and clinical guidelines, which are then concatenated into the LLM prompt. Evidence-Grounded RAG for Health Advice (Zhang et al., 2024) further

introduced a fact-verification layer that cross-checks the generated output against retrieved evidence, substantially reducing unsupported claims.

Large-scale commercial and open-source systems now operationalize RAG in health contexts. MedRAG (Thirunavukarasu et al., 2023) employs multi-source retrieval, including consumer-health forums, to personalize responses to lay users, while MedPrompt (OpenAI, 2024) integrates retrieval from dynamic health-web indexes to improve currency and safety of ChatGPT-based health assistants. In the research domain, RAG-based Med-PaLM 2 variants augment the original fine-tuned backbone with retrieval of PubMed and UpToDate snippets, combining parameterized reasoning with external evidence (Singhal et al., 2023). These hybrid designs represent a shift from memorization to grounded generation, allowing models to cite, verify, and update domain knowledge without full retraining.

Recent improvements employ retrieval-chain prompting, where multiple retrieval-generation steps iteratively refine the answer (Anil et al., 2023), and semantic caching, which stores prior retrieval-response pairs for efficiency (Wang et al., 2024). Some CHQA frameworks also embed a query-understanding layer before retrieval, using discourse or rhetorical segmentation to identify sub-questions, thereby retrieving more granular evidence. Despite these advances, challenges remain: retrieval quality still depends heavily on corpus coverage and index freshness, while evidence-injection may overload the model's context window or cause source-selection bias.

Overall, RAG-based CHQA architectures constitute a crucial evolutionary stage toward safe and verifiable health question answering. By externalizing domain knowledge, they achieve stronger factual consistency and adaptability than end-to-end or purely fine-tuned systems, paving the way for the multi-agent and hybrid designs discussed in the next section.

(3) Agentic Architecture

The concept of autonomous agents has long been recognised in artificial intelligence research. Classical definitions characterise an agent as a system that perceives its environment, maintains internal state, and acts autonomously in pursuit of goals (Franklin & Graesser, 1997). Building on these foundations, recent work has extended the notion of agency to LLM-based agents, which leverage an LLM to reason about tasks, construct multi-

step plans, and invoke external tools or APIs to execute those plans (Li et al., 2024; Varshney et al., 2024).

As illustrated in Figure 2-12, a typical LLM-based agent in a CHQA system operates in a continuous observe–decide–act loop. After receiving a user’s question, the agent first observes by interpreting the query and its conversational or clinical context. It then decides on the next step, informed by different forms of memory that store dialogue history, retrieved medical knowledge, and internal reasoning traces. Based on this internal state, the agent acts by, for example, generating a provisional answer, querying external medical resources such as PubMed or MedlinePlus, or asking the user or a clinician for clarification. The outcomes of these actions become new observations, closing the feedback loop and enabling the agent to iteratively refine its responses and maintain coherence over extended, multi-turn interactions (Wang et al., 2025; Yue, 2025).

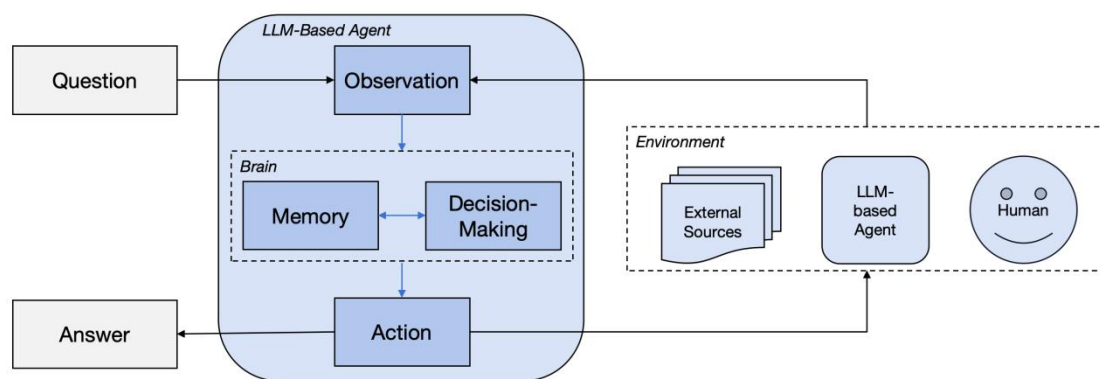


Figure 2-12. A Typical Agentic Architecture for LLM-based CHQA System.
Adapted from Yue (2025).

Recent advances in LLM-based agentic systems for healthcare are moving beyond monolithic, single-agent designs towards multi-agent architectures that more closely resemble collaborative, adaptive workflows within clinical teams (Li et al., 2024; Wang et al., 2025).

An expanding body of work investigates how multiple LLM agents can coordinate effectively, often achieving more robust performance than a single model (Du et al., 2023; Li et al., 2024). A widely adopted paradigm is role-playing, in which each agent is assigned a specialised role (e.g., an assistant agent versus a manager agent), allowing the overall task to be decomposed into sub-tasks that agents solve collaboratively (Tseng et al., 2024). While role-playing primarily emphasises cooperative, multi-step reasoning, multi-agent debate instead requires agents to generate independent solutions and

then critique one another’s arguments before reaching a final decision. This debate mechanism has been shown to improve factuality, mathematical reasoning, and general problem-solving robustness (Du et al., 2023). Other frameworks, such as consensus voting and round-table conferences that combine discussion with confidence-weighted aggregation (Chen et al., 2024), or more general architectural surveys of multi-agent collaboration strategies (Zhang et al., 2025), further explore how to balance diversity and convergence in LLM-based multi-agent systems.

MDAgents (Kim et al., 2024) provides an illustrative example of these ideas in a medical setting. The framework first classifies the complexity of an incoming medical query (low, moderate, or high), and then dynamically assembles either a single “primary care clinician” agent, a multidisciplinary team of specialist agents, or an integrated care team. The outputs of these agents are subsequently synthesised into a final recommendation. Across multiple medical benchmarks, MDAgents outperformed fixed solo- and fixed-group baselines, achieving up to 4.2% higher accuracy overall and an average improvement of 11.8% when moderator review and external medical knowledge were enabled (Kim et al., 2024).

Overall, the agentic paradigm marks an important step towards trustworthy, explainable, and collaborative health-AI systems. By embedding explicit reasoning loops, adaptive multi-agent coordination, and domain-specific safety or oversight agents, agentic CHQA architectures offer a promising foundation for next-generation digital health assistants that can balance autonomy with clinical accountability and human oversight (Yuan, 2025).

2.5.2 Datasets and Benchmarks

CHQA datasets constitute essential benchmarks for training and evaluating CHQA methods. From a communication-perspective, they can be broadly categorised into three settings: (a) peer-to-physician consultation data, where consumers pose questions to medical professionals and receive expert answers, typically through online health consultation services or moderated Q&A websites; (b) peer-to-peer communication data, where both questions and answers are contributed by laypersons (e.g., on Reddit, Twitter/X, or specialised health forums); and (c) peer-to-machine communication data, where consumers interact directly with information retrieval systems or automated agents, such as health-related search engine logs (e.g., query suggestions and click-through data). These complementary

settings collectively capture a wide range of consumer health information needs and discourse styles.

Beyond communication settings, CHQA corpora can also be distinguished by whether they provide downstream annotations (e.g., question focus, type, or socio-emotional labels). This thesis therefore groups existing datasets into two broad categories that align with different CHQA research objectives. The first category comprises original consumer–professional Q&A corpora without fine-grained annotations. These resources primarily support pre-training, language modelling, or retrieval-based methods but are less suitable for supervised learning of question understanding components (e.g., focus recognition or emotion classification). Representative examples include:

MedDialog (Zeng et al., 2020). A large-scale multilingual medical dialogue dataset containing a Chinese subset with approximately 3.4 million consultation conversations across 172 disease specialties and an English subset with about 0.26 million consultations across 96 specialties. Each sample consists of a patient’s description and a physician’s response, providing rich coverage of symptom narratives and medical advice.

HealthSearchQA (Singhal et al., 2023). A dataset of 3,173 commonly-searched English consumer health questions, created by crawling real-world search-engine suggestions from seed medical conditions and associated symptoms. It differs from many prior medical-QA corpora in that it is question-only (i.e., without paired expert answers publicly released) and is explicitly designed for evaluating long-form answer generation and expert assessment in the MultiMedQA benchmark.

Huatuo-26M (Wang et al., 2025). Currently the largest Chinese medical question-answer dataset, comprising more than 26 million QA pairs. It is constructed by unifying Q&A pairs collected from an OHC website, medical encyclopaedias, and structured medical knowledge bases. The dataset has been shown to be useful for fine-tuning medical LLMs, serving as an external knowledge source in retrieval-augmented generation, and providing a large-scale pre-training corpus for biomedical models.

The second category consists of annotated CHQA datasets that provide critical labels on question attributes such as decomposition structures, focus entities, question categories, and emotional states. These datasets are indispensable for developing and evaluating question understanding modules,

such as type classifiers, focus recognisers, and socio-emotional intent detectors, and for analysing the linguistic and cognitive complexity of consumer health questions (Kilicoglu et al., 2018; Welivita & Pu, 2023). Key resources include:

GARD (Roberts et al., 2014a) contains 1467 consumer-generated requests available on the GARD website regarding disease conditions. Each request is decomposed into subquestions and are annotated with 13 different question types (e.g., Anatomy, Cause, Complication, and Diagnosis). Additionally, each request is annotated with one or more focus diseases. This process has resulted in 2937 annotated subquestions altogether. The dataset is intended to help train and evaluate automatic techniques for decomposing complex medical questions and recognizing question focus and type.

TREC LiveQA 2017-MQA (Ben Abacha et al., 2017) was introduced at the TREC LiveQA 2017 medical subtask. It consists of two training datasets and a testing dataset with questions and their respective reference answers. The two training datasets contain 634 QA pairs altogether, which are constructed from FAQs on trusted websites of the U.S. National Institute of Health (NIH). The (sub)questions in the dataset are annotated with one of four question foci and 23 question types. Candidate answers for the questions in the first training dataset were retrieved using automatic matching between the CHQs and the FAQs based on the question focus and type. But only the manually validated QA pairs were retained for training. answers in the second training dataset were retrieved manually by librarians using PubMed and Ib search engines. The testing dataset contains 104 questions received by the U.S. National Library of Medicine (NLM) along with reference answers, which are manually collected from trusted sources such as NIH websites. The (sub)questions span across five question foci and 26 question types.

NLM-CHQA (Kilicoglu et al., 2018) is a two-part corpus containing semantically annotated consumer health questions. The first part, CHQA-email, consists of 1740 email requests received by the NLM customer service regarding consumer health. The second part, CHQA-web, consists of 874 relatively shorter questions posed to the MedlinePlus search engine as queries. Each (sub)question is manually annotated with its named entities, question focus and category, and question type and trigger . This information is arranged in a representation called a question frame. Each (sub)question is allowed to have more than one focus. The questions in CHQA-email are associated with 33 question types, whereas the questions in CHQA-web are

associated with 26 question types (created by merging some of the question types in CHQA-email that do not frequently occur in CHQA-web).

MedQuAD (Ben Abacha & Demner-Fushman, 2019) contains 47,457 medical QA pairs generated from 12 trusted NIH websites. Handcrafted patterns specific to each website were used to automatically generate Q&A pairs based on the document structure and the section titles. Each question in the dataset is annotated with one of three question foci and 37 question types. The questions are annotated additionally with synonyms of the question focus, its UMLS Concept Unique Identifier, and UMLS semantic type.

Medication QA (Ben Abacha et al., 2019) is a gold standard corpus for Answering CHQs about medications. It consists of 674 real consumer questions received by the U.S. NLM regarding medications and associated answers extracted from websites such as MedlinePlus, DailyMed, MayoClinic, NIH or U.S. government websites, academic institutions' websites, and other websites returned by Google search. Each question is manually annotated with the question focus (name of the drug the question is about) and type.

MASH-QA (Ming et al., 2020) is a benchmark dataset designed for long-document medical question answering. Unlike typical machine reading comprehension datasets that contain short, single-span answers, MASH-QA utilises lengthy health articles (e.g., from WebMD and PubMed) as contexts, and many of its approximately 35 000 question–answer pairs require multi-span extraction across several distant parts of a document. This makes it particularly suitable for evaluating models' long-range reasoning and information aggregation capabilities.

MedRedQA (Nguyen et al., 2023) extends CHQA research to social media contexts. It comprises roughly 51 000 consumer-expert interactions collected from the Reddit community r/AskDocs, where verified clinicians respond to real user queries. The dataset defines three benchmark tasks—candidate answer ranking, open-ended answer generation, and evidence generation—providing a unified evaluation framework for generative and retrieval-based CHQA systems.

RedHot (Wadhwa et al., 2023) is another Reddit-derived corpus containing about 22 000 posts annotated for claims, experiences, and questions across 24 health conditions. Each medical claim is further labelled with its PIO (Population, Intervention, Outcome) structure, allowing investigation of

misinformation detection and evidence retrieval in user-generated health discourse. Although not a traditional QA dataset, RedHot captures how laypersons express experiential and evidential reasoning in online discussions.

CHQ-SocioEmo (Alasmari et al., 2023) introduces the first dataset explicitly targeting the socio-emotional dimension of consumer health queries. It contains 1 500 community-posted questions annotated for basic emotion categories (e.g., anxiety, fear, sadness) and social support needs (emotional, esteem, network, and tangible support). The dataset also records question focus, demographic indicators, and presence of evidence statements, thus providing a unique resource for modelling empathy and social support understanding in CHQA.

Taken together, these datasets span a broad spectrum of consumer health communication, from factual information seeking to socio-emotional support, across clinical websites, community Q&A forums, social media, and search logs. They thus provide the empirical foundation for developing multi-faceted CHQA systems that can address both informational and affective user needs.

2.5.3 Evaluation Methodology

Evaluation of CHQA systems has evolved substantially from early task-oriented assessments to multidimensional frameworks that reflect both clinical accuracy and human-AI interaction quality. Traditional CHQA systems primarily focused on a single criterion—accuracy, defined as whether a system’s response is correct and verifiable against established medical knowledge. For instance, in the medical subtask of TREC LiveQA 2017, human assessors from the National Institute of Standards and Technology (NIST) evaluated single-turn responses using a four-point Likert scale (1 = incorrect, 2 = incorrect but related, 3 = correct but incomplete, 4 = correct and complete) (Ben Abacha et al., 2017). With the advent of LLMs, evaluation has extended beyond correctness to encompass dimensions crucial to healthcare—namely, safety, hallucination detection, fairness, personalisation, and empathy. This section synthesises these developments into five interrelated dimensions. Representative LLM-based CHQA systems and their evaluation strategies are summarised in Table 2-4, illustrating the field’s reliance on BLEU/ROUGE-like automatic metrics and the gradual shift toward clinician-based evaluations (e.g., Med-PaLM2, HuatuoGPT).

Table 2-4. Representative LLM-based CHQA Models and Their Evaluation Strategies.

	Base Model & Alignment Strategies	Evaluation Dataset	Evaluation Method (Metrics)
ChatDoctor (Li et al., 2023)	RAG + Fine-tuning	iCliniq (long-form, real-world consultation)	BERTScore (Precision, Recall, F1 Scores)
ClinicalGPT (Wang et al., 2023)	BLOOM-7B + Fine-tuning + RLHF	MedDialog (long-form, real-world consultation)	BLEU, ROUGE, and GLEU
HuatuoGPT (Wang et al., 2023)	GPT-4 + SFT + RLMF	Chinese medical consultation datasets (e.g., Huatuo-26M, long-form dialogues)	Case-based clinician evaluation, medical safety evaluation, knowledge consistency scoring
Med-PaLM2 (Singhal et al., 2023)	PaLM2 + Fine-tuning	Multi-domain medical benchmarks (e.g., MedQA (USMLE), HealthSearchQA, Long-Form clinical consultations)	Expert panel evaluation (10-dimension rubric: correctness, safety, alignment, empathy, reasoning clarity, etc.)
ChiMed-GPT (Tian et al., 2024)	Ziya-13B-v2 + Pre-training + SFT + RLHF	ChiMed (Chinese consumer health QA)	BLEU, ROUGE, Fluency, Completeness, and Precision

Note: HuatuoGPT and ChiMed-GPT are Chinese medical dialogue systems, while ChatDoctor, ClinicalGPT, and Med-PaLM2 primarily target English or multilingual CHQA tasks.

Accuracy. While accuracy remains foundational, its definition in the LLM era has expanded to include clinical safety and hallucination resilience. Unsafe responses may inadvertently encourage harmful actions (e.g., discontinuing medication without consultation), provide illegal or unethical advice, violate privacy, or contravene medical ethics (Han et al., 2024). Hallucinations, plausible yet unsupported or false statements (Huang et al., 2025), pose particular risks in health communication. Contemporary benchmarks therefore employ evidence-grounded verification, ensuring that each factual claim can be traced to peer-reviewed literature or authoritative databases such as PubMed or MedlinePlus. Adversarial benchmarks like Med-HALT (Pal et al., 2023) further stress-test models with “false-confidence” prompts to quantify hallucination tendencies.

Automatic metrics used in current LLM-based CHQA systems largely operationalise accuracy as surface- or embedding-level similarity between model outputs and reference answers. ChatDoctor, ClinicalGPT and ChiMed-GPT, for instance, report BLEU, ROUGE, GLEU or BERTScore on long-form consultation datasets, reflecting how closely generated responses match gold-standard answers in wording or semantic space. However, such metrics

are insensitive to many aspects of clinical correctness, for example, whether a safer alternative is recommended, whether critical red-flag symptoms are recognised, or whether advice is appropriately cautious. Because medical factuality can also exist in a “grey zone” where clinical guidelines evolve or expert consensus varies, automated metrics alone are insufficient (Landsheer, 2018). Hence, expert human evaluation remains the gold standard, with clinicians rating system outputs on accuracy, completeness, and adherence to medical consensus, as exemplified by the multi-dimensional expert assessments employed in Med-PaLM2 (Singhal et al., 2023).

Robustness. Robustness denotes the model’s stability under linguistic and contextual perturbations. A robust CHQA system should sustain performance across paraphrased, incomplete, or adversarial inputs (Ye et al., 2024; Goyal et al., 2023). Robustness can be quantified by perturbing real queries—altering symptom descriptions, inserting spelling errors, or reordering clauses—and evaluating whether the resulting outputs remain semantically consistent. For instance, Ness et al. (2024) introduced MedFuzz, a framework that systematically perturbs medical questions to expose model brittleness; their findings show that minor lexical variations can disrupt reasoning and factual retrieval.

Personalisation. Personalisation evaluates how well the system tailors its responses to an individual user’s context—this includes prior medical knowledge, health-literacy level, cultural and linguistic background, current health status, risk profile, and related preferences. A truly personalized system response is contextually adaptive: it is not only relevant to the user’s specific query and background, but also readable, actionable, and minimally burdensome in terms of cognitive load. Previous research (Kocaballi et al., 2019; Liu et al., 2022) has emphasised that personalisation in health-care conversational agents demands both user-perceived relevance and equitable performance across diverse user segments. As such, modern evaluations increasingly conceptualize personalisation as a composite metric that synthesises these dual dimensions.

Fairness. Fairness in CHQA concerns whether a system’s performance is equitable across diverse user groups and contexts, avoiding biased or stereotypical responses. In medicine, fairness concerns involve patient demographics, health conditions, or socioeconomic factors (Gallegos et al., 2024). An unfair system may provide inconsistent answers based on demographic attributes or reflect biases from training data (Li et al., 2023).

Crucially, fairness evaluation must distinguish harmful social biases from medically justified, evidence-based demographic differences (Jones et al., 2024). For instance, the higher prevalence of sickle cell anemia in individuals of African descent is a clinically relevant demographic pattern that should be preserved, not a form of algorithmic bias to be mitigated.

Empathy. Empathy represents an emerging but essential criterion for CHQA evaluation. Beyond informational adequacy, empathetic systems demonstrate awareness of emotional cues and provide reassurance, encouragement, and clarity in their responses. Empathy has been shown to enhance user trust, adherence to medical advice, and satisfaction with digital consultations (de Lima & Osório, 2021). Evaluation frameworks for empathy typically combine automated sentiment analysis with user-rated scales of emotional appropriateness, contextual alignment, and perceived empathic engagement. Recent LLM studies further propose rubric-based scoring for empathy that integrates tone control, affective relevance, and context-sensitive phrasing (Sharma et al., 2023).

Overall, while quantitative evaluation metrics such as accuracy, the F-measure, AUROC, BLEU, ROUGE, GLEU, and BERTScore provide convenient statistical indicators and are widely adopted in representative systems like ChatDoctor, ClinicalGPT, and ChiMed-GPT, they cannot fully capture the generative and interactive nature of LLMs, nor do they directly assess the clinical utility, safety, and empathic quality required for deployment in healthcare (Singhal et al., 2023; Sivarajkumar et al., 2024). Recent suggestions of using LLMs to evaluate LLM outputs are problematic (Chiang & Lee2023), particularly considering the questionable quality of summarization and the presence of misinformation in LLMs.

These limitation has prompted a growing emphasis on comprehensive assessments by human evaluators to ensure that LLMs are reliable, accurate, safe and ethical for use in healthcare. Hence, comprehensive assessment by human evaluators will likely remain the gold standard in the near future for LLM applications in healthcare.

2.5.4 Research Gap

Clinical communication research has long recognised that patients' questions extend beyond factual information seeking to encompass affective needs such as empathy, reassurance and validation. Recent analyses of

online CHQs further reveal their multi-intent nature, typically combining informational needs (e.g., diagnosis, treatment, prognosis) with emotional and social support needs (e.g., fear, uncertainty, loneliness) (see Section 2.3). From an IS perspective, this underscores a socio-technical gap between consumers' multidimensional support needs and current CHQA systems, which predominantly operationalise these needs in a purely informational form. A truly user-centred, LLM-based CHQA architecture should therefore emulate clinicians' holistic response style by jointly identifying and addressing informational and emotional support needs within a single, integrated interaction pipeline. However, existing work on CHQA datasets, system architectures and evaluation frameworks has not yet converged on such a holistic paradigm, leading to three interrelated research gaps.

Dataset Scope. Most existing CHQA benchmarks annotate only a single user intent and rely on short, information-centric queries. This simplification fails to represent the intertwined informational and emotional dimensions of real-world CHQs. Although datasets such as CHQ-SocioEmo (Alasmari et al., 2023) explicitly capture socio-emotional factors by annotating emotions and social support needs in consumer health questions, they remain limited in size and are not jointly integrated with detailed informational intent annotations or discourse-level structures. Conversely, psychological counselling datasets often focus exclusively on emotional support and counselling strategies, neglecting the informational dimension that is central to CHQA tasks. As a result, no benchmark currently provides multi-intent annotations that jointly encode informational and emotional support needs within a coherent symbolic discourse structure that mirrors the way users articulate multiple, overlapping concerns. This absence of a discourse-aware, multi-intent benchmark constrains the design, training and evaluation of LLM-based CHQA architectures that aim to jointly understand informational and emotional support needs.

System Design. From the perspective of system design, existing CHQA systems remain largely confined to the informational dimension of user needs. Earlier studies, which predominantly handled short and single-intent CHQs, focused on improving information retrieval or extraction accuracy as a proxy for overall system performance (e.g., Demner-Fushman et al., 2019). In contrast, research on question understanding, the critical stage for intent parsing and discourse-level reasoning, has been limited to syntactic or semantic analyses (see Section 2.2). A few attempts have explored multi-intent understanding, yet they typically operate within a single informational dimension, relying on shallow summarisation or decomposition that fails to capture the hierarchical discourse relations and affective cues inherent in

complex CHQs. Consequently, there is still no CHQA architecture that treats informational and emotional support needs as first-class, jointly modelled constructs throughout the question-understanding and response-generation pipeline.

Although contemporary LLMs are capable of processing multi-dimensional, context-rich inputs, existing LLM-based CHQA frameworks remain primarily oriented toward improving factual retrieval accuracy. They therefore encounter a critical architectural bottleneck at the question-understanding stage, the entry point for parsing, disambiguating and structurally representing multi-intent consumer health questions. Among widely adopted designs, RAG architectures explicitly integrate external knowledge to enhance informational accuracy, yet their focus remains largely unidimensional, addressing users' informational intents while overlooking affective understanding. Agentic architectures, particularly multi-agent frameworks, have demonstrated notable performance gains and are theoretically well-suited for multi-dimensional reasoning. However, they introduce substantial computational cost and latency, alongside additional limitations such as coordination fragility, run-to-run variability, non-deterministic planning and privacy or auditability challenges (Kim et al., 2024; Wang et al., 2025). End-to-end architectures, by contrast, typically concentrate on two main strategies: (a) optimising chain-of-thought prompting to strengthen retrieval and reasoning; and (b) fine-tuning LLMs on curated QA pairs. Yet most of these pairs are single-intent and lack explicit annotation for intent parsing or structured representation. This approach also entails high supervision cost, as it depends on domain experts for high-quality reference answers, and suffers from rapid obsolescence given the fast-paced evolution of medical knowledge.

Across these paradigms, symbolic knowledge, such as discourse trees, medical ontologies, socio-emotional taxonomies and clinical communication frameworks, is rarely integrated into LLM-based CHQA architectures in a systematic way. As a result, current systems have limited capacity to constrain or guide LLMs' reasoning about users' informational and emotional support needs using explicit, inspectable structures, particularly in safety-critical contexts. These architectural limitations at the question-understanding and reasoning stages collectively weaken downstream modules' capacity to integrate factual reasoning with empathetic expression. They also reveal a second, more specific gap: the absence of a unified neural-symbolic CHQA architecture in which symbolic knowledge is explicitly integrated with LLMs to enhance safe and empathetic joint understanding and response.

Evaluation Methodology. Although recent studies have expanded evaluation dimensions beyond accuracy to include aspects such as personalisation, empathy and safety for health-related conversational agents, existing metrics remain predominantly task-specific and fail to capture a system’s joint performance in handling overlapping informational and affective intents. Current evaluation practices typically assess these dimensions in isolation, measuring factual correctness or emotional appropriateness separately, thereby overlooking their interdependence in shaping the perceived quality and safety of CHQA interactions. In particular, there is limited work on evaluation schemes that explicitly mirror multi-intent question structures and jointly account for whether a system has (a) correctly identified and prioritised co-occurring informational and emotional needs and (b) produced responses that appropriately address both. What remains absent is a unified, multidimensional evaluation framework capable of simultaneously quantifying informational reliability, emotional appropriateness and safety, and of assessing how effectively a system aligns its responses with users’ informational and emotional support needs in an integrated and context-sensitive manner. Such a framework is essential for evaluating not only what information is delivered, but also how it is conveyed, thereby providing a rigorous basis for assessing whether an LLM-based CHQA architecture truly succeeds in jointly understanding and responding to consumers’ multidimensional support needs.

Taken together, these three gaps: (a) the lack of discourse-aware, multi-intent CHQA benchmarks that jointly encode informational and emotional support needs; (b) the absence of a unified neural-symbolic LLM-based CHQA architecture that integrates symbolic knowledge to support safe and empathetic joint reasoning; and (c) the lack of an integrated evaluation framework for jointly assessing informational and emotional performance, highlight the need for a new architectural approach to LLM-based CHQA in OHC settings. Accordingly, this study is guided by the following central research question:

RQ: *How can a large language model-based consumer health question answering architecture be designed to jointly understand and respond to informational and emotional support needs embedded in consumer health questions within online health consultation settings?*

This overarching question is further decomposed into two sub-questions:

Sub-RQ1: *How can a symbolic knowledge representation be designed to effectively support the joint understanding of informational and emotional*

support needs in large language model-based consumer health question answering systems?

Sub-RQ2: *How can symbolic knowledge be integrated with large language models to enhance a consumer health question answering system's capability to jointly understand and respond informational and emotional support needs?*

2.6 Theoretical Framework

Building on the research gaps identified in Section 2.5 and the central research question and sub-questions, this section articulates the theoretical foundations that guide the design of the proposed LLM-based CHQA architecture. Whereas the preceding literature review has shown what is missing in current CHQA systems, the present section clarifies which bodies of theory this study draws on to address these deficits. Specifically, it mobilises three complementary domain-level lenses to model the informational and emotional support needs embedded in consumer health questions and to inform the design of a system that can jointly understand and respond to these needs. The subsequent subsections introduce and synthesise these theoretical perspectives, which together provide the conceptual scaffolding for the research framework developed in Section 2.6.4.

2.6.1 Neural-Symbolic Integration (NSI)

Within the broader field of AI, neural-symbolic integration (also referred to as neural-symbolic AI or neural-symbolic computing) represents a hybrid paradigm that reconciles two historically distinct traditions: symbolic, rule-based AI and neural, data-driven AI (Garcez et al., 2019; De Raedt et al., 2020; Wang et al., 2022). Symbolic systems emphasise explicit knowledge representation, logical reasoning and interpretability, whereas neural or connectionist systems, grounded in Parallel Distributed Processing (PDP) theories (Rumelhart & McClelland, 1986), excel at learning distributed representations, pattern recognition and generalisation from data. These traditions therefore exhibit complementary strengths: symbolic approaches provide structure, explanation and normative constraints; neural approaches provide flexibility, robustness and scalability.

The intellectual roots of neural-symbolic integration can be traced to the foundational connection between logic and neural computation established

by McCulloch and Pitts (1943), who demonstrated that finite automata could be modelled by neural networks, thereby linking formal reasoning with biologically inspired computation. Subsequent research deepened this bridge between symbolic and connectionist representations. In the following decades, however, research in AI focused predominantly on purely symbolic approaches, and the potential of artificial neural networks remained under-recognised until the 1980s, when the backpropagation algorithm (Rumelhart et al., 1986) made connectionist learning practically feasible. Building on these advances, Pollack (1990) proposed Recursive Auto-Associative Memories (RAAM) as a mechanism to encode structured symbolic terms within distributed neural representations, enabling neural systems to handle hierarchical symbolic information. Similarly, Shastri and Ajjanagadde (1993) developed the SHRUTI system, a pioneering model that supported rapid, reflex-like reasoning over symbolic structures using synchrony-based connectionist computation. Hölldobler and Kalinke (1994) subsequently introduced the Connectionist Model Generation method, further demonstrating that logic programs could be systematically represented and executed within neural architectures. Collectively, these approaches laid the theoretical foundations for contemporary work in neural-symbolic learning and reasoning, in which neural networks perform pattern recognition and generalisation while symbolic components maintain logical consistency and interpretability.

Building on this historical trajectory, Bader and Hitzler (2005) systematised a neural-symbolic learning cycle that many integrated systems implicitly follow. As depicted in Figure 2-13, a front-end symbolic system is used to encode (partial) expert knowledge and feed it into a neural or connectionist system, which is then trained on raw data while taking the internally represented symbolic knowledge into account. Knowledge acquired through learning can be extracted back into the symbolic system (which now also acts as a back-end) and made available for further processing in symbolic form. Individual studies often address only parts of this neural-symbolic learning cycle (such as knowledge representation or extraction), but they can nonetheless be regarded as contributing to the overall investigation of the cycle.

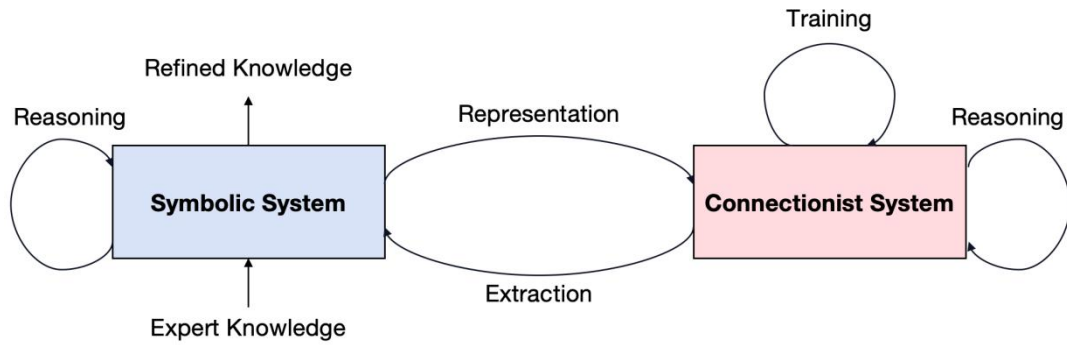


Figure 2-13. Neural-Symbolic Learning Cycle. Adapted from Bader & Hitzler (2005).

Neural-symbolic systems can be broadly organised into two canonical architectural paradigms: hybrid and unified (Hilario, 1995; Bader & Hitzler, 2005; Besold et al., 2017). In the hybrid paradigm, neural and symbolic components operate as distinct modules coordinated by a controller, exchanging representations through explicit interfaces. In contrast, unified approaches embed symbolic structures directly within the neural substrate, achieving an end-to-end flow of representation and reasoning (see Figure 2-14).

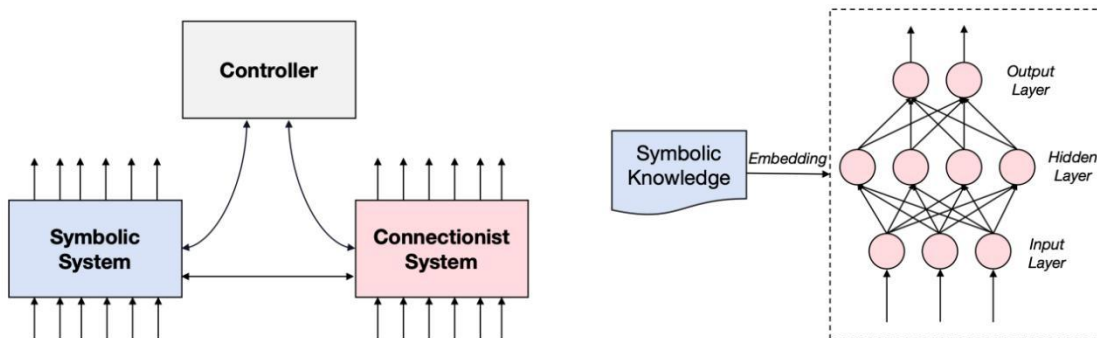


Figure 2-14. Hybrid (left) versus Unified (right) Neural-Symbolic Architectures. Adapted from Bader and Hitzler (2005).

Beyond this high-level dichotomy, Bader and Hitzler (2005) proposed a three-dimensional classification framework for neural-symbolic systems, structured along the axes of interrelation, language and usage. The interrelation axis captures the degree of coupling between symbolic and neural components. At one extreme, loosely coupled hybrid approaches maintain separate symbolic and neural modules that coexist or cooperate, typically exchanging information at predefined interfaces. At the other extreme, unified approaches seek a tight, intrinsic integration in which symbolic and subsymbolic representations are embedded within a single computational architecture. The language axis concerns the representational formalism

used to encode knowledge, ranging from purely logical or propositional languages to hybrid representations that combine discrete symbols with continuous vector spaces. The usage axis describes the functional role of the symbolic component within the neural-symbolic system, from knowledge representation and reasoning (symbolic content guiding neural inference) to knowledge extraction and explanation (deriving symbolic structures from trained networks).

With the rise of deep learning and renewed emphasis on explainable and robust AI in the 2010s, these earlier efforts were consolidated into more explicit taxonomies. Surveys such as *Neural-Symbolic Learning and Reasoning: A Survey and Interpretation* (Besold et al., 2017) and *Towards Data- and Knowledge-Driven Artificial Intelligence: A Survey on Neural-Symbolic Computing* (Wang et al., 2022) document how neural-symbolic systems combine distributed learning with symbolic reasoning to achieve both flexibility and transparency.

Building upon this convergence, Kautz (2022) proposed a widely cited taxonomy that systematically categorises major approaches to neural-symbolic integration. His taxonomy delineates six primary types of integration architectures, ranging from loose coupling, where neural and symbolic components interact via data exchange, to tightly coupled and fully unified systems, in which symbolic reasoning is embedded directly within neural architectures. The six types can be summarised as follows:

Type 1: Symbolic Neural Symbolic. Symbolic inputs (e.g., tokens) are mapped to vector representations, processed by a neural network and then decoded back into symbolic outputs. This is the predominant architecture in contemporary NLP models, including LLMs such as BERT, GPT-3 and GPT-4, and thus most current LLMs (e.g., GPT-4, DeepSeek, Qwen-2.5) can be interpreted as belonging to this category.

Type 2: Symbolic[Neural]. A hybrid but overall symbolic system that calls neural subroutines as components within a symbolic problem solver. For example, AlphaGo's use of deep networks within a Monte-Carlo tree search framework (Silver et al., 2016). Multi-agent, tool-augmented CHQA systems that employ LLMs for pattern recognition or planning within an otherwise symbolic workflow can be viewed as instances of this type.

Type 3: Neural | Symbolic. A neural architecture interprets perceptual inputs and produces symbolic representations that are then processed by a symbolic reasoning system (e.g., Neural Concept Learner). Many information-retrieval-based CHQA systems and RAG-style LLM-based CHQA architectures exemplify this pattern, where neural encoders map text to embeddings that are used to retrieve or construct symbolic knowledge, which then guides downstream reasoning.

Type 4: Neural: Symbolic → Neural. Symbolic knowledge is compiled into training data or constraints for a neural network. Logical rules or knowledge graphs are used to generate or label examples, which are then learned by a purely neural model. End-to-end LLM-based CHQA systems fine-tuned on medical knowledge graphs can be interpreted as Type 4 architectures.

Type 5: *Neural*_{Symbolic}. Symbolic rules are directly mapped into neural network structures, so that a neural net is generated from symbolic knowledge (e.g., Neural Theorem Provers, Logic Tensor Networks). In such tightly coupled systems, symbolic constraints act as soft or hard regularisers on the network's loss function, enabling the incorporation of domain rules into neural learning. LLM-based CHQA systems that integrate rule-based rewards or symbolic consistency constraints into reinforcement learning can be seen as moving towards this type.

Type 6: Neural[Symbolic]. A neural model invokes a symbolic reasoning engine as an external tool (e.g., a neural controller calling a theorem prover or a symbolic calculator). In principle, a fully fledged Type 6 system would support rich combinatorial reasoning within a neural controller that orchestrates symbolic modules. Kautz (2022) argues that although no complete Type 6 system yet exists, recent progress in tool-augmented LLMs suggests a path towards this goal.

This typology provides a comprehensive lens through which to analyse how different methodologies operationalise the bridging of neural learning and symbolic reasoning. It thus serves as an organising framework for understanding the design choices and theoretical orientations underpinning neural-symbolic systems in both AI and IS research.

Building on the classification scheme proposed by Bader and Hitzler (2005) and updated to reflect recent development trends in the AI and IS fields, Wang et al. (2024) conceptualises NSI as comprising three tightly coupled

operational components: knowledge representation, knowledge embedding, and functionality.

Knowledge Representation. Understanding symbolic knowledge serves as the cornerstone of a Neural-Symbolic system (Bader & Hitzler, 2005). Existing representation approaches for symbolic knowledge can be classified into five main groups: (a) Knowledge graphs. Knowledge graphs are typically directed labeled graphs, formed by representing entities as nodes, and relations between entities as edges. They contain facts that are represented as “SPO” triples: (Subject, Predicate, Object) where Subject and Object are entities and Predicate is the relation between them. Edges are directed from subject to object, and edge labels represent different types of relations. (b) Propositional logic. Propositional logic also known as boolean logic or sometimes zeroth-order logic, is the simplest form of logic where all the statements are made by propositions. (c) First-order logic. First-order logic can express the relationship between objects by allowing variables in predicates bound by quantifiers. Specifically, first-order logic augments propositional logic with two new linguistic features, viz. variables and quantifiers. (d) Programming language: A family of formal language used for writing computer programs and communicating with machines. Typically, they consist of syntax and semantics, where syntax represents rules that define the combinations of symbols and semantics assigns computational meaning to valid strings formulated with respect to the syntax. (e) Symbolic expression. Symbolic expression here roughly refers to other types of knowledge representation other than those mentioned above. Representative examples include mathematical expressions and specific symbolic sequences generated from some informal symbolic systems with self-defined rules.

Knowledge Embedding. Knowledge embedding concerns the question of where symbolic knowledge resides within connectionist, neural-network-based solutions. Clarifying this locus is crucial for understanding how symbolic knowledge and neural representations are integrated in contemporary neural-symbolic systems. Drawing on prior survey work, knowledge embedding can be organised into four levels (Wang et al., 2025): (a) Data. A natural strategy to embed knowledge into connectionist approaches is to straightforwardly embed it in the structure of data. A prominent approach is to translate symbolic expressions into a structured (typically tree-/graphorganized) symbolic sequence, with respect to the corresponding grammars, semantics, and/or the relational structure of the knowledge. The advantage of this knowledge embedding strategy is the

tremendous relief of burden on the engineering of network architecture and training objective. However, such simple strategy has its limits for embedding complex knowledge. (b) Sub-symbolic representation. A common way of building such knowledge-specialized training objectives is to make discrete symbolic operations differentiable. Designing appropriate loss functions for distributed encoding of symbolic knowledge is appealing as it does not require architectural change to the connectionist models or extra load of preprocessing the input data. However, it appears to be particularly challenging and there is no guarantee that distributed knowledge embedding can always lead to valid outputs that are coherent with the symbolic knowledge. (c) Network architecture. It is another common way to integrate knowledge into DNNs that design the network architecture to reflect the structure of the knowledge. This strategy requires significant engineering efforts in neural architecture design. (d) Neural inference. Embedding symbolic knowledge into network feedforward inference is also a feasible way, which imposes explicit constraints to force the final hypothesis to agree with the knowledge.

Functionality. Conceptually, the goal of NSI is to realise AI systems that combine the strengths of data-driven learning and knowledge-driven reasoning. In practice, however, most existing neural-symbolic systems still lean predominantly towards either statistical learning or symbolic reasoning. This functional dimension therefore captures whether a given system primarily uses symbolic components for knowledge representation and constraint enforcement, or whether it additionally exploits neural-symbolic mechanisms for knowledge extraction, explanation, and higher-order reasoning (Bader & Hitzler, 2005; Garcez et al., 2019).

Within the IS field, NSI has increasingly been recognised as a theoretical and technical bridge between data-driven analytics and rule-governed organisational logics. This synthesis is particularly salient for the design of complex decision-support, knowledge-management, and socio-technical systems that must satisfy both algorithmic efficiency and institutional accountability (Gregor & Hevner, 2013; Rai, 2017). As modern IS artefacts embed AI capabilities more deeply, the traditional dualism between symbolic reasoning and statistical learning is being reconceptualised as a continuum rather than a strict dichotomy (Yoo, Henfridsson, & Lyytinen, 2010). NSI provides a unifying meta-theoretical lens for this transformation: it explains how systems can simultaneously learn from empirical data while remaining constrained by domain-specific rules, ethical norms, and organisational semantics.

At the same time, NSI remains an evolving paradigm rather than a fully formalised theory. Current limitations include the absence of unified evaluation metrics for symbolic–neural co-operation, difficulties in guaranteeing consistency between symbolic and sub-symbolic layers, and limited empirical validation in socio-technical contexts. Accordingly, this study employs NSI primarily as a meta-theoretical lens to guide system design and analysis, rather than as a prescriptive or predictive theory. Future research may further operationalise neural-symbolic principles into measurable constructs within IS evaluation frameworks and draw on them to structure hybrid human–AI governance mechanisms.

2.6.2 Rhetorical Structure Theory (RST)

RST, introduced by Mann and Thompson (1988), is a linguistic and discourse-analytic framework that explains how coherence and communicative intent emerge from the structural organisation of a text. RST assumes that discourse is composed of hierarchically related spans of text connected by rhetorical relations such as Elaboration, Cause, Contrast, and Evidence. Each relation links a nucleus (the central informational unit) and a satellite (a supporting or qualifying unit) forming a network of functional dependencies that reflect the writer’s communicative goal (Mann & Thompson, 1988).

The key insight of RST is that textual coherence arises not merely from lexical cohesion, but from an underlying intentional and relational organisation. This makes RST both a descriptive theory of discourse structure and a computational model of text organisation, capable of representing how human communication achieves clarity, persuasion, and coherence through structured relations (Mann & Thompson, 1988; Mann, Matthiessen, & Thompson, 1992). Building on its original formulation, Mann et al. (1992) extended RST to cross-linguistic contexts, while Taboada and Mann (2006) refined its operational definitions and annotation procedures. The construction of the RST Discourse Treebank (RST-DT) by Carlson and Marcu (2001) enabled large-scale empirical validation and computational modelling, positioning RST as a cornerstone of discourse parsing research.

Early RST research was primarily concerned with micro discourse analysis, which examines intra-textual coherence within a single document or dialogue. In this setting, RST is used to identify how individual clauses or sentences are connected through rhetorical relations to express a unified

communicative intent (Mann & Thompson, 1988). Micro-level RST analysis has been widely applied in text summarisation (Marcu, 2000), argument mining, and dialogue act segmentation, where the goal is to infer the functional role of each textual span (e.g., justification, elaboration, or contrast) and its contribution to local meaning.

By contrast, macro-level discourse analysis extends RST to cross-document or multi-episode contexts, analysing how multiple discourse units, such as paragraphs, documents, or conversation turns, combine to express global rhetorical and thematic structures (Taboada, 2016; Hovy & Lavid, 2010). Macro-level RST analysis captures phenomena such as topic development, argumentation flow, and emotional framing across long texts or corpora, enabling applications in organisational communication, scientific writing, and large-scale QA or summarisation systems. These developments have motivated the construction of macro-level RST corpora and parsers that operate above the clause level (Jiang et al., 2018, 2021).

RST is typically operationalised through three stages:

Stage 1: Elementary Discourse Units (EDUs) Segmentation. EDUs are the minimal spans of text that convey a single propositional meaning or communicative act. Segmentation into EDUs constitutes the first stage of discourse parsing, where each EDU functions as a node in the rhetorical structure. During this stage, each EDU is annotated with its communicative function (e.g., nucleus or satellite), enabling fine-grained analysis of how meaning and informational salience are distributed across the text (Carlson & Marcu, 2001). Classical EDU segmentation in RST follows syntactic and semantic criteria: an EDU generally corresponds to a clause bounded by coordinating or subordinating conjunctions, punctuation marks, or discourse connectives such as *because*, *however*, or *although* (Mann & Thompson, 1988; Carlson & Marcu, 2001). Subsequent works refined this rule-based approach using syntactic parse trees and dependency structures to identify clause boundaries more consistently (Taboada & Mann, 2006; Feng & Hirst, 2012).

Beyond the sentence level, researchers have extended EDU segmentation to capture macro-level discourse units that operate above clause granularity. Jiang et al. (2018) proposed the Macro-level Chinese Discourse Treebank (MCDTB), where entire paragraphs are treated as elements of a macro-discourse unit to model long-range discourse dependencies. Building on this

idea, Jiang et al. (2021) introduced a hierarchical topic-segmentation mechanism that detects implicit topic boundaries across multiple paragraphs, allowing document-level parsers to construct discourse trees in a top-down manner. Other approaches employ hierarchical or neural segmentation models to integrate lexical cohesion, topic continuity, and contextual embeddings into EDU boundary detection (Li et al., 2016; Lin et al., 2019).

Stage 2: Rhetorical Relations Identification. In the second stage, adjacent or hierarchically related EDUs are linked through predefined rhetorical relations, such as Elaboration, Evidence, Contrast, Condition, or Cause. Each relation defines the logical and pragmatic dependency between a nucleus and its satellite, specifying how one span supports or constrains the interpretation of another (Mann & Thompson, 1988). The identification of these relations is typically guided by linguistic cues, semantic roles, and discourse markers (Taboada & Mann, 2006). According to the original formulation of RST, Mann and Thompson (1988) enumerated approximately twenty-three relation types, grouped into two broad classes: subject-matter relations and presentational relations, reflecting whether the focus is on content organisation or on authorial intention and effect. Subject-matter relations include those like Elaboration, Evidence, Background, Cause, Condition, Enablement, Means, and Purpose. The presentational relations include those such as Evaluation, Motivation, Summary, Restatement, and Contrast.

Stage 3: Discourse Structure Tree (DST) Formation. Finally, identified relations are recursively combined to form a DST, representing the hierarchical and functional organisation of the text. This tree models both local coherence (through immediate rhetorical connections) and global coherence (through higher-level groupings). It provides a computationally tractable abstraction of human reasoning about how ideas are related and mutually supportive, forming the theoretical basis for applications such as discourse parsing, text summarisation, question understanding, and response generation (Marcu, 2000; Joty et al., 2015).

Within IS, RST has been increasingly adopted as a cognitive and analytical framework to examine knowledge representation, human-computer communication, and information design. For instance, Pérez-Quiñones and Shneiderman (1998) applied RST principles to interface documentation and help-system design, showing that rhetorical coherence facilitates user comprehension and decision making. Delin et al. (2002) used RST to analyse

document coherence in information retrieval systems, demonstrating that rhetorical relations can enhance query relevance and document summarisation.

In computational linguistics and QA research, RST provides a robust basis for modelling discourse-level coherence. Studies such as Marcu (2000) and Joty et al. (2015) employed RST-based discourse parsing to improve text summarisation, dialogue analysis, and answer generation. By mapping rhetorical relations between user questions and candidate answers, systems can infer the functional role of each clause (e.g., request, justification, reassurance) and produce contextually appropriate responses. This makes RST particularly relevant to CHQA tasks, where multi-intent queries often combine informational and emotional components (Delin et al., 2002; Jiang et al., 2018). In such contexts, RST-enabled discourse representations can serve as symbolic inputs to LLM-based models, facilitating more coherent, personalised, and empathetic responses in consumer health question answering.

2.6.3 Appraisal Framework for Clinical Empathy (AFCE)

AFCE (Pounds, 2011) conceptualises empathy in healthcare encounters not merely as affective resonance but as a process of recognising, interpreting, and aligning with a patient's evaluative stance. Building on Martin and White's (2005) Appraisal Theory within Systemic Functional Linguistics, Pounds argues that patients' communicative expressions encode attitudinal meaning, encompassing their emotions, moral judgements, and evaluations of events, which clinicians must first interpret before delivering biomedical information. Within this model, empathy involves three interconnected dimensions: Attitude (subdivided into Affect, Judgement, and Appreciation), Engagement (the patient's voice and positioning towards alternative views), and Graduation (the intensification or attenuation of stance).

AFCE specifies three major dimensions of patients' attitudinal expression, with particular emphasis on negative evaluations that are especially salient in clinical encounters. First, Feeling refers to explicit expressions of negative emotion (e.g., fear, sadness, anger) or implicit expressions of distress through references to adverse experiences. Second, Judgement captures explicit or implicit negative evaluations of oneself or others, including perceived blame, inadequacy, or wrongdoing. Third, Appreciation concerns explicit or implicit negative evaluations of things, events, or actions, such as

dissatisfaction with treatments, services, or institutional arrangements. These categories provide a systematic, linguistically grounded lens for detecting how patients encode emotional needs, moral concerns, and evaluative stances in their language.

Central to Pounds' framework is the conceptualisation of clinical empathy as stance recognition followed by alignment. On the basis of her discourse analysis of clinical consultations, Pounds (2011) identifies three prototypical stance-alignment strategies that clinicians can deploy after appraisal-based recognition. Understanding involves explicitly acknowledging or validating patients' feelings and perspectives. Sharing consists of aligning with the patient's stance by expressing agreement or similar experiences. Acceptance entails offering unconditional positive regard or neutral support in response to patients' explicit, implicit, or anticipated negative self-judgements. These strategies do not replace biomedical explanation, rather, they constitute a prerequisite interpersonal layer that creates emotional safety and trust, thereby enabling patients to better process clinical information and engage in shared decision-making (Suchman et al., 1997).

Recent work has extended AFCE beyond face-to-face consultations into digital and information-systems-oriented contexts. Lahnala et al. (2024), for example, applied the appraisal framework to annotate breaking-bad-news dialogues and to develop NLP-enabled tools for training empathic communication, demonstrating that AFCE categories can be operationalised as computational labels for empathic opportunities and responses. Velasco et al. (2022) combined empathy assessment with Systemic Functional Linguistics to analyse patient cues in asynchronous health-coaching messages, illustrating how appraisal-based categories capture subtle emotional signals in text-based interactions. Together, these studies highlight the potential of AFCE for Health Information Systems (HIS): by mapping patient utterances onto structured attitudinal categories, HIS modules can be designed to detect emotional cues, trigger clinician-facing prompts, log empathic responses, and support longitudinal analytics of empathic behaviour. In this thesis, AFCE is therefore adopted as a language-based framework for modelling how consumer health queries encode affective and evaluative meanings, and for guiding the design of CHQA components that aim to recognise and respond to patients' emotional and social-support needs.

2.6.4 Research Framework

Building on the theoretical foundations outlined in the previous sections and guided by the central research question, this study develops an integrative research framework that aligns theoretical insight with system design and evaluation. The framework synthesises multiple theoretical lenses drawn from medicine, linguistics, and information systems, ensuring that both the design and evaluation of the system are grounded in human-centred, explainable, and adaptive principles.

The framework integrates two complementary theoretical models: the AFCE as a linguistic framework for detecting and aligning with patients' emotional stance, and RST as a discourse framework for capturing the hierarchical and relational organisation of information units. These theories collectively provide the conceptual scaffolding for designing and evaluating a multi-intent CHQA system that integrates medical reasoning and emotional understanding. To operationalise this integration, the framework adopts NSI as the overarching meta-theoretical foundations, bridging symbolic interpretability with neural adaptability and linking cognitive theory with IS design.

NSI as the Cognitive and Engineering Foundation. NSI reconciles symbolic, rule-based reasoning with neural, data-driven learning by embedding explicit knowledge structures within connectionist architectures. In this study, NSI provides the meta-theoretical foundation for coupling symbolic representations of medical and emotional knowledge with LLMs in a single CHQA architecture. Symbolic components, such as RST-based discourse structure trees, and AFCE-based stance labels, are treated as structured knowledge resources that constrain and explain neural outputs. The neural components, instantiated by LLMs, supply flexibility in language generation, pattern recognition over heterogeneous CHQA corpora, and adaptive generalisation across user populations. This hybrid design reflects NSI's core learning cycle: expert knowledge is encoded in symbolic form, translated into neural representations, refined through data-driven learning, and partially re-extracted or inspected to maintain transparency and alignment. Within an IS design science perspective, NSI thus specifies how the artefact's architecture operationalises a combination of symbolic and neural capabilities, supporting rigorously justified design decisions and knowledge contribution.

RST for Discourse-Level Contextualisation. RST conceptualises texts as hierarchically organised structures of nuclei and satellites connected by rhetorical relations such as Elaboration, Evidence, Cause and Contrast (Mann & Thompson, 1988; Taboada & Mann, 2006). Within the proposed framework, RST is used to model the discourse structure of consumer health questions and associated background narratives. EDUs are segmented from the query and organised into a DST that distinguishes core informational intentions (nuclear spans) from contextual satellites (e.g., symptoms, prior treatments, psychosocial concerns). Macro-level extensions of RST, including paragraph-level discourse units and topic segmentation, further support the analysis of long, narrative CHQs that span multiple episodes or problems.

By aligning each nuclei (problem) with its corresponding satellites (context), the framework encodes not only “what” the patient is asking but also “how” and “why” the questions are framed. This discourse-level representation is then exposed to the NSI-based architecture: the symbolic DST acts as structured input to the LLM (e.g., via prompts or intermediate representations), guiding retrieval, reasoning, and response ordering, while the neural model ensures fluent, contextually appropriate language generation. In this way, RST operationalises the question-understanding dimension of the research, enabling systematic examination of how discourse structure affects CHQA performance.

AFCE for Modelling Clinical Empathy. AFCE builds on Appraisal Theory (Martin & White, 2005) and conceptualises empathy as a process of recognising and aligning with a patient’s evaluative stance across the dimensions of Attitude, Engagement and Graduation. Pounds (2011) shows that patients’ utterances encode affective states (Affect), moral or behavioural evaluations (Judgement), and evaluations of events or objects (Appreciation), as well as the degree of commitment to these positions and their intensification. In this framework, AFCE provides the conceptual and operational basis for detecting these attitudinal patterns in CHQs and for specifying stance-alignment strategies (understanding, sharing, acceptance) that the system should emulate. This ensures that system responses are not only medically accurate but also emotionally attuned and patient-centred.

In line with IS design science principles, this integrated framework thus links the overarching research question to three interrelated concerns: (a) how to represent and operationalise complex medical and emotional needs in symbolic form; (b) how to embed these representations within a NSI-based

architecture that leverages LLMs for flexible reasoning and generation; and (c) how to evaluate the resulting artefact in terms of clinical accuracy and empathic quality. The framework consequently serves not only as a theoretical lens but also as a practical blueprint for the design, implementation and assessment of the CHQA system detailed in Chapter 5.

2.7 Chapter Summary

Chapter 2 established the theoretical and conceptual foundations for this research, synthesising interdisciplinary insights to guide the design of the proposed LLM-based CHQA system. The chapter's key contribution lies in synthesizing the theoretical gap and articulating a composite theoretical foundation that unites three complementary perspectives: Neural-Symbolic Integration (NSI), Rhetorical Structure Theory (RST), and the Appraisal Framework for Clinical Empathy (AFCE). Together, these frameworks form an integrated lens through which the informational, affective, and contextual dimensions of CHQs can be jointly modelled and operationalised.

By the end of this chapter, the literature review has established a robust theoretical and technological basis for artefact construction. The insights developed here directly inform Chapter 3, which details the research methodology and explains how the DSR process operationalises this foundation through iterative design, implementation, and evaluation cycles.

Chapter 3 Research Methodology

3.1 Chapter Introduction

Chapter 3 presents the research methodology underpinning this study. While the preceding chapters established the theoretical background and problem space (Chapter 2), this chapter articulates the methodological foundations that guided the design, development, and evaluation of the proposed Med-Emo CHQA system. The objective is to demonstrate methodological rigour by explicating how the study’s philosophical assumptions, design framework, and research methods collectively ensure validity, reproducibility, and theoretical contribution.

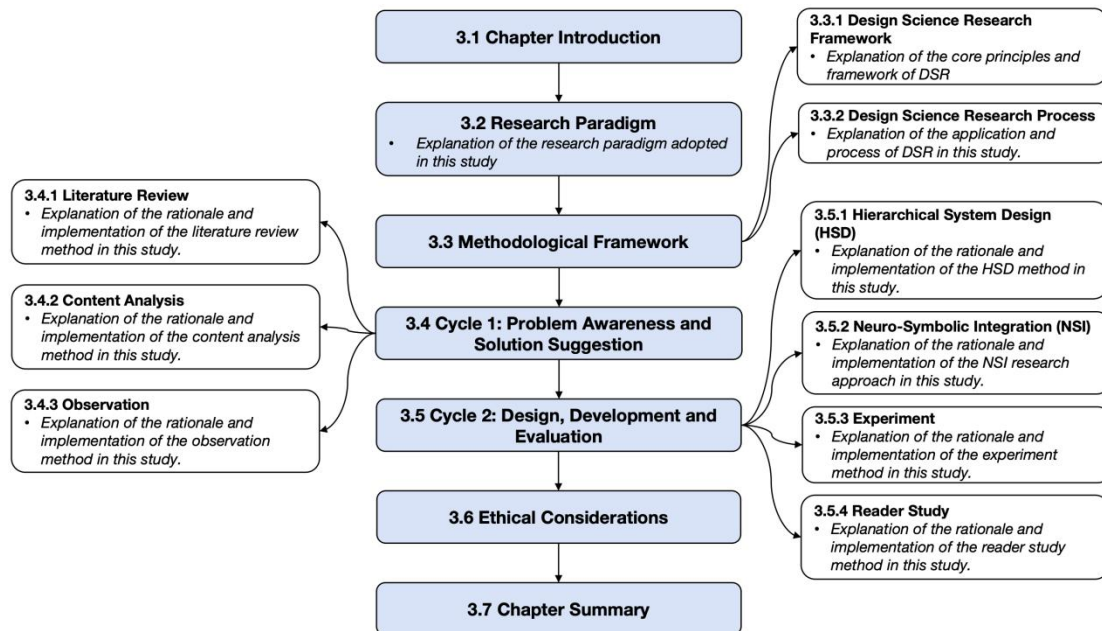


Figure 3- 1. Structure of Chapter 3.

As illustrated in Figure 3-1, the chapter begins by introducing the research paradigm that frames the study’s epistemological and ontological stance (Section 3.2), followed by an explanation of the DSR methodological framework (Section 3.3). DSR is adopted as the overarching framework through which the artefact is iteratively designed and evaluated across two research cycles. Cycle 1 (Section 3.4) focuses on problem awareness and solution suggestion, combining three complementary methods—literature review, content analysis, and observation—to identify the research gap, extract empirical insights from real-world OHC records, and derive system requirements. Cycle 2 (Section 3.5) concerns design, development, and evaluation, detailing the systematic application of Hierarchical System Design

(HSD) and NSI approaches, followed by experimental validation and a reader study to assess system performance from both computational and human-centred perspectives. Subsequent sections (Sections 3.6 and 3.7) address ethical considerations in data handling and participant involvement, and conclude the chapter with a summary that links the methodological design to the system implementation described in Chapter 5.

3.2 Research Paradigm

The selection of an appropriate research paradigm is a critical step in constructing the foundational framework of a study. A research paradigm provides the philosophical and methodological underpinnings that guide researchers in investigating specific research questions or problems (Guba & Lincoln, 1994). It shapes the researcher's worldview, influencing how questions are formulated, how data are collected, and how findings are interpreted (Okesina, 2020). According to Orlikowski and Baroudi (1991), a research paradigm is shaped by three foundational assumptions:

- **Ontology**, concerning the nature of reality (e.g., objective or subjective).
- **Epistemology**, determining what constitutes knowledge (e.g., empirical validation or contextual understanding).
- **Methodology**, defining how research is conducted (e.g., quantitative or qualitative approaches).

While numerous paradigms exist, such as pragmatism and the critical paradigm, positivism and interpretivism remain the two dominant paradigms most widely recognised and applied in IS research (Orlikowski & Baroudi, 1991). Each paradigm is suitable for addressing different types of research problems.

Positivism assumes that reality is objective, measurable, and independent of the researcher. It is grounded in the scientific method and aims to uncover universal laws through empirical observation and experimentation (Creswell, 2014). Positivist IS research therefore tends to focus on measurable, hypothesis-driven studies that identify patterns, test theories, and establish causal relationships (Davis, 1989; Venkatesh et al., 2003).

In contrast, interpretivism asserts that reality is socially constructed through human interpretation (Walsham, 1993; Myers, 1997). This paradigm emphasises understanding rather than prediction or hypothesis testing (Guba

& Lincoln, 1994), and typically employs qualitative methods such as case studies, interviews, and ethnographies. Interpretivist IS research often explores themes such as organisational change, user resistance, or implementation processes by documenting how various stakeholders perceive and interpret technology differently (Orlikowski, 1992; Walsham, 1995).

It has been consistently emphasised that no single research methodology is universally superior. The choice of methodology depends on the nature of the research problem and the researcher's objectives (Benbasat, 1984; Kaplan & Duchon, 1988; Pervan, 1994). Given the objectives of this study, namely, the design, implementation, and evaluation of a LLM-CHQA system, a positivist research paradigm is adopted as the primary philosophical foundation. This paradigm is appropriate because the research aims to develop and empirically evaluate a technological artefact designed to address identifiable informational and emotional support needs in consumer health questions. The evaluation process involves systematic experimentation and quantitative performance assessment, which aligns closely with positivist principles of objective measurement and empirical validation.

To operationalise this paradigm, the study adopts the Design Science Research (DSR) framework as the primary methodological approach. DSR is widely recognised in IS research for its focus on the creation and evaluation of innovative artefacts intended to solve real-world problems (Hevner et al., 2004; Peffers et al., 2007). Within the DSR framework, research progresses through iterative cycles of problem identification, artefact design, development, and rigorous evaluation.

By integrating a positivist paradigm with the DSR framework, this study ensures both methodological rigour and practical relevance. The positivist paradigm provides the philosophical basis for objective evaluation of the proposed system, while DSR offers a structured methodology for systematically designing and validating the CHQA artefact. Together, these approaches support the development of a robust, empirically grounded solution capable of effectively identifying and responding to both informational and emotional support needs in consumer health questions.

3.3 Methodological Framework

Building upon the philosophical foundations outlined in Section 3.2, this section introduces the DSR framework, which provides the methodological foundation for this study. DSR is particularly well-suited for research that aims not only to understand a phenomenon but also to design and evaluate artefacts that address identified problems in a systematic and scientifically rigorous manner. In the context of this study, DSR offers a structured pathway for bridging theoretical inquiry with practical implementation through iterative cycles of design, evaluation, and refinement.

3.3.1 Design Science Research Framework

DSR is a problem-solving research framework that seeks to create and evaluate artefacts designed to address identified real-world problems while advancing theoretical understanding (Hevner et al., 2004; Gregor & Hevner, 2013). Rooted in Herbert Simon's notion of the sciences of the artificial (Simon, 1969/1996), DSR views the design and evaluation of artefacts as legitimate scientific activities capable of generating prescriptive and explanatory knowledge. Rather than focusing solely on the description or interpretation of existing phenomena, DSR aims to build innovative solutions that are both effective in practice and grounded in theory (March & Smith, 1995).

March and Smith (1995) proposed that design science activities can be categorised into two complementary processes: build and evaluate. The artefacts produced through these activities can take four forms: (a) Constructs, which provide a conceptual vocabulary for defining the problem domain; (b) Models, which capture relationships among constructs; (c) Methods, which offer procedural guidelines or algorithms for problem-solving; and (d) Instantiations, which implement artefacts in a real or simulated environment.

Hevner et al. (2004) later expanded this into the three-cycle view of DSR, integrating the relevance cycle, design cycle, and rigour cycle, which together ensure a continuous interaction between theory and practice:

- **The relevance cycle** connects the research environment with the design process by translating practical problems and stakeholder requirements

into design objectives and evaluation criteria. It ensures that the artefact remains contextually meaningful and addresses real-world needs.

- **The design cycle** represents the iterative process of building, testing, and refining artefacts, where feedback from evaluation is continually incorporated to improve design quality and utility.
- **The rigour cycle** anchors the research in existing knowledge by drawing upon the theoretical foundations, models, and methods of the scientific knowledge base to inform and justify design decisions.

These three interrelated cycles collectively ensure that design science research achieves both practical relevance (through environmental grounding) and scientific rigour (through theoretical integration), allowing new knowledge to emerge from the interplay between problem context, artefact construction, and evaluation outcomes.

Besides, Hevner et al. (2004) have laid down seven widely acknowledged guidelines for conducting and evaluating effective DSR in the realm of information systems:

- **Design as an artefact.** Research must produce a viable artefact in the form of a construct, model, method, or instantiation.
- **Problem relevance.** The research must address an important and real problem within its environment, ensuring that the artefact provides a meaningful solution.
- **Design evaluation.** The utility, quality, and efficacy of the artefact must be rigorously demonstrated through well-established evaluation methods.
- **Research contributions.** The study must provide clear and verifiable contributions to the existing knowledge base, either through novel artefacts, improved methods, or generalisable design principles.
- **Research rigour.** The research must apply rigorous methods for both construction and evaluation, drawing upon the extant theoretical and empirical knowledge base.
- **Design as a search process.** The design process should be iterative and explorative, involving the search for an optimal solution through successive refinements.
- **Communication of research.** The results must be communicated effectively to both technical and managerial audiences, ensuring that findings are accessible, interpretable, and actionable.

Collectively, these seven guidelines provide a comprehensive framework for ensuring that DSR maintains scientific robustness while producing artefacts of genuine practical value. They are widely accepted as the benchmark for evaluating methodological soundness in IS-oriented design research and form the guiding principles for the present study.

In this study, the DSR framework serves as the overarching methodological foundation guiding the iterative design and evaluation of the system. By following these established principles and frameworks, the research ensures both the rigour of its scientific reasoning and the relevance of its practical outcomes. The next subsection elaborates on the specific DSR process adopted in this study, detailing how the framework was implemented across two research cycles.

3.3.2 Design Science Research Process

The Design Science Research Process (DSRP) model proposed by Vaishnavi and Kuechler (2007), which builds upon the General Design Theory (GDT) introduced by Takeda et al. (1990), is adopted in this study to guide the overall research process. Unlike prescriptive linear methodologies, the DSRP model explicitly supports iterative cycles of problem awareness, artefact design, evaluation and refinement. This cyclical nature is particularly appropriate for the complex socio-technical research problems to be solved in this study.

The DSRP model also facilitates continuous interaction between the theoretical knowledge base and the practical problem environment. Through this bidirectional knowledge exchange, the study is able to draw upon existing theories, including the NSI, the AFCE, and RST, while contributing new design principles and frameworks derived from the proposed system. Furthermore, the model's emphasis on systematic iteration and evaluation ensures that each design decision is both theoretically justified and empirically validated, thereby enhancing the overall rigour and relevance of the research.

The model consists of five iterative and interrelated stages that together ensure both rigour and relevance in design-oriented research:

- **Problem Awareness.** Identifying and analyzing research questions that emerge from new discoveries, practical application needs, or challenges within the field.
- **Solution Suggestion.** Proposing potential solutions or design ideas derived from existing knowledge or empirical insights.
- **Design and Development.** Refining and implementing tentative design solutions to meet the identified requirements.
- **Evaluation.** Assessing the artifact's performance using predefined criteria and analyzing its behaviour in the intended context.
- **Conclusion.** Reflecting on the research process to articulate new knowledge and formalize research conclusions.

Besides, as suggested by Takeda et al. (1990), the DSRP model incorporates three primary reasoning approaches, namely abduction, deduction and circumscription, which facilitate iterative knowledge generation throughout the research process (see Figure 3-2). These reasoning approaches enable researchers to move dynamically between theory and empirical observation. Abduction supports the formulation of tentative design propositions and hypotheses based on observed phenomena; deduction allows the derivation of logical consequences and design rules from existing theories; and circumscription provides the mechanism for refining and validating these propositions through iterative testing and reflection. Together, these approaches ensure a comprehensive understanding of the problem domain and the continuous refinement of artefacts within the design science cycle.

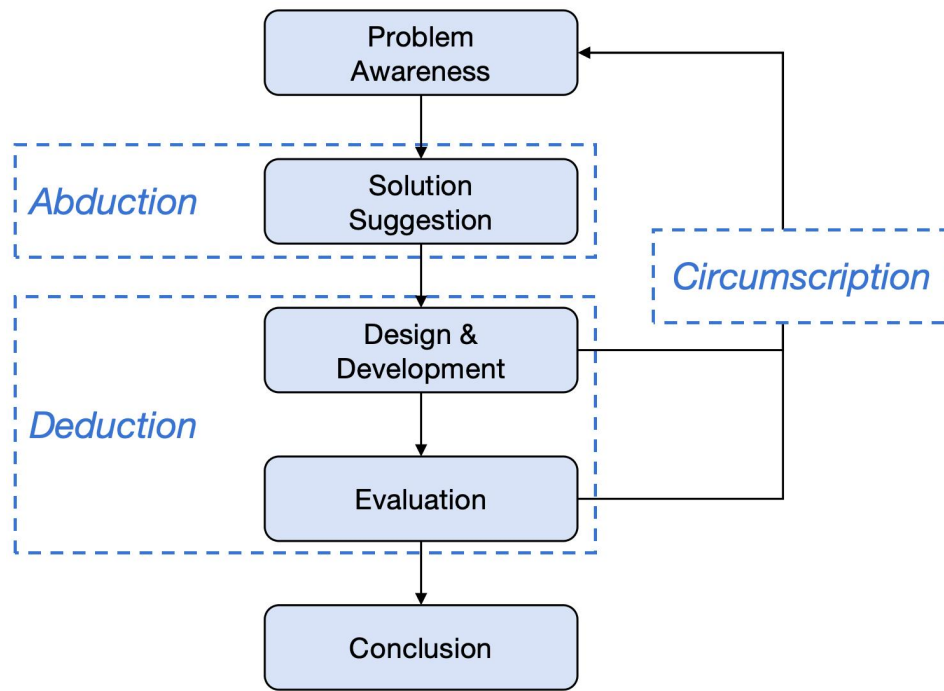


Figure 3-2. The Reasoning Approaches in DSRP. Adapted from Takeda et al. (1990)

An explanation of each reasoning approach and its implementation in the DSR process is presented below.

Abductive reasoning. Coined by Charles Sanders Peirce, abduction was first introduced as a cognitive process of “inference to the best explanation” (Peirce & Houser, 1998). It involves generating plausible hypotheses to explain surprising or anomalous observations (Cohen et al., 2011). In contrast to induction, which depends on repeated observation for validation, abduction produces tentative explanations that can later be tested. Within DSR, abductive reasoning is central to the suggestion stage, where potential design solutions are proposed based on observed problems. Hypotheses are evaluated for their simplicity, feasibility and relevance, guided by existing theories and empirical insights. This approach enables the creation of innovative artefacts while remaining grounded in established domain knowledge (Takeda et al., 1990).

Deductive reasoning. Deduction refers to a logical process in which conclusions necessarily follow from a set of premises (Sternberg, 2011). In social science research, it is commonly used to test theoretical propositions through empirical investigation (Cohen et al., 2011). In DSR, deduction operates primarily during the design, development and evaluation stages. Following the abductive formulation of hypotheses, deductive reasoning helps structure artefact design according to established theories and design

principles. The artefact is then implemented and evaluated against these theoretical expectations. However, unlike traditional natural or behavioural science approaches, conclusions in DSR are often provisional and context-dependent (Vaishnavi et al., 2007). Deductive reasoning thus ensures that artefacts are systematically tested for theoretical consistency while allowing for necessary contextual adaptation.

Circumscription. Circumscription complements abduction and deduction by enabling iterative refinement and learning within the DSR process. The initial artefact rarely meets all contextual requirements, and evaluation often reveals inconsistencies or gaps in the theoretical assumptions. These are not treated as failures but as opportunities for learning and theory development (Vaishnavi et al., 2004). As illustrated in Figure 3-2, circumscription occurs when empirical observations challenge existing knowledge, prompting iterative refinement of both artefacts and theoretical understanding (Takeda et al., 1990). Introduced by McCarthy (1980) as a “rule of conjecture,” circumscription limits the range of possible explanations within theoretical boundaries, helping identify discrepancies between theory and practice. In DSR, this reflective mechanism allows researchers to “learn or discover when things do not work according to theory” (Vaishnavi & Kuechler, 2004). Each cycle of circumscription feeds insights back into the abductive and deductive processes, leading to new hypotheses, refined artefacts and, ultimately, strengthened theoretical contributions.

Generated from the preceding discussion, the implementation of the DSRP model in this research results in the research process shown in Figure 3-3. This diagram illustrates the research activities related to the DSR processes, along with the intended research outcomes and knowledge contribution.

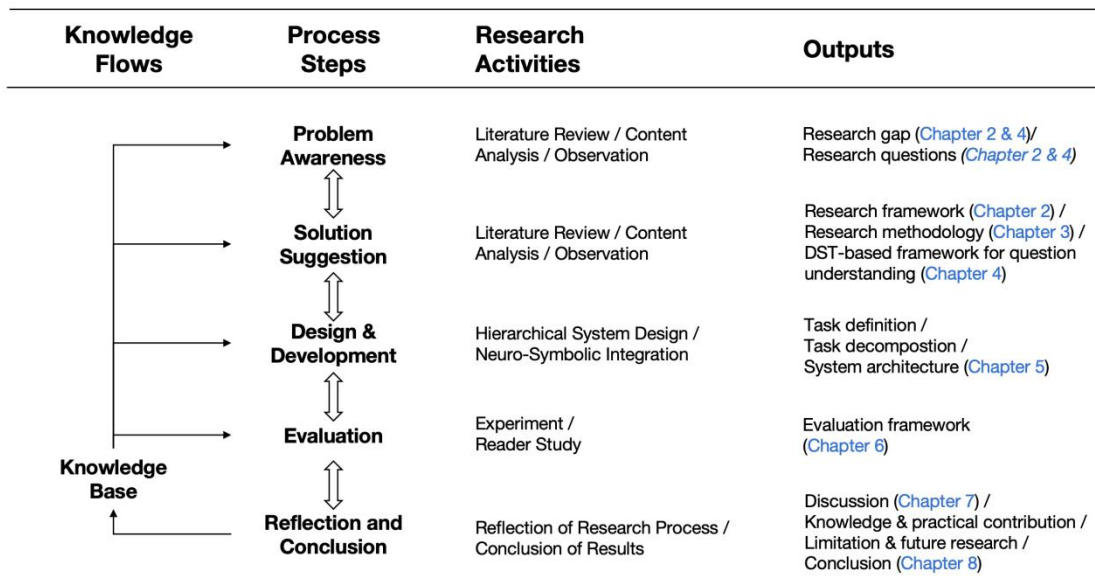


Figure 3-3. Research Process with Knowledge Flows, Activities and Intended Outputs.

3.4 Cycle 1: Problem Awareness and Solution Suggestion

3.4.1 Literature Review

A literature review is a fundamental research method for establishing the theoretical, empirical, and contextual foundations of a study. As Webster and Watson (2002) argue, a well-structured literature review constitutes the “building block of high-quality research” by synthesising prior knowledge, identifying gaps, and defining the conceptual boundaries of inquiry. Similarly, vom Brocke et al. (2015) emphasise that the literature review is not merely descriptive but analytical and generative—it develops an informed understanding of what is known, what is missing, and what new knowledge is required. In DSR, the literature review thus serves as a knowledge-building mechanism that informs both the problem identification and the solution design phases (Hevner et al., 2004; Gregor & Hevner, 2013).

Accordingly, in this study the literature review was employed as a systematic analytical method to support the Problem Identification and Situation Awareness stages. In the problem awareness phase, it enabled a structured diagnosis of limitations in existing CHQA systems and a synthesis of theoretical insights from related domains, including consumer health communication and LLM research. Through iterative review and coding, the literature revealed that current CHQA systems predominantly address

informational accuracy while neglecting emotional support needs, and that evaluation practices remain narrowly focused on single-intent tasks.

In the solution suggestion phase, the literature review provided the conceptual foundation for designing a new research artefact by identifying candidate theories, design principles, and computational paradigms capable of bridging the informational–emotional divide. Specifically, the theoretical integration of AFCE, and RST was informed by this analytical synthesis, forming the backbone of the system architecture, evaluation framework, and dataset design outlined in subsequent sections.

Through this iterative literature-driven process, the study moved from problem identification to solution articulation, whereby suitable theoretical constructs and computational strategies were selected to inform artefact design. Hence, the literature review functioned as both a diagnostic and a generative method, transforming theoretical synthesis into actionable design knowledge.

3.4.2 Content Analysis

Content analysis was employed in this study to complement the literature-based diagnosis by empirically examining how informational and emotional support needs co-occur in real CHQs. As Krippendorff (2018) defines, content analysis is a systematic, replicable method for making valid inferences from textual data to their context of use. Within the DSR process, it serves as a bridge between problem identification and artefact design (Hevner et al., 2004), providing grounded evidence of user behaviours, unmet needs, and contextual factors that inform requirement specification and theoretical refinement.

Data Source. The CHQs used in this study were collected from a widely recognized Chinese OHC platform, Haodaifu Online (www.haodf.com), for the following reasons: (a) Widely recognized platform. Haodaifu Online, founded in 2006, is one of China's earliest internet healthcare platforms. With 18 years of stable operation, it has accumulated a vast number of real patient consultation cases. As of July 2023, the platform has served over 84 million patients. (b) Large and credible database. As of September 2024, aggregated over 940,000 certified physicians from more than 10,000 accredited hospitals, and has served over 89 million patients across a diverse range of medical specialties (BCC Global, 2024). Of these, 280,000 doctors are verified with

real names and provide Direct online medical services to patients. Among the active doctors, 73% are from top-tier hospitals, offering a high level of authoritative medical services. (c) Comprehensive data. Compared to other well-known Chinese internet medical platforms (e.g., Chunyu Doctor, Ping An Good Doctor), Haodaifu Online provides richer consultation data. This includes, but is not limited to: detailed descriptions of patients' conditions, height, light, disease name, duration, hospitals and departments previously visited, allergy history, past medical history, desired help from the consultation, and doctors' case summaries, preliminary diagnoses, and management suggestions. (d) Wide range of specialties. Haodaifu Online offers consultation data covering 27 medical departments, nearly encompassing all disease types. This makes it an ideal source to assess the LLM's performance in generating diagnostic suggestions for various types of cases.

Data Collection. To ensure a balanced, representative, and ethically sound dataset, two complementary data sources were integrated. First, an existing open source clinical OHC datasets crawled from haodf.com, MedDialog-CN (He et al., 2020), is included. The consultations cover 29 broad categories of specialties including internal medicine, pediatrics, dentistry, etc. and 172 fine-grained specialties including cardiology, neurology, gastroenterology, urology, etc. The consultations are conducted from 2010 to 2020. Each consultation consists of three parts: (a) description of patient' s medical condition and history; (b) conversation between patient and doctor; (c) diagnosis and treatment suggestions given by the doctor. In the description of patient' s medical condition and history, the following fields are included: present disease, detailed description of present disease, what help is needed from the doctor, how long the disease has been, medications, allergies, and past disease.

The MedDialog-CN dataset is a highly suitable choice for this research due to its relevance, accuracy, and accessible, which are main metrics in data source selection assessment (Wang & Strong, 1996; Malhotra, 2010; Bryman, 2016):

- *Data Relevance.* The MedDialog-CN dataset directly aligns with this research focus on doctor-patient communication in OHC systems. Its 1.1 million dialogues and 4 million utterances, extracted from the Chinese platform Haodf.com, provide rich, domain-specific data for analyzing empathic communication and identifying challenges in doctor-patient

interactions. The data structure, consisting of multi-turn conversations, is particularly relevant for training and refining LLMs designed for conversational and empathic tasks. MedDialog-CN is continuously updated, ensuring data relevance to contemporary communication trends.

- *Data Quality.* The dataset contains real-world dialogues, originate from a professional and widely-used OHC platform, ensuring credibility and authenticity in content. The dataset covers multiple medical departments (27 categories), representing a broad spectrum of scenarios and communication styles. Its large volume minimizes noise through statistical averaging, allowing reliable insights even with variations in individual records. Furthermore, the dialogues originate from a professional and widely-used OHC platform, ensuring credibility. Table 3-1 shows a comparison of MedDialog-CN dataset with several other Chinese online medical dialogue datasets. The number of dialogs and diseases in this dataset are both much larger than those in other datasets.
- *Data Accessible.* MedDialog-CN is publicly available for research purposes, and its structured format simplifies integration into LLM pipelines. The licensing terms are clear, ensuring legal and ethical compliance.

Table 3-1. Comparison of MedDialog-CN Datasets with Other Chinese OHC Datasets.

Dataset	Data Sources	Data Size	Departments
MedDialog-CN (He et al., 2020)	haodf.com	1.1 million dialogues	27
Chinese Medical Dialogue Data (Toyhom, 2019)	not mention	792,099 dialogues	6
WebMedQA (He et al., 2019)	Multiple OHC websites, such as Baidu Doctor, 120Ask	63,284 dialogues	23

However, this dataset only covers OHC records from 2010 to 2020. To enhance the timeliness of this dataset, this study further augment the dataset with a new dataset of recent OHC records collected from haodf.com platform. A Python-based web crawler is used to randomly collect 5400 cases from 27 departments on haodf.com. Align the format of MedDialog-CN, for each case, researcher collected: (a) case information (including disease description, height, light, disease name, duration, hospitals and departments previously visited, allergy history, past medical history, and the help sought), (b) consultation recommendations (including preliminary diagnoses and management suggestions), and (c) dialogue between physician and patient.

This inclusion served to verify that observed intent patterns were not platform-specific but generalisable across Chinese-language OHC contexts. The integration of these two sources ensured both data depth (through authentic and detailed consultation narratives) and breadth (through domain diversification), thereby strengthening the internal validity and external generalisability of the analysis.

Data Sampling. A total of 5,600 consumer health questions were randomly sampled from public consultation threads across 14 medical departments. Each question represents a unique patient query seeking advice on symptoms, diagnosis, treatment, or psychosocial support. Personally identifiable information was removed, and data usage followed ethical guidelines for secondary textual analysis. This multi-source, multi-specialty dataset provides a robust empirical foundation for understanding how medical informational and emotional needs co-occur in real-world CHQs. Its size, authenticity, and heterogeneity make it particularly effective for supporting the joint intent modelling, evaluation framework design, and dataset construction phases of this research.

Analytic Framework. Grounded in the Rhetorical Structure Theory (RST) and the Appraisal Framework for Clinical Empathy (AFCE) introduced in Chapter 2, the analytical framework for this study was designed to capture how informational and emotional support needs are hierarchically structured and contextually expressed in CHQs. These theories jointly guided the development of the coding schema and ensured that the analysis moved beyond surface-level lexical patterns toward a deeper understanding of discourse organisation, problem structuring, and affective stance representation.

- RST provided the foundation for modelling the discourse-level relations between EDUs, enabling the identification of rhetorical linkages, such as Elaboration, Evidence, Condition, and Justification, that connect informational and emotional segments within each CHQ. This helped to operationalise how users linguistically structure complex, multi-intent questions.
- AFCE offered the conceptual lens for identifying and categorising emotional expressions within each CHQ. It enabled systematic annotation of Affect (emotions), Judgement (moral or capability evaluations), and Appreciation (evaluations of experiences or outcomes),

thereby grounding the emotional dimension of the analysis in a well-validated linguistic theory.

Coding Procedure. The coding strategy adopted a multi-layer, hybrid inductive-deductive approach. Deductively, the initial coding frame was developed based on the three theoretical pillars above and refined through iterative team discussions. Inductively, emergent patterns and new subcategories were added to capture nuances specific to Chinese OHC discourse. Each CHQ was segmented into EDUs following the RST segmentation guidelines (Carlson & Marcu, 2001), then assigned to one or more problem units (e.g., symptom, diagnosis, treatment, emotional state). Within each problem unit, AFCE-based categories were used to code affective stance and intensity.

This layered analytical strategy served three core purposes within the DSR process:

- Problem comprehension. To empirically characterise the intertwined informational and emotional dimensions of CHQs, providing a grounded basis for artefact requirements.
- Design input. To translate observed rhetorical and emotional patterns into design specifications for the system’s joint intent-recognition module and evaluation framework.
- Dataset construction. To inform the annotation schema and data-label hierarchy used in building the multi-intent benchmark dataset.

Overall, the analytical framework operationalised the integration of clinical reasoning, discourse coherence (via RST), and empathetic stance detection (via AFCE). This ensured that the subsequent system design was not only technically informed but also theoretically grounded in models of medical reasoning, discourse structure, and human empathy. key elements in developing a human-centred, LLM-based CHQA system.

3.4.3 Observation

Observation was adopted as a complementary qualitative evaluation method to contextualise and enrich the findings obtained from the preceding literature review and content analysis. While textual analysis elucidates what kinds of informational and emotional needs exist in CHQs, observation enables the researcher to understand how these needs are expressed,

interpreted, and operationalised within real-world systems. Within the DSR paradigm, observational inquiry contributes to the problem awareness phase by grounding theoretical insights in empirical reality (Hevner et al., 2004). This approach provides evidence of actual workflow patterns, stakeholder interactions and system limitations that cannot be fully revealed through secondary data alone.

Compared with interviews or surveys that rely on self-reported behaviour, direct observation captures the authentic sequence of actions and situational context as they occur, thereby reducing retrospection bias. It enables the identification of mismatches between intended design and actual user behaviours on digital platforms. Observation also helps uncover implicit task dependencies. For example, how physicians interpret patient narratives or how system interfaces prioritise informational over emotional cues.

In relation to the study's overall framework, observation provides empirical grounding for the NSI perspective by revealing how human reasoning, discourse structure, and affective exchange unfold in natural CHQA environments. These insights then inform subsequent system modelling and dataset construction by clarifying how symbolic structures (e.g., rhetorical relations, empathy cues) manifest in practice.

In IS research, observation is defined as a systematic, purposeful, and contextual examination of human–technology interactions (Yin, 2018; Miles et al., 2014). The goal is not merely to describe behaviour but to interpret patterns of meaning—how actors sequence, justify, and coordinate their actions within socio-technical systems. This approach allows triangulation between qualitative insight and system-design inference, aligning with the problem identification and requirement formulation stages of DSR.

The Observation method was applied to publicly available patient–physician consultation threads on a Chinese OHC platform (Haodaifu Online). A purposive sampling strategy selected threads spanning multiple medical specialties, each representing a complete interaction from user question to clinician response. The researcher observed how user queries are structured, how physicians interpret them, and how clinical reasoning evolves. Through a reverse-engineering technique, the researcher further reconstructed and visualised the workflow of the OHC platform based on its observable interaction records and interface structures. This artefact-based observation allowed for the identification of implicit process logics embedded in the

consultation flow, complementing behavioural observations with a structural perspective.

3.5 Cycle 2: Design, Development and Evaluation

3.5.1 Hierarchical System Design (HSD)

The Hierarchical System Design (HSD) approach was adopted to manage the inherent complexity of the proposed system, whose objective is to interpret and respond to multi-intent consumer health questions that encompass both medical and emotional dimensions. HSD provides a structured way to decompose a cognitively rich process—spanning perception, interpretation and reasoning—into manageable layers while maintaining coherence across the overall architecture. This hierarchical abstraction aligns with the research goal of designing an end-to-end system that reflects the staged nature of human clinical communication, in which the system first understands the user’s needs and then generates an informed and empathetic response.

In contrast to flat or strictly pipeline-based designs, HSD supports functional specialisation within layers while maintaining vertical information flow across them. This layered configuration mitigates the fragmentation often seen in modular CHQA frameworks, enabling integrated reasoning without compromising interpretability. The hierarchical structure also facilitates the definition of learning objectives at different levels, such as local intent recognition and global discourse inference, thereby enhancing both controllability and transparency of model behaviour.

Within the overarching DSR paradigm, HSD serves as the system design strategy that operationalises the artefact-building process. It provides the architectural logic through which the theoretical NSI framework (see Section 3.5.2) is instantiated. In this study, the upper layers of reasoning correspond to symbolic abstraction, whereas the lower layers represent neural representation learning. In this way, HSD functions as a structural bridge between DSR’s design cycles and NSI’s cognitive unification principle, ensuring that the system’s design remains both theoretically grounded and empirically verifiable.

HSD is rooted in classical systems engineering and cognitive architecture research (Simon, 1962; Newell & Simon, 1972; Miller et al., 1960), which propose that complex intelligent behaviour emerges from layered interactions

among specialised subsystems. Modern AI design extends this principle to hierarchical modelling, where each layer encapsulates a specific granularity of representation—from low-level data processing to high-level reasoning and decision-making. Key theoretical tenets underpinning HSD include:

- **Decomposition and abstraction:** breaking a global task into interrelated sub-tasks arranged in logical or cognitive order.
- **Layered integration:** ensuring that outputs of lower layers provide semantically enriched inputs to higher ones, supporting emergent reasoning.
- **Top-down coherence and bottom-up learning:** upper-layer constraints guide interpretation, while lower-layer learning adapts from data.

These principles are consistent with hierarchical cognitive processing observed in natural language understanding and are therefore well suited for CHQA tasks that require both factual precision and affective sensitivity.

Within this study, task decomposition was employed as a core design technique under the HSD framework to translate the meta-requirements into implementable and interdependent sub-tasks. This process enabled the systematic structuring of the CHQA problem into two functional tiers: Question Understanding and Answer Generation. Each tier was further articulated into cognitive sub-tasks, such as Medical Needs Recognition, Emotional Needs Recognition. Although these sub-tasks are conceptually distinct, they were operationalised within a single end-to-end LLM rather than as isolated modules. The hierarchy therefore functions at both conceptual and representational levels, where each layer's objectives are embedded through fine-tuned datasets, structured prompts and evaluation criteria corresponding to its cognitive role.

The hierarchical mapping guided the entire implementation process, including data annotation (medical and emotional labels with contextual spans), model training (stage-wise supervised fine-tuning) and evaluation (both component-level and holistic performance metrics). This design ensured methodological consistency, interpretability and reproducibility, thereby fulfilling the DSR requirements of rigour and relevance.

3.5.2 Neural-Symbolic Integration (NSI)

This study adopts the NSI paradigm to reconcile the complementary strengths of LLM-based contextual learning and symbolic reasoning grounded in structured knowledge. Within the CHQA domain, neural components provide data-driven adaptability and contextual inference, while symbolic structures ensure interpretability, logical coherence, and safety in responses. NSI thus enables a unified cognitive mechanism capable of jointly supporting informational accuracy and emotional empathy, two dimensions that are often interdependent yet traditionally modelled in isolation.

By coupling neural adaptability with symbolic constraints, NSI bridges the epistemic gap between statistical learning and logical inference. This dual capability is particularly critical for CHQA, where users' informational needs and emotional states co-evolve within the same communicative context. In such environments, purely neural systems risk factual inconsistency or affective insensitivity, whereas purely symbolic systems lack generalisation. NSI provides a balanced paradigm that sustains both expressive flexibility and reasoning transparency.

The NSI approach originates from the long-standing effort to integrate connectionist and symbolic paradigms (Bader & Hitzler, 2005). It aims to combine the robustness and generalisation of neural learning with the structure and reasoning capabilities of symbolic logic. As reviewed in Section 2.6.3, Neural-Symbolic systems can be organised into two canonical paradigms (Bader & Hitzler, 2005; Wang et al., 2025): hybrid and unified architectures.

This study deliberately adopts the unified architecture, wherein symbolic knowledge of discourse parsing in consumer health question are internalised within the neural model via fine-tuning, rather than coupled as an external reasoning module. This architectural decision eliminates representational discontinuities between symbolic and neural layers, reduces system interface complexity, and fosters a continuous learning–reasoning process. Such integration aligns with contemporary NSI principles that emphasise end-to-end differentiability, semantic continuity, and mutual interpretability between data-driven and rule-based knowledge systems.

Within the unified paradigm, three design subtypes have been identified (see Section 2.6.3): Neuro \rightarrow Symbolic \rightarrow Neuro (Type 4), Symbolic \rightarrow Neuro

(Type 5), and Neuro[Symbolic] (Type 6). The present study adopts the Neuro: Symbolic \rightarrow Neuro (Type 4) configuration, which operationalises symbolic reasoning as an internalised intermediate representation within the neural flow (see Figure 3-4). This choice contrasts with prior CHQA research that typically unified medical knowledge graphs or clinical ontologies with LLMs, focusing solely on the informational dimension during the answer generation stage. In contrast, the present system unifies an RST-based Discourse Structure Tree (DST) that encodes both intents, contexts and rhetorical relations during the QU stage.

The rationale for this design is threefold. (a) First, the CHQs targeted in this study are structurally and semantically more complex than those addressed in previous research: they contain dual-dimensional intents (medical informational and emotional support needs) and are often long-form and multi-clausal. Consequently, strengthening symbolic reasoning at the Question Understanding (QU) stage is essential rather than relying solely on the LLM's implicit parsing. (b) Second, compiling logical knowledge at the QU stage is more cost-effective, stable, and reliable than performing logic integration at the answer-generation stage. The logical relations governing question understanding tend to be more timeless and domain-invariant, whereas those governing answer generation are context-sensitive and require frequent updates to remain clinically valid. (c) Third, the QU-level integration ensures a lower cumulative error rate and enhances interpretability, as the reasoning process remains grounded in explicit rhetorical and affective structures before response generation.

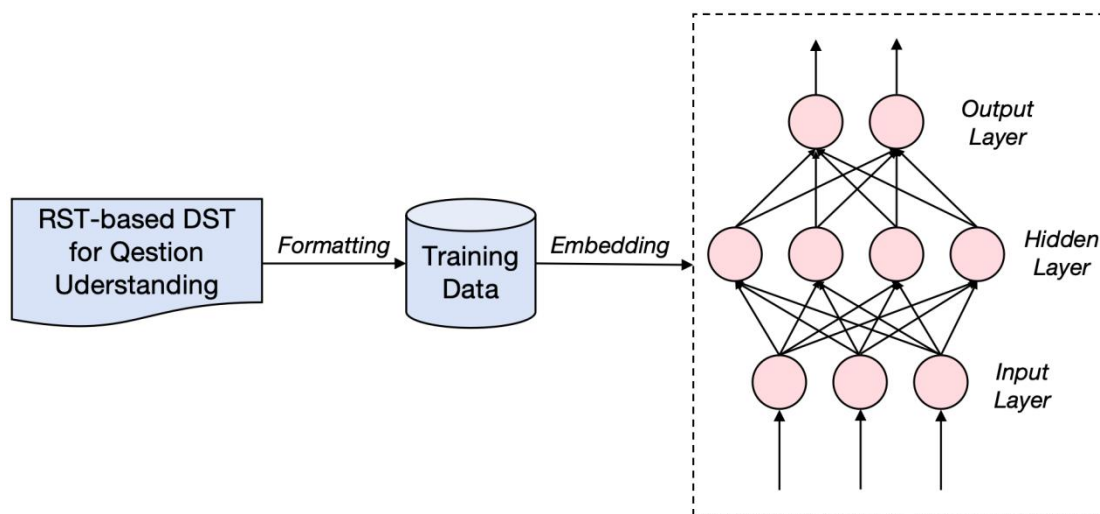


Figure 3-4. Neuro: Symbolic \rightarrow Neuro Pathway for System Architecture Design in this Study.

This design choice is grounded in two key considerations: (a) First, the CHQs targeted in this study are more complex than those addressed in previous work, containing both medical informational and emotional support dimensions, often expressed in long-form, multi-intent discourse. Such complexity necessitates a strengthened logical compilation at the question-understanding stage, enabling the LLM to interpret symbolic discourse structures autonomously. (b) Second, compiling logical knowledge at the Question Understanding (QU) stage is more cost-effective, stable, and efficient than compiling it at the Answer Generation (AG) stage. Unlike AG stage logic, which requires frequent updating and incurs higher error rates, the logical and affective knowledge encoded at the QU stage remains semantically stable and reusable across queries. Hence, this approach provides a more efficient, low-error, and sustainable pathway for integrating symbolic reasoning into neural architectures, making it a promising direction for further exploration in health-domain NSI research.

The proposed system implements NSI through three tightly coupled stages, knowledge representation, knowledge embedding, and functional integration, which collectively embody a learning–reasoning continuum.

Knowledge Representation. A RST-based DST is employed to formalise the logical and affective organisation of consumer health questions. Each question is segmented into EDUs connected through predefined rhetorical relations (e.g., Elaboration, Background, Contrast, Condition). This representational design integrates the hierarchical readability of symbolic logic with the abstraction power of first-order expressions, allowing the neural model to internalise structured discourse while maintaining explicit interpretability and semantic traceability, which are key properties for ensuring safety and explainability in health-domain AI systems.

Knowledge Embedding. Symbolic knowledge is embedded primarily at the data level via Supervised Fine-Tuning (SFT). The training corpus is augmented with explicit rhetorical and affective labels, enabling the LLM to learn symbolic dependencies through exposure rather than through architectural modification. During inference, lightweight neural-inference constraints, including structured decoding and discourse consistency prompts, preserve rhetorical and emotional coherence. This “data-centric embedding with inference refinement” strategy ensures high compatibility with mainstream LLMs, scalability across domains, and minimal engineering overhead.

Functionality. Functionally, the NSI-based CHQA system establishes a two-stage reasoning continuum that unifies learning and reasoning within a single neural framework. In the first stage-Implicit QU (Learning)-the model internalises symbolic and affective regularities from annotated discourse data, forming latent representations that support the comprehension of multi-intent consumer health queries. In the second stage-Explicit AG (Reasoning)-these internalised structures are leveraged to produce responses that are both factually reliable and rhetorically coherent, while simultaneously demonstrating emotional attunement to users' needs. By integrating both processes within a unified architecture, the system achieves reasoning-through-representation, a defining feature of modern NSI designs. This approach preserves the adaptability and generalisation capacity of neural models while enforcing the logical consistency and interpretability characteristic of symbolic reasoning, thereby bridging the long-standing divide between statistical pattern recognition and structured logical inference in health-domain AI systems.

3.5.3 Experiment

This study adopts the experiment method as the primary evaluation method under the DSR paradigm. The research aims to validate whether the proposed Neural-Symbolic CHQA architecture, which integrates neural contextual learning with symbolic reasoning, can effectively support both informational accuracy and emotional empathy in CHQA. Given that this study develops a conceptual artefact rather than a deployable clinical system, experimentation provides a controlled, replicable, and ethically sound environment for testing its functional and theoretical validity.

Compared with other evaluation methods such as field studies or case analyses, experimentation offers causal transparency and internal control. It allows the researcher to isolate the contribution of each architectural component (e.g., neural, symbolic, integrated) and examine how their interaction influences reasoning coherence and empathy expression. Simulation-based experimentation also mitigates the ethical and practical risks associated with clinical trials, such as misinformation, patient privacy breaches, or unverified medical outputs.

Within the DSR cycle, experimentation represents the ex post evaluation phase that tests whether the artefact achieves its design objectives (Hevner

et al., 2004). In this study, it operationalises the NSI paradigm (Section 3.5.2) into a testable prototype. The experiment thereby acts as the empirical counterpart of the conceptual design process, linking theoretical construction to verifiable system performance.

The experimental method has long served as a core mechanism for evaluating artefacts in information systems design and AI-based system development. According to March and Smith (1995), experiments in DSR constitute artificial evaluation environments, in which artefacts are tested under simulated yet realistic conditions to assess efficacy, utility, and consistency.

In AI system research, particularly in NSI and Human-AI Interaction, experiments are not purely hypothesis-driven but design-driven. Their purpose is to verify whether the system exhibits the intended reasoning behaviour and architectural logic under predefined conditions. This study follows this principle by treating experimentation as a design validation mechanism, examining whether the NSI-based CHQA system achieves reasoning-through-representation—the hallmark of integrated NSI architectures (Bader & Hitzler, 2005).

The experiment was operationalised through a two-stage simulation pipeline, ensuring both architectural and system-level validation of the proposed framework.

Stage 1: Architecture-Level Simulation. This stage examines the internal validity of the Neural-Symbolic architecture itself. Three types of configurations were implemented for comparison: (a) Baseline LLM (purely neural, without symbolic grounding); (b) Hybrid NSI model (neural and symbolic modules coupled via prompting interfaces); (c) Integrated NSI model (symbolic discourse and affective structures embedded through fine-tuning).

Each configuration was evaluated on the CHQA-MedEmo dataset using a sample of questions that varied in clinical intent and emotional tone. The analysis compared the three architectures along four dimensions—clinical accuracy, personalisation, empathy, and latency—using human-coded metrics. Rather than merely checking whether symbolic knowledge can be embedded, this stage assesses the relative benefits of different neural-symbolic designs, examining how much performance improves when moving

from a purely neural baseline to a hybrid NSI model, and further to an integrated NSI architecture in which symbolic discourse and affective structures are more tightly coupled to the LLM.

Stage 2: System-Level Validation. Building upon the validated architecture, the experiment extended to the Joint Medical-Emotional Question Answering (JMEQA) framework, which decomposes CHQA into six subtasks: (a) Medical Need Recognition (MNR); (b) Emotional Need Recognition (ENR); (c) Medical Need Context Extraction (MNCE); (d) Emotional Need Context Extraction (ENCE); (e) Medical Need Answering (MNA); (f) Emotional Need Answering (ENA).

For question understanding subtasks (MNR, ENR, MNCE, ENCE), span-level and relation-level F1 scores were used to assess precision and recall. For answer generation subtasks (MNA, ENA), human evaluations were conducted to assess factual reliability, empathy, and contextual appropriateness. Finally, extrinsic evaluation measured the overall ability of the system to deliver patient-centred medical support and context-specific emotional support.

This two-stage implementation ensures that: (a) the NSI mechanism is empirically validated as a functional architecture; and (b) the JMEQA decomposition is proven both theoretically motivated and empirically justified.

3.5.4 Reader Study

To ensure a rigorous, reproducible, and human-centred evaluation of the system, this study employed a reader study-based evaluation framework. This method was inspired by large-scale assessments of LLM-generated outputs in clinical and biomedical domains (Singhal et al., 2025; Thirunavukarasu et al., 2023). In the context of CHQA, evaluating system responses involves not only factual correctness but also empathy, appropriateness, and contextual relevance—dimensions that cannot be fully captured by automatic metrics alone. Therefore, a reader study design, where qualified human experts independently assess model outputs, was selected to evaluate the clinical and emotional adequacy of system responses in a structured and statistically reliable way.

Compared with automatic scoring or user-based surveys, the reader study method offers several advantages: (a) Clinical validity, ensuring expert-level judgment in assessing nuanced medical and affective content; (b) Controlled

reliability, allowing for blinded and independent assessment under predefined metrics, reducing bias; (c) Statistical robustness, supporting quantitative reliability analysis (e.g., inter-rater agreement) and feedback loops for calibration. These properties make it particularly suited for verifying the human-perceived validity of LLM-based systems that operate in sensitive healthcare contexts.

Within the DSR cycle, the reader study complements the experimental evaluation (Section 3.5.3) by providing human-centred validation. It bridges the gap between artefact performance (what the system does) and perceived quality (how experts interpret its outputs). Furthermore, the design aligns with the NSI paradigm by ensuring that the system’s symbolic reasoning and empathetic responses are verified through expert interpretation—a necessary step for demonstrating interpretability and safety.

In clinical AI research, a reader study refers to a controlled evaluation in which multiple qualified human readers independently review and rate system-generated outputs (e.g., diagnostic findings, clinical summaries, or empathetic messages) according to predefined metrics. Such designs have been widely used for validating radiology and pathology models, and more recently, for assessing LLM generations in healthcare communication tasks (Singhal et al., 2025).

A reader study typically incorporates three key theoretical principles: (a) Blinded evaluation. Readers assess model outputs without knowing which system produced them, ensuring unbiased comparison. (b) Standardised criteria. All assessments follow consistent scoring guidelines derived from predefined evaluation dimensions. (c) Reliability measurement. Inter-rater agreement quantifies the consistency of human judgments and the objectivity of the framework. These principles form the foundation for the Reader Study Evaluation Framework used in this study.

The reader study was implemented as a three-stage human-evaluation pipeline, integrating quantitative metrics and expert annotation to ensure methodological transparency and reproducibility.

Stage 1: Evaluation Criteria and Metrics Definition. Evaluation dimensions and scoring rubrics were first defined based on the tri-dimensional evaluation framework introduced in Section 2.5.3—Accuracy, personalisation, and Empathy. Each dimension was further operationalised into multi-level criteria.

The finalized criteria and rating scales were compiled into training materials used in reader orientation sessions.

Stage 2: Reader Study Protocol. Qualified readers with backgrounds in clinical medicine and health communication were recruited and trained using a benchmark dataset of representative CHQs and sample LLM outputs. Training included calibration exercises and inter-annotator discussions to align interpretation standards. During the blinded assessment phase, readers independently rated anonymized system outputs according to the predefined criteria. Assessments were performed via a secure evaluation interface that automatically logged ratings and timestamps.

Stage 3: Reliability Analysis. Upon completion of the blinded assessments, inter-rater agreement was computed to quantify reliability. Statistical measures such as Cohen's κ and Krippendorff's α were applied to evaluate both pairwise and overall consistency among readers. Discrepancies were analysed to identify ambiguous items and refine the evaluation rubric. The feedback loop from this stage informed updates to both training materials and metric definitions in Stage 1, ensuring continual calibration and methodological consistency.

3.6 Ethical Consideration

This study acknowledges the potential ethical implications associated with its research activities and has taken multiple measures to address them. Firstly, the data utilized in this work consist of publicly accessible online health consultation records that have been de-identified prior to collection, with all personally identifiable information removed to protect user privacy. Recognizing the risk of bias inherent in AI systems, particularly the possibility of perpetuating or amplifying healthcare disparities, the researcher made deliberate efforts to enhance the diversity and representativeness of the medical record dataset used in this study. Additionally, rigorous ethical protocols were consistently followed throughout all stages of the research process. Ethical approval for this study (Ethical Application: HREC (Health) 2025#13) was obtained from the University of Waikato Human Research Ethics Committee, ensuring that the research fully complies with internationally recognized ethical standards and guidelines.

3.7 Chapter Summary

Chapter 3 outlined the methodological foundations of this research, demonstrating how a rigorous DSR framework guided the iterative design, development, and evaluation of the proposed system. The chapter's key methodological contribution lies in establishing a multi-method design-evaluation process that integrates quantitative rigour with qualitative interpretability, thus ensuring both internal validity and external relevance within the DSR paradigm. Ethical considerations in data management and participant involvement were also addressed to ensure compliance with research integrity standards and the ethical expectations of human-centred AI research.

By the end of this chapter, the research design has been fully specified at the philosophical, methodological, and procedural levels, forming a robust foundation for artefact implementation. The methodological blueprint developed here directly informs Chapter 5, where the Med-Emo CHQA system is instantiated through architectural design, technical integration, and system prototyping, followed by empirical evaluation in Chapter 6.

Chapter 4 Situation Awareness

4.1 Chapter Introduction

Building upon the methodological foundation established in Chapter 3, this chapter develops situation awareness for the research by empirically analysing real-world OHC. In the context of Cycle 1 of the DSR framework, this chapter serves as the problem contextualisation and solution grounding stage, where empirical insights from the OHC environment are synthesised into design implications for the subsequent system development in Chapter 5.

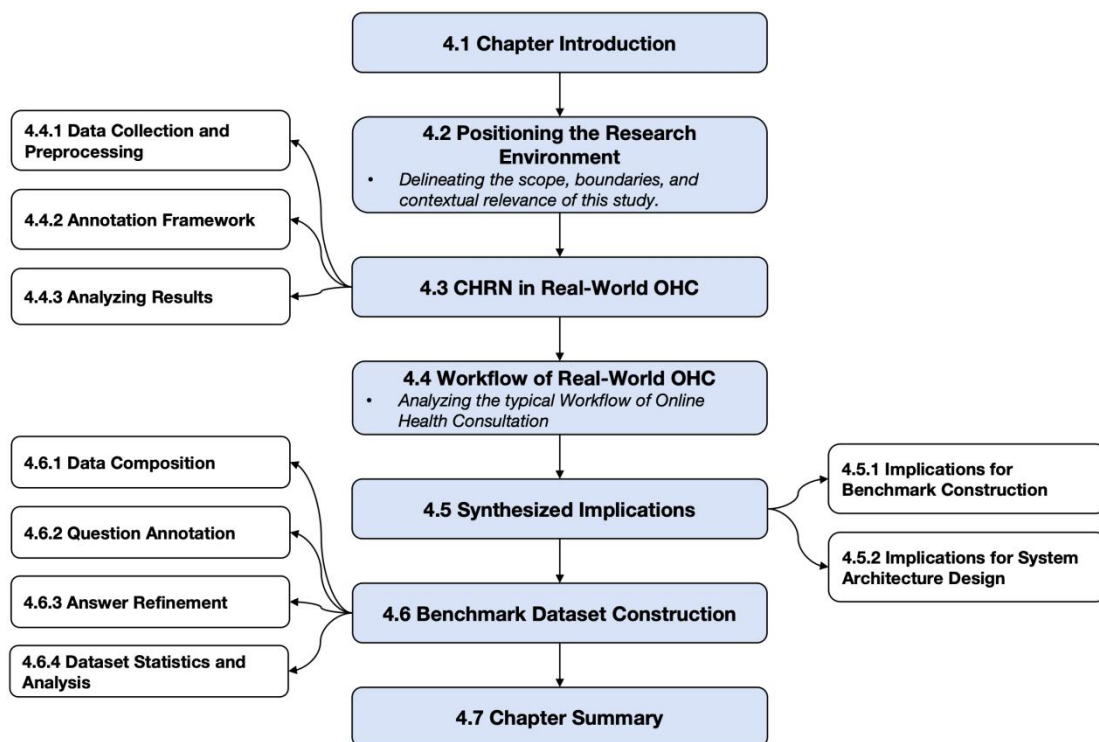


Figure 4-1. Structure of Chapter 4.

As illustrated in Figure 4-1, the chapter proceeds through six interrelated sections that collectively establish a multi-layered understanding of the CHQA context. First, Section 4.2 positions the research environment by defining the scope, boundaries, and contextual relevance of Chinese OHC platforms, particularly Haodf Online. Next, Section 4.3 analyses CHRN as manifested in real-world OHC records, examining how patients articulate informational and emotional needs and how physicians interpret and respond to them in natural interactions. Following this, Section 4.4 explores the workflow of real-world OHC, identifying the structural and rhetorical patterns underlying patient–clinician dialogues. Building on these findings, Section 4.5

synthesises the observed communication dynamics into design implications for intelligent CHQA systems. Subsequently, Section 4.6 details the benchmark dataset construction process, including data collection, annotation, and statistical analysis. The dataset developed in this section not only serves as a training and evaluation foundation for the proposed system but also contributes a reproducible empirical resource for future CHQA research.

4.2 Positioning the Research Environment

In line with the relevance cycle of DSR, this section positions the research within the real-world environment of OHC, delineating its scope, boundaries, and contextual relevance to subsequent artefact design.

This study situates its research environment within the domain of OHC, which refers to asynchronous, text-based clinical interactions in which consumers seek medical advice from licensed physicians or certified health professionals through regulated, closed platforms (Al-Mahdi et al., 2015). Unlike open health forums, OHC platforms operate as formalised clinical micro-environments—characterised by structured communication protocols, regulated participation, and clear accountability for the accuracy and safety of medical advice.

This environment represents a boundary object between medical and informational systems: it preserves the professional norms of healthcare practice while adopting the interactional affordances of digital communication platforms. Such characteristics make OHC an ideal context for investigating how intelligent systems can mediate both informational reliability and emotional empathy—two dimensions often separated in traditional AI-driven communication systems.

Within this boundary, the proposed LLM-based CHQA system is conceptualised as an automated OHC agent capable of augmenting, or partially substituting, the interpretive and empathetic functions of human physicians. Modelling the OHC problem space is therefore essential for ensuring the relevance, ethical safety, and evaluability of the proposed system within real-world constraints.

Existing CHQA studies have predominantly sourced data from online health communities (OHCm), which are open platforms (e.g., health forums, Q&A

boards, Reddit health subforums) where users exchange health-related experiences and information with peers (see Section 2.2 & 2.5). While OHCm data offer large volumes and rich lay perspectives, they exhibit several characteristics that limit their suitability for designing and testing CHQA systems intended to perform health consultation compare with OHC data (see Table 4-1).

Table 4-1. Comparison between Online Health Communities (OHCm) and Online Health Consultation (OHC) Contexts.

	OHCm	OHC
Access Model	Open-access forums; anyone can answer	Closed platforms; only licensed physicians respond
Interaction Structure	Multi-party, informal peer discussions	One-to-one, structured consultation dyads
Communication Style	Often conversational and narrative	Task-oriented, concise, and focused
Response Quality	Variable response quality, no accountability	Professionally curated, medically accountable
Regulation and Safety	No explicit duty of care	Implicit duty of care and safety obligations
Primary Communication Goal	Emotional support dominant, medical accuracy inconsistent	Medical accuracy primary, emotional support complementary

Because OHCm lacks professional accountability and structured workflows, they do not reflect the operational conditions under which a CHQA system would ultimately be deployed. In contrast, OHC environments closely mirror the functional requirements of CHQA, making them a more appropriate and ecologically valid setting for system design and evaluation.

OHC platforms involve three core stakeholder groups:

- Consumers (patients): lay users seeking professional medical advice, often presenting multi-intent queries containing medical descriptions, personal context, and implicit or explicit emotional expressions (e.g., anxiety, fear, reassurance-seeking).
- Providers (physicians): licensed medical professionals who respond with medically accurate, safe, and context-sensitive advice under explicit or implicit professional accountability.
- Platform operators: entities that manage the system infrastructure, enforce safety protocols, moderate content quality, and ensure compliance with data protection and ethical guidelines.

The typical OHC interaction follows a single-turn, asynchronous pattern: consumers submit queries; providers respond after a delay; the platform archives and moderates the consultation. This closed-loop consultation structure makes OHC data particularly suitable for constructing, training, and evaluating automated CHQA systems.

Defining OHC as the research environment establishes the relevance basis for this design science study. It ensures that the proposed LLM-based CHQA system is designed to operate under realistic constraints, stakeholder expectations, and safety requirements comparable to those faced by professional physicians. By contrast, using data from open online health communities would risk misaligning system capabilities with its intended operational context.

To maintain methodological focus and ensure analytical tractability, this study adopts three key delimitations:

- Modality constraint. The analysis is limited to text-based interactions. Audio or multimodal consultations are excluded to avoid the confounding influence of paralinguistic cues that require different modelling paradigms.
- Platform and language scope. The empirical corpus is drawn from Chinese OHC platforms, representing a culturally and linguistically coherent environment with well-defined clinical communication norms.
- Interaction granularity. The study focuses on single-turn question–answer pairs, which constitute the fundamental unit of consumer health communication and enable clear mapping between informational and emotional intents. Multi-turn dialogues are reserved for future research once the core JMEQA framework is validated.

Through these boundaries, the research maintains alignment with DSR’s rigour–relevance balance, ensuring that artefact design is grounded in a realistic problem domain while remaining methodologically manageable for evaluation and replication. Subsequent sections of this chapter will focus on analyzing real-world OHC data to identify the situation elements, structures, and dynamics that must be incorporated into the CHQA system’s design.

4.3 CHRN in Real-World OHC

To design a robust LLM-based joint medical-emotional CHQA system, a deeper discourse-level content analysis of real OHC records is therefore

essential. Unlike well-structured biomedical queries, CHQs on Chinese OHC platforms are characterised by linguistic informality, multi-intent complexity, and affective contextualisation. Patients frequently narrate personal histories, describe emotional distress, and seek reassurance in addition to medical advice. Understanding these intertwined discourse patterns enables the system to model CHQs as multi-layered communicative acts rather than isolated question units, providing a more accurate, human-centred foundation for question understanding, dataset annotation, and subsequent system design. As shown in Section 2.3, prior research on CHRN has made important progress in characterising the informational dimensions of CHQs. However, several persistent limitations restrict its applicability to real-world OHC contexts.

First, most existing studies have drawn their data from open-access online health communities, such as expert-curated Q&A sites (e.g., WebMD Answers) or community-based platforms (e.g., Yahoo! Answers). While these sources provide large volumes of user-generated content, they lack the structural and regulatory characteristics of closed, physician-mediated OHC platforms (e.g., Haodf Online, Chunyu Doctor), where consultations occur in formal clinical micro-environments. Such environments more closely resemble the conditions under which CHQA systems are expected to operate, making them more suitable for system-oriented design and evaluation.

Second, prior studies have often focused on disease-specific inquiries in isolation, such as cancer (Oh et al., 2016; Johnson & Meischke, 2016) or diabetes (Park et al., 2016), rather than examining the broader communicative structures through which health information and emotions co-evolve in general consultations. This narrow focus limits our understanding of how users frame multi-intent questions that span medical, emotional, and contextual concerns simultaneously.

Third, the majority of research has concentrated on identifying informational needs, with comparatively little attention paid to emotional or psychosocial dimensions of user expression. Yet in real OHC settings, patients frequently articulate feelings of anxiety, uncertainty, or reassurance-seeking alongside factual questions about symptoms or treatments, which is an intertwined pattern that purely informational analyses fail to capture.

Finally, existing analyses of CHQs have predominantly adopted lexical or sentence-level semantic approaches (Neha et al., 2024), which provide useful

micro-level insights but overlook macro-level discourse organisation, that is, how informational and emotional intents are hierarchically structured, sequentially related, and contextually embedded within a complete consultation narrative.

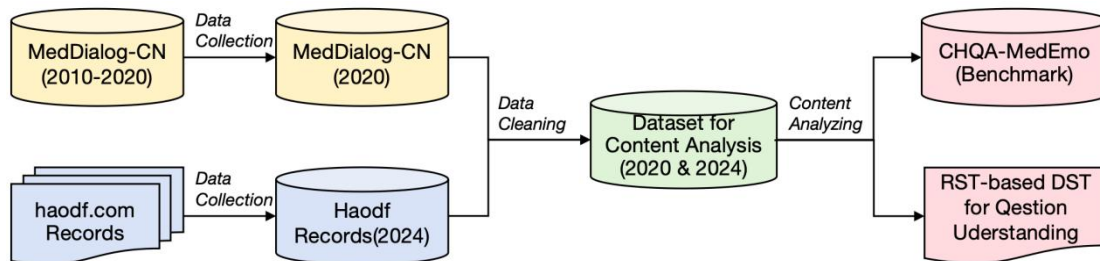


Figure 4-2. Content Analyzing Process in This Section.

To achieve this discourse-level understanding, this section conducts a multi-source content analysis based on both open-domain and platform-specific OHC datasets, as illustrated in Figure 4-2. After data cleaning and preprocessing, representative samples from both sources were integrated into a dataset for content analysis. Through this process, Section 4.3 contributes both empirical grounding and methodological innovation. It transforms fragmented observations of user needs into a structured, discourse-level understanding that bridges human consultation dynamics and machine reasoning mechanisms, laying the theoretical and data-driven foundation for the design principles synthesised in Section 4.5 and the system implementation detailed in Chapter 5.

4.3.1 Data Collection and Preprocessing

As illustrated in Section 3.4.2, the CHQs analysed in this study were sourced from Haodaifu Online (www.haodf.com), a widely recognised Chinese OHC platform. The detailed procedures for data collection and preprocessing are outlined as follows:

Data Collection. To ensure adequate sample size and period coverage, two complementary data sources were integrated. First, an automated web scraping script in Python was developed to collect classic consultation records from Haodf Online. Each sub-department provides approximately ten pages of publicly accessible records (around 900 entries per department). From this pool, 400 cases were randomly selected per department. Second, to supplement departments where additional samples were required, I incorporated data from the MedDialog-CN (2020) dataset. From this corpus,

an additional 400 eligible consultation records per department were randomly sampled. After combining both sources, each department contributed up to 800 records, provided sufficient cases were available. A total of 14 departments met this threshold, while other departments fell short due to data scarcity. For each consultation case, I extracted nine key components from the CHQ, including: (a) case title, (b) medical description, (c) height and weight, (d) diagnosis, (e) duration of wellness, (f) medical institutions visited, (g) allergy history, (h) past medical history, and (i) requested support. All extracted data were organized into a structured Excel file for subsequent annotation and analysis.

Data Preprocessing. From the 400 CHQs initially collected for each eligible department, the following cases were excluded: (a) incomplete entries, primarily those lacking detailed disease descriptions; (b) follow-up consultations rather than initial inquiries; (c) highly similar or duplicate entries; and (d) non-text-only formats, where questions were accompanied by multimedia elements such as images or audio. The collection and filtering process spanned from August 1 to September 1, 2024. Ultimately, 14 departments contributed 400 anonymized user queries each (200 from 2020, 200 from 2024), yielding a total dataset of 5,600 CHQs for subsequent content analysis (see Table 4-2).

Table 4-2. Overview of Collected Consumer Health Questions

Clinical Specialty	# Number	Clinical Specialty	# Number
Traditional Chinese Medicine	400	Andrology	400
Pediatrics	400	Burn Medicine	400
Stomatology	400	Dermatology	400
Gynecology	400	Ophthalmology	400
General Internal Medicine	400	Otolaryngology	400
Rehabilitation Medicine	400	Medical Oncology	400
General Surgery	400	Psychiatry and Psychology	400
Total Records	5600		

4.3.2 Annotation Framework

Annotation Schema. To design an annotation schema that faithfully represents CHQs, researcher first drew a stratified sample of 240 CHQs—20 from each clinical department. Grounded in the RST, POMR, and AFCE presented in Section 2.6 and guided by an exploratory analysis of this sample, researcher drafted the initial schema to capture its predominant

discourse patterns. As researcher annotated additional queries, researcher continuously refined the schema to reflect emerging insights, ultimately producing a framework that balances conciseness with descriptive power for all discourse structures observed in the test set.

RST views coherent discourse as the result of functional relationships between text spans, categorized as nuclei (the primary intent) and satellites (supporting information such as reasons, elaborations, or conditions). In the CHQ context, I treat each user need as a nucleus and classify it into one of two main categories: medical information needs and emotional support needs.

Medical information needs (M-N). Given a predefined set of medical needs categories (derived after the manual evaluation of 100 CHQs), annotators label a medical question with one of them. Below are the included medical need categories along with their definitions:

- General medical information (M-N-GMI): Requests broad or non-specific medical information, such as general knowledge about a condition, medication, or procedure (e.g., “Could you tell me more about how this medication works?”).
- Etiology (M-N-ETI): Requests the cause or origin of a specific symptom or disease (e.g., “What causes my headache?”).
- Diagnosis (M-N-DIA): Requests clarification of the likely diagnosis based on specific symptoms (e.g., “What on earth is the wellness of mine?”).
- Treatment (M-N-TREAT): Requests specific treatment advice, medication recommendations, or therapy options (e.g., “How can I treat this condition effectively?”).
- Prognosis (M-N-PROG): Asks about the expected course of a disease, possible outcomes, or recovery timeline (e.g., “will this wellness resolve on its own?”).

Medical needs related context (M-C). For each recognized medical need, annotate zero or more related context segments as satellites linked to that need. Each satellite is labeled with its rhetorical function:

- Elaboration (M-C-ELA): Adds further detail or explanation to the nucleus (e.g., “I am currently experiencing dizziness, nausea, and fatigue.”)

- Background (M-C-BACK): Provides information necessary to understand the user's current concern (e.g., "I have been taking Duloxetine (Cymbalta) for one year.")
- Condition (M-C-CON): Specifies the circumstances under which the need applies (e.g., "I no longer wish to continue medication.")

Emotional support needs (E-N). Emotional need categories are derived from Pounds' (2011) appraisal framework for language-based clinical empathy, where patients' expression of feelings and views are categorized as empathy opportunities, representing key moments for empathetic engagement. Categories include:

- Feel (E-N-FEEL). Seek acknowledgment or understanding by explicitly or implicitly expressing their feelings, indicating a need for empathy or support (e.g., "Would you please help me? It's so hard!").
- View (E-N-VIEW). Seek acknowledgment or understanding of their personal viewpoint or attitude regarding their situation or treatment (e.g., "I'm worried that I can't quit this medication anytime soon").

Specifically, annotators may infer and summarize implicit emotional expressions and implicit attitudinal views by considering contextual cues. In the case of implicit emotions, the underlying affective state should be categorized according to Plutchik's(1980) psychoevolutionary model of basic emotions, which includes: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation.

Emotional needs related context (E-C). This category is defined based on recent research indicating that leveraging the underlying causes of emotions (CAUSE) enhances the empathetic response generation (Li et al., 2021; Wang, et al., 2021). For each recognized emotional need, annotate zero, one or more related context segments as cause:

- Cause (E-C-CAUSE): Offers a reason or trigger for the emotional need, clarifying why the patient experiences the stated feeling or viewpoint (e.g., "I've tried numerous medications with no relief, and the side effects are severe").

Annotation Process. Three Chinese postgraduate students with medical backgrounds were recruited to participate in the annotation and revision of the corpus. Each user query was independently annotated by two annotators.

In cases of disagreement, a third annotator was assigned to adjudicate and finalize the labels.

To ensure consistency and high-quality annotation, this study developed a comprehensive guideline booklet, synthesizing relevant theoretical foundations and providing practical annotation examples. This guideline served both as mandatory pre-annotation training material and as reference documentation during the annotation process. Prior to formal annotation, all annotators completed a practice batch of ten queries to familiarize themselves with the task specifications.

Annotation tasks were performed using Label Studio, an open-source data labeling platform that supports customized text span and relational annotations, audit logging, and JSON-format exports. As no existing template in the template gallery met the specific requirements of this study, a custom annotation interface was designed using XML configuration (see Appendix 1). This tailored configuration ensured the annotation process aligned closely with the multi-layered structure of the proposed framework.

An example annotation interface in Label Studio is attached in Appendix 2, showcasing multi-level structured labeling of a consumer health query. The interface supports span-level tagging of user needs and contextual information, as well as the manual linking of related spans using relation arrows. Global attributes, such as disease complexity and asker-patient relationship, are also recorded.

3. Inter-Annotator Agreement Analysis

This study adopts overall agreement as the primary measure of inter-annotator agreement (IAA). The annotation task requires annotators to freely extract and label text spans from long, unstructured queries, with each annotation unit defined by a partial match between the span boundaries and the assigned label. Agreement is counted only when both elements are identical between annotators. This design yields an extremely large combinatorial space of possible span-label pairs, making the probability of achieving agreement by chance negligibly low. As Krippendorff (2018) notes, the rationale for chance-corrected coefficients such as Cohen's κ (Cohen, 1960) and Krippendorff's α is to compensate for non-trivial expected agreement by chance (P_e); however, when P_e is close to zero, raw agreement provides an equivalent and more interpretable measure. In our

setting, the strict matching criterion and high variability of extracted spans ensure that random alignment is rare, and thus the numerical difference between chance-corrected statistics and raw agreement is minimal. Therefore, overall agreement serves as a sufficient and transparent metric for quantifying annotation reliability in this context, aligning with prior recommendations that raw agreement is appropriate when the likelihood of coincidental matches is vanishingly small (Krippendorff, 2018; Artstein & Poesio, 2008).

This study adopts a semi-automated, manual adjudication protocol to quantify IAA for the multi-layer annotation scheme. The first two annotators' files were compared with PyCharm Compare Files, which automatically highlighted divergent spans. The third adjudicator inspected each highlight and declared the spans matching when (a) their labels were identical and (b) the spans shared $\geq 50\%$ semantic overlap, operationalized as $\geq 50\%$ token overlap or clear paraphrastic equivalence. To assess adjudication bias, 10% of the corpus was doubly adjudicated by an independent reviewer (one of the authors); the adjudicator-to-adjudicator Cohen's kappa coefficient (Banerjee et al., 1999) was 0.81, indicating high internal consistency.

Agreement was then computed hierarchically. At the need level. A pair of units was counted as a match only when their labels were identical and their text spans satisfied the 50% rule. At the context level. Within each matched need, supporting text spans were compared under the same rule, contingent on identical labels.

For each of the 13 (label + span) units, this study computed overall agreement. I then produced frequency-lighted macro averages across four dimensions, medical needs, medical needs related context, emotional needs, and emotional needs related context relations.

For the annotations of medical needs and related context units, the overall agreement is 87.91% and 92.62%, respectively. For the annotations of emotional needs and related context units, the overall agreement is 84.31% and 84.40%. Emotional annotation is an open-ended task, therefore moderate agreement as in the other open-ended tasks is acceptable (Alasmari et al., 2023). Overall, these results show that the consistency between the annotators is satisfactory (see Table 4-3).

Table 4-3. Overall agreement in 4 dimensions, 11 label units.

Label Units	Overall Agreement (%)
Medical Needs	87.91
- General medical information (M-N-GMI)	86.35
- Etiology (M-N-ETI)	94.96
- Diagnosis (M-N-DIA)	85.62
- Treatment (M-N-TREAT)	93.09
- Prognosis (M-N-PROG)	87.91
Medical Needs Related Context	92.62
- Elaboration (M-C-ELA)	87.84
- Background (M-C-BACK)	90.69
- Condition (M-C-CON)	90.38
Emotional Needs	84.31
- View (E-N-VIEW)	78.80
- Feel (E-N-FEEL)	84.67
Emotional Needs Related Context	84.40
- Cause (E-C-CAUSE)	

4.3.3 Analyzing Results

The dataset comprises 5,600 clinical Chinese CHQs, including 2,800 originating from 2020 and 2,800 from 2024. Table 4-4 presents the average number of Chinese characters per query across 15 clinical specialties for the two data origin years. Overall, the mean query length for 2020-origin queries is 145.57 characters, for 2024-origin queries is 152.66 characters, with an overall average of 149.12 characters across all specialties. The longest queries were observed in Psychiatry and Psychology (225.92) and Medical Oncology (212.27), while Stomatology (106.68) and Dermatology (116.73) had the shortest, suggesting substantial domain-specific variation in the detail and complexity of patient descriptions.

Table 4-4. Number of Chinese Characters per Query.

Clinical Specialty	2020	2024	Average
Psychiatry and Psychology	<u>215.13</u>	<u>236.70</u>	<u>225.92</u>
Medical Oncology	<u>229.41</u>	<u>195.13</u>	<u>212.27</u>
Gynecology	<u>188.22</u>	153.25	<u>170.74</u>
Rehabilitation Medicine	162.45	<u>171.72</u>	167.08
Traditional Chinese Medicine	185.61	136.32	160.97
General Surgery	144.61	146.85	145.73
General Internal Medicine	113.29	171.27	142.28
Otolaryngology	121.89	141.51	131.70
Pediatrics	122.50	140.68	131.59
Andrology	124.45	135.74	130.10
Burn Medicine	121.69	127.77	124.73
Ophthalmology	109.20	133.07	121.13
Dermatology	105.38	128.07	116.73
Stomatology	94.17	119.18	106.68
Total	145.57	152.66	149.12

Table 4-5 reports the average number of medical needs per query across the same 15 clinical specialties, comparing CHQs originating from 2020 and 2024. The overall average decreased slightly from 1.75 in 2020-origin queries to 1.67 in 2024-origin queries, resulting in an overall mean of 1.71. Medical Oncology (1.90) and Burn Medicine (1.80) exhibited the highest averages across both years, indicating that patient queries in these specialties often involve multiple concurrent medical needs, possibly due to complex disease management scenarios. In contrast, Pediatrics (1.63) and Psychiatry and Psychology (1.53) showed the lowest averages, suggesting more focused or single-need consultations in these domains.

Table 4-5. Number of Medical Needs per Query.

Clinical Specialty	2020	2024	Average
Psychiatry and Psychology	1.48	1.58	1.53
Medical Oncology	<u>1.88</u>	<u>1.92</u>	<u>1.90</u>
Gynecology	<u>1.84</u>	1.59	1.71
Rehabilitation Medicine	1.81	1.73	1.77
Traditional Chinese Medicine	1.59	<u>1.79</u>	1.69
General Surgery	1.81	<u>1.77</u>	<u>1.79</u>
General Internal Medicine	1.81	1.60	1.70
Otolaryngology	1.70	1.61	1.66
Pediatrics	1.64	1.62	1.63
Andrology	1.82	1.58	1.70
Burn Medicine	<u>1.89</u>	1.72	<u>1.80</u>
Ophthalmology	1.76	1.56	1.66
Dermatology	1.74	1.62	1.68
Stomatology	1.76	1.67	1.71
Total	1.75	1.67	1.71

Figure 4-3 summarizes the frequency distribution of five medical needs categories—Diagnosis (M-N-DIA), Etiology (M-N-ETI), Prognosis (M-N-PROG), Treatment (M-N-TREAT), and General Medical Information (M-N-GMI)—across CHQs originating from 2020 and 2024. Treatment-related needs (M-N-TREAT) overwhelmingly dominated the dataset, accounting for 5,981 instances (51.2% of all annotated medical needs), with a slight increase from 2,890 in 2020 to 3,091 in 2024. Diagnosis-related needs (M-N-DIA) were the second most frequent, totaling 1,787 instances (15.3%), showing a small decrease from 2020 (921) to 2024 (866). In contrast, Etiology (M-N-ETI, 696 instances, 6.0%) and Prognosis (M-N-PROG, 564 instances, 4.8%) were comparatively infrequent, and both experienced declines from 2020 to 2024, suggesting a reduced emphasis on causal inquiries and outcome predictions in more recent CHQs. General Medical Information (M-N-GMI) had the lowest frequency (541 instances, 4.6%), with a marked drop from 369 in 2020 to 172 in 2024, indicating a decreasing tendency for patients to request broad, non-specific medical information over time. These patterns highlight that patient information-seeking behaviour in CHQs is predominantly oriented toward treatment guidance, followed by diagnostic clarification, while interest in etiology, prognosis, and general medical information remains relatively limited. The temporal shifts suggest a gradual narrowing of focus toward actionable, case-specific clinical advice in recent years.

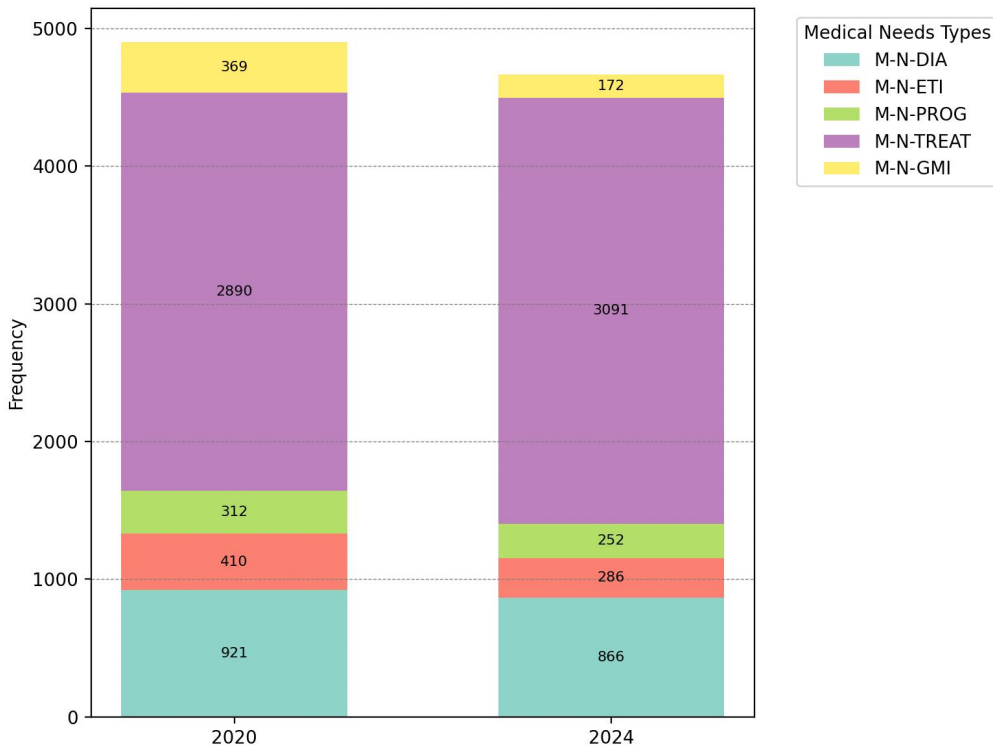


Figure 4-3. Distribution of Medical Need Types.

Table 4-6 presents the average number of emotional needs per query across 15 clinical specialties, comparing CHQs originating from 2020 and 2024. The overall average decreased from 1.06 in 2020-origin queries to 0.97 in 2024-origin queries, yielding a combined mean of 1.01. Among all specialties, Psychiatry and Psychology exhibited by far the highest average (1.70), reflecting the inherently emotion-rich nature of mental health consultations, where patient concerns often extend beyond purely medical issues. Medical Oncology (1.11) and Traditional Chinese Medicine (1.11) also demonstrated relatively high emotional needs counts. The relatively high emotional needs in Medical Oncology may be strongly associated with the severity and life-threatening nature of the disease, which often elicits heightened psychological distress alongside medical concerns. In contrast, specialties such as Ophthalmology (0.83), Dermatology (0.86), and General Surgery (0.90) recorded the lowest averages, indicating that consultations in these fields tend to focus more on technical or procedural aspects rather than emotional support.

Table 4-6. Number of Emotional Needs per Query.

Clinical Specialty	2020	2024	Average
Psychiatry and Psychology	<u>1.60</u>	<u>1.81</u>	<u>1.70</u>
Medical Oncology	<u>1.18</u>	<u>1.04</u>	<u>1.11</u>
Gynecology	<u>1.07</u>	1.01	1.04
Rehabilitation Medicine	<u>1.07</u>	0.92	0.99
Traditional Chinese Medicine	1.02	<u>1.20</u>	<u>1.11</u>
General Surgery	0.98	0.83	0.90
General Internal Medicine	1.03	0.80	0.91
Otolaryngology	1.04	0.81	0.92
Pediatrics	1.00	0.88	0.94
Andrology	1.02	1.03	1.02
Burn Medicine	1.00	0.85	0.92
Ophthalmology	0.90	0.77	0.83
Dermatology	0.95	0.77	0.86
Stomatology	0.96	0.86	0.91
Total	1.06	0.97	1.01

A cross-dimensional comparison reveals that high query length does not necessarily correspond to a high number of medical needs (see Figure 4-4). For instance, Psychiatry and Psychology had the longest average query length but one of the lowest medical needs counts, instead showing the highest emotional needs count—reflecting detailed, narrative-rich descriptions focused on fewer clinical but more emotionally nuanced concerns. Conversely, Burn Medicine combined moderate query length with high medical needs and relatively low emotional needs, indicating concise yet multifaceted requests primarily centered on treatment and recovery. These findings highlight the heterogeneity of patient information-seeking behaviour and emphasize the necessity of modeling both narrative complexity and intent density—medical and emotional—when developing domain-adaptive natural language processing systems for clinical question understanding. Notably, 87.9% of the queries present both medical and emotional demands, underscoring the necessity of integrating evidence-based clinical guidance with empathic support in the design of downstream QA or dialogue agents.

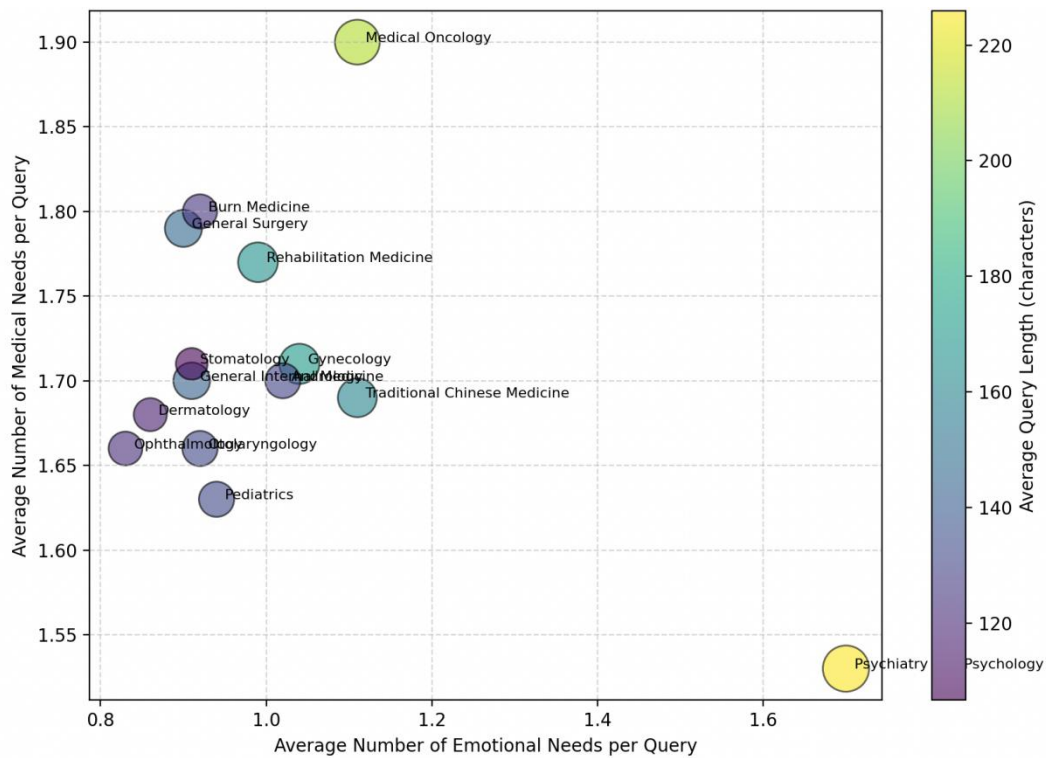


Figure 4-4. Cross-Dimensional Bubble Chart of Emotional Needs (X), Medical Needs (Y), and Query Length (Size & Color)

4.4 Workflow of Real-World OHC

To construct a foundational understanding of OHC interactions, this study adopts a reverse engineering technique and observations at the level of single-turn exchanges. Each consultation turn can be decomposed into a three-step process flow.

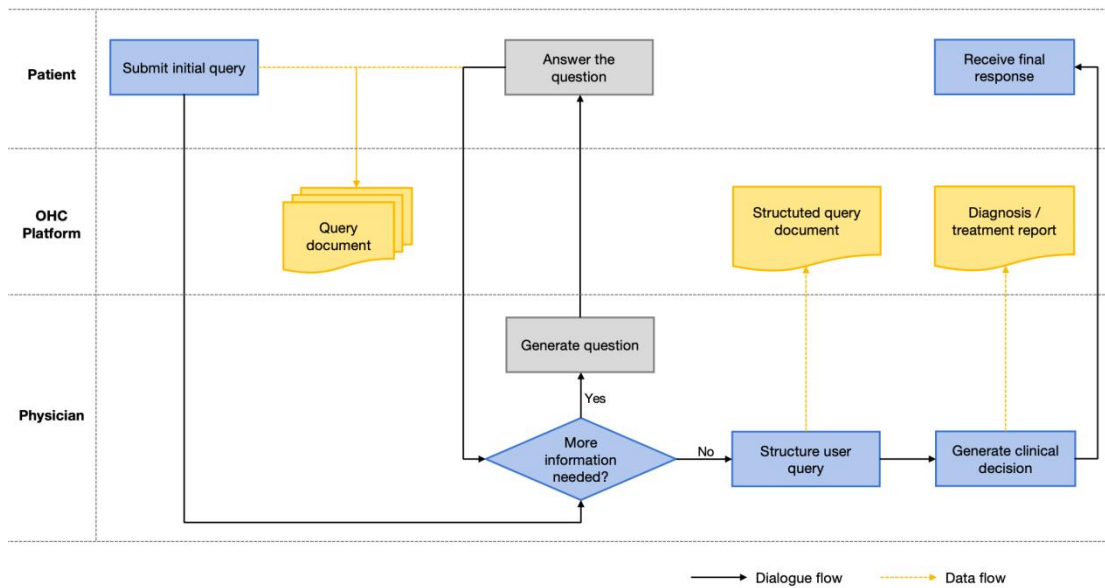


Figure 4-5. Typical Workflow of Text-Based Online Health Consultation.

Figure 4-5 illustrates a simplified abstraction of OHC at the level of a single-turn exchange. The flow distinguishes the roles of patient, physician, and OHC platform. Patients initiate the process by submitting an unstructured query and, when necessary, providing clarifications in response to physician requests. Physicians interpret and encode the patient’s narrative into clinically meaningful structures, determine whether additional information is required, and subsequently generate clinical decisions that may include diagnostic reasoning, treatment advice, or reassurance. The OHC platform serves as an intermediary layer that stores and transmits patient queries and physician outputs, enabling the transition from unstructured input to structured representation. The model highlights how effective OHC relies on physicians’ cognitive ability to transform heterogeneous patient input into structured, actionable reasoning, while simultaneously addressing patients’ informational and emotional concerns.

Submit Initial Query (Patient). This step represents the first stage of the OHC workflow, which is typically initiated by the patient, a family member, or a close friend. At this stage, the information provider submits a structured description of the case to the platform, serving as the foundation for subsequent physician assessment and advice. Queries usually include: (a) Medical description: a narrative outlining the patient’s current symptoms, treatment process, and recent changes in physical condition; (b) Height and weight: basic biometric data such as stature and body mass, usually measured within the past few months; (c) Diagnosis: the formal medical condition or surgical outcome recorded by healthcare providers (e.g., right

hemicolectomy); (d) Duration of illness: the period since symptom onset or diagnosis, indicating chronicity or progression of the disease; (e) Medical institutions visited: a list of hospitals or clinics where the patient has previously received treatment or examinations; (f) Allergy history: any known allergic reactions to medication or food; (g) Past medical history: relevant comorbidities, surgical procedures, or long-term conditions such as hypertension or diabetes; and (h) Requested support: the patient's explicit inquiry or request for advice, typically regarding follow-up treatment options, prognosis, or hospitalization eligibility.

Generate Question (Physician). This step occurs when the physician reviews the user's initial query and determines whether additional information is required to make a clinical judgement. Based on the completeness, clarity, and relevance of the submitted information, the physician may generate supplementary questions to clarify symptoms, diagnostic history, medication use, or prior treatments. These follow-up questions serve to refine the understanding of the case and ensure diagnostic accuracy. However, the present study focuses exclusively on single-turn consultations, in which the patient's initial query and the physician's response form a complete interaction. Multi-turn consultations involving iterative question generation and clarification are therefore beyond the scope of this analysis.

Structure User Query (Physician). Physicians interpret and internally structure the patient's narrative into medically meaningful categories, a process akin to structured clinical documentation (Weed, 1968; Bodenreider & McCray, 2003). Rather than treating the patient's description as free text, physicians summarize and decompose it into discrete clinical fields that align with diagnostic reasoning and electronic health record (EHR) conventions. According to the platform's requirements, the user query is systematically organised into the following categories: (a) Chief complaint. The primary symptom or problem that prompted the consultation; (b) History of present illness. A chronological account detailing the onset, duration, and progression of the current condition; (c) Past medical history. Previously diagnosed diseases, surgeries, or chronic conditions that may influence current symptoms; and (d) Auxiliary examinations. Laboratory results, imaging findings, or other diagnostic tests that provide objective evidence. This structured representation enables physicians to apply clinical reasoning frameworks such as the SOAP (Subjective – Objective – Assessment – Plan) model, facilitating both standardised documentation and efficient case comprehension.

Generate Clinical Decision (Physician). This step involves the physician’s synthesis of structured information to formulate diagnostic reasoning and corresponding medical advice. Drawing upon the categorised data obtained in the previous stage—chief complaint, history of present illness, past medical history, and auxiliary examinations—the physician evaluates potential differential diagnoses, determines the disease stage, and recommends follow-up actions. This process reflects the Assessment and Plan phases of the classical SOAP framework, where subjective and objective information are integrated to support evidence-based decision-making.

4.5 Synthesized Implications

This section synthesises the empirical findings derived from Sections 4.3 and 4.4 to generate design implications that inform the subsequent benchmark construction and system architecture in Sections 4.6 and 5.2, respectively.

4.5.1 Implications for Benchmark Construction

The content analysis of Haodf Online records presented in Section 4.3 reveals that approximately 87.9% of patient-generated CHQs contain implicit or explicit emotional cues interwoven with medical informational needs. This empirical finding reinforces the dataset-scope research gap identified earlier in Section 2.5.4, confirming that existing CHQA corpora fail to capture the multidimensional nature of real-world health consultations. To date, no publicly available dataset systematically encodes dual-dimensional annotations that jointly represent informational and affective intents within a coherent discourse structure. This absence has constrained both model training and benchmark comparability, preventing current CHQA systems from generalising effectively to the long, multi-layered, emotionally nuanced queries typical of Chinese OHC environments.

The analysis in Section 4.3 further quantified the linguistic and structural properties of CHQs, such as average character length, distribution of informational versus emotional intents, and variance across fifteen clinical specialties, thereby providing a data-driven rationale for benchmark sampling. These results indicate that real CHQs are substantially longer and semantically richer than those found in open-community datasets, and that emotional expressions frequently occur in proximity to key diagnostic or

treatment concerns. Such characteristics must therefore be reflected in the benchmark design to ensure representational fidelity and training validity.

4.5.2 Implications for System Design

The discourse-level content analysis in Section 4.3 produced a comprehensive framework for understanding long-form, multi-intent CHQs through RST. This framework identifies the hierarchical and relational structures that connect medical informational needs with emotional support needs, thereby providing a symbolic representation of question logic and affective cues. Such symbolic structures serve as essential knowledge inputs for the NSI approach adopted in the system design. In particular, the RST-based DST functions as an intermediate reasoning layer that translates unstructured text into structured, machine-interpretable symbolic knowledge, facilitating the integration of LLMs with explicit reasoning components during question understanding.

Insights from Section 4.4, which examined the workflow of real-world OHCs, further inform the overall architectural design of the proposed system. Rather than intervening at a single operational step, such as response generation or retrieval, the designed artefact aims to simulate the physician's holistic role, automating the full consultation logic within a controlled, single-turn interaction setting. This conceptualisation positions the artefact not as an auxiliary tool but as an early-stage, full-cycle consultation agent, capable of performing both cognitive reasoning and affective communication within the OHC environment.

As illustrated in Figure 4-5, the typical OHC workflow, in which physicians often engage in iterative clarification cycles, requesting supplementary information before formulating diagnostic decisions. In contrast, the proposed LLM-based CHQA system is intentionally scoped as a single-turn consultation prototype, designed to emulate the two most critical cognitive stages of physician reasoning: (a) Structuring the user query, and (b) Generating the clinical decision.

This design choice deliberately constrains the system's scope to focus on core reasoning and empathy modelling before extending to multi-turn dialogue in future iterations. The simplification enables more controlled evaluation of the artefact's internal logic, particularly its ability to balance

factual precision with affective sensitivity, a balance often neglected in both real-world OHC interactions and current CHQA systems.

Observations from the Haodf dataset further revealed that physicians typically organise their question-understanding process into four canonical components: (a) chief complaint, (b) history of present illness, (c) past medical history, and (d) auxiliary examinations. However, emotional expressions from patients were seldom integrated into this cognitive structuring process. Consequently, the absence of emotional context at the question-understanding stage often propagated to the clinical-decision stage, leading to emotionally detached replies. This empirical insight underscores the critical importance of enhancing question-understanding capabilities within CHQA systems, not merely for factual interpretation but for empathetic comprehension.

4.6 Benchmark Dataset Construction

4.6.1 Dataset Composition

Building on the analysis presented in Section 4.4, this study constructed a benchmark dataset named CHQA-MedEmo, designed to facilitate the development and evaluation of LLM-based CHQA systems capable of understanding and responding to both medical informational needs and emotional support needs. The dataset was derived from the large-scale Chinese online health consultation corpus introduced in Section 4.4, with a specific focus on the Psychiatry and Psychology specialty.

The selection of the Psychiatry and Psychology specialty was guided by two main considerations: (a) Complexity of patient queries. Psychiatry and Psychology consultations contained the longest patient queries, averaging 225.92 Chinese characters per case. Such length not only reflects richer contextual descriptions but also suggests a higher likelihood of patients articulating multiple medical needs alongside emotional concerns. This characteristic makes these specialties particularly suitable for evaluating systems designed to process multi-intent and multi-dimensional consultations. (b) Diversity of emotional needs. Psychiatry and Psychology also exhibited the highest number of emotional needs per query (1.70 on average). Including these specialties therefore ensures that the dataset encompasses both informationally complex and emotionally intensive consultations. This dual coverage allows the benchmark to reflect the

intertwined informational and emotional dimensions inherent in consumer health queries.

From the selected Psychiatry and Psychology specialties, an initial pool of 400 user queries was expanded by identifying 1,439 psychiatry-related user queries from MedDialog’s publicly available records (Zeng et al., 2020). To capture more recent consultation styles, a Python-based web crawler was subsequently employed to randomly extract an additional 300 queries from the Haodaifu online platform’s 2024 mental-health consultation logs.

From the combined 2,139 cases, a semi-automated screening procedure, integrating manual review with Python scripts, was applied to select high-quality records that satisfied all of the following inclusion criteria: (a) Each patient query had to contain both medical and emotional needs, ensuring multi-dimensional consultation intents; (b) Among eligible queries, preference was given to those with longer textual descriptions and a greater number of distinct needs, as such cases are more suitable for evaluating system performance on complex inputs; (c) Physician responses were required to be comprehensive and contextually detailed; and (d) A majority of cases were drawn from the subspecialty of substance dependence, which typically integrates both medical and emotional needs.

Finally, the resulting CHQA-MedEmo dataset comprises 703 online health consultation QA pairs, of which 235 were collected in 2024 and 468 were sourced from 2020.

4.6.2 Question Annotation

All 703 questions in the CHQA-MedEmo dataset were annotated following the annotation schema developed in Section 4.4.2. Consistent with the previous annotation process, three Chinese postgraduate students with medical backgrounds were again recruited to participate in the annotation and revision of the corpus, thereby eliminating the need for additional training. Each user query was independently annotated by two annotators. In cases of disagreement, a third annotator was assigned to adjudicate and finalise the labels.

Following the same IAA computation method described in Section 4.4.2, the overall agreement was calculated for each of the 11 units. As shown in Table 4-7, the overall agreement for medical needs and their related context units

reached 80.17% and 80.95%, respectively. For emotional needs and their corresponding context units, the overall agreement was 78.00% and 85.45%, respectively. Overall, these results indicate a satisfactory level of consistency among annotators, demonstrating the reliability of the annotation process and the robustness of the designed schema.

Table 4-7. Overall agreement in 4 dimensions, 11 label units

Label Units	Overall Agreement (%)
Medical Needs	80.17
- General medical information (M-N-GMI)	85.23
- Etiology (M-N-ETI)	90.91
- Diagnosis (M-N-DIA)	87.08
- Treatment (M-N-TREAT)	73.23
- Prognosis (M-N-PROG)	92.65
Medical Needs Related Context	80.95
- Elaboration (M-C-ELA)	80.41
- Background (M-C-BACK)	80.62
- Condition (M-C-CON)	84.03
Emotional Needs	78.00
- View (E-N-VIEW)	76.42
- Feel (E-N-FEEL)	80.96
Emotional Needs Related Context	85.45
- Cause (E-C-CAUSE)	

4.6.3 Answer Refinement

Specifically, 141 user queries were selected from this dataset—corresponding to those used in Section 5.4 for evaluating the fine-tuned question understanding module. These samples were not included in the fine-tuning process and were therefore independent of any model training data. This separation ensured that the benchmark evaluation remained unbiased, preventing potential data leakage and preserving the integrity of the performance comparison.

For each query, the original physician-authored responses were carefully reviewed, refined, and extended to develop 141 standardized question-answer pairs. Preliminary examination revealed that many of the original responses lacked sufficient empathetic expression and personalisation when addressing users’ emotional and medical needs. To enhance the benchmark’s representativeness and to align it with the dual objectives of clinical accuracy and empathetic communication, each response was rewritten by qualified psychiatrists in accordance with guideline-concordant and patient-centered communication principles.

A structured three-step workflow was implemented to operationalize these expert-authored responses into an evaluative benchmark (see Figure 4-6). This process systematically translated physician expertise and empathetic intent into standardized, fine-grained evaluative units that could be directly compared with model-generated outputs. The resulting benchmark thus served as a gold-standard reference for both reader-study-based assessments in subsequent sections. The process consisted of:

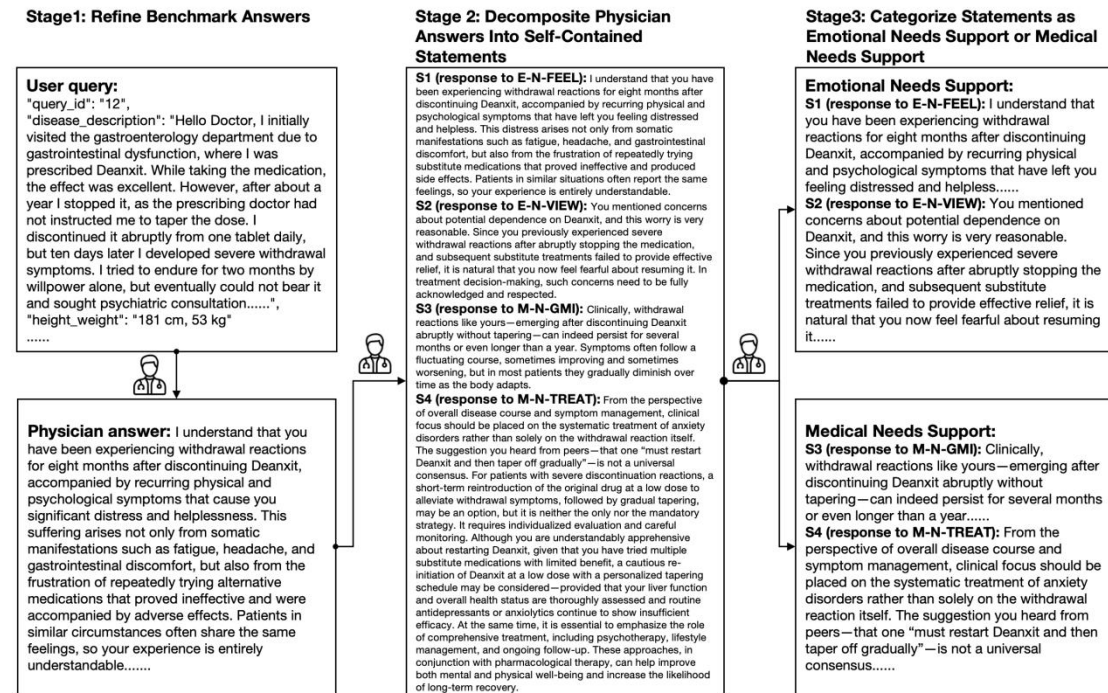


Figure 4-6. An Example of the Benchmark Generate and Structure. Including (a) generate physician benchmark responses, (b) decompose them into self-contained statements, and (c) categorizing them into emotional versus medical needs support.

Generate Benchmark Answers. For each user query, reference answers were authored by licensed physicians with clinical experience in psychiatry and psychosomatic medicine. Two professional psychiatrists were recruited from the verified online health consultation platform Haodaifu Online, a nationally recognized telemedicine service in China. Both physicians have more than five years of clinical experience, as well as over three years of online health consultation experience. These benchmark responses were explicitly designed to align with established clinical guidelines (National Health Commission of China Clinical Practice Guidelines for Mental Disorders [2020 Edition]) and to incorporate empathetic communication strategies appropriate for online mental health consultations. Each draft was independently reviewed and revised by both participating psychiatrists to ensure informational correctness, emotional appropriateness, and linguistic

clarity. The finalized benchmark answers served as gold-standard references for subsequent model evaluation.

Decompose Answers into Self-Contained Statements. Each physician-authored response was segmented into discrete Self-Contained Statements (SCSs)—the minimal, semantically independent units that preserve complete meaning and can be directly mapped to specific need categories. This decomposition enhanced interpretability, minimized semantic overlap, and enabled fine-grained alignment with user needs. For example, empathetic acknowledgments, symptom explanations, and treatment recommendations were extracted as separate units (e.g., S1-S4 in Figure 4-6), allowing each to be evaluated independently.

Categorize Statements. The decomposed statements were subsequently mapped to two primary support categories: (a) Emotional Needs Support, encompassing responses addressing patients' feelings (E-N-FEEL) and viewpoints (E-N-VIEW); and (b) Medical Needs Support, covering general medical information (M-N-GMI), Etiology (M-N-ETI), Diagnosis (M-N-DIA), Prognosis (M-N-PROG), treatment-related guidance (M-N-TREAT). This hierarchical taxonomy allowed evaluators to assess not only the factual accuracy of medical content but also the empathy and relevance of emotional engagement.

Through this three-stage process, the benchmark was operationalized into fine-grained, interpretable evaluative units, providing a transparent and reproducible foundation for reader-study-based assessments. It thereby ensured that subsequent evaluations could capture the dual objectives of the proposed Med-Emo CHQA framework—clinical correctness and empathetic alignment. Appendix 3 provides a complete benchmark example.

4.6.4 Dataset Statistics and Analysis

To further characterize the benchmark dataset and situate it within the broader landscape of annotated CHQA resources, this study conducted both quantitative analysis of CHQA-MedEmo itself and a comparative review against existing datasets.

Internal statistics. The benchmark dataset comprises 703 online health consultation QA pairs—235 collected in 2024 and 468 sourced from 2020—annotated with 1,346 medical-need instances and 1,155 emotional-need

instances. Medical requests are predominantly treatment-seeking intents (61%), followed by diagnostic clarification needs (21%). Emotional needs are distributed between expressions of feeling (56%) and expressions of viewpoint (44%). Among them, 141 queries' original physician-authored responses were carefully reviewed, refined, and extended.

Each query, on average, contains 1.92 annotated medical needs and 1.64 annotated emotional needs, supported by 3.68 medical-context spans and 1.65 emotional-context spans that capture clinical history, symptom evolution, and emotion-eliciting causes (see Table 4-8). Notably, 89.9% of the queries present both medical and emotional demands, underscoring the necessity of integrating evidence-based clinical guidance with empathic support in the design of downstream QA or dialogue agents.

Table 4-8. Statistics of CHQA-MedEmo

Statistics	Average
# of Chinese characters per query	254
# of Chinese characters per need (annotated)	15
# of medical needs per query(annotated)	1.92
# of emotional needs per query (annotated)	1.64
# of medical needs related context per query (annotated)	3.68
# of emotional needs related context per query (annotated)	1.65

To examine the diagnostic spectrum represented in the corpus, we performed a semi-automated clinical entity normalization on all user-generated text labeled as `medical_needs`. This process aimed to map free-text mentions of illnesses to standardized diagnostic categories defined by the International Classification of Diseases, 10th Revision (ICD-10). Specifically, we used a prompt-based large language model (GPT-4o) to extract candidate clinical expressions and infer the most likely ICD-10 code for each. Prompts were designed to simulate clinical reasoning and included few-shot examples to improve consistency across diverse linguistic inputs. For quality assurance, all mappings flagged as low confidence by the model were manually reviewed by an author. The final normalized corpus comprises 582 unique disease-related expressions, mapped to 16 distinct ICD-10 codes. The five most prevalent diagnostic categories were: depressive episode (77 cases), anxiety disorder (67), non-organic insomnia (32), obsessive-compulsive disorder (15), and bipolar affective disorder (12). This broad diagnostic coverage enables downstream models to be evaluated across a realistic range of psychiatric, psychological, and comorbid clinical scenarios, enhancing the ecological validity and generalizability of model performance.

To examine the distribution of emotional expressions within the emotional_needs dataset, this study conducted a semi-automated fine-grained emotion classification on the subset labeled E-N-FEEL. This approach combined large language model – assisted extraction with manual validation to balance scalability and annotation reliability. Specifically, this study employed GPT-4o guided by structured prompts informed by psychological context to automatically extract emotion-related terms and map them to the most relevant of Plutchik's eight basic emotions (Plutchik, 1980), based on semantic proximity and emotional taxonomy. A stratified sample of extracted terms was manually reviewed and validated by a domain expert to ensure semantic accuracy and category alignment. As a result of this process, the five most frequently occurring emotion-related terms were: anxious (275 cases), fearful (81), depressed (74), afraid (36), and worried (32), reflecting the predominance of negative affective states in mental health consultations. To visualize the emotional distribution, I generated a multi-colored word cloud using Python's matplotlib and word cloud libraries. Terms were grouped and color-coded based on their associated Plutchik category to enhance interpretability (see Figure 4-7). Fear and sadness emerged as the dominant emotional categories, underscoring the emotional burden frequently expressed in consumer mental health questions.

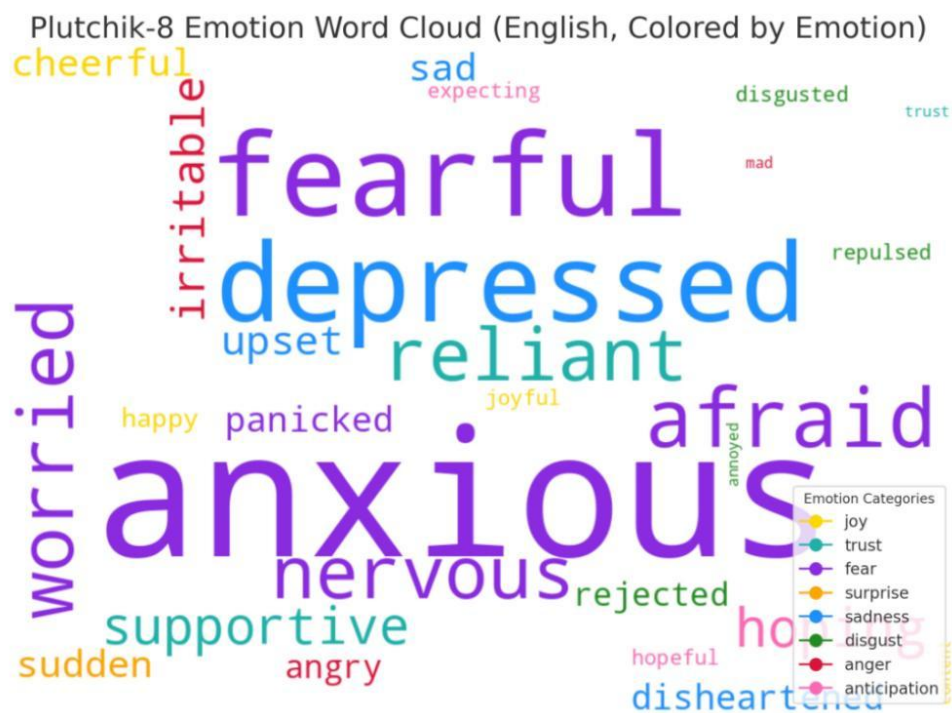


Figure 4-7. Emotion Word Cloud Categorized by Plutchik's Eight Basic Emotions in E-N-FEEL Entries. Each category is color-coded, with a legend indicating the mapping between colors and emotion types. Note: Original text translated from Chinese into English for clarity.

To better understand the subjective perspectives expressed in the E-N-VIEW subtype of emotional needs, this study conducted a semi-automated thematic classification of each need_text entry. This process combined GPT-4o assisted theme-based labeling with manual validation to balance processing efficiency and interpretive accuracy. During the manual evaluation phase, researcher observed that GPT-4o frequently misclassified conceptually specific instances into the residual Other category. To mitigate this issue and improve thematic reliability, this study performed focused reannotation of all entries initially assigned to Other. As a result (see Figure 4-8), the corpus was organized into six distinct themes: (a) Opinions on treatment approaches: Reflections on medical interventions, medications, or treatment plans (249 cases). (b) Opinions on disease etiology: Subjective interpretations regarding the causes or perceived triggers of the condition (23). (c) Opinions on diagnostic results: Judgments about the accuracy, credibility, or meaning of diagnostic outcomes (31). (d) Opinions on disease prognosis: Views related to expectations or concerns about the future course of the wellness (9). (e) Cognitive expressions: Statements reflecting patients' beliefs, reasoning patterns, or interpretations about themselves, others, or the broader world (104). (f) Other: Expressions that could not be clearly assigned to the above categories (92).

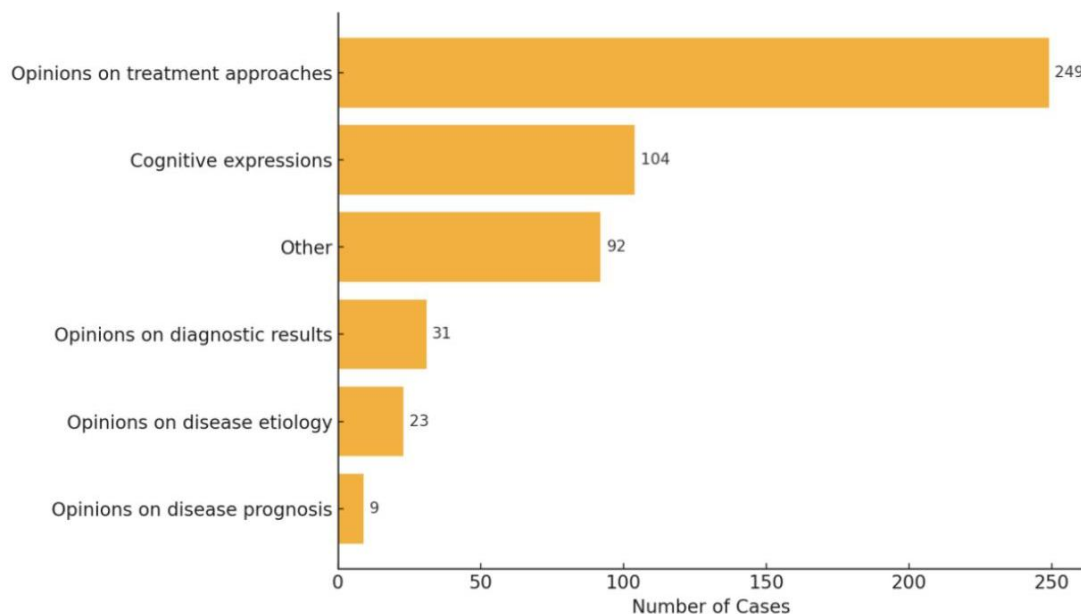


Figure 4-8. Thematic Distribution of Subjective Perspectives in E-N-VIEW Entries.

Comparative positioning. A comparative analysis of our proposed CHQA-MedEmo dataset against existing annotated CHQA datasets is presented in Table 4-9. As shown in Table 4-9, existing CHQA datasets differ widely in data sources, languages, and annotation objects. Most prior datasets are derived from health websites, PubMed abstracts, or self-generated corpora, often focusing on medical information needs alone. A few recent resources, such as CHQ-SocioEmo (2023) and PsyQA (2021), incorporate emotional dimensions, but they remain limited either in domain coverage or annotation depth. In contrast, CHQA-MedEmo offers three distinctive contributions:

- **Authentic consultation data:** Unlike self-generated or abstract-based corpora, CHQA-MedEmo is grounded in real patient-physician interactions from an online health consultation platform, ensuring realism and contextual richness.
- **Dual-layer annotation:** Both medical needs (MN, MNC) and emotional needs (EN, ENC) are systematically annotated, along with physician answer strategies (ANS)-a scope rarely integrated in prior datasets.
- **Chinese language coverage:** While most existing CHQA datasets are English-dominant, CHQA-MedEmo expands the resource space to Chinese consultations, addressing the need for culturally and linguistically diverse benchmarks.

Table 4-9. Comparison Between CHQA-MedEmo and Other CHQA Dataset

Dataset	Data source	Lang.	Annotation objects				
			MN	MNC	EN	ENC	ANS
GARD (2014)	Health website	EN	✓	✓	✓		
CHQA-email (2018)	Health website	EN	✓	✓			
MedRedQA (2023)	Reddit posts	EN	✓				✓
RedHot (2023)	Reddit posts	EN	✓	✓			
CHQ-SocioEmo (2023)	Online health communities	EN			✓	✓	
MentalQA(2024)	Online mental health communities	AR	✓				✓
MHQA (2025)	PubMed abstracts	EN	✓				✓
PsyQA (2021)	Online psychology service	CH			✓		✓
ESConv (2021)	Self-generated (crowd-sourced)	EN			✓	✓	✓
CAMS (2022)	Reddit posts	EN			✓	✓	
ExTES (2023)	Self-generated (LLM)	EN				✓	✓
EHD (2024)	Self-generated (LLM)	EN					✓
CHQA-MedEmo (This Study)	Online health consultation platform	CH	✓	✓	✓	✓	✓

Note: EN, CH are short for English and Chinese, and MN, MNC, EN, ENC, ANS are short for medical information need (e.g., diagnosis), medical information need related context (e.g., symptom), emotional support need (e.g., emotion state), emotional support need related context (emotion cause), answer (e.g., response strategy)

4.7 Chapter Summary

Chapter 4 developed the situation awareness foundation of this study by empirically analysing real-world OHC environments and synthesising insights into actionable design implications. The chapter began by defining the research environment within the Chinese OHC context, focusing on Haodf Online as a representative platform. Through discourse-level content analysis and workflow examination, it explored how consumers express informational and emotional needs in authentic consultation settings, and how physicians interpret and respond to such multi-intent questions. These analyses collectively established a nuanced understanding of the communicative, cognitive, and affective dynamics underlying consumer health interactions.

The key contribution of this chapter lies in translating these empirical findings into two sets of design implications that directly inform the artefact development process. First, the content analysis of Haodf Online records

provided quantitative and structural evidence for constructing a benchmark dataset that captures both informational and emotional dimensions of consumer health questions. Second, the workflow analysis revealed critical insights into the system architecture design, highlighting the need to model both cognitive reasoning and empathetic communication through NSI and RST. Together, these insights advance the methodological transition from theoretical framing to system conceptualisation.

By the end of this chapter, the problem awareness and solution suggestion stages of the DSR Cycle 1 were fully realised. The outputs, namely, the benchmark design principles, discourse framework, and workflow-based architectural insights, form a robust foundation for the next stage of the study. Chapter 5 builds upon these empirical and conceptual outcomes to translate them into the system design and development of the Med-Emo CHQA architecture, thereby initiating Cycle 2: System Design, Development, and Evaluation.

Chapter 5 System Design and Development

5.1 Chapter Introduction

Chapter 5 presents the design of the Med-Emo CHQA system, translating the theoretical foundations and problem definition established in earlier chapters into a structured system specification. Whereas Chapters 4 identified the dual challenge of addressing medical informational needs and emotional support in consumer health queries, this chapter is concerned with developing a coherent architectural framework capable of supporting a unified question–answering process. The aim is to formalise the system’s functional components, define its operational tasks, and establish the mechanisms through which user queries are interpreted and transformed into meaningful responses.

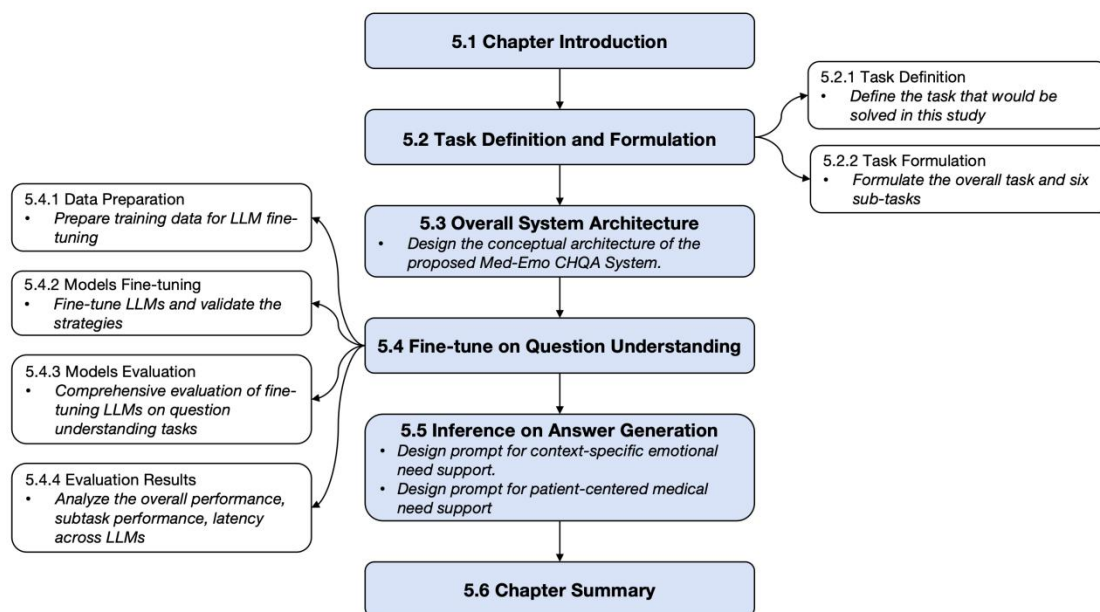


Figure 5-1. Structure of Chapter 5.

As shown in Figure 5-1, this chapter begins by outlining the overall system objectives, defining the overall task and decomposing it into six sub-tasks (Section 5.2), clarifying the role of question understanding and answer generation within the integrated framework. Section 5.3 presents the overall system architecture, outlining how the Med-Emo CHQA pipeline aligns implicit reasoning with explicit response generation. Section 5.4 details the fine-tuning process for question understanding, including data preparation, model adaptation, and subtask-level evaluation. Section 5.5 designs the

inference strategy for answer generation, specifying prompt mechanisms that guide both clinical relevance and patient-centred empathy.

5.2 Task Definition and Formulation

5.2.1 Task Definition

Drawing on the empirical analysis in Section 4.3 of real-world OHC data, it is observed that CHQs often express both medical informational needs and emotional support needs, and that these two dimensions are tightly interdependent rather than separable.

On the one hand, elements in the emotional layer, such as the user’s view may encode implicit medical questions and treatment preferences. For example: “Dailixin is so addictive that I’m now afraid to resume it.” This statement implicitly seeks information about whether Dailixin is addictive and signals a preference to avoid resuming it—constraints that should be accounted for when offering treatment advice.

On the other hand, expressions of medical need frequently carry implicit affective states. Consider: “The psychiatrist kept prescribing substitute medications to alleviate my symptoms, but after trying many different drugs, none proved effective and all had significant side effects—in fact, I ended up switching medications almost weekly.” Beyond its informational content, this narrative conveys distress and fatigue arising from repeated, ineffective substitutions and adverse effects.

Therefore, when interpreting and responding to these dual demands, they cannot be treated as independent and processed in parallel. This study introduces Joint Medical-Emotional Question Answering (JMEQA), a novel task specifically designed to jointly understand and generate responses that address both medical informational needs and emotional support needs in complex consumer health queries.

In contrast to traditional Medical Question Answering (MQA), which primarily emphasizes factual correctness (Ben Abacha et al., 2017; Demner-Fushman et al., 2019; Welivita et al., 2023), or Emotional Support Conversation (ESC), which focuses on affective support but typically lacks clinical decision-making capabilities (Liu et al., 2021; Zheng et al., 2023), the recently proposed Knowledge-Grounded Empathy-Enhanced Medical Dialogue (KGE-

MD) does consider both medical informativeness and empathy (Wu et al., 2024). However, its essence remains that of an enhanced Emotional Support Conversation task, where medical knowledge is leveraged only to strengthen the provision of emotional support, rather than to fully integrate medical reasoning with empathetic communication.

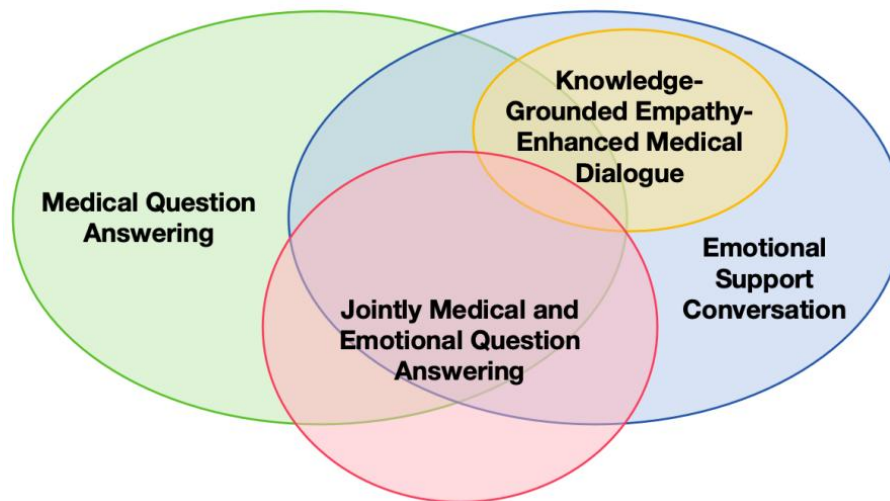


Figure 5-2. Illustration of Relationship among JMEQA Task and Other Three Tasks.

JMEQA explicitly requires the integration of both dimensions, but not as a simple additive combination. Instead, by modeling the interactions between medical informational needs and emotional support needs, JMEQA seeks to achieve two interrelated objectives:

- **Patient-Centered Medical Need Support.** Beyond providing factually correct medical information, JMEQA emphasizes tailoring responses to the individual’s unique situation, concerns, and treatment preferences, thereby aligning clinical reliability with patient-centered care. In this study, patient-centered medical need support is achieved by explicitly linking users’ medical needs with their corresponding view.
- **Context-Specific Emotional Need Support.** Emotional reassurance is not offered in isolation but is grounded in the medical context of the query. This approach ensures that empathetic responses are personalized rather than templated. In this study, context-specific emotional need support is realized by explicitly associating users’ emotional needs with their corresponding medical background.

This dual-objective design reflects realistic use cases in online health consultation platforms, where patients simultaneously seek accurate medical

guidance and empathetic acknowledgment of their struggles. By positioning itself at the intersection of clinical rigor and human-centered empathy, JMEQA defines a new paradigm for question answering in healthcare. It extends beyond traditional task boundaries to integrate trustworthy medical knowledge delivery with emotionally attuned support, thereby offering a more holistic, context-aware response framework for consumer health applications.

5.2.2 Task Formulation

Formally, Let $Q_i = \{q_1, q_2, \dots, q_i\}$ be a consumer health query. For each Q_i , the object is to generate a response R_i that jointly response both its medical informational needs and emotional support needs refer to their associated contextual information. Grounded in RST and Pounds (2011)' AFCE, further supported by the empirical analysis of real-world consumer health queries in Section 4.2, JMEQA is decomposed into six sub-tasks (see Figure 5-3):

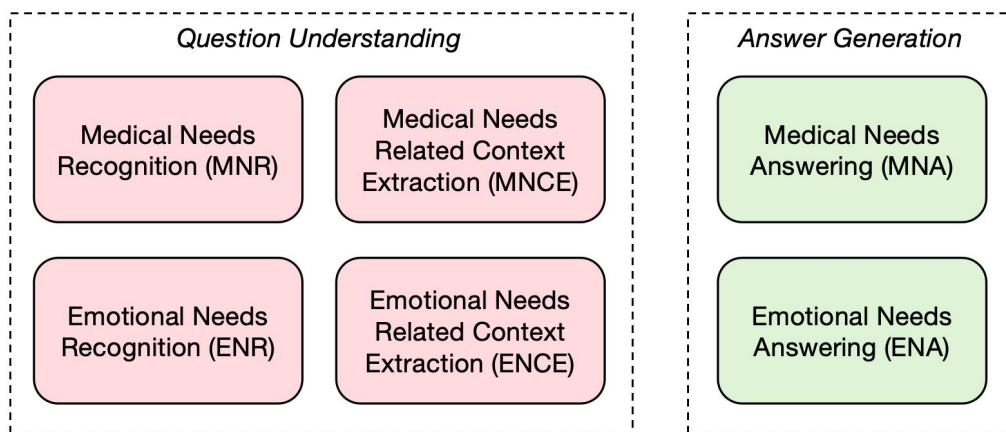


Figure 5-3. Six Sub-tasks of JMEQA.

Emotional Needs Recognition (ENR). ENR locates spans that convey emotional support requests and classifies them as either expressions of feeling or viewpoint, based on Pounds (2011)' appraisal model. Accurate identification of these emotional nuclei enables dialogue systems to adapt their counseling tone appropriately.

Formally, identify or generate all emotional-need span $e_i \subset Q_i$, and assign each to one of two categories $d_i \in \{\text{View, Feel}\}$, corresponding to viewpoint expressions or feeling expressions.

Emotional Needs-related Context Extraction (ENCE). For each emotional need, ENCE extracts causal satellites that describe the underlying triggers of

the expressed emotions. Incorporating such causal information has been shown to enhance empathetic response generation (Li et al., 2021; Wang et al., 2021).

Formally, for each pair (e_i, d_i) , extract or generate every related context span $f_i \subset Q_i$, and label it uniformly as CAUSE.

Emotional Needs Answering (ENA). ENA is formulated as a conditional generation task that conditions on emotional needs and their causal contexts to generate context-specific emotional need support. Its design is grounded in two complementary theoretical frameworks: (a) Pounds' (2011) AFCE, which conceptualizes empathy as a multi-level cognitive-affective process involving three communicative strategies: acknowledgment, alignment, and alignment acceptance. (b) The NURSE protocol, which operationalizes empathy into a sequence of verbal actions that healthcare professionals employ to provide emotional needs support: naming \rightarrow understanding \rightarrow respecting \rightarrow supporting \rightarrow exploring.

Specifically, this study interprets acknowledgment as Naming, alignment as Understanding, and alignment acceptance as either Respect or Support, depending on the type of emotional nucleus being addressed. For View-type emotions, which express patients' cognitive stances or opinions about their illness or treatment, the response should aim to validate the speaker's reasoning rather than their affective state. Hence, the corresponding communicative strategies are: Naming \rightarrow Understanding \rightarrow Respect. This sequence acknowledges the patient's viewpoint, shows comprehension of its rationale, and conveys respect for the patient's evaluative judgment. For Feeling-type emotions, which convey affective distress or subjective suffering, the response must provide emotional containment and reassurance. Accordingly, the corresponding strategies are: Naming \rightarrow Understanding \rightarrow Support. This sequence first labels the emotion, then demonstrates understanding of its underlying causes, and finally provides supportive language that normalizes the patient's affective experience.

Formally, for each triplet (e_i, d_i, f_i) , generate a response R_{e_i} guided by the strategy set M_{d_i} :

$$M_{d_i} = \begin{cases} \{\text{Naming } (e_i), \text{ Understanding } (f_i), \text{ Respect}\}, & \text{if } d_i = \text{View} \\ \{\text{Naming } (e_i), \text{ Understanding } (f_i), \text{ Support}\}, & \text{if } d_i = \text{Feel} \end{cases}$$

$$R_{e_i} = \text{Generate } (e_i, d_i, f_i \mid M_{d_i})$$

Medical Needs Recognition (MNR). In RST terms, MNR identifies the nucleus of a consumer health query: the patient’s explicit informational request. Specifically, MNR identifies all spans that express medical informational requests and assigns them to one of five clinically motivated categories: general medical information, etiology, diagnosis, treatment, or prognosis. Detecting these nuclei provides downstream systems with a structured understanding of the patient’s evidence-based medical needs.

Formally, identify all medical-need spans $m_i \subset Q_i$, and assign each span to one of five identified categories $c_i \in \{\text{General Medical Information, Etiology, Diagnosis, Treatment, Prognosis}\}$, according to their request medical information support category.

Medical Needs-related Context Extraction (MNCE). For each identified medical need, MNCE extracts supporting satellites labeled as elaboration, background, or condition, following the nucleus–satellite relations defined by RST. These supporting spans supply information on symptom progression, prior treatments, or prerequisite conditions that are critical for tailoring clinical advice. Each satellite is explicitly linked to its medical nucleus via RST labels.

Formally, for each pair (m_i, c_i) , extract every related context span $s_i \subset Q_i$, and label it with a identified relation $r_i \in \{\text{Elaboration, Background, Condition}\}$, where condition is applied only if $c_i = \text{Treatment}$.

Medical Needs Answering (MNA). Similar to ENA, MNA is formulated as a conditional generation task, producing clinically grounded and guideline-consistent answers to the identified medical nuclei and their contextual satellites derived from MNR and MNCE. This task emphasizes factual correctness, and personalized response, ensuring that the medical nucleus is addressed directly, while contextual satellites refine and personalize the response. In practice, MNA integrates domain-specific knowledge sources such as clinical guidelines, best-practice protocols, and evidence-based medical literature to ensure accuracy and applicability.

Each response is generated through a two-step reasoning process designed to emulate expert-level clinical reasoning: (a) Condition appraisal. For each identified medical need, this task first examines any patient preferences or subjective views encoded within the Condition field. These are appraised against authoritative clinical guidance to determine their reasonableness and safety. Then generates a brief justification for each appraisal, distinguishing between clinically acceptable preferences (e.g., preferring gradual medication tapering) and contraindicated or risky preferences (e.g., abrupt discontinuation of antidepressants). (b) Contextualised answer generation. Based on the medical nucleus, this task integrates relevant contextual satellites, including Elaboration, Background, and Condition (subjective preferences deemed acceptable), to produce a guideline-aligned, and personalized response.

Formally, for each tuple (m_i, c_i, s_i, r_i) , generates a response R_{m_i} guided by the strategy set M_{c_i} .

$$M_{c_i} = \begin{cases} \{\text{Appraise (CON), Integrate (ELA, BACK, Appraised CON)}\}, & \text{if } c_i = \text{TREAT} \\ \{\text{Integrate (ELA, BACK)}\}, & \text{if } c_i = \{\text{GMI, ETI, DIA, PROG}\} \end{cases}$$

$$R_{m_i} = \text{Generate}(m_i, c_i, s_i, r_i \mid M_{c_i})$$

5.3 Overall System Architecture

To operationalise the six subtasks defined in Section 5.2, this study designs Med-Emo CHQA, an end-to-end architecture that embeds a symbolic question-understanding enhancement mechanism inside a Large Language Model. As shown in Figure 5-4, the architecture is conceptualised as an understanding-enhanced end-to-end framework that directly maps user queries to guideline-consistent and empathetic responses, while making the underlying reasoning process transparent enough to support evaluation and design learning.

From an implementation perspective, the Med-Emo CHQA system behaves as a single unified reasoning engine rather than a traditional multi-module pipeline: a consumer health query is processed through one inference call to an LLM-based service. From a design perspective, however, the processing flow follows the HSD principle outlined in Chapter 3 and can be interpreted as comprising two continuous reasoning stages that together instantiate the proposed neural-symbolic integration approach:

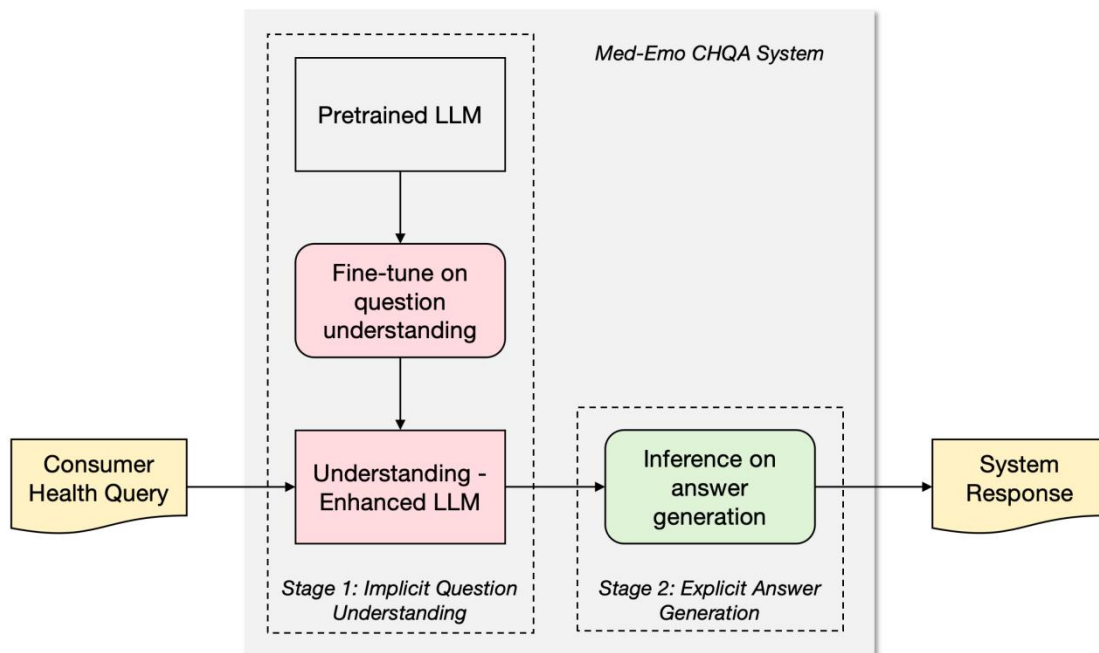


Figure 5-4. Conceptual Architecture of the Proposed Med-Emo CHQA System.

The system is executed as a single end-to-end process that implicitly encompasses two reasoning stages within the LLM: (a) question understanding and (b) answer generation. These stages are not separate modules but represent the internal cognitive flow of the model during inference. The outer boundary highlights the complete Med-Emo CHQA system as a unified architecture.

Stage 1: Fine-Tune on Question Understanding. Stage 1 corresponds to the question-understanding layer and directly operationalises the symbolic representation developed for Sub-RQ1. In this stage, the base LLM is supervised fine-tuned on the CHQA-MedEmo corpus, where each Consumer Health Question is annotated with an RST-based discourse structure tree. During fine-tuning, the model learns to perform four comprehension-oriented subtasks, MNR, MNCE, ENR, and ENCR, by aligning its latent representations with these structured annotations.

This stage unfolds entirely within the latent space of the LLM: the intermediate predictions of needs and contextual spans are not exposed as explicit outputs at inference time. Instead, they act as symbolic anchors that shape the model’s internal representations of CHQs and encode the joint informational-emotional structure of consumer health queries. In line with the NSI perspective, the symbolic discourse and intent labels are thus embedded into the LLM’s parameter space, enhancing its contextual awareness and alignment with real-world online consultation patterns without adding extra runtime components.

Stage 2: Inference on Answer Generation. Stage 2 corresponds to the answer-generation layer and implements the integration strategies studied under Sub-RQ2. Here, the LLM fine-tuned for question understanding in Stage 1 is reused in a frozen state, and task-specific constraint-based prompts are designed to bridge its learnt structured representations and open-ended generation objectives. In this stage, the model performs two response-oriented subtasks, MNA and ENA, within a single inference step.

The prompting templates explicitly encode the dual objectives of the Med-Emo CHQA system: (a) delivery of patient-centered, medically accurate, guideline-consistent information; and (b) provision of context-sensitive emotional support. By conditioning the frozen understanding-enhanced LLM on these structured instructions, the CHQA task is reformulated as a prompt-induced inference process, where the model leverages its internal representation of medical and emotional needs, acquired during supervised fine-tuning, to generate responses that are contextually coherent, empathetic, and clinically faithful.

By embedding the symbolic understanding-enhancement mechanism inside an end-to-end LLM service, the Med-Emo CHQA architecture balances conceptual interpretability (through the two-stage neural-symbolic reasoning design) with practical efficiency (through single-pass inference). It advances conventional CHQA paradigms by replacing explicit multi-module pipelines with an internally aligned, cognitively coherent LLM whose behaviour can still be analysed along distinct understanding and generation phases.

Moreover, by fine-tuning the model only on the question-understanding subtasks, rather than directly on answer generation as in most existing LLM-based CHQA systems, the proposed architecture substantially reduces annotation costs and enhances clinical feasibility. This design choice reflects the DSR principle of rigour-relevance balance: it leverages richly annotated, high-certainty sub-tasks (MNR, MNCE, ENR, ENCR) to achieve system-level gains in Accuracy, Personalisation, and Empathy.

5.4 Fine-Tune on Question Understanding

This section details the fine-tuning process designed to enhance question understanding within the proposed CHQA architecture. As illustrated in Figure 5-5, the procedure consists of three sequential stages: (a) data preparation, where the CHQA-MedEmo dataset is structured into training,

validation, and testing splits; (b) model fine-tuning, in which the pre-trained LLM is adapted to interpret multi-intent health queries through supervised learning; and (c) evaluation, where the fine-tuned model is benchmarked against baseline outputs to assess improvements in medical coverage, personalisation, and empathetic alignment. This pipeline operationalizes the implicit question understanding mechanism underpinning the overall system design.

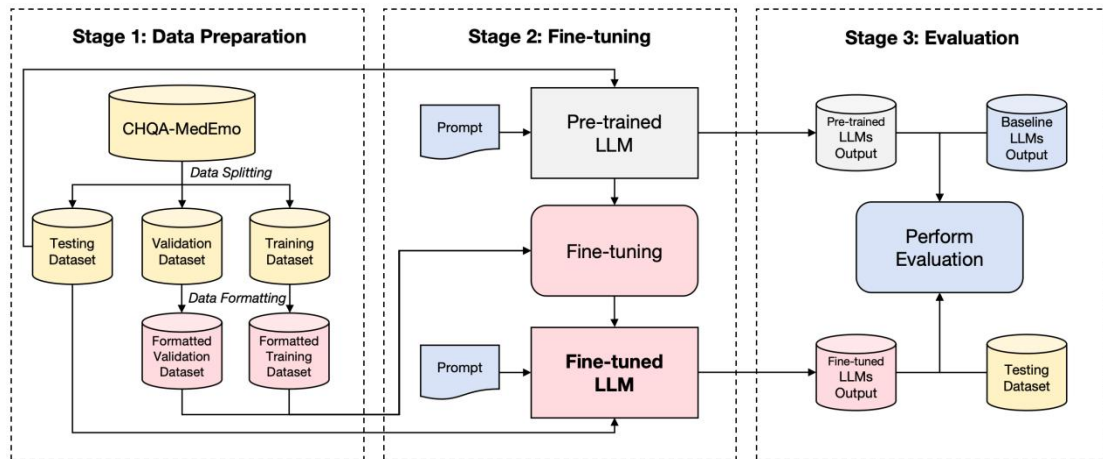


Figure 5-5. Fine-tuning Process on Question Understanding.

5.4.1 Data Preparation

To ensure both effective model training and robust evaluation, this study employed the CHQA-MedEmo dataset (presented in Section 4.5), which comprises annotated instances of medical informational needs, emotional support needs, and their associated contextual relations. To prevent temporal leakage and enhance generalizability across evolving online health communication practices, the dataset was stratified by both year of collection and task-label distribution prior to partitioning. A stratified 60-20-20 split was adopted, yielding 422 samples for training, 140 for validation, and 141 for testing. This stratification strategy preserved the dataset’s temporal diversity while maintaining balanced representation across the four subtasks - MNR, MNCE, ENR, and ENCE. The test set was held out exclusively for final evaluation, providing a reliable estimate of the system’s out-of-sample generalization performance and preventing information contamination between training and inference stages.

Each instance paired a consumer health query with its corresponding structured representation, thereby aligning the model’s generation behaviour

with the schema defined in Section 4.3. The 422 training and 140 validation samples were converted into the ChatML JSONL format.

```
{
  "messages": [
    {
      "role": "user",
      "content": "You are a professional clinical assistant. Please extract structured medical and emotional support needs, along with relevant contextual information, from the user query below. Return the results in standardized JSON format.\n\nUser query: ....."
    },
    {
      "role": "assistant",
      "content": "\n\n\"medical_needs\": [\n  {\n    \"need_id\": \"M1\",\n    \"need_type\": \"M-N-TREAT\",\n    \"need_text\": \"How can I overcome my fear of hepatitis B?\",\n    \"related_contexts\": [\n      {\n        \"relation_type\": \"M-C-ELA\",\n        \"context_text\": \".....\n      }\n    ]\n  }\n]"
    }
  ]
}
```

Figure 5-6. An Example of JSONL Training Data in ChatML Schema.

5.4.2 Models Fine-tuning

Backbone Candidates. To ensure both diversity and feasibility, this study selected three representative LLMs that differ in architecture and accessibility while supporting efficient fine-tuning (see 5-1):

- **GPT-4o (OpenAI, 2024):** As OpenAI’s flagship multimodal model, GPT-4o demonstrates strong reasoning and cross-domain generalization, with an Massive Multitask Language Understanding (MMLU) benchmark score of 88.7%. Its robust alignment and contextual comprehension make it an appropriate proprietary reference for assessing the upper bound in joint medical–emotional question understanding.
- **Qwen2.5-72B (Yang et al., 2024):** This 72-billion-parameter dense transformer, trained on over 15 trillion tokens, achieves a balanced trade-off between performance and efficiency. Its strong instruction-following and multilingual QA capabilities make it a high-capacity open-source baseline capable of domain adaptation via QLoRA-based parameter-efficient fine-tuning.
- **Qwen3-32B (Yang et al., 2025):** A medium-scale successor in the Qwen3 family, trained on 20 trillion tokens of multilingual and code-augmented data, it integrates improved positional encoding and safety-aligned instruction tuning. The model maintains reasoning consistency while offering substantially lower computational cost, serving as a practical yet powerful alternative for academic-scale fine-tuning.

Table 5-1. Base LLMs Adopted for Fine-tuning and Their Specifications.

	Qwen2.5-72B	Qwen3-32B	GPT-4o
# Architecture	Dense	Dense	Dense (hybrid MoE, estimated)
# Total Params	72B	32B	Undisclosed (est. 1–2 T)
# Activated Params	72B	32B	Undisclosed (est. 1–2 T)
# Licensing	Open Source	Open Source	Proprietary
# Model Category	General	General	General

Note: Estimates parameters of GPT-4o based on OpenAI’s public model descriptions (2024–2025).

Although ultra-large open-weight models such as DeepSeek-V3, DeepSeek-R1, Qwen2.5-Max, and Qwen3-235B-A22B theoretically support Parameter-Efficient Fine-Tuning (PEFT), their enormous model scale and hardware requirements render Direct adaptation impractical under typical academic or moderate cloud budgets. Even PEFT approaches such as LoRA or QLoRA require multi-GPU distributed training and high-bandwidth memory configurations, which exceed the computational limits of this research setting.

Consequently, this study limits fine-tuning experiments to smaller yet representative open-source models (e.g., Qwen2.5-72B, Qwen3-32B) that balance architectural comparability with computational feasibility, while employing proprietary models (e.g., GPT-4o) through API-based incremental fine-tuning for complementary validation. This combination ensures that the proposed fine-tuning framework remains both methodologically rigorous and operationally scalable within the constraints of academic research infrastructure.

Fine-Tuning Strategies. To adapt general-purpose LLMs to the domain-specific task of joint medical-emotional question understanding, this study employs a Supervised Fine-Tuning (SFT) paradigm. Besides, rather than performing full-parameter fine-tuning, a partial and parameter-efficient adaptation is adopted to ensure both the feasibility of deployment and the continuity of downstream answer-generation tasks that built on the same base models. This design minimizes computational cost and mitigates catastrophic forgetting, while preserving the pretrained model’s general linguistic and reasoning capabilities. Most importantly, this configuration allows the inference stage of answer generation to be directly executed on the same fine-tuned models, thereby ensuring the operational feasibility and architectural consistency of the proposed system.

Specifically, QLoRA-based PEFT is applied to open-source backbones such as Qwen2.5-72B and Qwen3-32B. QLoRA enables low-rank adaptation on quantized model weights, allowing high-capacity models to be fine-tuned efficiently on limited hardware without sacrificing expressive power. This approach quantizes the frozen base weights to 4-bit precision and injects trainable low-rank adapters into each transformer layer, allowing only the adapter parameters to be updated during training. It is particularly suitable for multi-task alignment settings, where both medical reasoning and empathy-related understanding must be retained across tasks. Panel (a) in Figure 5-7 illustrates the open-source QLoRA-based PEFT, in which lightweight adapters are externally attached to the frozen base model.

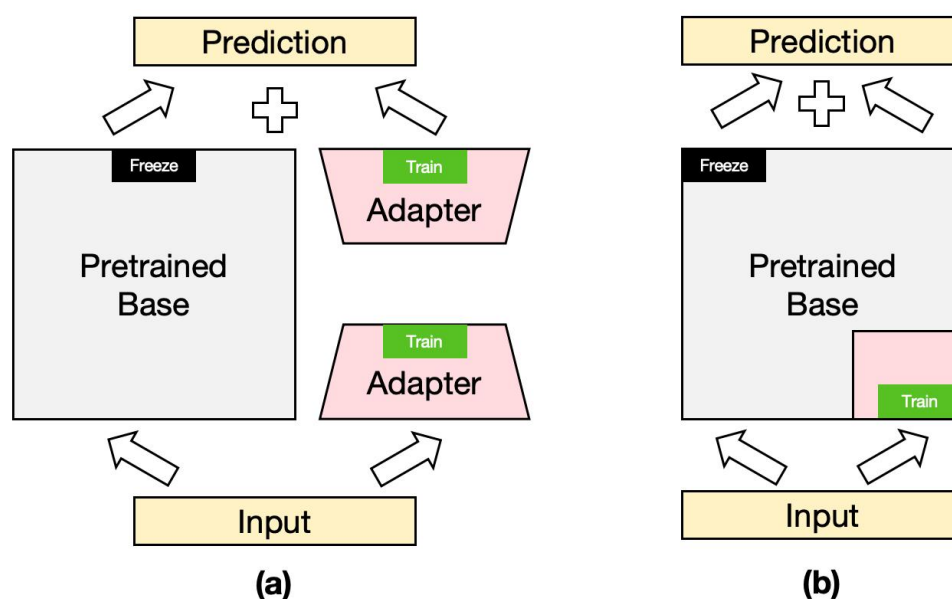


Figure 5-7. Illustration of (a) PEFT (b) Limited-Scope Adaptation. These two diagrams collectively conceptualize the methodological symmetry between proprietary and open-source partial fine-tuning paradigms, both update minimal parameters to achieve efficient domain alignment, yet differ in openness and implementation transparency.

For the proprietary models accessed through OpenAI’s fine-tuning API (e.g., GPT-4o), the internal mechanism remains undisclosed. However, based on official documentation and empirical characteristics—such as training efficiency, response latency, and model stability—the fine-tuning process can be reasonably categorized as a form of Hybrid Incremental Supervised Fine-Tuning (HISFT). This closed-source strategy appears to combine limited-scope parameter updates with incremental adaptation of higher-level model representations, thereby achieving efficient customization without exposing or retraining full model weights (OpenAI, 2025). Accordingly, the fine-tuned GPT-4o variant in this study is treated as a black-box instance of

incremental supervised adaptation, functionally parallel to PEFT but differing in transparency and controllability. Panel (b) in Figure 5-7 depicts the closed-source limited-scope adaptation used by proprietary models such as GPT-4o, where only a small subset of internal parameters is updated.

Fine-Tuning of GPT-4o. The fine-tuning was conducted via OpenAI’s managed fine-tuning API, in which all training hyper-parameters were predefined and controlled by the platform. The configuration included a batch size of 1, a learning rate multiplier of 2, and three training epochs. This managed approach provided an accessible, standardized, and reproducible pathway to adapt a frontier closed-source LLM to the joint medical-emotional understanding task. The fine-tuning job completed in approximately 51 minutes and 26 seconds under default optimization settings. The job completed in approximately 51 minutes 26 seconds. Figure 5-8 illustrates the training loss trajectory of GPT-4o fine-tuning. The sharp decline during early iterations, followed by stabilization at a low plateau, confirms successful convergence and the absence of training collapse.

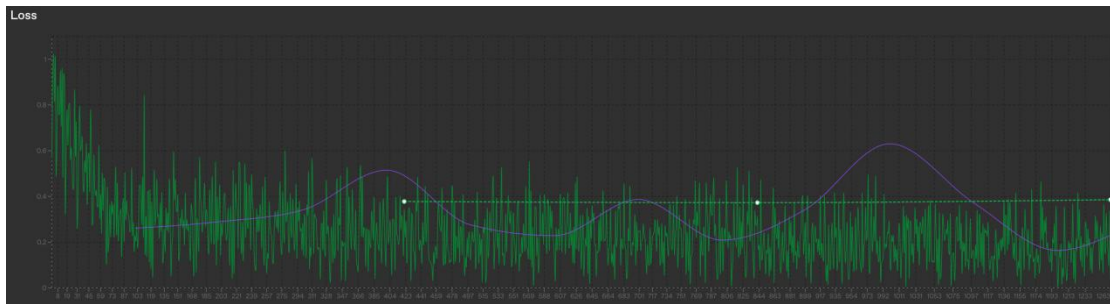


Figure 5-8. Training Loss Curve of GPT-4o Fine-Tuning.

Complementing the loss analysis, Figure 5-9 presents the accuracy trajectory of GPT-4o instruction-tuning. Accuracy improves rapidly in the initial epochs and stabilizes above 90%, suggesting that the model acquired robust understanding of both medical and emotional subtasks prior to evaluation on the held-out test set.

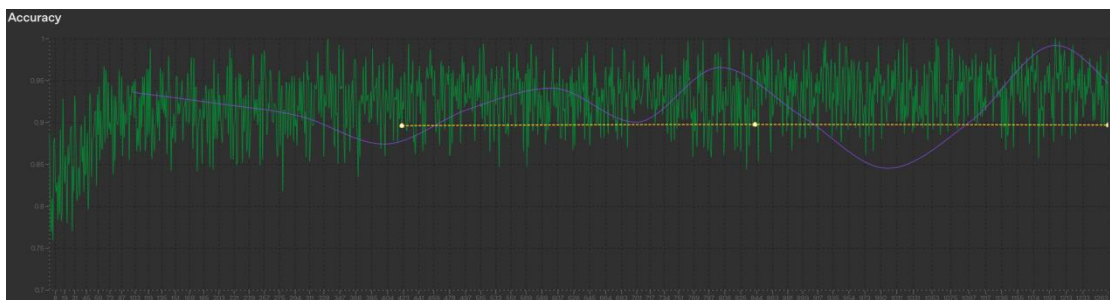


Figure 5-9. Training Accuracy Curve of GPT-4o Fine-Tuning.

Fine-tuning of Qwen2.5-72B. Training was performed on Alibaba Cloud Bairuan using the High-Efficiency SFT profile, which enables optimized GPU resource scheduling for large-scale parameter-efficient adaptation. The hyper-parameters were configured as: LoRA rank = 8, $\alpha = 16$, batch size = 16, learning rate = $3e^{-4}$. The process completed in 31 minutes 40 seconds, demonstrating the practicality of QLoRA for adapting 70B-class models under constrained computational budgets. As shown in Figure 5-10, the training loss decreased sharply during early iterations and stabilized around 0.2, indicating smooth convergence and the absence of overfitting. This confirms that QLoRA can effectively fine-tune high-capacity open-source models for the CHQA-MedEmo dataset while maintaining training stability.

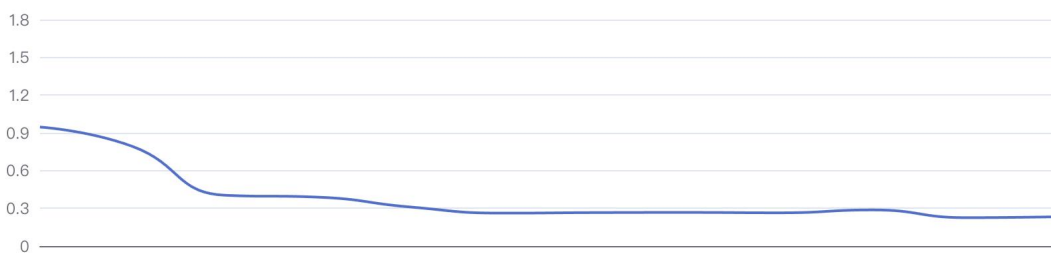


Figure 5-10. Training Loss Curve of Qwen2.5-72B Fine-Tuning with QLoRA.

Fine-Tuning of Qwen3-32B. Training was again executed on Alibaba Cloud Bairuan under the High-Efficiency SFT profile to ensure optimal GPU utilization. The hyper-parameter configuration mirrored that of Qwen2.5-72B (LoRA rank = 8, $\alpha = 16$, batch size = 16, learning rate = $3e^{-4}$). The fine-tuning completed in 29 minutes 22 seconds, illustrating the efficiency of QLoRA for 30B-scale models. As presented in Figure 5-11, the loss function exhibited a rapid initial decline followed by stabilization around 0.2, reflecting successful convergence and effective adaptation to the CHQA-MedEmo dataset without overfitting.

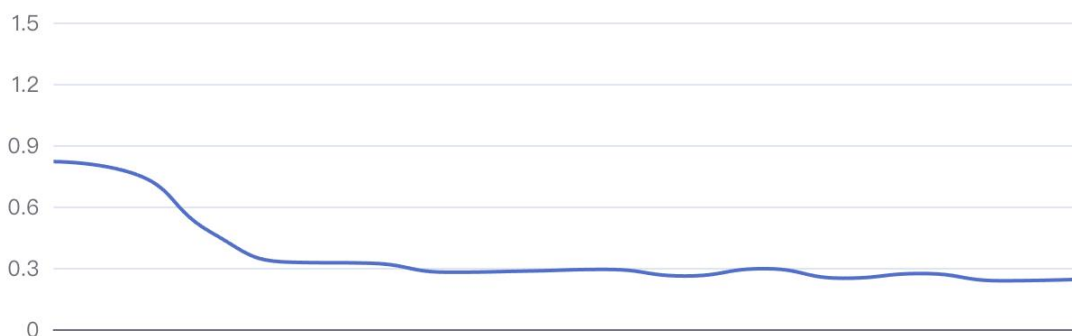


Figure 5-11. Training Loss Curve of Qwen3-32B Fine-Tuning with QLoRA.

5.4.3 Models Evaluation

Other Models Adopted for Evaluation. To ensure a comprehensive and fair assessment, this study additionally evaluated four non-fine-tuned baseline models that represent diverse architectures, parameter scales, and accessibility tiers (see Table 5-2). These baselines were included to establish a prompt-based performance benchmark, allowing comparison between fine-tuned models (GPT-4o, Qwen2.5-72B, Qwen3-32B) and few-shot inference under identical prompting conditions. Each model was tested solely through prompt-driven inference, without any fine-tuning or parameter updates, thereby reflecting its inherent generalization capacity across medical and emotional understanding tasks. These modules include:

- **DeepSeek-V3 (DeepSeek-AI, 2025a):** This 671-billion-parameter Mixture-of-Experts (MoE) model activates only 37 billion parameters per token, optimizing computational efficiency. Trained on 14.8 trillion tokens, DeepSeek-V3 matches or surpasses the performance of many closed-source models while significantly reducing training costs.
- **DeepSeek-R1 (DeepSeek AI, 2025b):** DeepSeek-R1 is a 260-billion-parameter dense language model trained entirely from scratch on a diverse 8.1 trillion token corpus. It features advanced instruction-following capabilities and excels in general-purpose reasoning, achieving strong results across academic, reasoning, and multilingual benchmarks.
- **Qwen3-235B-A22B (Yang et al., 2025):** Qwen3-235B is the largest model in Alibaba's Qwen3 series, built with 235 billion parameters with over 22 billion activated per token. It adopts a dense transformer architecture and is trained on 36 trillion high-quality tokens. Notably, it achieves top-tier performance on a wide range of open benchmarks including MMLU, GSM8K, and GPQA, and supports multilingual and long-context understanding.
- **Qwen2.5-Max (Qwen Team, 2024):** Qwen2.5-Max is a large-scale Mixture-of-Experts (MoE) model pretrained on over 20 trillion tokens, and further refined through curated SFT and Reinforcement Learning from Human Feedback (RLHF). It demonstrated leading performance on multiple benchmarks, including Arena-Hard, LiveBench, LiveCodeBench, and GPQA-Diamond.

Table 5-2. Other Base LLMs Adopted for Evaluation. Note: MoE is short for Mixture-of-Expert.

	DeepSeek-V3	DeepSeek-R1	Qwen2.5-MAX	Qwen3-235B-A22B
# Architecture	MoE	Dense	MoE	MoE
# Total Params	671B	260B	325B	235B
# Activated Params	37B	260B	22B	22B
# Licensing	Open Source	Open Source	Open Source	Open Source
# Model Category	General	Reasoning	General	Reasoning

Prompt Design for Question Understanding. To move beyond treating each subtask in isolation, this study embed RST’s nucleus–satellite relations and appraisal theory into a single structured prompt. This unified design enables the model to jointly reason about medical informational needs, emotional support needs, and their associated contextual information within a single inference pass. The prompt guides the model through a stepwise, semantically grounded reasoning process. It begins by instructing the model to identify medical needs, extract relevant context, recognize emotional needs, and interpret their underlying causes. Each need is explicitly linked to its contextual information, forming a coherent semantic unit. This mirrors human-like discourse reasoning, consistent with RST structures.

To enhance the model’s performance and reasoning transparency, and after multiple rounds of empirical experimentation and ablation analysis, two widely adopted prompting strategies are incorporated into the unified prompt design:

- **Chain-of-Thought (CoT) Prompting:** CoT prompting has been shown to enhance performance in multi-step reasoning tasks by making the model's intermediate reasoning processes explicit (Wei et al., 2022). Our prompt lays out an explicit multi-step reasoning path that reflects clinical decision-making. The model is prompted to: (a) Identify medical needs, with the instruction: “Identify and annotate all medical needs (medical_need). Each entry should contain only one specific need. The categories are ...”; (b) Extract contextual information such as elaboration, background, and conditional constraints, with the instruction: “For each identified medical need, extract and annotate relevant contextual information from the user query.” (c) Identify emotional needs (“Identify all emotional needs (emotional_need). Categories are...”) and Link them to their causes (“For each emotional need, extract and annotate the

associated context from the user query. Context category ...”). This structured process encourages the model to maintain semantic coherence across related elements, aligning with discourse-theoretic principles.

- **Few-Shot Prompting:** This approach leverages the model's ability to learn from limited examples, thereby improving performance on structured output tasks such as span classification and relation extraction (Brown et al., 2020). In this study, we include one labeled example per subtask to serve as a guiding template. For instance, Emotion Expression (E-N-FEEL) is illustrated with: “I feel like I’m falling apart and don’t know what to do.” and Causal Explanation (E-C-CAUSE) with: “I’ve tried many medications and none worked, and they all had strong side effects.” These examples help the model distinguish among task categories and learn the structure of the expected output. In addition, a full user query with corresponding output is included to illustrate how the semantic units are combined in the unified JSON format.

This structured integration of CoT and few-shot prompting (see Figure 5-12) ensures token-level clarity, semantic consistency across subtasks, and robust end-to-end parsing of complex mental-health queries. Details of the full prompt template for joint question understanding are provided in Appendix 4.

Role
You are a professional clinical assistant specializing in extracting medical and emotional needs and related contextual information in mental consumer health queries.

Task
For the given user query (`user_query`), follow the steps below to perform structured annotation and return the result in JSON format:

Step 1: Identify and annotate all medical needs (`medical_need`). Each entry should contain only one specific need. The categories are:
 - General Medical Information (M-N-GMI): A broad or non-specific request for medical knowledge, such as disease background, medication function, or treatment procedures. Example: "Can you tell me how this drug works?"
 ...

Step 2
 ...

Step 3
 ...

Step 4
 ...

Output
You are expected to return a structured JSON object, including all detected needs and their contextual relationships. A sample user query and expected output are shown below.

```

user_query:
{
  ...
}

output:
{
  ...
}
    
```

Figure 5-12. The Illustration of the Prompt Design with CoT and Few-Shot Prompt Strategies for Joint Question Understanding.

Evaluation Methodology. To comprehensively evaluate the performance of the four multi-level tasks (MNR, ENR, MNCE, and ENCE), considering both the accuracy of predicted labels and corresponding text spans, this study adopts a hierarchical evaluation framework.

For the evaluation of parent-level tasks (MNR and ENR), two need units are considered a match if they satisfy the following conditions: 1) they belong to the same query_id; 2) they share the same need_type; and 3) their need_text spans meet the defined span-matching criterion. For the evaluation of child-level tasks (MNCE and ENCE), we assess the recognition performance of related context units based on successfully matched parent need units. A related context unit is regarded as a match when: (a) it belongs to the same query_id; (b) its parent need unit has been successfully matched; (c) it has the same relation_type label; and (d) its context_text span satisfies the span-matching criterion.

In implementing the span-matching criterion, we recognize that exact boundary matching may be overly restrictive, particularly for complex and open-ended tasks such as ours. Therefore, we adopt a partial matching strategy, where a predicted span is considered correct if its cosine similarity with the corresponding ground-truth span exceeds 0.5. This threshold balances the need for precision with the understanding that minor discrepancies in span boundaries may not substantially affect the informational value of the extracted content. Such a nuanced evaluation approach acknowledges the inherent challenges of span-based tasks and aligns with best practices in the field, which advocate for flexible matching criteria to better capture real-world task demands (Chen et al., 2023).

In our implementation, we utilize OpenAI’s text-embedding-3-large model to generate high-dimensional (up to 3072 dimensions) semantic embeddings for textual data. This model demonstrates superior performance in capturing semantic relationships within long-form texts, outperforming traditional bag-of-words approaches and earlier embedding models such as text-embedding-ada-002 (OpenAI, 2024).

To address scenarios where multiple instances of the same need type exist within a single query, we implement an enhanced greedy matching algorithm based on the Kuhn-Munkres algorithm, also known as the Hungarian algorithm. This approach effectively resolves one-to-many matching

challenges by optimizing the assignment of predicted spans to ground truth annotations.

For performance evaluation, this study adheres to standard information retrieval metrics as outlined by Manning et al. (2008). This study defines True Positives (TP) as the number of predicted need or context units that semantically match the ground truth. False Positives (FP) are predicted units that do not correspond to any ground truth annotations, while False Negatives (FN) are ground truth units that the system fails to predict. Based on these definitions, we calculate precision, recall, and F1-score for each need type as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Performance is reported for each label type individually, as well as through aggregated metrics across medical needs, emotional needs, medical needs-related contexts, emotional needs-related contexts, and overall multi-dimensional understanding. To mitigate the effects of label imbalance, weighted averages were computed based on the frequency of each label type, providing a more representative measure of model performance than simple macro-averaging.

Considering the practical application, this study measured and logged the LLM's inference latency on a per-query basis to assess its real-time performance. Specifically, for 141 test queries, the total processing time was automatically recorded by our Python script, from which we derived the mean latency per query. All measurements were captured at millisecond resolution using the OpenAI Python SDK (Software Development Kit)'s built-in timing hooks.

5.4.4 Evaluation Results

Before presenting the comparative results, it is important to note that the fine-tuned Qwen3-32B model exhibited pronounced hallucination behaviour during validation. Specifically, the model generated spurious labels and schema elements that were not present in the prompt or the defined

annotation schema, indicating instability in instruction adherence and task boundary control. A likely explanation is that joint fine-tuning across four interdependent subtasks imposed excessive representational complexity for a medium-scale model (32B parameters). Such complexity may have led to over-generalization and semantic drift during structured output generation, causing the model to fabricate unseen label types. Given the severity and consistency of these hallucinations across multiple validation runs, Qwen3-32B was excluded from the final evaluation phase to preserve methodological validity and ensure fair cross-model comparison.

Accordingly, the final evaluation was conducted across eight LLMs, including both fine-tuned and prompt-based variants, as summarized in Figure 5-13. The subsequent analysis compares their accuracy (F1 score) and computational efficiency (processing time per query) under identical prompting and evaluation protocols.

Overall Performance vs. Latency Trade-Off. Most strikingly, the fine-tuned GPT-4o variant attained the highest overall F1-score of 0.699, with an average inference time of only 8.64 s per query, significantly outperforming all other LLMs. DeepSeek-R1 follows closely with an F1 of 0.693 but at a substantially higher latency (122.84 s/query). In contrast, prompt-only models such as Qwen3-235B-A22B (F1=0.651, 45.87 s/query) and Qwen2.5-max (F1=0.646, 36.5 s/query) strike a mid-range balance, and smaller fine-tuned variants like Qwen2.5-72B (F1=0.629, 18.96 s/query) demonstrate that targeted adaptation of more compact models can yield robust performance with moderate latency.

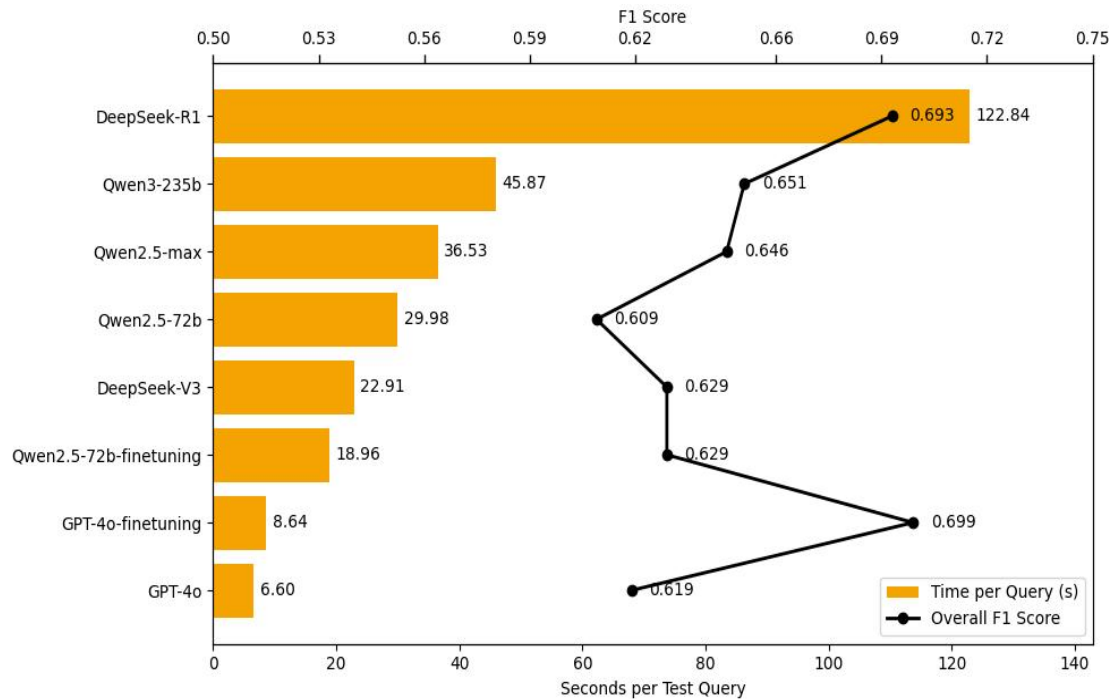


Figure 5-13. Overall F1 Score and Processing Time Per Test Query Across 8 LLMs.

Subtask Sensitivity Differs Across Models. MNR is dominated by models with strong domain alignment. Among the Qwen family, fine-tuned Qwen2.5-72B leads with an F1 of 0.728 (versus 0.686 for its base), outperforming larger MoE variants, such as Qwen2.5-MAX (0.682), Qwen3-235B (0.651). In the GPT and DeepSeek group, DeepSeek-V3 achieves the highest MNR F1 of 0.790, with fine-tuned GPT-4o close behind at 0.737. ENR remains challenging. Here the largest gains come from fine-tuning: GPT-4o-finetuning tops the chart at 0.528 F1 (versus 0.412 base), while Qwen3-235B outperforms its smaller kin among prompt-only Qwens (0.507 vs. 0.445 for Qwen2.5-MAX and 0.444 for finetuned 72B). MNCX sees MoE models excel in prompt-only settings, Qwen2.5-MAX leads with 0.710 F1, whereas among fine-tuned or dense models, DeepSeek-R1 achieves the best result at 0.737 F1. Fine-tuned Qwen2.5-72B also closes the gap (0.690). ENCX shows the greatest benefit from fine-tuning. GPT-4o-finetuning achieves the highest F1 overall (0.846), followed closely by Qwen2.5-72B-finetuning (0.816) and DeepSeek-R1 (0.790). Prompt-only MoE models lag (Qwen3-235B, 0.802; Qwen2.5-MAX, 0.771; DeepSeek-V3, 0.671).

Table 5-3. Performance Results of Models (Qwen2.5-72B, Qwen2.5-72B-finetuning, Qwen2.5-max & Qwen3-235B)

Tasks	Models											
	Qwen2.5-72B Base			Qwen2.5-72B-finetuning			Qwen2.5-max Base			Qwen3-235B-A22B Base		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MNR	0.691	0.681	0.686	0.741	0.714	0.728	0.649	0.718	0.682	0.646	0.656	0.651
DIA	0.539	0.724	0.618	0.796	0.603	0.686	0.563	0.776	0.652	0.667	0.759	0.710
ETI	0.429	0.600	0.500	0.391	0.900	0.546	0.348	0.800	0.485	0.350	0.700	0.467
GMI	0.471	0.500	0.485	0.438	0.438	0.438	0.533	0.500	0.516	0.300	0.375	0.333
PROG	0.333	0.294	0.313	0.222	0.235	0.229	0.200	0.294	0.238	0.222	0.353	0.273
TREAT	0.862	0.727	0.789	0.864	0.814	0.838	0.818	0.756	0.786	0.806	0.674	0.734
ENR	0.382	0.462	0.418	0.437	0.453	0.444	0.405	0.493	0.445	0.494	0.520	0.507
FEEL	0.460	0.512	0.485	0.430	0.422	0.426	0.474	0.520	0.496	0.584	0.594	<u>0.589</u>
VIEW	0.300	0.398	0.342	0.444	0.490	0.466	0.336	0.460	0.388	0.389	0.429	0.408
MNCX	0.644	0.648	0.646	0.710	0.672	0.690	0.709	0.711	0.710	0.658	0.726	0.690
BACK	0.593	0.509	0.549	0.729	0.556	0.631	0.702	0.656	0.678	0.632	0.655	0.643
CON	0.431	0.500	0.463	0.500	0.467	0.483	0.419	0.605	0.495	0.467	0.636	0.539
ELA	0.745	0.835	0.787	0.742	0.868	0.800	0.824	0.795	0.810	0.750	0.827	0.787
ENCX (CAUSE)	0.705	0.733	0.718	0.838	0.795	0.816	0.771	0.771	0.771	0.795	0.809	0.802
Overall	0.595	0.624	0.609	0.634	0.624	0.629	0.623	0.671	0.646	0.633	0.669	0.651

Table 5-4. Performance Results of Models (GPT-4o, GPT-4o-finetuning, DeepSeek-V3 & DeepSeek-R1)

Tasks	Models											
	GPT-4o Base			GPT-4o-finetuning			DeepSeek-V3 Base			DeepSeek-R1 Base		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
MNR	0.798	0.637	0.709	0.762	0.714	0.737	0.833	0.751	0.790	0.769	0.755	0.762
DIA	0.830	0.672	0.743	0.851	0.690	0.762	0.757	0.914	<u>0.828</u>	0.746	0.707	0.726
ETI	0.385	0.500	0.435	0.471	0.800	<u>0.593</u>	0.667	0.400	0.500	0.467	0.700	0.560
GMI	0.500	0.438	0.467	0.471	0.500	0.485	0.692	0.563	<u>0.621</u>	0.500	0.563	0.529
PROG	0.429	0.177	0.250	0.308	0.235	0.267	0.455	0.294	0.357	0.625	0.294	<u>0.400</u>
TREAT	0.876	0.698	0.777	0.833	0.785	0.808	0.918	0.779	<u>0.843</u>	0.837	0.837	0.837
ENR	0.471	0.367	0.412	0.502	0.557	0.528	0.373	0.398	0.385	0.500	0.430	0.462
FEEL	0.504	0.545	0.523	0.553	0.594	0.573	0.393	0.463	0.425	0.542	0.520	0.531
VIEW	0.359	0.143	0.204	0.443	0.510	<u>0.474</u>	0.341	0.316	0.328	0.431	0.316	0.365
MNCX	0.707	0.608	0.654	0.753	0.704	0.727	0.706	0.612	0.656	0.701	0.777	0.737
BACK	0.609	0.506	0.553	0.697	0.639	0.667	0.658	0.526	0.585	0.694	0.761	<u>0.726</u>
CON	0.867	0.289	0.433	0.550	0.468	0.506	0.537	0.449	0.489	0.574	0.700	<u>0.631</u>
ELA	0.779	0.815	0.796	0.856	0.836	<u>0.846</u>	0.790	0.750	0.770	0.749	0.817	0.781
ENCX (CAUSE)	0.707	0.716	0.712	0.846	0.846	0.846	0.671	0.671	0.671	0.790	0.790	0.790
Overall	0.681	0.569	0.619	0.706	0.692	0.699	0.652	0.608	0.629	0.690	0.696	0.693

In summary, all models performed better on structured, knowledge-based categories like MNR-TREAT (F1>0.75 in majority of models) or MNR-DIA (F1>0.65 in majority of models) compared to abstract or subjective ones such as ENR-FEEL (F1<0.55 in majority of models) or ENR-VIEW (F1 < 0.45 in majority of models). This performance gap highlights LLMs’ continued difficulty in handling ambiguous, discourse-level interpretation, particularly for emotion viewpoints that lack explicit lexical markers.

Impact of Fine-Tuning. Fine-tuning consistently boosts performance. Qwen2.5-72B's overall F1 improves from 0.609 to 0.706 after LoRA-based adaptation, driven by gains on emotionally complex subtasks (ENR+0.026 F1, ENCX+0.098 F1). GPT-4o's fine-tuned version rises from 0.619 to 0.699 overall, with particularly large improvements in ENR (from 0.412 to 0.528 F1). This suggests that domain-specific adaptation enhances LLMs' ability to reason about subjective emotional content and nuanced causality.

Architecture and scale effects. Model size is not the only determining factor. While DeepSeek-V3 has the largest parameter count (671B) among open-source models, its performance on ENR and ENCX is lower than that of smaller, fine-tuned models like Qwen2.5-72B-finetuning and GPT-4o-finetuning. This indicates that model alignment and domain adaptation may be more critical than sheer scale. MoE models like DeepSeek-V3, Qwen2.5-Max and Qwen3-235B show less stable performance across subtasks. For example, Qwen3-235B has strong performance on ENCX (F1 = 0.802) but lags behind in MNR (F1 = 0.651). Besides, DeepSeek-V3 attains the highest F1 on MNR but records the lowest F1 on ENR. Such inconsistency may stem from the models' expert-routing strategy: by preferentially activating a small subset of experts for sparse, domain-specific inputs, the mechanism promotes specialized processing but may fail to capture the complexity inherent in multi-intent queries.

5.5 Inference on Answer Generation

The answer generation stage performs inference on the fine-tuned model obtained from the previous stage, without any further parameter updates. Generation is governed by prompt-conditioned decoding, where task-adaptive templates trigger two coordinated reasoning pathways: (a) context-specific emotional-need support (corresponding to the ENA task) and (b) patient-centred medical-need responses (corresponding to the MNA task). The prompt design integrates two well-established prompt-engineering strategies, CoT guidance and few-shot exemplars, to scaffold transparent clinical reasoning and ensure stable, structured output formats. Importantly, the prompt templates were refined through multiple rounds of empirical experimentation and ablation analysis, ensuring both robust task alignment and consistent generative behaviour across different query types.

Prompt for Context-Specific Emotional Need Support. The prompt encodes an explicit, multi-step reasoning path that operationalises clinical empathy within the model’s generation process: (a) For each expression of feeling, the model is guided to first apply the understanding strategy which, in this design, merges the “naming” and “understanding” stages into a single coherent reasoning step. This fusion was empirically validated through multiple rounds of comparative experiments to ensure that the model produces natural, contextually grounded responses rather than mechanically layered statements. By integrating causal explanations that explicitly acknowledge the user’s emotional state within its contextual background (e.g., “I understand that you are feeling overwhelmed by these symptoms”), the prompt encourages smooth empathetic articulation. This is then followed by the supporting strategy, which broadly encompasses sharing, encouragement, and assistance (e.g., “Many people in similar situations also feel this way, which is completely understandable”). (b) For each expression of viewpoint, the model is instructed to first employ the understanding strategy, again implemented as a combined naming-understanding phase to achieve fluent and contextually integrated expression. This step incorporates causal reasoning that recognises the user’s perspective within its situational context (e.g., “I understand that you are concerned about developing medication dependence, especially after experiencing prolonged withdrawal symptoms and unsuccessful attempts with alternative treatments.”). It is then complemented by the respecting strategy, which conveys non-judgemental acknowledgement and balanced support (e.g., “Your concern is valid and respected, and it is important that we take it into account”).

As illustrated in Figure 5-14, Step 1 and Step 2 correspond directly to these two prompting components. Notably, no explicit exemplars were included in these steps, to preserve diversity and spontaneity in the model’s emotional-support expressions and to avoid over-templated outputs.

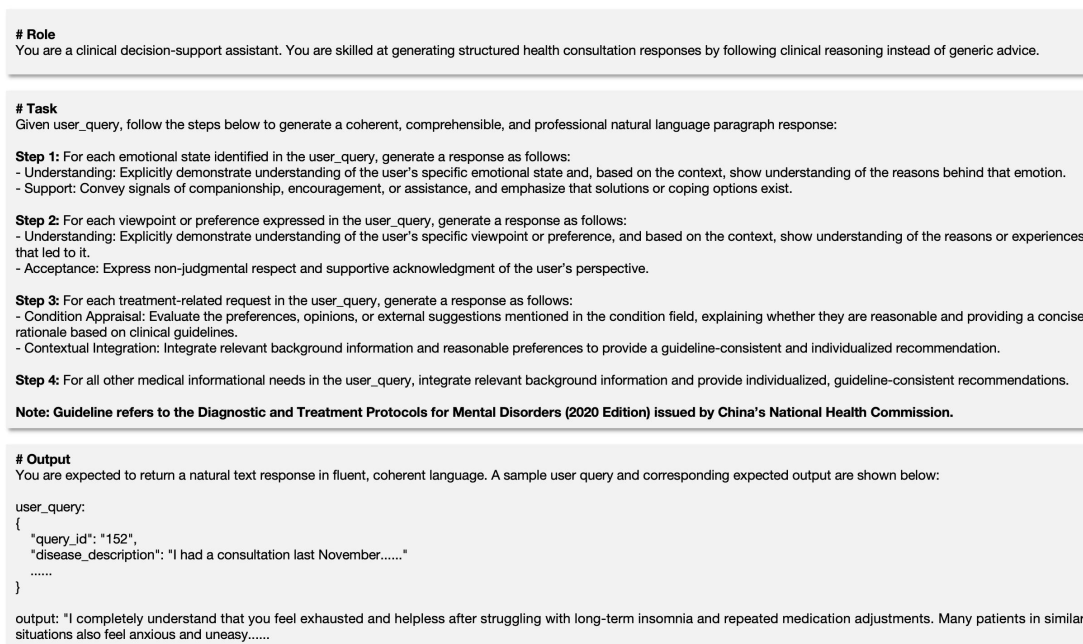


Figure 5-14. The Illustration of the Prompt Design for Joint Answer Generation.

Prompt for patient-centered medical need support. The prompt lays out an explicit multi-step clinical reasoning pathway while enforcing a China-first guidance policy. (a) For each medical need related to treatment, the model is guided to first appraise any patient preferences or views in the condition field against authoritative guidance, briefly justifies their reasonableness, and then generate an answer that integrates other relevant contextual satellites (elaboration, background) together with those preferences deemed acceptable. (b) For other types of medical needs, the model is guided to provide response directly integrating elaboration and background information.

Step 3 and Step 4 in Figure 5-14 correspond directly to these two prompting components. Notably, to ensure safety, regulatory compliance, and local clinical applicability, this study adopts a guideline-concordant prompting strategy: the prompt explicitly instructs the model to adhere to the most recent, locally authoritative clinical guidelines, which has been shown to substantially improve guideline adherence and clinical usefulness of LLM outputs in multiple studies (Krešević et al., 2024; Wang et al., 2024). In particular, the normative standards outlined in the Diagnostic and Treatment Protocols for Mental Disorders (2020 Edition), issued by the Chinese National Health Commission (NHC), are treated as the regulatory baseline for all treatment-related reasoning and response generation.

Considering that general-purpose LLMs often adopt a guideline-oriented yet decision-averse approach, it becomes necessary to employ targeted prompt design to shift their interaction style from a conversational mode toward a clinical reasoning and decision-support paradigm. This transition is illustrated in the role framework shown in Figure 5-14.

By integrating these two dimensions within a unified prompt, this stage yields holistic responses that combine clinical reliability with human-centered empathy. The full prompting templates and exemplar outputs are provided in Appendix 5.

5.6 Chapter Summary

Chapter 5 presented the complete design of the Med-Emo CHQA system, translating the research problem and theoretical foundations into a functional system blueprint. The chapter began by defining the overall task and decomposing it into six sub-tasks to address both medical informational needs and emotional support. A conceptual system architecture was then developed to integrate question understanding and answer generation within an understanding-enhanced end-to-end pipeline.

A key contribution of this chapter was the introduction of a dual-stage mechanism: supervised fine-tuning to improve the model's capability in multi-intent query interpretation, followed by constraint-based prompting to enforce clinical reliability and empathetic expression during response generation. The fine-tuning process, data preparation, and subtask-level evaluation established the internal reasoning capacity of the system, while the inference strategy defined how patient-centred responses would be produced.

By the end of this chapter, the system was fully specified at the architectural and procedural levels, forming a robust foundation for implementation. The design developed here directly informs the experimental realisation and empirical evaluation that follow in Chapter 6, where the system's effectiveness is tested in realistic health consultation scenarios.

Chapter 6 System Evaluation

6.1 Chapter Introduction

Chapter 6 presents the implementation and evaluation of the Med-Emo CHQA system, completing the transition from conceptual design to operational validation. While Chapter 5 introduced the system architecture, task formulation, and model mechanisms at a design level, this chapter brings the system into execution through experimental deployment. In the context of Design Science Research, implementation could be achieved through controlled experimentation, where the system operates on real queries and demonstrates its intended behaviour.

To validate both functional performance and clinical applicability, a reader study methodology is adopted. This expert-based evaluation involves clinical and health communication professionals, who assess response quality in terms of medical accuracy, personalisation, and empathy. In parallel, latency efficiency is automatically recorded to evaluate the system’s practical usability under real-world constraints. Taken together, these complementary approaches provide a holistic assessment, capturing both human-perceived quality and model-level robustness.

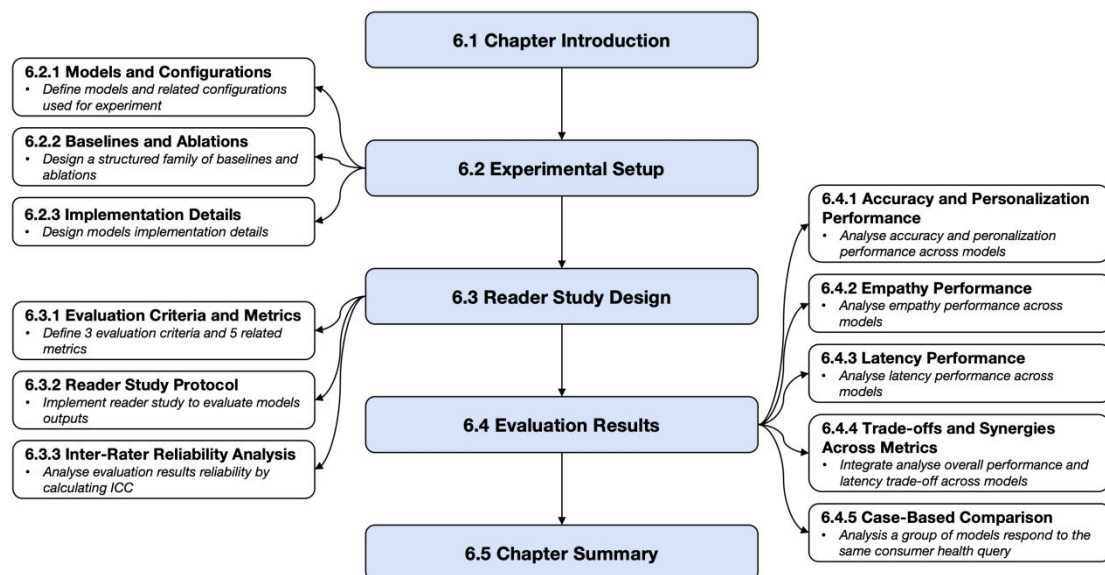


Figure 6-1. Structure of Chapter 6.

Figure 6-1 summarises the overall structure of this chapter. Section 6.2 outlines the experimental setup, detailing model configurations, baselines,

and implementation settings. Section 6.3 presents the reader study evaluation, including evaluation criteria, reader study protocol, and inter-rater reliability analysis using Intraclass Correlation Coefficient (ICC). Section 6.4 reports and compares results across accuracy, personalisation, empathy, and latency, followed by an integrated analysis of trade-offs and synergies across metrics, and qualitative exemplar cases. Finally, Section 6.5 concludes the chapter by synthesising key findings and positioning them for the broader discussion in Chapter 7.

6.2 Experimental Setup

6.2.1 Models and Configurations

The experiments employed the fine-tuned GPT-4o variant introduced in Section 5.4 as the production backbone for the proposed Med-Emo CHQA system. The model was partially fine-tuned on the CHQA-MedEmo dataset to enhance multi-intent understanding, contextual reasoning, and empathetic response generation. Among all candidate models benchmarked in Section 5.4, the fine-tuned GPT-4o achieved the highest overall macro-F1 score of 0.699 across the four question understanding subtasks, while maintaining an average end-to-end inference time of 8.64 seconds per query under standardized decoding parameters (temperature = 0.1, top-p = 0.9, max_tokens = 1024). Based on its superior balance of accuracy and efficiency, the fine-tuned GPT-4o variant was adopted as the primary backbone for subsequent benchmark and reader-study evaluations.

6.2.2 Baselines and Ablations

To attribute the contribution of each system component in a principled manner, this study systematically varied three orthogonal factors - question understanding, answer generation, and architecture - yielding a structured family of baselines and ablations centered around the proposed MedEmo CHQA method (see Table 6-1):

- Question understanding was configured at three levels: (a) none, representing the naïve condition without any prompting or fine-tuning; (b) prompting, in which the model received explicit instructions for multi-intent question understanding; and (c) SFT, a supervised fine-tuning setup designed to internalize consumer-health query understanding signals from the CHQA-MedEmo dataset.

- Answer generation was varied between none (i.e., using only the raw user query) and prompting (explicit answer-generation instruction).
- Architecture was tested in two modes: E2E (end-to-end), where understanding and generation occurred within a single model call, and Pipeline, where a first-stage model produced a structured JSON representation of understanding results that were subsequently consumed by a second-stage generator for final response production.

As the JMEQA task has not been previously explored, there is no established benchmark model for direct comparison. Accordingly, this study adopted a naïve GPT-4o configuration, responding directly to the user query without any prompting or fine-tuning, as the baseline. This setting represents the lower bound of general-purpose reasoning performance and serves as a neutral reference point against which all ablation arms are evaluated.

To systematically analyze the contribution of individual design components, this study implemented four ablation settings derived from the proposed MedEmo CHQA framework. Each ablation arm selectively removes or alters one functional element - question understanding, answer generation, or system architecture - while keeping all other configurations constant.

- **Abl-SFT**: introduces prompting for both question understanding and answer generation without fine-tuning. This configuration represents a strong prompting-only setup, allowing comparison between zero-shot prompting and fine-tuned understanding.
- **Abl-QU**: removes the question-understanding mechanism to examine its influence on the model' s ability to recognize user intent and contextual nuance during response generation.
- **Abl-AG**: omits the explicit answer-generation prompt, isolating the effect of fine-tuned understanding on the model' s spontaneous response formulation.
- **Abl-Arch**: replaces the end-to-end configuration with a two-stage pipeline, where the understanding model produces a structured JSON intermediate representation that is subsequently consumed by the generator. This setup evaluates the trade-off between interpretability and efficiency.

Table 6-1. Configurations of Baseline, Ablation, and Proposed Models

Arm	Question understanding	Answer generation	Architecture	Purpose
Baseline	none	none	E2E	Lower bound
Abl-SFT	prompting	prompting	E2E	Replace fine-tuned understanding with prompting
Abl-QU	none	prompting	E2E	Remove question understanding stage
Abl-AG	SFT	none	E2E	Remove explicit answer-generation prompt
Abl-Arch	SFT	prompting	Pipeline	Swap to two-stage architecture
MedEmo CHQA (Proposed)	SFT	prompting	E2E	Full design integration

6.2.3 Implementation Details

The same software environment was maintained across all experiment runs to ensure consistency and reproducibility.

Hardware Configuration. All local executions were performed on an Apple iMac equipped with a 3 GHz Quad-Core Intel Core i5 CPU, Radeon Pro 555 2 GB GPU, and 16 GB 2400 MHz DDR4 memory. As all model inference was executed via the OpenAI cloud infrastructure, the local hardware served only as a client node for scheduling, latency measurement, and logging, without performing any neural computation. This configuration provided sufficient computational stability and timing precision for the experimental workflow, ensuring that no local hardware bottlenecks affected network latency or measurement validity.

SDK and Software Stack. All inference experiments were conducted in Python 3.11 using the OpenAI Python SDK (v1.42.0) on macOS Monterey 12.7.6. Core dependencies included requests (v2.31.0), numpy (v1.26.4), and pandas (v2.2.3) for API request handling, latency measurement, and response logging. These versions were chosen to ensure compatibility with the most recent OpenAI API schema and to maintain consistent JSON serialization and timing behaviour across runs.

API Invocation Mode. All model calls were executed through the OpenAI Python SDK using the response-based invocation interface (`client.chat.completions.create`). The following standardized parameters were used for all model variants to ensure comparability:

- Temperature: 0.1, reducing stochasticity in clinical text generation while preserving natural variation.
- Top-p: 0.9, accommodating clinical language complexities.
- Max tokens: set to the provider's default maximum, ensuring no truncation of long clinical queries.
- Stream: False.

Inference Environment. Inference was conducted on the local workstation described above, acting as a client node dispatching API calls to the respective model endpoints. This client-server separation ensured that observed latency primarily reflected true remote inference time and network transmission cost, rather than local computational overhead. To ensure that the latency statistics captured genuine inference and transmission overheads rather than environmental noise, the following controls were applied:

- **Network:** All experiments were conducted over a stable research-grade broadband connection provided by Spark New Zealand (≈ 21.83 Mbps downlink / 64.32 Mbps uplink, average RTT ≈ 6 ms). The low latency and high throughput of this environment ensured that network variance did not confound timing measurements.
- **Execution Mode:** Requests were processed strictly sequentially to avoid artefacts caused by asynchronous or parallel execution.
- **Load Factor:** No concurrent requests were issued, eliminating queuing and rate-limiting effects.
- **Logging:** Start and end timestamps were recorded using Python's `time.perf_counter` with millisecond precision, and all results were automatically written to structured log files for reproducibility and auditability.

6.3 Expert-Based Evaluation

To ensure a rigorous and reproducible evaluation, this study developed a reader study-based evaluation framework (see Figure 6-2), inspired by recent large-scale assessments of LLM generations in clinical settings (Singhal et al., 2025; Thirunavukarasu et al., 2023). This approach was chosen because

evaluating medical and empathetic adequacy requires clinical expertise and contextual judgment that cannot be reliably captured by automatic metrics alone.

In clinical AI research, a reader study design refers to a controlled evaluation framework where multiple qualified human experts (“readers”) independently assess LLM-generated outputs, such as diagnostic findings or treatment suggestions, according to predefined metrics. This design allows for statistically robust comparisons, inter-rater reliability analysis, and validation of the system’s clinical utility under real-world decision-making conditions.

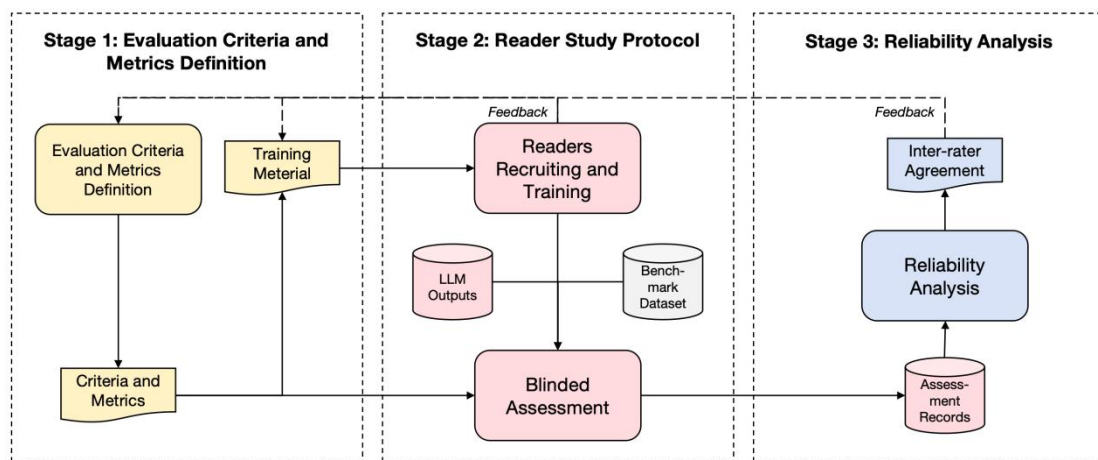


Figure 6-2. Overview of the Reader Study Evaluation Framework. The evaluation process was divided into three stages: (a) Evaluation Criteria and Metrics Definition, where the evaluation dimensions and scoring scales were defined; (b) Reader Study Protocol, where trained readers conducted blinded assessments of model outputs against benchmarks; and (c) Reliability Analysis, where inter-rater agreement and consistency were quantitatively analyzed. Dashed arrows indicate iterative calibration and feedback loops.

6.3.1 Evaluation Criteria and Metrics

The evaluation framework in this study was designed to align with the overarching goal of patient-centered medical need support and context-specific emotional need support (presented in Section 5.2), both of which extend beyond conventional measures of factual correctness or surface empathy.

In traditional CHQA systems, evaluation often focuses solely on the factual accuracy of responses. However, the JMEQA framework redefines this perspective by emphasizing individualized clinical relevance. Beyond providing medically reliable information, it prioritizes how well model-

generated responses are tailored to a patient's unique situation, concerns, and treatment preferences. In this study, patient-centered medical need support is operationalized by explicitly linking each identified medical need with its corresponding patient view, that is, how the response accommodates the user's expressed goals, constraints, or preferences.

Emotional reassurance in clinical communication cannot be isolated from the medical context in which it arises. Rather than offering generic or templated empathy statements, effective emotional support should be grounded in the user's medical situation and expressed experience. In this study, context-specific emotional need support is realized by explicitly associating each emotional need with its corresponding medical background, ensuring that empathy emerges from clinical understanding rather than detached sympathy.

Based on these principles, three complementary criteria were established for human evaluation: (a) Accuracy. Measuring the degree to which the response fulfills the patient's medical needs accurately; (b) Personalization. Measuring the degree to which the response fulfills the patient's medical needs in a patient-centered manner and (c) Empathy. Assessing the appropriateness, specificity, and contextual grounding of emotional support in relation to the medical scenario. These criteria together ensure that both informational reliability and emotional responsiveness are evaluated as integrated components of patient-centered communication, aligning with contemporary evaluation frameworks in AI-based health communication (Tam et al., 2024).

Accuracy. Accuracy was evaluated by comparing system outputs with benchmark annotations and authoritative clinical practice guidelines. Specifically, the Guidelines for the Diagnosis and Treatment of Mental Disorders (2020 Edition), issued by the National Health Commission of China, was adopted as the primary reference. This guideline provides standardized diagnostic criteria, pharmacological recommendations, and therapeutic principles for psychiatric and psychological disorders, ensuring that evaluation remains aligned with evidence-based clinical practice.

Accuracy was evaluated by three measurable metrics: comprehensiveness level, hallucination level, safety level. Each metric was rated using a 3-point ordinal scale (0-2).

- **Comprehensiveness level:** Measures how faithfully the model reproduces the benchmark medical needs. Annotators aligned each benchmark medical need and its corresponding reference answer with model outputs, rating each according to three levels: Omission (0): the response fails to respond to the benchmark medical need; Partial comprehensiveness (1): the response partially addresses the need but omits essential details (e.g., dosage, duration, contraindications); Full comprehensiveness (2): the response fully covers the benchmark need with semantically and clinically complete information.
- **Hallucination level:** Quantifies unsupported or fabricated medical statements that contradict benchmark evidence or established clinical guidelines. Ratings were assigned as: No hallucination (0); Mild hallucination (1); Severe hallucination (2).
- **Risk level:** Evaluates the presence of verifiable factual errors that could cause clinical harm or misinformation, such as incorrect pharmacological mechanisms, indications, contraindications, or dosage. Ratings follow: Safe (0); Low risk (1); High risk (2).

Personalisation. Beyond accuracy, the Personalisation criterion evaluates the system's ability to tailor its responses to the user's contextual and psychosocial characteristics, consistent with the patient-centered care paradigm. This includes consideration of comorbid conditions, medication history, and expressed treatment concerns. The underlying assumption is that effective medical communication requires adaptation to individual circumstances rather than generic, one-size-fits-all recommendations.

The corresponding metric, personalisation level, operationalizes this criterion by quantifying the extent to which the response aligns with user-specific factors. Based on prior work (Kocaballi et al., 2019; Liu et al., 2022), this study defined a response as personalized when it explicitly incorporates contextual elements such as prior treatment responses or conditional advice reflecting user preferences. This metric measures how effectively medical recommendations are individualized rather than generic, and is evaluated using a 3-point ordinal scale (0–2):

- **Omission** (score 0): the response is generic and does not incorporate user-specific information.
- **Partial personalisation** (score 1): the response acknowledges some user factors but lacks integration.

- **Full personalisation** (score 2): the response explicitly tailors advice or support to the individual's medical and psychosocial background.

Empathy. The Empathy criterion evaluates the model's capacity to demonstrate understanding, validation, and emotional attunement toward the user's expressed feelings or views, consistent with the principle of context-specific emotional need support. In patient-centered communication, empathy should not rely on generic or templated reassurance, but instead be grounded in the user's medical context and lived experience. This ensures that emotional understanding is both authentic and clinically relevant, rather than superficial acknowledgment.

The corresponding metric, empathy level, operationalizes this criterion by quantifying the degree of empathetic engagement expressed in model-generated responses. Building upon the computational empathy framework proposed by Sharma et al. (2020), which differentiates between weak and strong empathy in text-based mental health support, this study extends the construct into a structured ordinal scale that distinguishes between generic expressions of empathy and those that exhibit naming, context-specific understanding, and supporting - three cumulative indicators of context-specific empathetic communication (presented in Section 5.3):

- **Naming the emotion/view.** Explicitly identifies or paraphrases the emotional state or subjective view expressed in the user query, as annotated in the benchmark (including both explicit and implicit emotions). Example: For a query annotated with implicit anxiety, "It sounds like you are feeling anxious about your son's condition."
- **Context-specific understanding.** Integrates emotional recognition with the user's medical or situational context, as indicated by the annotated causal information. Example: "I can see you are worried because the treatment has not improved his symptoms."
- **Supportive expression.** Provides explicit emotional support, such as encouragement, reassurance, validation, or sharing. Informational or instrumental support (e.g., medical advice) is not considered sufficient for this category. Example: "Many parents experience similar worries. You are not alone, and your persistence in seeking help is important."

Empathy Level was evaluated on a 5-point ordinal scale (0 – 4) based on the cumulative fulfillment of the above three dimensions:

- **Omission** (Score 0): The response fails to acknowledge the emotional or attitudinal content expressed in the query, or incorrectly identifies unrelated emotions or viewpoints.
- **Generic empathy** (Score 1): The response includes only templated or vague expressions of empathy (e.g., “I understand,” “That must be hard”), without recognizing specific emotions or contextual factors.
- **Partial contextualised empathy** (Score 2): The response fulfills only one of the three elements (Naming, Understanding, or Supporting).
- **Contextualised empathy** (Score 3): The response fulfills two of the three elements, such as: Naming + Understanding (without Supporting); Understanding + Supporting (without Naming); or Naming + Supporting (without Understanding).
- **Full empathy** (Score 4): The response satisfies all three elements - Naming, Understanding, and Supporting - demonstrating comprehensive emotional attunement and contextual grounding.

The medical evaluation dimensions were assessed using a 3-point ordinal scale (0–2) reflecting discrete entailment levels, whereas empathy-related dimensions adopted a 5-point ordinal scale (0–4) to capture graded nuances of emotional attunement. This hybrid design follows prior clinical LLM evaluations (Singhal et al., 2023; Amini et al., 2024), enabling both categorical reliability analysis (Cohen’s κ) and continuous agreement analysis (ICC).

Table 6-2. Description and Ordinal Scale of Evaluation Criteria and Metrics

Medical Need		
Support Evaluation Criteria (Metric)	Description	3-point Ordinal Scale
Accuracy		
- Comprehensiveness Level	Extent to which the response addresses all core medical needs expressed or implied in the query.	Omission (0); Partial comprehensiveness (1); Full comprehensiveness (2)
- Hallucination Level	Presence of unsupported, irrelevant, or fabricated medical statements that are inconsistent with established clinical evidence.	No hallucination (0); Mild hallucination (1); Severe hallucination (2)
- Risk Level	Evaluation of potential harm or clinical risk caused by the information (e.g., unsafe advice, omission of critical warning).	Clinical safe (0); Low risk (1); High risk (2)
Personalisation		
- Personalisation Level	Degree to which the response is tailored to the user’s specific context, including medical history, psychosocial background, and expressed preferences.	Omission (0); Low personalisation (1); High personalisation (2)
Emotional Need		
Support Evaluation Criteria (Metric)	Description	5-point Ordinal Scale
Empathy		
- Empathy Level	The extent to which the response demonstrates understanding, validation, and supportive expression toward the user’s emotional state or concerns.	Omission (0); Generic empathy (1); Partial contextualised empathy (2); Contextualised empathy (3); Full empathy (4)

To enable a more precise quantification of model performance, this study calculates the following six quantitative evaluation metrics based on experts assessment:

Formally, let \hat{R} denotes the model-generated response, M represents the set of all benchmark medical needs annotated in the user query. Each medical need has been aligned with \hat{R} and assigned a Comprehensiveness Level $c_x \in \{0, 1, 2\}$, Hallucination Level $h_x \in \{0, 1, 2\}$, Risk Level $r_x \in \{0, 1, 2\}$. The Partial Comprehensiveness Rate (PCR) and Full Comprehensiveness Rate (FCR) are defined as follows:

$$PCR(\hat{R}) = \frac{|\{x \in M | c_x = 1\}|}{|M|}, PCR(\hat{R}) \in [0, 1]$$

$$FCR(\hat{R}) = \frac{|\{x \in M | c_x = 2\}|}{|M|}, FCR(\hat{R}) \in [0, 1]$$

The Mild Hallucination Rate (MHR) and Severe Hallucination Rate (SHR) are defined as follows:

$$MHR(\hat{R}) = \frac{|\{x \in M | h_x = 1\}|}{|\{x \in M | c_x \geq 1\}|}, MHR(\hat{R}) \in [0, 1]$$

$$SHR(\hat{R}) = \frac{|\{x \in M | h_x = 2\}|}{|\{x \in M | c_x \geq 1\}|}, SHR(\hat{R}) \in [0, 1]$$

The Low Risk Rate (LRR) and High Risk Rate (HRR) are defined as follows:

$$LRR(\hat{R}) = \frac{|\{x \in M | r_x = 1\}|}{|\{x \in M | c_x \geq 1\}|}, LRR(\hat{R}) \in [0, 1]$$

$$HRR(\hat{R}) = \frac{|\{x \in M | r_x = 2\}|}{|\{x \in M | c_x \geq 1\}|}, HRR(\hat{R}) \in [0, 1]$$

Let \hat{R} denotes the model-generated response, C represents the set of all benchmark medical needs-related condition annotated in the user query. Each medical need-related condition has been aligned with \hat{R} and assigned a Personalisation Level $p_x \in \{0, 1, 2\}$. The Low Personalisation Rate (LPR) and High Personalisation Rate (HPR) are defined as follows:

$$LPR(\hat{R}) = \frac{|\{x \in C | p_x = 1\}|}{|C|}, LPR(\hat{R}) \in [0, 1]$$

$$HPR(\hat{R}) = \frac{|\{x \in C | p_x = 2\}|}{|C|}, HPR(\hat{R}) \in [0, 1]$$

6.3.2 Reader Study Protocol

Readers Recruiting and Training. To ensure professional and unbiased human evaluation, a separate group of psychiatrists, distinct from those

involved in the benchmark construction phase, were recruited to serve as readers. The reader panel consisted of three certified Chinese psychiatrists, each with over five years of clinical practice and at least three years of experience in online health consultation. This combination ensured both clinical rigor and familiarity with digital consultation discourse.

Prior to formal assessment, all readers underwent a structured training and calibration process. The training materials included (a) annotated benchmark samples illustrating standardized evaluation across the defined criteria (accuracy, personalisation, and empathy) and (b) reference scoring scales and clarity checklists for each criterion to ensure shared understanding of the evaluation framework.

To minimize subjective bias and enhance inter-rater reliability, training sessions incorporated guided discussions and consensus exercises on borderline cases, followed by a pilot rating session. Inter-rater agreement from this pilot phase was reviewed before the official evaluation commenced. During the formal assessment, all readers worked independently and in a blinded manner, without access to model identifiers or each other's ratings.

Blinded Assessment. To ensure objectivity and minimize evaluator bias, a blinded assessment protocol was implemented. The three certified psychiatrists who completed the training stage independently evaluated a total of 141 groups of responses. Each group comprised six model-generated answers produced by different LLM-based models and one benchmark reference answer constructed by clinical experts. During the assessment, all outputs were randomized and anonymized to remove any identifiable information regarding model provenance or generation order. Readers were explicitly instructed that all responses should be judged solely on their observable content, without speculation about the underlying model type, brand, or parameter size. This blinding procedure was designed to eliminate expectation bias and to preserve the integrity of cross-model comparisons. Figure 6-3 illustrates a single evaluation group presented to readers under the blinded assessment protocol.

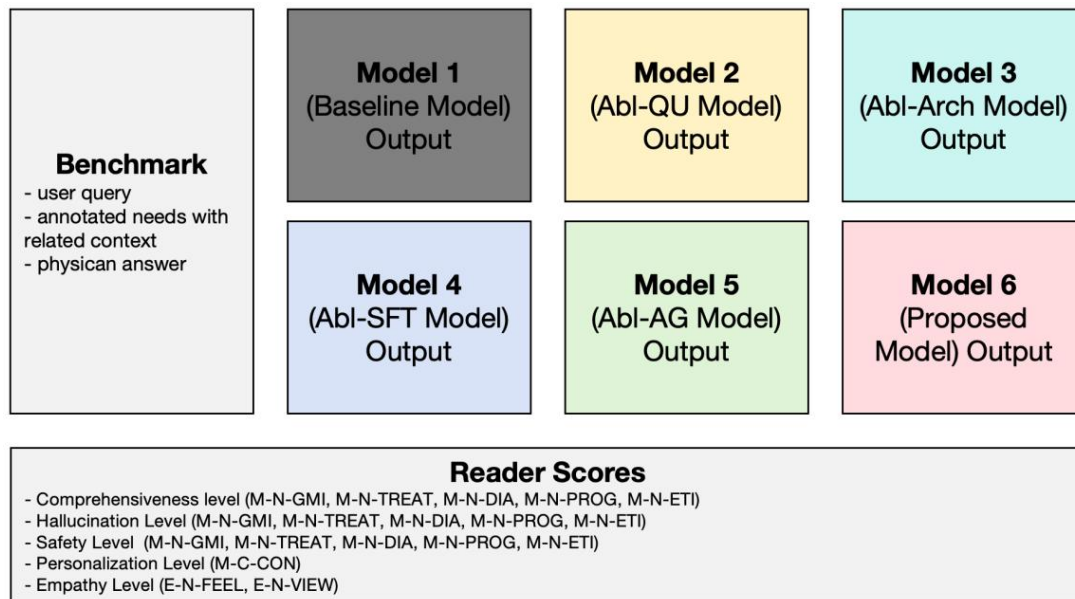


Figure 6-3. Illustration of a Single Evaluation Group Presented to Readers under the Blinded Assessment Protocol. Note: All outputs were randomized and anonymized before being presented to different readers.

Each psychiatrist conducted the blind evaluation in two sequential stages, as illustrated in Figure 6-4. In Stage 1, readers examined each model-generated response and identified medical and emotional need support segments based on the initial user query, annotated emotional and emotional needs in benchmark. In Stage 2, these identified segments were compared against benchmark physician statements and clinical guidelines to judge whether the model output adequately addressed the underlying needs. Following this comparative process, each response was rated across three evaluation criteria operationalized through five ordinal-scale metrics, as summarized in Table 6-2. To ensure methodological consistency and minimize subjective variability, all psychiatrists used a standardized reference guide developed during the training phase, including annotated examples, scoring rubrics, and decision rules for borderline cases.

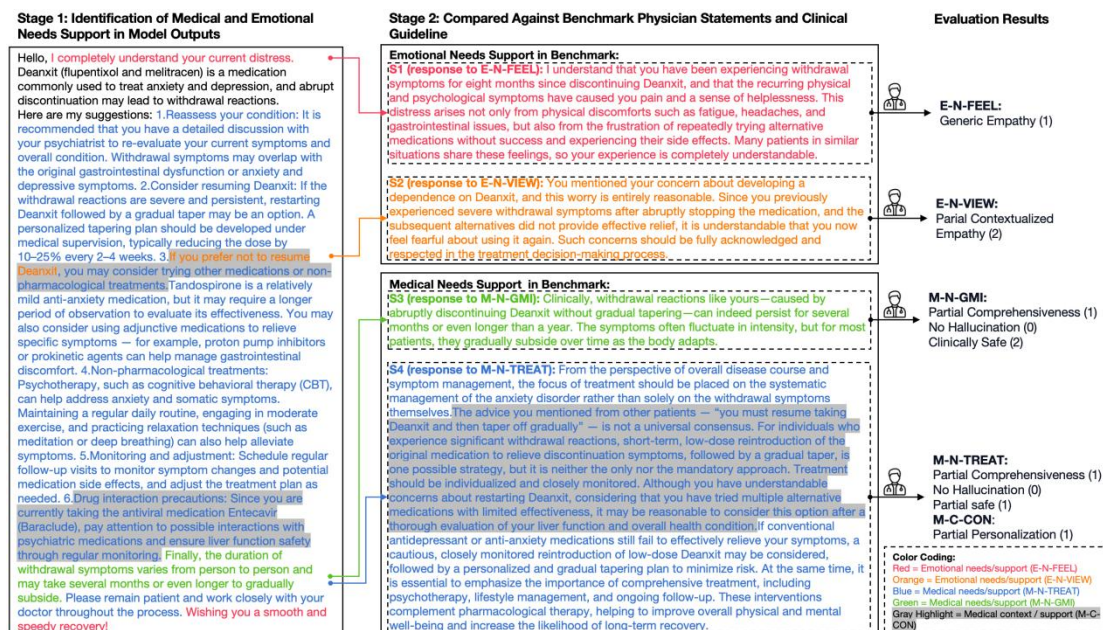


Figure 6-4. An Example of the Blinded Assessment under the Reader Study Protocol.

All ratings were recorded in standardized evaluation sheets and subsequently cross-checked for completeness and alignment with the established evaluation framework. These blinded, criterion-based assessments constitute the core human evaluation dataset for validating the Med-Emo CHQA system’s ability to produce clinically accurate, personalized, and empathetic responses.

6.3.3 Inter-Rater Reliability Analysis

To ensure robustness, inter-rater agreement was calculated using an established reliability metric, which is Intraclass Correlation Coefficient (ICC). ICC, first introduced by Bartko (1966), quantifies the proportion of total score variance that is attributable to actual differences between evaluated items rather than random variation among raters. A higher ICC indicates more reliable and consistent human evaluation.

ICC was selected over categorical agreement measures such as Cohen’s or Fleiss’ Kappa because the evaluation scores in this study were based on ordinal clinical scales (0-2, 0-4) rather than nominal labels. Unlike Kappa, which only detects exact match versus mismatch, ICC captures the magnitude of score differences (e.g., empathy rated as 2 vs. 3), which carries substantive meaning in clinical and patient-centered assessment. This makes

ICC particularly appropriate for evaluating nuanced dimensions such as comprehensiveness, personalisation, and empathy.

ICC values were computed using the *pingouin* statistical library in Python, which implements the standard Shrout and Fleiss (1979) formulations, equivalent to the *irr* package in R (Gamer et al., 2019). Python was selected to ensure seamless integration with the LLM evaluation pipeline and to enable reproducibility across computational experiments.

For reliability analysis, the unit of evaluation (item) was defined at the output level as the Cartesian product of an annotated need instance and a model-generated response (i.e., each need instance yields six items, one per model). Separate items × raters matrices were constructed for each construct: medical adequacy (Comprehensiveness, Hallucination, Safety; $n=1,566=261 \times 6$), personalisation ($n=756=126 \times 6$), and empathy ($n=1,650=275 \times 6$). For each construct, this study computed an overall ICC(3,k). ICC(3,k) denotes a two-way mixed-effects model with absolute agreement and average measures, appropriate for fixed expert raters. Point estimates, 95% CIs, F, df, and p-values were reported.

As shown in Table 6-3, all five evaluation metrics achieved high ICC, indicating substantial to excellent agreement among the expert raters. This confirms that the evaluation framework was applied consistently and that the human assessments provide a reliable basis for subsequent model comparison.

Table 6-3. Intraclass Correlation for the 5 Metrics Among the 3 Readers.

Metric	Intraclass Correlation (95% CI)	P value
Comprehensive level	0.84 (0.81 – 0.87)	<.001
Hallucination level	0.88 (0.85 – 0.91)	<.001
Safety level	0.83 (0.80 – 0.86)	<.001
Personalisation level	0.77 (0.73 – 0.81)	<.001
Empathy level	0.75 (0.71 – 0.79)	<.001

Note: All ICC values are based on a two-way mixed-effects model with absolute agreement (ICC(3,k)).

Besides, considering the practical application, this study measured and logged the LLM’s inference latency on a per-query basis to assess its real-time performance. Specifically, for 141 test queries, the total processing time was automatically recorded by a Python script, from which the mean

latency per query was derived. All measurements were captured at millisecond resolution using the OpenAI Python SDK’s built-in timing hooks.

6.4 Evaluation Results

6.4.1 Accuracy and Personalisation Performance

Table 6-4 and Table 6-5 summarise the overall performance of all models in terms of comprehensiveness, hallucination, risk, and personalisation. While comprehensiveness is originally annotated across five medical-need categories (M-N-DIA, M-N-ETI, M-N-GMI, M-N-PROG, M-N-TREAT), some categories contain only a limited number of instances (e.g., M-N-PROG = 11, M-N-ETI = 10), which makes category-level statistical comparison unreliable. Therefore, this study reports aggregated counts across all categories to ensure robustness and interpretability. Similarly, hallucination and risk annotations are sparse at the per-category level. Thus, we analyse their overall frequency rather than per-type distributions.

Table 6-4. Comprehensiveness-, Hallucination-, Risk-, and Personalisation-Level Counts for All Models.

	Models																	
	Baseline			Abl-QU			Abl-AG			Abl-SFT			Abl-Arch			MedEmo CHQA (Proposed)		
	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
Comprehensiveness Level	20	104	138	8	97	157	4	52	206	7	90	165	6	86	170	11	76	175
M-N-DIA	2	32	12	1	29	16	1	16	29	1	26	19	1	26	19	1	24	21
M-N-ETI	2	6	2	1	3	6	0	2	8	2	0	8	0	4	6	0	3	7
M-N-GMI	4	11	10	2	9	14	1	6	18	2	10	13	2	8	15	4	8	13
M-N-PROG	0	10	1	0	5	6	0	3	8	0	10	1	1	3	7	1	6	4
M-N-TREAT	12	45	113	4	51	115	2	25	143	2	44	124	2	45	123	5	35	130
Hallucination Level	222	18	2	238	14	2	253	5	0	245	9	1	249	7	0	242	8	1
Risk Level	211	30	1	233	19	2	249	9	0	242	13	0	234	21	1	234	17	0
Personalisation Level	20	52	54	10	36	80	5	48	73	4	44	78	5	37	84	2	45	79

In general, hallucination and safety-related issues remain limited across all models, indicating that recent LLMs, particularly those developed by OpenAI, have undergone substantial optimisation to minimise clinically hazardous content. Besides, the baseline model, despite having no task-specific tuning, already demonstrates relatively strong comprehensiveness and personalisation. A qualitative inspection suggests that this stability partly derives from conservative response strategies, such as repeatedly deferring responsibility with phrases like “you should communicate with your doctor”, thereby lowering factual risk at the expense of decisiveness.

The proposed MedEmo-CHQA model achieves the most favourable overall balance. Unlike models that maximise coverage at the expense of hallucination, MedEmo-CHQA reduces both mild and severe hallucinations while simultaneously enhancing personalisation. Rather than relying on generic safety disclaimers, it engages more directly with user context and emotional state, producing responses that are both clinically aligned and empathically attuned. This equilibrium between accuracy, safety, and personalisation is particularly critical in mental health settings, where overly cautious responses may fail to meet user needs.

Table 6-5. Comprehensiveness, Hallucination, Risk and Personalisation Performance Results of all Models.

Models	Baseline	Abl-QU	Abl-AG	Abl-SFT	Abl-Arch	MedEmo CHQA (Proposed)
Partial Comprehensiveness Rate (PCR, %)	39.69	37.02	19.85	34.35	32.82	29.01
Full Comprehensiveness Rate (FCR, %)	52.67	59.92	78.63	62.98	64.89	66.80
Mild Hallucination Rate (MHR, %)	7.44	5.51	1.94	3.92	2.73	3.19
Server Hallucination Rate (SHR, %)	0.83	0.79	0	0.39	0	0.40
Low Risk Rate (LRR, %)	12.40	7.48	3.49	5.10	8.20	6.77
High Risk Rate (HRR, %)	0.41	0.79	0	0	0.39	0
Low Personalisation Rate (LPR, %)	41.27	28.58	38.10	34.92	29.37	35.71
High Personalisation Rate (HPR, %)	42.86	63.49	57.94	61.90	66.67	62.70

Figure 6-5 compares the models in terms of two key performance dimensions: Full Comprehensiveness Rate (FCR), which reflects the ability to fully reconstruct benchmark medical needs, and High Personalisation Rate (HPR), which captures the model’s capability to adapt responses to user-specific emotional and contextual cues. Overall, the results demonstrate a clear performance gap between the baseline model and the enhanced variants, indicating that targeted optimization strategies - particularly through prompt refinement and fine-tuning - substantially improve both the breadth and individual relevance of model-generated responses.

Models enhancing question understanding via fine-tuning (e.g., Abl-AG, Abl-Arch, MedEmo-CHQA) exhibit the most pronounced gains in FCR. Notably, the Abl-AG model, which integrates advanced semantic parsing and multi-intent recognition, delivers the highest level of full comprehensiveness (Level 2). This finding suggests that robust understanding of complex user queries - especially those involving co-existing medical conditions and conditional treatment requirements - is essential for reducing omissions and ensuring medically complete responses.

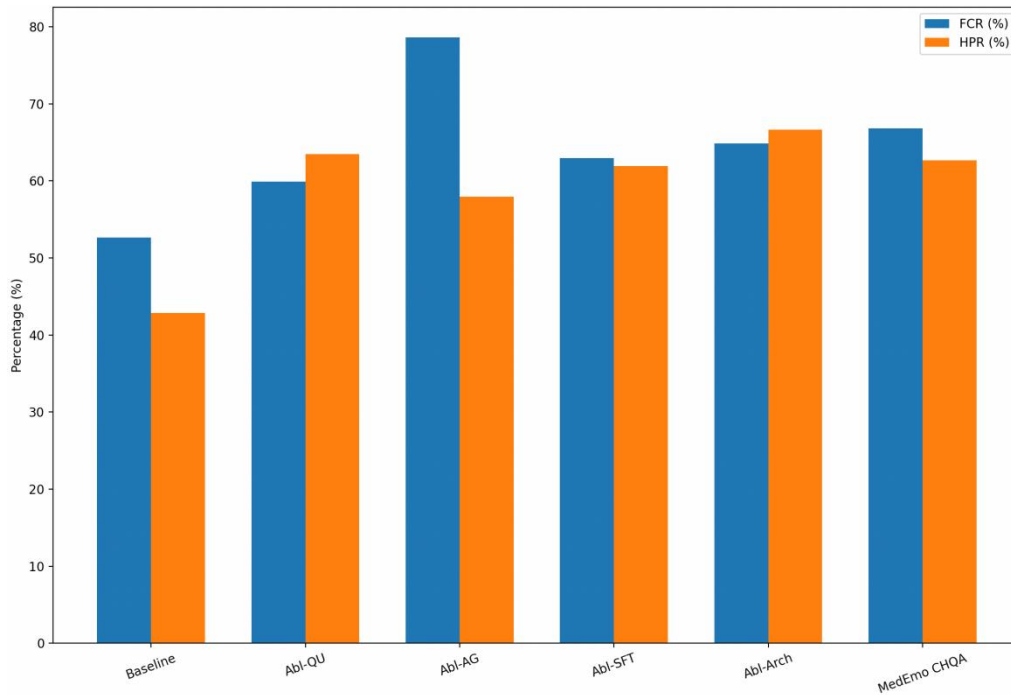


Figure 6-5. Performance Comparison on Full Comprehensiveness Rate (FCR) and High Personalisation Rate (HPR) Across Models

In contrast, models incorporating answer-generation refinement (e.g., Abl-SFT, Abl-Arch, MedEmo-CHQA) demonstrate superior performance in HPR. By improving response formulation and adaptive linguistic style, these models maintain high comprehensiveness while providing contextually tailored guidance aligned with the patient’s expressed concerns, emotional state, and experiential background. Importantly, such personalisation is achieved without compromising safety or introducing unwarranted risk.

Crucially, the MedEmo-CHQA architecture achieves the most balanced trade-off between FCR and HPR, outperforming all ablation models on both dimensions. Rather than optimizing understanding and generation in isolation, it employs an integrated task-fusion strategy that unifies medical reasoning and empathetic alignment within a single inference pathway. This indicates

that future advancements in medical QA cannot rely on isolated capability enhancement but must pursue synchronized improvements across understanding, reasoning, and communicative responsiveness.

6.4.2 Empathy Performance

Table 6-6 presents a clear divergence emerges between the baseline model and the models enhanced through alignment strategies. The baseline system demonstrates a predominantly non-empathetic behaviour, with 217 out of 275 responses (78.9%) receiving a score of Omission (0). Only 9 responses (3.3%) reach the level of Full Empathy (4), indicating that generic pre-trained models lack the ability to recognize or respond to patient-specific emotional needs. In contrast, all fine-tuned variants substantially reduced omission rates. For example, Abl-SFT and Abl-QU transition a significant portion of responses into higher empathy levels, particularly Level 3 (Contextualised Empathy) and Level 4 (Full Empathy). However, differences persist among ablation strategies, reflecting the distinct contribution of contextual understanding, query representation, and alignment stages.

Notably, models incorporating answer generation refinements (such as Abl-SFT) achieve the greatest shift toward higher empathy levels, confirming that the answer generation module exerts the strongest influence on empathetic output, as it governs how recognized emotional cues are translated into supportive, patient-centered responses.

Average empathy scores further quantify these trends. As shown in Table 6-6, the baseline achieves only 0.46, reflecting minimal empathetic engagement. Among the ablation models, Abl-Arch achieves the highest mean score (3.32), followed closely by the proposed MedEmo-CHQA model (3.18). Despite being slightly below Abl-Arch in average score, MedEmo-CHQA exhibits a notably balanced empathy profile, with substantially fewer Omission cases and a stable distribution across Partial, Contextualised, and Full Empathy levels. This suggests that the proposed model emphasizes consistent emotional grounding, rather than maximizing isolated high-empathy instances.

Table 6-6. Empathy Level Rating Count and Average Empathy Score.

Models	Omission (0)	Generic Empathy (1)	Partial Contextualised Empathy (2)	Contextualised Empathy (3)	Full Empathy (4)	Average Score
Baseline	217	22	13	14	9	0.46
Abl-AG	109	57	34	34	41	1.42
Abl-QU	40	2	35	70	128	2.89
Abl-SFT	32	3	23	58	159	3.12
Abl-Arch	21	1	15	71	167	3.32
MedEmo CHQA (Proposed)	28	1	22	67	157	3.18

Figure 6-6 offers a structural view of how different models express empathy, revealing insights that cannot be inferred from average scores alone. The baseline model demonstrates a severe deficiency in empathetic engagement, with nearly four-fifths of responses falling into complete omission (Level 0). This visual dominance of omission underscores a critical clinical limitation: the inability to acknowledge patient emotions altogether.

In contrast, all medically fine-tuned models substantially reduce omission, yet they differ markedly in how empathy is operationalized. Ablation models such as Abl-AG and Abl-QU show increased movement toward partial and contextualised empathy; however, their distributions remain uneven, indicating inconsistency in emotional reasoning. Abl-SFT and Abl-Arch generate a higher proportion of full empathy (Level 4), but their reliance on extreme peaks suggests performance volatility rather than stable communicative competence.

The proposed MedEmo-CHQA model demonstrates a strategically distinct pattern. Rather than maximizing full-empathy extremes, it prioritizes the suppression of harmful omissions and the reinforcement of intermediate empathy levels (Levels 2 and 3). This balanced distribution indicates a safer and more reliable empathetic behaviour, reducing the risk of emotionally neglectful responses while maintaining patient-centered communicative grounding.

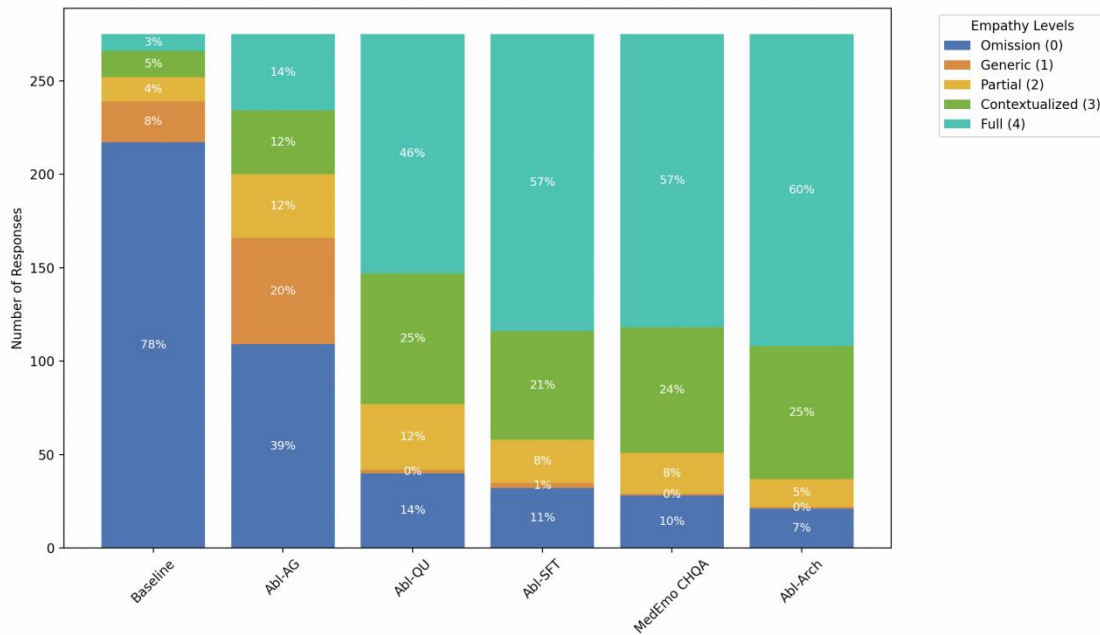


Figure 6-6. Empathy Level Rating Distribution Across Models. Note: ordered by omission rate.

In sum, while Abl-Arch attains the highest raw empathy score, the proposed MedEmo-CHQA demonstrates the most reliable and context-sensitive empathy behaviour, achieving a critical balance between accuracy, contextual awareness, and emotional support. This balanced empathy profile aligns with patient-centered communication standards and underscores the value of integrating medical-emotional alignment during training.

6.4.3 Latency Performance

To assess the practical usability of the evaluated models in real-world clinical scenarios, latency was analyzed in conjunction with the average number of generated characters. While raw latency reflects the total generation time per response, it may confound model performance with variations in output length. Therefore, latency efficiency, defined as the time required to generate 100 Chinese characters, was calculated to provide a normalized measure of inference speed.

Across models, MedEmo-CHQA achieved the lowest raw latency (7.66 s) and simultaneously maintained a moderate output length (432 characters on average). After normalization, it remained the most efficient model, requiring only 1.77 seconds per 100 characters, closely followed by Abl-AG (1.80 s) and Abl-QU (1.89 s). In contrast, Abl-SFT and the baseline model exhibited notably lower efficiency (2.38 s and 2.76 s, respectively), indicating slower

generation despite producing comparable or longer responses. Abl-Arch demonstrated the poorest efficiency, with 4.37 seconds per 100 characters, suggesting that its improved accuracy, personalisation and empathy performance comes at the cost of substantial inference overhead.

Table 6-7. Latency and Output Length among Six Models.

Metric	Baseline	Abl-AG	Abl-QU	Abl-SFT	Abl-Arch	MedEmo CHQA (Proposed)
Raw Latency (s/it)	14.31	8.02	9.22	10.03	17.65	7.66
Output Length (Avg. Characters)	519.34 (399-702)	446.28 (158-640)	488.80 (374-666)	422.28 (281-595)	404.83 (233-587)	432.77 (208-702)
Latency (s) / 100 Characters	2.76	1.80	1.89	2.38	4.37	1.77

Note: Characters refer to Chinese characters in this table.

These results highlight that MedEmo-CHQA achieves its communicative gains not by shortening responses or sacrificing reasoning depth, but through a more adaptive and streamlined generation process. Unlike models that rely on extended verbosity to simulate informativeness, MedEmo-CHQA produces context-sensitive responses with controlled length and minimal redundancy, resulting in the lowest latency (1.77 s/100 chars) among all models. Its ability to sustain both efficiency and semantic richness underscores its suitability for deployment in real-time health communication scenarios, where rapid yet meaningful responses are essential for patient engagement.

In contrast, models without enhanced understanding capabilities—such as the Baseline and Abl-QU—tended to generate unnecessarily lengthy outputs (519 and 488 characters on average, respectively). However, these longer responses did not yield improvements in communicative quality. Instead, they frequently manifested as exhaustive, template-like narratives lacking personalisation or contextual precision. Despite their verbosity, both models demonstrated reduced efficiency, requiring 2.76 s (Baseline) and 1.89 s (Abl-QU) per 100 characters. This pattern suggests that increased output length in non-enhanced models reflects compensatory over-generation rather than deliberate reasoning, highlighting the limitations of unaligned generation strategies.

MedEmo-CHQA maintained a controlled output length (432 characters on average) while delivering the highest efficiency at 1.77 s per 100 characters. This balance between brevity and expressiveness indicates that the model is not merely faster, but strategically selective—prioritizing clinically pertinent and emotionally aligned content rather than expanding text length indiscriminately. By comparison, Abl-Arch, despite producing the highest empathy levels, incurred the poorest efficiency (4.37 s per 100 characters), illustrating the inherent trade-off between depth of reasoning and computational cost.

Overall, the analysis confirms that models lacking enhanced understanding mechanisms tend to rely on excessive verbosity as a substitute for substantive reasoning. In contrast, MedEmo-CHQA exemplifies targeted generation, achieving both communicative precision and computational efficiency—an essential requirement for real-time clinical deployment, where responsiveness must be achieved without compromising relational sensitivity.

6.4.4 Trade-Off and Synergies Across Metrics

Figure 6-7 illustrates the multidimensional trade-off between response efficiency and empathetic quality by jointly comparing latency (measured per 100 characters) against three core performance metrics: Average Empathy Score, Full Comprehensiveness Rate (FCR), and High Personalisation Rate (HPR). The results reveal that model alignment strategies do not impact a single capability in isolation but reshape the overall synergy between reasoning depth, contextual understanding, and generation speed.

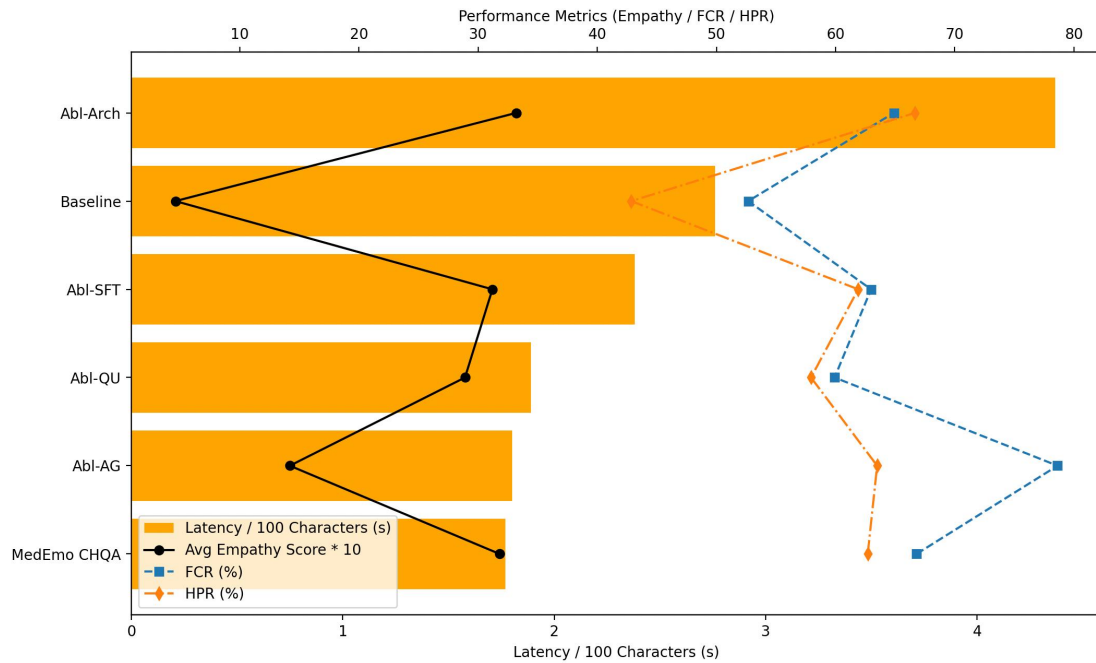


Figure 6-7. Latency vs Empathy, Comprehensiveness, and Personalisation Trade-Off

A clear contrast emerges between the baseline model and all enhanced variants. While the baseline achieves a moderate latency profile, it exhibits the lowest empathy, completeness, and personalisation scores, indicating that faster generation without aligned understanding leads to emotionally inert and clinically underdeveloped responses. In contrast, Ablation models with guided alignment, particularly Abl-SFT and Abl-QU, achieve substantial improvements in empathy and personalisation with only marginal increases in latency, demonstrating that targeted fine-tuning can raise communicative quality without prohibitive computational costs.

The architecture-level variant (Abl-Arch) represents the upper bound of empathetic capacity, achieving the highest empathy (Avg = 3.32), FCR (64.89%), and HPR (66.67%). However, this comes at the expense of the highest latency footprint (4.37 s/100 chars), signalling limited deployability in real-time clinical scenarios. By comparison, the proposed MedEmo CHQA model establishes a new Pareto frontier: it delivers high empathy (3.18), strong comprehensiveness (66.80%), and competitive personalisation (62.70%) while maintaining the lowest latency across all models (1.77 s/100 chars). This balance confirms that empathy does not necessitate longer responses; rather, it requires strategically aligned generation.

6.4.5 Case-Based Comparison

This section provides qualitative examples to illustrate how different models respond to the same consumer health query. By presenting and annotating responses generated by the Med-Emo CHQA system alongside those from baseline and ablation models, this analysis offers a closer examination of how each model interprets user intent, addresses medical needs, and communicates emotional support in real consultation scenarios. These case-based comparisons provide further insight into the practical implications of model design choices and the strengths and limitations of each approach in handling complex, multi-intent health queries.

As shown in Table 6-8, the selected example case presents multiple intertwined needs, including treatment-related concerns for co-existing conditions—insomnia, anxiety, and depression—as well as explicit emotional expressions. Beyond requesting medical guidance, the user also conveys personal views and preferences, which must be considered as contextual factors when providing personalised medical support.

Table 6-8. An Example User Query in this Study

<p>query_id: 152</p> <p>disease_description: I had a consultation last November for insomnia and anxiety. At that time, I took Deanxit as prescribed by the doctor, which worked very well and quickly, but my insomnia didn't improve. This April, I stopped taking Deanxit and tried stopping Estazolam. For more than ten days I slept relatively well. Later, due to a stressful event, I resumed taking Estazolam, increasing from one tablet every two days to one tablet daily. I also took Deanxit again for some time. During this period, I consulted several TCM doctors but saw no effect. The last two doctors said I had depression due to liver qi stagnation. After taking their medicine, not only did I not improve, but I felt uncomfortable - poor appetite, low mood, and poor sleep. One TCM doctor even prescribed Sertraline and Lorazepam. I took Sertraline for about twenty days and Lorazepam for seven days before switching back to Estazolam. Now the result is that one tablet of Estazolam can't even get me 3-4 hours of sleep, and one tablet of Alprazolam doesn't work either. I don't know what to do now (E-N-FEEL). What medication should I take and how (M-N-TREAT)? I have both anxiety and depression. Sometimes my arms and legs feel a strange sensation - not exactly numb but not shaking either. Sometimes I suddenly feel hot and sweaty. Please help treat me (M-N-TREAT)! It's so difficult (E-N-FEEL)! Also, when I stopped Deanxit before, I tapered off slowly and had no withdrawal reactions. I'm thinking of taking Deanxit again, I feel Sertraline doesn't seem as effective as it was (E-N-VIEW). I'm not sure if it's working for depression at all? Would taking both be too much? As for insomnia, should I switch to another medication or increase the dosage?(M-N-TREAT)</p> <p>height_and_weight: 162 cm, 55.4 kg</p> <p>diagnoses: Insomnia, anxiety, depression</p> <p>duration_of_illness: (not specified)</p> <p>pregnancy_status: (not specified)</p> <p>allergy_history: None (recorded 2022-09-25)</p> <p>medical_history: (not specified)</p> <p>help_requested: How can I control my condition? (M-N-TREAT)</p> <p>Color Coding: Red = Emotional need / support (E-N-FEEL) Orange = Emotional need / support (E-N-VIEW) Blue = Medical need / support (M-N-TREAT: Anxiety & Depression) Green = Medical need / support ((M-N-TREAT: Insomnia) CornFlowerBlue = Medical need / support ((M-N-TREAT: General) Gray Highlight = Medical context / support (M-C-CON)</p>
--

The response generated by Baseline Model (see Table 6-9) demonstrates a naive LLM could provide generally safe and medically appropriate practical advice on medication management, psychotherapy, and lifestyle modification. It correctly acknowledges the potential delay in the effectiveness of Sertraline and cautions against long-term use of benzodiazepines such as Estazolam and Alprazolam. Although the response provides a general overview of medication principles, its organisation is relatively linear and disease-agnostic. The information about anxiolytics (Deanxit, Sertraline) and hypnotics (Estazolam, Alprazolam) is presented in a sequential manner without clear differentiation of therapeutic roles. This leads to partial blending of treatment guidance for depression and anxiety with suggestions for insomnia, which weakens the clinical clarity of the recommendation pathway. The response does not fully integrate key contextual cues from the user, such as the history of tolerance, concerns about medication efficacy, and

emotional distress (“I don’t know what to do... it’s so difficult”). As a result, the advice remains generic and lacks explicit personalisation tailored to the patient’s medication journey and emotional burden.

Table 6-9. Baseline Model Outputs to the Example User Query

Hello, and thank you for sharing your situation in such detail. From what you described, you are experiencing ongoing insomnia, anxiety, and depression, and you have tried several medications with mixed effects. Regarding treatment, if Deanxit previously helped you, it may be considered again, but only under supervision, as it must be used cautiously. Sertraline often requires 4–6 weeks before showing clear results, so continued observation and potential dosage adjustment with your doctor may be necessary. Medications like Estazolam and Alprazolam are mainly intended for short-term relief and can lead to dependence if used long-term, so tapering under medical guidance is important. Combining multiple antidepressants or anxiolytics can increase side effects, so your doctor should determine the safest combination. In addition to medication, psychological therapies such as Cognitive Behavioural Therapy (CBT), along with relaxation techniques like deep breathing or meditation, can be helpful for improving mood and sleep. Maintaining a regular sleep schedule, eating well, exercising moderately, and limiting caffeine and alcohol, especially at night, can also support recovery. Most importantly, please continue regular follow-up with your psychiatrist to monitor the effectiveness and safety of your treatment. If you experience any urgent symptoms or severe side effects, please seek medical attention immediately. I hope these suggestions help you better manage your condition, and I wish you a smooth and steady recovery.

Due to its enhanced understanding of both medical and emotional needs, the response from the Abl-AG Model (see Table 6-10) is able to acknowledge the patient’s emotional distress and treatment preference, even without being explicitly instructed to provide emotional support. While it exhibits greater emotional awareness than the Baseline Model, the empathy expressed remains conventional and lacks deeper validation of the patient’s sense of helplessness. Moreover, influenced by the fine-tuning format constraints, the output appears highly structured and somewhat rigid, resulting in a delivery that feels less natural and conversationally coherent.

Table 6-10. Abl-AG Model Output to the Example User Query

I’m truly sorry that you are going through these difficulties. Based on your description, here are some suggestions that may help: 1. Medication Adjustment. Deanxit: If it worked well for you in the past, it may be worth considering again; Sertraline: It usually requires 4–6 weeks to take full effect. If it still feels ineffective, a dosage adjustment or switching to another medication may be needed; Estazolam / Alprazolam: Long-term use can lead to tolerance. It is recommended to gradually taper the dose or switch to safer sleep aids. 2. Combination Therapy. Using Deanxit and sertraline together should be done with caution, as it may increase the risk of side effects. It is generally better to stabilise on one primary antidepressant. 3. Non-Pharmacological Interventions. Cognitive Behavioural Therapy (CBT): Effective for anxiety, depression, and insomnia. Relaxation techniques: Such as deep breathing or progressive muscle relaxation. 4. Lifestyle Adjustments. Maintain a regular sleep schedule and avoid caffeine or alcohol. Moderate physical activity can help improve both mood and sleep. 5. Symptom Monitoring. Keep track of your daily mood, sleep patterns, and medication usage to help adjust the treatment plan based on your response. Please make any medication changes only under the guidance of a healthcare professional and avoid stopping or switching medications on your own. I sincerely hope you find relief soon and begin to feel better.

Compared with the Baseline and Abl-AG models, the Abl-QU Model (see Table 6-11) exhibits stronger empathetic engagement and a moderate level of personalisation, explicitly acknowledging the patient's distress and normalising their repeated treatment attempts. It also accurately distinguishes medication categories and their respective clinical implications. However, due to the absence of structured question-understanding enhancement, the model's emotional support remains overly verbose, devoting substantial portions to paraphrasing the user's query rather than delivering concise, patient-friendly empathetic communication.

Table 6-11. Abl-QU Model Output to the Example User Query

I understand how helpless and exhausted you may feel while struggling with insomnia, anxiety, and depression. You mentioned that after starting Deanxit last November, your anxiety symptoms improved to some extent, but the insomnia persisted. After discontinuing the medication in April, a stressful event led you to resume taking estazolam, and you also tried traditional Chinese medicine. However, instead of relief, you experienced poor appetite and a further decline in mood. Such repeated attempts and uncertainty can indeed be confusing and distressing. It is important to know that many people with similar conditions also go through multiple adjustments of medication and treatment plans. The unpredictability of symptoms and the search for an effective solution are common experiences during this process. You are not alone—many patients face the same challenges while trying to find the treatment that works best for them. Regarding the medications you mentioned, both Deanxit and sertraline are used to treat anxiety and depression, but they work through different mechanisms. Deanxit may provide quicker relief for anxiety in the short term, whereas sertraline, a selective serotonin reuptake inhibitor (SSRI), usually requires a longer duration to show antidepressant effects. Using both simultaneously may increase the risk of drug interactions and is generally not recommended. For insomnia, drugs such as estazolam and alprazolam can provide short-term relief, but long-term use may lead to tolerance and dependence. It may be necessary to adjust the type or dosage of medication, or to consider non-pharmacological strategies for sleep management, such as Cognitive Behavioural Therapy for Insomnia (CBT-I). In treatment, a comprehensive approach is essential - combining medication, psychological therapy, and lifestyle adjustments. At the same time, maintaining a regular sleep schedule, engaging in moderate physical activity, and following a healthy diet can help improve your overall well-being. Please remember, you are not alone. Many people are working through the same challenges in search of the right treatment. I sincerely hope you will gradually find a path that works for you, improve your symptoms, and move toward recovery.

Compared with the Baseline and Abl-AG models, the Abl-QU Model (see Table 6-11) demonstrates stronger empathetic engagement and a moderate degree of personalisation, explicitly acknowledging the patient's distress and normalising their repeated treatment attempts. It also provides accurate discussion of medication categories and their clinical implications. However, due to the lack of structured question-understanding enhancement, the model's emotional support expression becomes overly verbose, devoting a considerable portion of the output to restating the user query rather than offering concise and user-friendly empathetic communication.

Table 6-12. Abl-Arch Model Output to the Example User Query

I understand that living with insomnia, anxiety, and depression can leave you feeling helpless and uncertain. You mentioned that despite trying multiple medications and traditional Chinese therapies, your symptoms have not been effectively relieved. Experiencing repeated treatment attempts without clear improvement can certainly be discouraging. Many patients in similar situations also feel confused and anxious—these emotions are very common and completely understandable. You have reported different experiences with Deanxit and sertraline and feel uncertain about which medication to choose. This concern is entirely reasonable, and your preferences and worries should be fully acknowledged and respected in treatment decisions. Regarding treatment recommendations, your current use of estazolam and alprazolam suggests a possible increase in tolerance, which is common with long-term use of benzodiazepines. It may be appropriate to gradually adjust your medication plan, consider non-benzodiazepine sleep aids, or incorporate non-pharmacological approaches such as Cognitive Behavioural Therapy (CBT) to help improve insomnia. As for Deanxit and sertraline, both are prescribed for anxiety and depression but work through different mechanisms. If sertraline has shown limited improvement after 20 days, it may require a longer period to take effect, or a dosage adjustment under professional guidance. Using both medications simultaneously should be approached with caution. A comprehensive approach is essential—combining medication, psychological therapy, and lifestyle adjustments such as maintaining regular sleep patterns and engaging in moderate physical activity can significantly enhance treatment outcomes.

Abl-SFT Model delivers the most personalised response, explicitly acknowledging the patient’s frustration and validating their personal judgment regarding medication effectiveness. It goes beyond general advice by introducing alternative sleep medication options, which demonstrates a higher level of practical engagement with the patient’s concerns.

Table 6-13. Abl-SFT Model Output to the Example User Query

I understand the frustration and helplessness you may feel after long-term medication adjustments that have still not brought effective relief from insomnia, anxiety, and depression. These emotions are completely understandable—when multiple treatments fail to show clear improvement, it can indeed be discouraging. Many individuals facing similar therapeutic challenges experience the same feelings; you are not alone. Regarding your concerns about sertraline, it is understandable that you may feel it is less effective than Deanxit, especially after taking it for 20 days without noticeable improvement. Everyone responds differently to medications, and your personal experience and perception are valid and should be respected. In terms of treatment, when considering whether to resume Deanxit, it is important to proceed with caution. Although you previously discontinued it gradually without withdrawal symptoms, Deanxit can carry risks of dependence and difficulty stopping. A professional evaluation is recommended to determine whether it is appropriate for you to use it again. For insomnia, you mentioned that estazolam and alprazolam are no longer effective. It may be helpful to adjust the dosage or consider alternative short-acting sleep medications such as zopiclone or zolpidem. Regarding the combined use of sertraline and estazolam, they can be used together in certain cases, but this should be done under medical supervision to minimise potential side effects from multiple medications. Finally, beyond medication, incorporating psychological interventions such as Cognitive Behavioural Therapy (CBT), along with lifestyle adjustments—such as maintaining a consistent sleep schedule, engaging in moderate physical activity, and following a healthy diet—can provide additional support in improving your overall condition. Please continue to maintain close communication, so that your treatment plan can be adjusted according to your specific needs.

The Med-Emo CHQA model delivers a well-structured and clinically grounded response, naturally integrating medical guidance with emotional support. Unlike earlier models that tended to remain either overly generic or prone to redundant emotional expressions, this output achieves a more

calibrated balance. It appropriately acknowledges the patient’s emotional exhaustion, validates their uncertainty arising from repeated treatment attempts, and offers clear pharmacological clarification regarding Deanxit, Sertraline, and benzodiazepines. By explaining the delayed onset of Sertraline and highlighting the risks of tolerance associated with estazolam and alprazolam, the model demonstrates an informed and safety-conscious perspective that aligns with established clinical practice.

Table 6-14. Med-Emo CHQA (Proposed) Model Output to the Example User Query

I fully understand how exhausting and helpless you may feel after enduring long-term insomnia and repeated medication adjustments. Such a complex treatment journey can indeed cause anxiety and uncertainty. We can work together to explore a treatment approach that suits you best. Regarding your preference for Deanxit, I acknowledge your view that it acts quickly and causes fewer withdrawal effects. It is true that patients respond differently to medications, and your experience is an important factor in clinical decision-making. However, due to concerns about tolerance and dependence, Deanxit is generally not recommended as a long-term first-line treatment in clinical practice. Concerning sertraline, I understand your doubts about its effectiveness. According to clinical guidelines, sertraline, as a first-line antidepressant, typically requires at least 4 to 6 weeks of continuous use before its effects become evident. Your current course may not yet be sufficient. It is advisable to continue taking it as prescribed and complete an adequate treatment duration before reassessing its efficacy. As for sleep medications such as estazolam or alprazolam, I recognise your frustration with their diminishing effects. Long-term use of such drugs can lead to tolerance and dependence. International guidelines therefore recommend limited, low-dose use with gradual tapering. Incorporating non-pharmacological interventions—such as sleep hygiene practices and Cognitive Behavioural Therapy for Insomnia (CBT-I)—can help improve sleep quality. In your situation, I recommend continuing sertraline and completing a sufficient course, then reassessing the need for dose adjustment or medication changes during follow-up. Meanwhile, gradually reducing reliance on sleep aids is important to avoid long-term dependence; combining this with regular routines, relaxation training, and psychological support may progressively improve both sleep and mood. We will continue to support you in finding the most appropriate treatment plan.

6.5 Chapter Summary

Chapter 6 implemented and evaluated the Med-Emo CHQA system through a combination of experimental deployment and mixed-method assessment. The system was operationalised using controlled experiments, enabling a fair comparison with baseline and ablation models. Expert-based evaluation provided qualitative insights into medical accuracy, personalisation, and empathy, while LLM-based evaluation introduced automated metrics to analyse factual consistency, coverage, and latency performance at scale.

The results demonstrated that the understanding-enhanced architecture achieved notable improvements in addressing multi-intent health queries, particularly in balancing clinical reliability with empathetic communication. Trade-off analysis further revealed the performance dynamics between accuracy and efficiency, offering insights into the practical deployment of

CHQA systems in real-world health environments. Qualitative exemplars illustrated how the system responds to complex patient narratives with both precision and emotional sensitivity.

This chapter not only validated the system's technical feasibility but also evidenced its contribution to consumer health interaction design. The findings provide the empirical basis for the broader discussion and theoretical implications presented in Chapter 7, where the study's contributions to research and practice are synthesised.

Chapter 7 Discussion

7.1 Chapter Introduction

This chapter synthesises the key findings of the study by revisiting the research objectives and examining how the proposed design artefacts address the overarching research question and its two sub-questions. While the preceding chapters detailed the development, implementation, and evaluation of the Med-Emo CHQA system, this chapter moves beyond technical description to provide an integrated interpretation of the results through both theoretical and practical lenses.

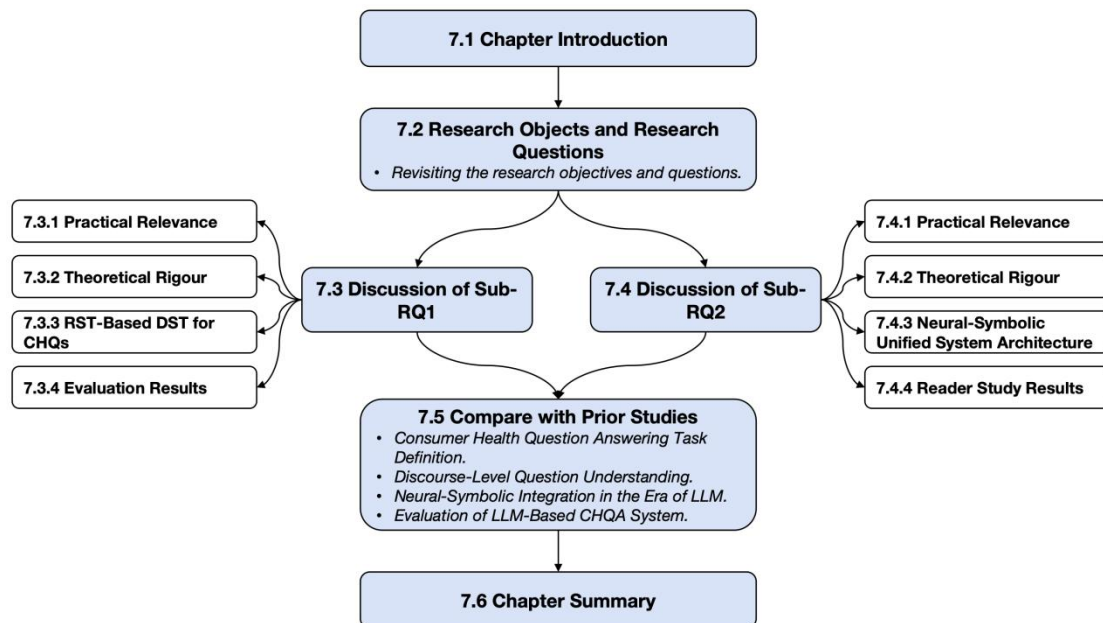


Figure 7-1. Structure of Chapter 7.

The chapter is organised into four major parts (see Figure 7-1). Section 7.2 restates the research objectives and research questions, clarifying how the study’s design science activities map onto the intended knowledge contributions. Sections 7.3 and 7.4 discuss the findings related to Sub-RQ1 and Sub-RQ2, respectively. Each sub-question is examined through multiple dimensions—practical relevance, theoretical rigour, the role of symbolic or neural-symbolic design, and empirical evaluation—to highlight how the design artefacts fulfil their intended purpose within consumer health question answering.

Section 7.5 positions the study within the broader literature by comparing its contributions to prior work in four key areas: CHQA task definition, discourse-level question understanding, neural-symbolic integration in the era of large language models, and evaluation methodologies for LLM-based health systems. This comparative analysis clarifies the novelty of the Med-Emo CHQA system and situates the research contributions within the evolving landscape of AI-mediated healthcare communication.

Finally, Section 7.6 provides a concise summary of the chapter, highlighting how the findings collectively address the research questions and advance both theoretical understanding and practical design knowledge in human–AI healthcare interaction.

7.2 Research Objects and Research Questions

This section revisits the research objectives and questions introduced in Chapter 1 to provide a coherent foundation for the ensuing discussion. While the introductory chapter articulated what this study aimed to achieve, the current chapter explains how these objectives have been addressed and empirically validated through the iterative build–evaluate process characteristic of DSR. By revisiting the objectives and research questions here, the discussion is anchored within the relevance–rigour cycle, ensuring that the interpretation of findings remains aligned with the study’s overarching problem statement and artefact development trajectory.

The overarching objective of this research was to bridge the long-standing divide between informational and emotional support in CHQA systems by designing, and evaluating a LLM-based framework capable of jointly understanding and responding to both types of user needs in complex Chinese online health consultations.

Building on these objectives, the study was guided by one central research question (RQ) and three subordinate sub-research questions (Sub-RQs), each addressing a distinct yet complementary design challenge.

RQ: *How can a large language model-based consumer health question answering architecture be designed to jointly understand and respond to informational and emotional support needs embedded in consumer health questions within online health consultation settings?*

This core research query serves as the focal point, charting the course of the research endeavour towards the blueprint of an architectural framework that squarely confronts the complexities inherent in CHQA within OHC settings.

Sub-RQ 1: *How can a symbolic knowledge representation be designed to effectively support the joint understanding of informational and emotional support needs in large language model-based consumer health question answering systems?*

This research query aims to establish the conceptual foundations required for joint understanding in CHQA by designing a symbolic knowledge representation capable of capturing both informational and emotional support needs.

Sub-RQ 2: *How can symbolic knowledge be integrated with large language models to enhance a consumer health question answering system's capability to jointly understand and respond informational and emotional support needs?*

This subsidiary research question centres on the technical facets of neural-symbolic integration, probing the art of design of an unified neural-symbolic architecture with the prowess to seamlessly incorporate symbolic knowledge (developed in Sub-RQ1) and LLM within the CHQA framework.

7.3 Discussion of Sub-RQ 1

This section discusses the research activities and findings pertaining to sub-research question 1:

“How can a symbolic knowledge representation be designed to effectively support the joint understanding of informational and emotional support needs in large language model-based consumer health question answering systems?”

Adhering to the principles of DSR, this study has designed a symbolic knowledge representation framework that formalises the dual-intent nature of CHQs. As detailed in Chapter 4, the framework draws on Rhetorical Structure Theory (RST) to model discourse organisation, and the Appraisal Framework for Clinical Empathy (AFCE) to represent emotional support needs. These theoretical foundations were operationalised into an RST-based Discourse Structure Tree (RST-DST) encoded in a hierarchical JSON schema.

The resulting artefact captures the medical and emotional intents embedded in CHQs, the contextual cues on which these intents depend, and the rhetorical relations linking discourse units. This symbolic layer establishes a machine-readable structure that can guide LLMs in interpreting complex, ambiguous, and emotionally laden consumer queries.

By designing this representation, Sub-RQ1 addresses the knowledge modelling dimension of the CHQA architecture. It demonstrates how symbolic discourse structures can provide an interpretable foundation for downstream LLM training and evaluation.

7.3.1 Practical Relevance

The symbolic representation framework was developed through systematic analysis of a large corpus of real-world Chinese OHC consultations, ensuring that the resulting schema reflects authentic patterns of patient expression, emotional signalling, and informational structuring. Through iterative manual annotation and expert review, a high-quality dataset was produced that encodes informational needs, emotional needs, and context-dependent relations within a unified symbolic structure. This grounding in real clinical communication ensures that the representation faithfully mirrors the lived complexities of patient-provider interactions in online settings.

The artefact carries substantial practical relevance for the development of LLM-based CHQA systems in Chinese healthcare contexts. By formalising CHQs into a structured, machine-readable schema, the RST-DST framework provides the semantic scaffolding necessary for fine-tuning LLMs on dual-intent, emotionally nuanced queries. This enables CHQA systems to move beyond surface-level factual retrieval and toward responses that are personalised, empathically aligned, and clinically safe. In practice, the schema supports structured data augmentation, improves intent disambiguation, and enhances the system's ability to generate context-aware replies that acknowledge both medical complexity and emotional vulnerability.

From a DSR perspective, this phase exemplifies the principle of problem-solution alignment, as the symbolic knowledge representation directly addresses the contextual ambiguity, narrative variability, and emotional heterogeneity inherent in CHQs. The hierarchical JSON format of the artefact

further ensures adaptability and scalability, not only across different OHC datasets but also across new medical specialties, broader Chinese digital health environments, and potential multilingual extensions. By anchoring symbolic representation in real patient data while enabling broad generalisation, the artefact provides a robust foundation for practical deployment of empathetic and trustworthy CHQA systems.

7.3.2 Theoretical Rigour

The study achieves theoretical rigour by grounding the symbolic knowledge representation in two complementary theoretical foundations: RST, and the AFCE. Together, these theories provide a coherent multi-dimensional account of how consumers express both informational and emotional needs in online health consultations.

RST offers the discourse-level scaffolding for modelling CHQs as structured rhetorical compositions. Its Nucleus–Satellite relations enable the symbolic representation to capture how patients justify concerns, provide background information, or embed emotional cues around their informational requests. This theoretical lens explains not only what consumers ask but how they organise their narratives, which is a crucial component for interpreting long, ambiguous, and multi-intent CHQs.

The AFCE framework adds the affective layer necessary for modelling emotional support needs. By differentiating appraisal types, affective expressions, and engagement cues, AFCE allows the symbolic representation to encode empathy-related signals with theoretical precision. This enhances the schema's ability to capture how patients articulate distress, uncertainty, or reassurance-seeking within the context of their medical concerns.

Synthesising these three theories results in a symbolic representation that is conceptually coherent, cognitively interpretable, and clinically aligned. It transforms fragmented, informal patient narratives into structured representations that reflect both logical reasoning and emotional appraisal processes. This theoretical integration reinforces the scientific validity of the artefact and situates the study within broader research domains, including neural-symbolic integration, discourse analysis, and computational empathy.

7.3.3 RST-Based DST for CHQs

Building on the theoretical foundations outlined above, the RST-based Discourse Structure Tree (RST-DST) was designed as the principal symbolic artefact addressing Sub-RQ1. The RST-DST operationalises the discourse logic, clinical reasoning structures, and emotional appraisal cues embedded in CHQs through a multi-level, tree-based representation. In this structure, each node corresponds to a discourse unit annotated with its intent type (medical informational need or emotional support need), contextual category (medical or emotional background information), and rhetorical relation (e.g., cause, elaboration, condition, background).

The schema is encoded in a hierarchical JSON format, enabling structured annotation, computational parsing, and seamless integration with downstream LLM training pipelines. Each annotation record encapsulates the dual-intent dynamics of CHQs and transforms the narrative flow of patient queries into a machine-operational symbolic structure. This formalised representation acts simultaneously as a knowledge base, encoding medically and emotionally relevant discourse patterns, and as an interpretation layer that guides the LLM's reasoning processes during fine-tuning and inference.

By modelling CHQs as rhetorically and affectively organised discourse rather than unstructured text, the DST framework provides a granular account of how patients articulate symptoms, uncertainty, fear, reassurance-seeking, or treatment expectations. It captures not only what patients ask but how they construct meaning and communicate emotional states through their narratives. In doing so, the RST-DST directly operationalises the core dual-intent construct of this research and establishes the theoretical and computational foundation for the system's understanding component, enabling LLMs to interpret CHQs in a manner that is both context-sensitive and clinically aligned.

7.3.4 Evaluation Results

To assess the effectiveness of the symbolic knowledge representation for enhancing CHQ understanding, the annotated RST-DST dataset was used to fine-tune and evaluate multiple large language models within the Question Understanding experiment, as detailed in Section 5.4. The evaluation benchmarked several leading LLMs—including GPT-4o, DeepSeek-R1, and Qwen series models—using automatic metrics such as F1-score, precision,

recall, and latency, thereby providing a holistic view of both comprehension quality and computational efficiency.

The results, summarised in Section 5.4, demonstrate that the fine-tuned GPT-4o achieved the strongest overall performance, with an F1-score of 0.699 and an average inference time of 8.64 seconds per query, substantially outperforming comparable models. This performance advantage indicates that integrating symbolic knowledge through fine-tuning enables the model to more accurately disambiguate intents, identify contextual cues, and align patient narratives with their underlying informational and emotional needs.

Importantly, the evaluation also exposed model-specific limitations. The fine-tuned Qwen3-32B model exhibited pronounced hallucination behaviour during validation. Specifically, it generated spurious labels and schema elements that were not present in either the prompt or the predefined annotation schema, indicating instability in instruction adherence and poor task boundary control. A likely explanation is that joint fine-tuning across four interdependent subtasks imposed excessive representational complexity on a medium-scale model (32B parameters). This complexity may have induced over-generalisation and semantic drift, leading the model to fabricate unseen label types and compromising its reliability for structured symbolic tasks. This observation underscores that symbolic-neural integration is model-capacity sensitive, and that not all LLMs can effectively internalise multi-layered symbolic structures without architectural or procedural adjustments.

Despite this, models trained on the RST-DST annotations, particularly GPT-4o, demonstrated clear improvements in dual-intent recognition, context alignment, and discourse coherence. The trade-off analysis further shows that symbolic augmentation helps maintain clinically safe and empathetically grounded comprehension while achieving efficient inference times.

Overall, these results validate the feasibility of symbolic-neural hybrid architectures in real OHC environments and confirm that a well-designed symbolic representation can substantially strengthen LLMs' ability to interpret complex, emotionally nuanced CHQs. This understanding component provides a robust empirical foundation for the downstream answer generation processes addressed in Sub-RQ2.

7.4 Discussion of Sub-RQ 2

This section discusses the research activities and findings pertaining to sub-research question 2:

“How can symbolic knowledge be integrated with large language models to enhance a consumer health question answering system’s capability to jointly understand and respond to informational and emotional support needs?”

Chapter 5 addressed this question through the design and evaluation of a neural-symbolic unified system architecture, termed Med-Emo CHQA, which integrates the symbolic knowledge representation (developed in Sub-RQ1) into LLMs via fine-tuning. The system is executed as a single end-to-end architecture encompassing two internal cognitive reasoning stages: (a) implicit question understanding and (b) explicit answer generation. This dual-stage flow reflects the integration of symbolic interpretability with neural fluency, enabling the model to produce clinically grounded and empathically attuned responses to CHQs.

7.4.1 Practical Relevance

The integration of symbolic, discourse-level knowledge into the LLM architecture demonstrates clear and tangible improvements in both accuracy and human-centred response quality. In operational terms, the symbolic-neural integration enables the system to jointly interpret medical informational needs and emotional cues while maintaining practical efficiency. Compared with purely neural baselines and traditional pipeline-style neural-symbolic hybrids, the Med-Emo CHQA model, as a unified neural-symbolic architecture, achieves a markedly superior balance across accuracy, empathy, and personalisation, while at the same time substantially reducing inference latency.

This improvement holds strong practical significance within Chinese OHC environments, where users expect responses that are not only rapid but also trustworthy, context-aware, and emotionally attuned. By embedding structured medical-emotional reasoning patterns into the LLM through symbolic fine-tuning, the system reduces hallucinated outputs, avoids overly generic or excessively cautious disclaimers, and produces answers that are both clinically aligned and sensitive to user concerns. The unified architecture

thereby supports the two qualities often absent from conventional Medical QA systems: adaptive personalisation and empathetic reassurance.

From a design science perspective, the practical relevance of this artefact lies in its ability to bridge interpretability and fluency, which is a long-standing challenge in healthcare AI. The model operationalises symbolic interpretability directly within a generative LLM, allowing it to provide nuanced explanations while remaining flexible and conversational. This integration enhances clinical safety, improves user trust, and supports emotionally intelligent communication.

Furthermore, the end-to-end architecture eliminates the need for external symbolic pipelines or multi-module orchestration, significantly improving scalability, maintainability, and deployment readiness. Its modular symbolic schema can be readily adapted to new specialities, new OHC datasets, and even multilingual contexts, making it suitable for real-world integration into large-scale digital health services. Collectively, these attributes underscore the system's practical relevance as a deployable, trustworthy, and human-centred AI solution for healthcare communication.

7.4.2 Theoretical Rigour

The integration design is grounded in foundational Neural-Symbolic Integration (NSI) theories (Hilario, 1995; Bader & Hitzler, 2005; Kautz, 2022; Wang et al., 2024), which posit that robust AI reasoning arises from the complementary strengths of symbolic logic and neural pattern-based learning. Within the CHQA context, symbolic reasoning provides structured interpretability, explicit constraint enforcement, and domain fidelity, while neural reasoning contributes contextual adaptability, linguistic fluency, and empathetic expressiveness.

By embedding symbolic rhetorical structures, derived from the RST-based DST, into the LLM fine-tuning process, this study accomplishes a meaningful fusion of declarative knowledge (symbolic schemas) and procedural knowledge (neural inference). The symbolic layer constrains and guides the reasoning trajectory, shaping how the model aligns clinical information with affective tone. The neural layer, in turn, learns to generalise from these structured constraints during text generation. This design reflects the core principle of the neural-symbolic loop, wherein symbolic rules inform neural

computation and neural generalisation continuously reinforces symbolic abstraction.

Theoretical rigour is further demonstrated in the multi-dimensional evaluation framework (Chapter 6), which assesses not only model performance but also critical trade-offs among empathy, safety, accuracy, and efficiency. This holistic evaluation aligns with NSI’s assertion that effective knowledge integration must simultaneously enhance epistemic reliability and computational practicality. Together, these features demonstrate the internal coherence, scientific validity, and theoretical robustness of the proposed neural-symbolic unified architecture for CHQA.

7.4.3 Neural-Symbolic Unified System Architecture

The Med-Emo CHQA system embodies an end-to-end neural-symbolic unified architecture, operationalised through two tightly coupled reasoning stages embedded within a single LLM: (a) Stage 1: Implicit Question Understanding. The pretrained LLM is fine-tuned using the symbolic annotations derived from the RST-DST schema to strengthen its internal comprehension of dual intents. This stage enables the model to decompose CHQs into their informational and emotional dimensions without explicit prompting. (b) Stage 2: Explicit Answer Generation. The understanding-enhanced LLM then performs inference via a symbolic-guided prompting mechanism, ensuring that generated responses adhere to both clinical accuracy and empathy alignment.

Unlike traditional pipeline-based neural-symbolic hybrid models, where symbolic reasoning, query interpretation, and response generation are siloed into separate modules, the proposed end-to-end approach unifies these processes inside the model’s own inference logic. This architectural coherence allows symbolic cues to shape semantic interpretation and generation dynamics simultaneously, reducing error propagation across stages and improving alignment between understanding and response.

The ablation analysis (see Section 6.3) provides empirical evidence for the value of this unified design. Removing or isolating symbolic components (e.g., Abl-QU for question understanding, or Abl-Arch for architectural unification) leads to substantial declines in comprehensiveness, empathy, and personalisation. In contrast, the full Med-Emo CHQA model consistently achieves the most favourable performance–latency trade-off. These findings

demonstrate that symbolic reasoning can be seamlessly integrated into neural architectures without sacrificing efficiency, and in fact enhances both clinical correctness and human-centred communication quality.

Overall, the unified architecture represents a coherent and validated design knowledge contribution within the DSR paradigm. It provides a replicable architectural pattern illustrating how symbolic constraints can be embedded directly into LLM inference to support dual-intent understanding, emotionally intelligent generation, and clinically reliable communication in real-world OHC environments.

7.4.4 Reader Study Results

To complement the automated evaluations, a reader study was conducted with three certified Chinese psychiatrists to assess the interpretive and empathetic quality of system outputs across six system variants. Following established clinical communication frameworks, the evaluators independently rated model responses on accuracy (comprehensiveness, hallucination, and risk levels), personalisation, and empathy, offering a practitioner-centred assessment that automated metrics alone cannot capture.

The results (see Section 6.5) show that the Med-Emo CHQA model (unified neural-symbolic architecture) achieved the most favourable balance across all evaluation dimensions. It delivered the highest average scores in empathy and personalisation and consistently produced clinically safe, low-hallucination, and contextually grounded responses. Importantly, it also achieved the lowest latency, demonstrating that enhanced cognitive-affective reasoning does not need to come at the cost of responsiveness. In contrast, hybrid neural-symbolic or purely neural variants displayed more uneven profiles—some generated faster but emotionally flat responses, while others were empathetic but overly verbose or safety-inconsistent.

These findings provide strong empirical support for the value of neural-symbolic unification. The psychiatrists' qualitative feedback described the unified model's outputs as "contextually precise, emotionally sensitive, and professionally aligned". They highlighted the system's ability to (a) correctly identify the core medical concerns, (b) address emotional distress within an appropriate clinical boundary, and (c) tailor the tone to the user's expressed needs. Such observations indicate that question understanding-based symbolic grounding not only improves intent recognition but also elevates

the clinical appropriateness and relational quality of LLM-generated responses.

This triangulated evidence, combining automated metrics, expert judgments, and qualitative assessments, reinforces that symbolic knowledge integration meaningfully enhances both the cognitive reasoning and affective communication capacities of LLMs. These outcomes confirm that the unified Med-Emo CHQA architecture successfully fulfils the central design objective of Sub-RQ2: enabling LLMs to jointly understand and respond to informational and emotional support needs in a clinically safe, empathetic, and context-sensitive manner.

7.5 Compare with Prior Studies

This section positions the findings of the Med-Emo CHQA system within the broader body of work on medical question answering, empathetic dialogue, and neural-symbolic integration. By comparing the proposed architecture with prior studies, this discussion highlights how the present research extends existing approaches across four dimensions: task framing, methodological design, symbolic integration, and evaluation scope.

7.5.1 Consumer Health Question Answering Task Definition

The definition of the CHQA task has evolved significantly over the past decade, reflecting broader shifts in how AI systems are expected to engage with patients' informational and emotional needs. Earlier Medical Question Answering (MQA) studies (Ben Abacha et al., 2017; Demner-Fushman et al., 2019; Welivita et al., 2023) primarily focused on factual correctness and diagnostic reasoning. These systems were designed to emulate clinical knowledge retrieval, optimising for precision and reliability but largely neglecting the affective dimensions of patient communication. Conversely, Emotional Support Conversation (ESC) studies (Liu et al., 2021; Zheng et al., 2023) concentrated on affective reassurance and empathetic dialogue, often at the expense of medical rigour. While these conversational agents successfully modelled emotional alignment and compassion, they typically lacked grounding in clinical reasoning or health-specific factual knowledge.

More recent research efforts, such as Knowledge-Grounded Empathy-Enhanced Medical Dialogue (KGE-MD) models (Wu et al., 2024), have sought to bridge this gap by integrating medical knowledge bases into emotionally

aware dialogue systems. However, despite their use of structured knowledge to strengthen affective communication, their primary objective remains to enhance empathy rather than to achieve a truly joint reasoning process that unifies medical and emotional understanding. In essence, these models still treat empathy as a supplementary feature to factual QA, rather than as a co-equal reasoning dimension.

In response to these limitations, this study introduces the Joint Medical-Emotional Question Answering (JMEQA) task, which is a reformulation of CHQA that explicitly models the interdependence between medical informational and emotional support needs. Drawing on the empirical analysis of real-world OHC data (Section 4.2), JMEQA redefines QA in healthcare as a dual-dimension-intent reasoning task, wherein the system must jointly identify, interpret, and respond to both informational and emotional components within a single consumer query. Rather than combining the two dimensions additively, JMEQA conceptualises their interaction as a dynamic process in which informational clarity and emotional resonance are mutually reinforcing.

This dual-objective task formulation directly reflects the communication dynamics of OHC platforms, where users simultaneously seek accurate medical guidance and empathetic acknowledgment of their concerns. By positioning itself at the intersection of clinical rigour and human-centred empathy, JMEQA transcends the limitations of prior MQA, ESC, and KGE-MD paradigms. It establishes a new conceptual foundation for CHQA—one that integrates trustworthy medical knowledge delivery with emotionally attuned, context-aware understanding. In doing so, it bridges the longstanding divide between informational reasoning and affective understanding, providing the theoretical and methodological basis for the Med-Emo CHQA system developed in this research.

7.5.2 Discourse-Level Question Understanding

A major limitation of prior CHQA and MQA studies lies in their syntactic-level or semantic-level understanding of user queries. Traditional systems typically decomposed questions into surface intents (e.g., symptom query, diagnosis advice, treatment suggestion) or extracted entity-relation pairs using shallow semantic parsing. Although such methods improved retrieval precision, they often failed to capture the rhetorical and affective structures through which consumers actually communicate their health concerns. In practice, OHC

messages are highly narrative, interweaving background stories, self-diagnoses, anxiety, and emotional appeals, rendering linear intent classification insufficient for deeper comprehension.

Recent discourse-aware approaches have begun to address this gap. For instance, discourse-based summarisation models (Cohan et al., 2018; Xu et al., 2020) and coherence-driven QA frameworks (Miao et al., 2024) introduced rhetorical relations into document-level reasoning. However, these efforts were mainly applied to news or encyclopaedic texts and seldom incorporated affective or interpersonal dimensions. Within healthcare, most QA systems still treat patient utterances as flat, independent turns rather than as structured discourse units that convey layered informational and emotional meanings.

To overcome these limitations, this study adopts a discourse-level perspective grounded in RST and operationalised through the RST-DST introduced in Chapter 4. The proposed symbolic layer represents each CHQ as a hierarchical rhetorical tree, in which clauses or sentences serve as discourse units connected by relations such as Elaboration, Cause, Motivation, and Contrast. Each unit is annotated with dual-intent attributes, its informational function (medical need type) and emotional function (affective need type), as well as its contextual dependencies. This representation enables the system to reason not only about what the user is asking, but also why and how those questions are framed emotionally.

Compared with prior discourse-based QA work, the RST-DST framework extends beyond textual coherence modelling to achieve dual-intent reasoning. It integrates symbolic cues from both RST and the AFCE, linking cognitive discourse roles with affective appraisals. In doing so, it bridges the cognitive-affective divide that previous semantic or syntactic parsers could not represent. The resulting hierarchical JSON schema transforms unstructured patient narratives into interpretable knowledge structures that can guide LLMs toward clinically consistent and empathetically aligned comprehension.

Empirical evidence from the question understanding experiment (Section 5.4) confirms the value of discourse-level modelling: LLMs fine-tuned with RST-DST annotations achieved notably higher F1-scores and stronger context alignment than their non-symbolic counterparts. The improvement demonstrates that rhetorical-discourse cues act as interpretive scaffolds that

enhance both comprehension accuracy and emotional awareness. Moreover, the symbolic representation contributes to explainability that each prediction can be traced to specific discourse relations or affective cues, thereby increasing transparency and trust in clinical AI systems.

7.5.3 Neural-Symbolic Integration in the Era of LLM

The integration of symbolic reasoning and neural learning has long been a central theme in AI, tracing back to early NSI frameworks that sought to reconcile the interpretability of symbolic systems with the adaptability of neural networks (Hilario, 1995; Bader & Hitzler, 2005). Traditional NSI paradigms were typically characterised by hybrid architectures, in which symbolic logic provided explicit reasoning rules or constraints that guided downstream neural models. While this separation enhanced transparency, it often came at the cost of efficiency and contextual adaptability, which are two capabilities increasingly essential in complex, language-rich domains such as OHC.

Recent advances in LLMs have transformed the landscape of NSI research. LLMs exhibit powerful implicit reasoning abilities through large-scale pattern induction, yet their internal processes remain opaque and prone to hallucination, particularly in safety-critical applications like consumer health communication. This tension has prompted a paradigm shift from hybrid symbolic-to-neural toward unified neural-symbolic architectures, where symbolic knowledge is embedded directly within the neural model's representational space. Studies such as Kautz (2022) and Wang et al. (2025) have argued that the next generation of NSI must enable mutual reinforcement: symbolic structures constrain neural reasoning, while neural representations enrich symbolic abstraction.

Building on this theoretical foundation, the Med-Emo CHQA system (Chapter 5) exemplifies how NSI can be re-conceptualised in the era of LLMs. Instead of treating symbolic knowledge as an external reasoning module, this study operationalises integration through fine-tuning and prompt-guided inference, allowing the model to internalise symbolic patterns derived from the RST-based DST. The symbolic layer, comprising rhetorical, clinical, and empathic relations, was used as supervised signals during model fine-tuning, thereby embedding interpretable discourse structures within the model's latent space. This design enables the LLM to perform implicit symbolic reasoning

during inference, generating responses that are simultaneously medically accurate, empathetically aligned, and contextually grounded.

Compared with previous NSI implementations, which often relied on logic programming or external knowledge graphs, the proposed unified architecture offers several advantages. First, it achieves end-to-end integration—symbolic knowledge influences every stage of understanding and generation without interrupting the neural processing flow. Second, it enhances representational coherence: symbolic constraints such as “emotion-linked justification” or “cause-driven elaboration” are internalised as part of the model’s reasoning priors rather than post-hoc adjustments. Third, it substantially reduces inference latency by eliminating the need for external symbolic reasoning engines, making real-time empathetic dialogue feasible within OHC platforms.

This approach thus redefines NSI not as a hybrid of two isolated paradigms but as a continuum of reasoning within LLMs, where symbolic interpretability and neural generalisation coexist dynamically. The empirical findings in Chapter 6 support this claim: models with neural-symbolic unified architecture demonstrated the optimal trade-off between empathy, accuracy, personalisation and latency, outperforming both purely neural and symbolic-neural hybrid alternatives. These results substantiate that embedding symbolic knowledge within LLMs leads to measurable gains in both cognitive reliability and affective intelligence.

7.5.4 Evaluation of LLM-Based CHQA System

Evaluation practices in prior CHQA research have predominantly emphasised factual accuracy or semantic similarity. Although these evaluations are important for ensuring safety and factual reliability, they overlook the affective, personalised, and context-sensitive dimensions that characterise real-world OHC. Similarly, emotional support dialogue systems (Liu et al., 2021) typically evaluate empathy or emotional alignment, but rarely assess clinical quality or safety. This narrowness in evaluation scope reflects a broader limitation in the field: existing frameworks tend to measure either what an AI system says or how it says it, but seldom integrate both perspectives.

Even recent hybrid initiatives such as knowledge-grounded empathy-enhanced medical models (Wu et al., 2024) primarily assess empathy or

linguistic coherence. Although these works acknowledge the need for medical grounding, their evaluation methods still largely treat medical informativeness as a supportive quality rather than as a core dimension requiring systematic measurement. In short, prior studies do not offer a multi-dimensional, clinically aligned evaluation framework capable of capturing the full spectrum of performance required for safe and reliable CHQA.

In contrast, this study introduces a comprehensive multi-dimensional evaluation framework tailored to the dual-intent nature of consumer health questions. As presented in Chapter 6, the evaluation spans four key dimensions:

- Accuracy, assessed through Comprehensiveness Level (coverage of both medical and emotional needs), Hallucination Level (degree of factual deviation), and Risk Level (clinical safety of recommendations). This expands beyond conventional factual accuracy to reflect clinical responsibility within patient-facing contexts.
- Personalisation, capturing the extent to which responses adapt to user-specific circumstances rather than generic templates.
- Empathy, measured through the context alignment of emotional support and the appropriateness of empathetic expressions—dimensions rarely integrated into previous CHQA evaluations.
- Latency, which is critical for real-time OHC deployment but largely neglected in existing CHQA studies.

Furthermore, this study incorporates a reader study with certified psychiatrists, providing an expert-based evaluation that benchmarks the clinical realism, safety, and emotional appropriateness of system outputs. Prior CHQA work seldom includes domain experts, and almost none simultaneously evaluate empathy, accuracy, personalisation, and latency through human judgement. This inclusion of clinician evaluators ensures that results are not only computationally valid but also clinically meaningful.

Empirical results demonstrate that the Med-Emo CHQA (unified neural-symbolic model) achieves the most favourable balance across accuracy, personalisation, empathy, and latency, outperforming purely neural and hybrid neural-symbolic systems (Section 6.4). The model not only produced responses rated as more comprehensive and empathetic but also maintained lower hallucination and risk levels, confirming the safety advantages of the unified neural-symbolic design. Importantly, despite its enhanced reasoning

capabilities, Med-Emo CHQA showed shorter average latency, supporting feasibility for deployment in time-sensitive OHC environments.

Taken together, this study's evaluation approach extends prior CHQA assessments in three significant ways. First, it treats factual accuracy, emotional alignment, and safety as co-equal dimensions, consistent with the dual-intent nature of CHQs. Second, it integrates expert-based assessment, grounding AI performance interpretation in clinical interpretive practices. Third, it explicitly evaluates the trade-off between latency and empathic comprehensiveness, a dimension critical for real-world adoption yet rarely addressed in previous work.

7.6 Chapter Summary

This chapter synthesised the key findings of the study by revisiting the research objectives and examining how the proposed design artefacts addressed the two sub-research questions. Through the discussions of Sub-RQ1 and Sub-RQ2, the chapter demonstrated how symbolic knowledge representation and neural-symbolic integration collectively enable a CHQA system to jointly understand and respond to consumers' medical and emotional support needs. The analysis highlighted both the practical relevance and theoretical grounding of the design choices, supported by empirical evidence from automated evaluation and expert reader studies. A comparison with prior studies further clarified the study's contributions relative to existing CHQA, discourse analysis, neural-symbolic reasoning, and LLM evaluation literature. Overall, the chapter established how the Med-Emo CHQA system advances the design science knowledge base and offers new pathways for developing trustworthy, empathetic, and context-aware healthcare AI systems.

Collectively, the discussions in this chapter consolidate the core findings of the research and articulate how the Med-Emo CHQA system fulfils its intended research objectives. These insights form the analytical foundation for the next chapter, which integrates the findings into broader research implications, articulates practical and theoretical contributions, and outlines the limitations and future directions of this work.

Chapter 8 Conclusion

8.1 Research Implications

This study provides several significant implications for both research and practice in the field of CHQA system. Guided by the DSR Knowledge Contribution Framework (Gregor & Hevner, 2013), the findings demonstrate a progression from Exaptation to Invention, reflecting both theoretical advancement and practical impact.

8.1.1 Theoretical Implications

In the DSR knowledge contribution framework, the nature of a contribution is determined by the maturity of the problem domain and the maturity of the solution knowledge base (Gregor & Hevner, 2013). As reflected in the DSR knowledge contribution framework, this research demonstrates a transition from Exaptation to Invention (see Figure 8-1).

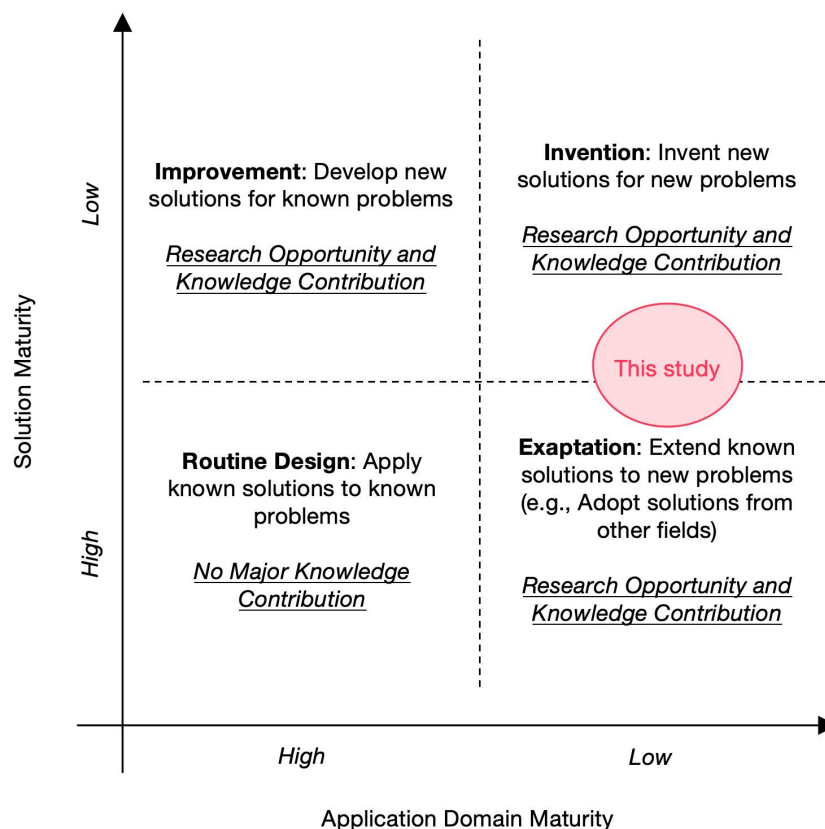


Figure 8-1. the DSR Knowledge Contribution Framework. Adapted from Gregor & Hevner (2013).

This research reframes CHQA from a single-intent analytical task into a dual-intent reasoning problem, formalised as JMEQA. Prior research on Medical QA and Emotional Support Conversation has treated informational and emotional needs separately, resulting in systems that excel in one dimension but fail in the other. This study contributes a unified theoretical model of CHQ structure, explaining how informational and emotional intents interact at discourse level and how they can be jointly modelled through symbolic-neural mechanisms.

Within this emergent problem space, the research begins by contributing on the level of exaptation, repurposing established theories—Rhetorical Structure Theory (RST), and the Appraisal Framework for Clinical Empathy (AFCE)—to conceptualise and operationalise the dual-intent nature of CHQs. These theories, traditionally used in discourse analysis, clinical reasoning, and psychological appraisal, are extended to explain how users simultaneously encode informational and emotional needs in online health consultations.

However, the study advances beyond theoretical adaptation toward Invention, producing design artefacts that did not previously exist in the literature. These include: the CHQA Med-Emo dataset, the RST-DST symbolic knowledge representation, the JMEQA task definition and formulation, the Med-Emo CHQA unified neural-symbolic architecture, and the multi-dimensional evaluation framework incorporating accuracy, safety, empathy, personalisation, and latency. Together, these artefacts represent new solution knowledge for a previously unaddressed design space. They extend DSR’s theoretical boundary conditions by demonstrating how empathetic, safety-aware reasoning can be embedded within large language models for health-related applications.

Furthermore, the research advances NSI theory by demonstrating how symbolic discourse structures can be internalised within LLMs via fine-tuning, resulting in an integrated reasoning process that is both interpretable and adaptive. The findings exemplify how NSI can evolve in the LLM era—from pipeline architectures into end-to-end unified systems that offer both contextual fluency and clinical coherence.

8.1.2 Practical Implications

Beyond its theoretical contributions, this research offers several practical implications for the design, deployment, and governance of healthcare AI systems.

First, the Med-Emo CHQA architecture demonstrates how accuracy, empathy, and safety can be jointly achieved in AI-mediated health communication. By reducing hallucinations, mitigating clinical risk, and grounding emotional expressions in medically relevant context, the system directly addresses patient concerns around trust and transparency. These characteristics align with emerging guidelines for responsible and explainable AI in healthcare.

Second, the unified neural-symbolic architecture achieves this enhanced quality without incurring substantial latency, making it suitable for large-scale deployment in online health consultation platforms where responsiveness is essential. The symbolic knowledge representation, implemented through a hierarchical JSON schema, can be readily adapted to new medical specialties, languages, or cultural environments, supporting broader internationalisation and domain transfer.

Third, the system provides a practical foundation for AI tools that augment rather than replace clinicians. It can assist with triage, provide empathetic pre-consultation guidance, and alleviate communication burdens in high-volume or resource-constrained digital health settings. Such augmentation is particularly relevant in mental health and primary care, where timely emotional support plays a critical role.

Fourth, the design principles derived from this research—symbolic grounding for dual-intent understanding, unified neural-symbolic reasoning, and multi-dimensional health AI evaluation—offer actionable guidance for system developers. These principles are transferable to other domains requiring simultaneous factual and affective reasoning, including chronic disease self-management, behavioural health support, and patient education systems.

Finally, by demonstrating how safety, empathy, personalisation, and accuracy can be evaluated in an integrated manner, the study contributes to the governance of conversational health AI. The proposed evaluation framework provides concrete, operationalisable metrics that can inform

regulatory audits, risk assessments, quality assurance processes, and certification standards for AI-mediated health communication.

8.2 Limitations and Future Directions

Despite the theoretical and practical contributions outlined above, this study has several limitations that provide opportunities for future research. These limitations relate to dataset scope, research boundaries, system architecture, and evaluation methodology. Consistent with DSR, acknowledging these boundaries not only clarifies the contextual validity of the proposed artefacts but also opens new pathways for extending the Med-Emo CHQA system into broader human-AI healthcare contexts.

8.2.1 Dataset Scope

Despite the theoretical and practical contributions outlined above, this study has several limitations that provide opportunities for future research. These limitations relate to dataset scope, research boundaries, system architecture, and evaluation methodology. Consistent with DSR, acknowledging these boundaries not only clarifies the contextual validity of the proposed artefacts but also opens new pathways for extending the Med-Emo CHQA system into broader human-AI healthcare contexts.

The fine-tuning and evaluation experiments were conducted using datasets from a single medical speciality. Prior work has shown that consumer health information needs vary substantially across clinical domains (Zeng et al., 2004; Zhang, 2018), and emotional expressions also differ based on disease type, severity, and patient demographics (Pennebaker et al., 2003). Future research should extend the JMEQA framework and symbolic-neural architecture to multiple OHC specialities, enabling more generalisable insights into dual-intent reasoning.

This study focused exclusively on Chinese-language OHC data. Although this ensures ecological validity within the Chinese digital health ecosystem, cultural psychology and communication studies suggest that emotional disclosure, uncertainty expression, and help-seeking behaviours vary significantly across cultures (Hall, 1976; Hofstede, 2001). Future studies should replicate the approach in other languages and cultural contexts, thereby testing the cross-cultural robustness of the symbolic annotation schema and Med-Emo CHQA architecture.

OHC platforms represent a specific form of computer-mediated communication characterised by asynchronous exchange, role asymmetry, and institutional trust (Rice & Love, 1987; Walther, 1996). As generative AI becomes a primary channel for health information seeking, AI-mediated communication will foster new patterns of information needs and emotional articulation (Hancock et al., 2020). Because this study trained and evaluated models using OHC-originated queries, the system may not fully capture the emerging forms of AI-prompted CHQs, which may be shorter, more direct, or influenced by prior AI responses. Future research should therefore collect dedicated AI-mediated communication health question datasets and investigate how symbolic and emotional structures evolve in these emerging environments.

8.2.2 Research Boundaries

The Med-Emo CHQA system developed in this study is intentionally scoped to single-turn, text-based consumer health questions. This design boundary provides a clear and manageable problem space suitable for early-stage system development and rigorous evaluation. However, it does not fully capture the complexity of real-world clinical interactions or AI-mediated health communication, both of which are inherently longitudinal, multi-modal, and context-evolving.

Extensive findings from conversation analysis (Sacks et al., 1974) and health communication research indicate that empathy, personalisation, rapport, and shared understanding tend to unfold progressively across multiple conversational turns. Emotional needs may intensify or attenuate over time, informational needs may shift as users receive new explanations, and trust develops iteratively rather than instantaneously. Consequently, limiting CHQA to a single turn constrains the system's ability to model intent trajectories, conversational grounding, and dynamic emotional adaptation.

Future research should therefore extend both the symbolic knowledge representation and the neural-symbolic integration mechanisms to multi-turn CHQA, allowing the system to track changing user intents, maintain discourse coherence, and sustain empathetic alignment across interactions. Such advances would bring the system closer to the natural flow of clinical communication and to the expectations users have when engaging with AI health assistants.

In addition to temporal boundaries, the system is limited to text-only input. Yet contemporary healthcare decision-making increasingly relies on multimodal information, including diagnostic images, laboratory reports, wearables-generated signals, and patient-uploaded artefacts. Prior studies in multimodal clinical AI (Chen et al., 2019; Esteva et al., 2021) demonstrate that integrating visual, textual, and physiological data substantially improves diagnostic reasoning and personalisation. The symbolic frameworks developed in this study, RST-based discourse structures and medical-emotional intent schemas, could be extended to serve as cross-modal grounding layers, enabling seamless integration between vision-language models, structured clinical data, and CHQA dialogue.

Future development should therefore explore multimodal CHQA architectures, where symbolic structures guide cross-modal alignment and interpretability. Such systems would be able to reason over both linguistic content and non-linguistic cues, ultimately supporting richer, more context-aware, and clinically realistic AI-mediated health consultations.

8.2.3 System Architecture

The present architecture integrates symbolic cues into LLMs through fine-tuning. While this approach has demonstrated clear benefits, it represents an early stage of neural-symbolic integration in the LLM era. The symbolic elements in this study primarily support question understanding, enabling the system to interpret informational and emotional needs more accurately. However, the answer generation stage remains largely governed by neural inference alone. This separation limits the system's ability to ensure fully reliable, controllable, and safety-aligned response generation.

Future research should therefore explore architectures in which the entire reasoning chain is symbolically represented, including both upstream interpretation and downstream generation. Symbolic structures, such as discourse relations, clinical reasoning patterns, and empathy appraisal frameworks, should not merely guide the model externally but instead be internalised within the neural reasoning pathway. Such integration would allow the system to combine logic-driven inference with pattern-driven generation, resulting in higher robustness, explainability, and consistency across medical and emotional communication tasks.

From a theoretical perspective, this study's architecture aligns with Type-3 neural-symbolic systems, which inject symbolic knowledge into neural models via training data and prompts (see Section 2.6.3). While Type-3 approaches are practical and effective, Kautz (2022) emphasises that the long-term trajectory of NSI points toward Type-6 neural-symbolic systems—fully unified architectures in which symbolic reasoning is embedded directly within neural computation (see Section 2.6.3). Such systems are expected to support combinatorial reasoning, explicit rule manipulation, and hybrid cognition, corresponding to Kahneman's (2011) distinction between System 1 (fast, intuitive) and System 2 (slow, deliberative) thinking.

Recent advances in differentiable logic (Rocktäschel & Riedel, 2017) and neural-symbolic transformers (Ying et al., 2023) demonstrate the feasibility of integrating symbolic operators directly into transformer layers. These developments point to several promising directions for next-generation CHQA systems, including:

- Differentiable discourse parsers that encode RST or dialogic relations within the model's internal structure;
- Symbolic constraint or verification modules to enforce clinical safety, empathy appropriateness, and logical coherence during generation;
- Logic-aware attention mechanisms that enable the model to reason explicitly over symbolic structures rather than relying solely on statistical patterns.

Advancing toward such unified architectures would allow future CHQA systems to reason with both precision and empathy, ensuring that medically relevant logic, emotional context, and safety constraints are upheld throughout the complete inference pipeline. This trajectory marks a substantive opportunity for future design science research in AI-mediated health communication.

8.2.4 System Evaluation

This study introduced a multi-dimensional evaluation framework and conducted a reader study involving three certified psychiatrists. While this provides clinically grounded insight into system performance, several limitations warrant consideration.

First, reader studies in healthcare AI typically rely on small expert groups due to the high cost of specialist participation. Although such evaluations possess strong domain validity, their limited sample size restricts generalisability (Friedman & Wyatt, 2010). Future research should engage larger and more diverse expert panels, potentially adopting structured consensus-building methods such as the Delphi technique (Linstone & Turoff, 2002).

Second, expert-based evaluations primarily reflect clinical appropriateness, but they do not necessarily capture consumer experience, which may differ substantially from professional judgement (Greene et al., 2011). To address this gap, future evaluations should incorporate patient-centred methods such as usability testing, real-world A/B experiments within OHC platforms, and deployment-based behavioural analytics to better understand how end users perceive accuracy, empathy, and trustworthiness.

Third, as recommended in evaluation science (Shadish, Cook & Campbell, 2002), methodological triangulation is essential for robust validation. Reliance on a single evaluation modality, whether expert judgement or automated metrics, may overlook important performance dimensions. Future research should therefore complement clinician assessment with large-scale automated evaluation, conversational quality benchmarks, real-user feedback, and longitudinal safety monitoring.

Combining these approaches would enhance both the internal validity and external generalisability of the Med-Emo CHQA system's performance claims, ensuring that the model meets the expectations of both clinical experts and end users in real-world AI-mediated healthcare settings.

8.3 Overall Conclusion

This research set out to address a fundamental challenge in the design of CHQA system: enabling a system to jointly understand and respond to the informational and emotional support needs embedded in CHQs. Existing MQA and ESC systems have traditionally treated factual reasoning and empathetic communication as separate design problems, resulting in systems that excel in one dimension but underperform in the other.

Drawing on real-world OHC data, the research demonstrates that consumer queries are inherently dual-intent expressions, combining medical and

emotional support needs. To capture this complexity, the study synthesises Rhetorical Structure Theory (RST), and the Appraisal Framework for Clinical Empathy (AFCE) into the design of the RST-DST symbolic knowledge representation, offering a discourse-level perspective that formalises how informational and emotional needs co-occur and influence one another.

Building on this representational foundation, the study proposes the Med-Emo CHQA system, an end-to-end neural-symbolic unified architecture that internalises symbolic discourse structures within a LLM. This integration enables the system to reason simultaneously about medical content and emotional context, supporting the generation of responses that are accurate, empathetically aligned, personalised, and safe. Through supervised fine-tuning and prompt-guided inference, symbolic knowledge becomes directly embedded in the model's latent reasoning mechanisms, advancing the field toward the next generation of neural-symbolic integration suited to LLMs.

A multi-dimensional evaluation framework was developed to assess accuracy, empathy, personalisation, and latency of the proposed system. Results from the clinician-led reader study demonstrate that the unified Med-Emo CHQA architecture achieves the best overall trade-off among these dimensions, outperforming both hybrid neural-symbolic systems and purely neural baselines. This evidence confirms the value of combining symbolic interpretability with neural adaptability and highlights the practical viability of deploying such systems in real-world health communication environments.

References

- Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., . . . Zhou, X. (2024, April 22). *PHI-3 Technical Report: A highly capable language model locally on your phone*. arXiv.org. <https://arxiv.org/abs/2404.14219>
- Ahn, K., Bos, J., Kor, D., Nissim, M., Webber, B. L., & Curran, J. R. (2005). Question answering with QED at TREC 2005. In *Proceedings of TREC 2005*. National Institute of Standards and Technology. Retrieved from https://trec.nist.gov/pubs/trec14/t14_proceedings.html
- Alasmari, A., Kudryashov, L., Yadav, S., Lee, H., & Demner-Fushman, D. (2023). CHQ-SocioEmo: Identifying social and emotional support needs in consumer-health questions. *Scientific Data*, 10, 329. <https://doi.org/10.1038/s41597-023-02203-1>
- Alasmari, A., Zhou, Y., Liu, X., & Wang, F. (2021). Share to seek: The effects of disease complexity on health information-seeking behavior. *Journal of Medical Internet Research*, 23(3), e21642. <https://doi.org/10.2196/21642>
- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). *Publicly available clinical BERT embeddings for clinical natural language processing*. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72-78. <https://aclanthology.org/W19-1909/>
- Alto, V. (2024). *Building LLM powered applications: Create intelligent apps and agents with large language models*. Packt Publishing.
- Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., . . . Saeta, B. (2023). *PALM 2 Technical Report*. arXiv. <https://doi.org/10.48550/arxiv.2305.10403>
- Aronson, A. R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program*. Proceedings of the AMIA Symposium, 17-21. <https://doi.org/10.1136/jamia.1994.95236136>
- Atanasova, S., & Petrič, G. (2019). Collective empowerment in online health communities: Scale development and empirical validation. *Journal of Medical Internet Research*, 21(11), e14392. <https://doi.org/10.2196/14392>
- Athenikos, S. J., & Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1), 1-24. <https://doi.org/10.1016/j.cmpb.2009.10.003>
- Athenikos, S. J., Han, H., & Brooks, A. D. (2009). A framework of a logic-based question-answering system for the medical domain (LOQAS-Med). In *Proceedings of the 2009 ACM Symposium on Applied Computing* (pp. 847–851). Association for Computing Machinery. <https://doi.org/10.1145/1529282.1529462>
- Bader, S., & Hitzler, P. (2005). Dimensions of neural-symbolic integration: A structured survey. In S. Artemov, H. Barringer, A. S. d'Avila Garcez, L. C. Lamb, & J. Woods (Eds.), *We will show them: Essays in honour of Dov Gabbay* (Vol. 1, pp. 167–194). King's College Publications.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*. <https://doi.org/10.48550/arXiv.1409.0473>
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., . . . Zhu, T. (2023, September 28). *Qwen Technical Report*. arXiv.org. <https://arxiv.org/abs/2309.16609>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., . . . Kaplan, J. (2022).

- Constitutional AI: Harmlessness from AI Feedback*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2212.08073>
- Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19(1), 3-11.
- BCC Global. (2024, September 24). *Ant Group acquires Haodf: A new wave of mergers in internet healthcare?* BCC.Global Media. <https://bccmedianews.com/?p=1043>
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3615-3620). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1371>
- Ben Abacha, A. B., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1), 511. <https://doi.org/10.1186/s12859-019-3106-0>
- Ben Abacha, A. B., & Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing & Management*, 51(5), 570-594. <https://doi.org/10.1016/j.ipm.2015.04.006>
- Ben Abacha, A. B., Roberts, K., & Demner-Fushman, D. (2019). *Bridging the gap between consumers' medication questions and trusted answers*. MEDINFO 2019: Proceedings of the 17th World Congress on Medical and Health Informatics.
- Ben Abacha, A., & Demner-Fushman, D. (2017). *Recognizing question entailment for medical question answering*. AMIA Annual Symposium Proceedings, 2017, 310-318. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977632/>
- Ben Abacha, A., & Demner-Fushman, D. (2019a). A question-entailment approach to question answering. *BMC Bioinformatics*, 20, 511. <https://doi.org/10.1186/s12859-019-3119-4>
- Ben Abacha, A., & Demner-Fushman, D. (2019b). On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2228-2234). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1215>
- Ben Abacha, A., Agichtein, E., Pinter, Y., & Demner-Fushman, D. (2017). Overview of the medical question answering task at TREC 2017 LiveQA. In *Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017)*. National Institute of Standards and Technology.
- Ben Abacha, A., Shivade, C., & Demner-Fushman, D. (2019). *Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering*. Proceedings of the BioNLP 2019 Workshop, 370-379. <https://doi.org/10.18653/v1/W19-5039>
- Benbasat, I. (1984). An analysis of research methodologies. In F. W. McFarlan (Ed.), *The information systems research challenge* (pp. 47-85). Boston, MA: Harvard Business School Press.
- Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly*, 11(3), 369-386. <https://doi.org/10.2307/248684>
- Besold, T. R., Garcez, A. D., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K., Lamb, L. C., Lowd, D., Lima, P. M. V., De Penning, L., Pinkas, G., Poon, H., & Zaverucha, G. (2017). *Neural-Symbolic Learning and Reasoning: A Survey and interpretation*. arXiv. <https://doi.org/10.48550/arxiv.1711.03902>
- Besta, M., Blach, N., Kubicek, A., von Rueden, L., & Hoefler, T. (2024). *Graph of Thoughts: Solving compositional problems with large language models*. arXiv. <https://doi.org/10.48550/arXiv.2308.09687>
- Bestsenny, O., Gilbert, G., Harris, A., & Rost, J. (2021, July 9). *Telehealth: A quarter-trillion-dollar post-COVID-19 reality?* McKinsey & Company.

<https://www.mckinsey.com/industries/healthcare/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality>

- Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue), D267-D270. <https://doi.org/10.1093/nar/gkh061>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., . . . Wang, W. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv. <https://doi.org/10.48550/arxiv.2108.07258>
- Bosma, W. (2005). *Extending answers using discourse structure*. In *Proceedings of RANLP 2005* (pp. 135-139). Retrieved from <https://research.utwente.nl/en/publications/query-based-summarization-using-rhetorical-structure-theory>
- Bouziane, A., Bouchiha, D., Doumi, N., & Malki, M. (2015). Question answering systems: Survey and trends. *Procedia Computer Science*, 73, 366–375. <https://doi.org/10.1016/j.procs.2015.12.005>
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). *A large annotated corpus for learning natural language inference*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 632-642. <https://doi.org/10.18653/v1/D15-1075>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33, 1877–1901. <https://doi.org/10.48550/arXiv.2005.14165>
- Bryman, A. (2016). *Social research methods (5th ed.)*. Oxford University Press.
- Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual (Tech. Rep. ISI-TR-545)*. University of Southern California, Information Sciences Institute.
- Chakraborty, N., Lukovnikov, D., Maheshwari, G., Trivedi, P., Lehmann, J., & Fischer, A. (2021). Introduction to neural network-based question answering over knowledge graphs. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3), e1389. <https://doi.org/10.1002/widm.1389>
- Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: What does it mean? (Or, it takes at least two to tango). *Social Science & Medicine*, 44(5), 681-692. [https://doi.org/10.1016/S0277-9536\(96\)00221-3](https://doi.org/10.1016/S0277-9536(96)00221-3)
- Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storlie, C. B., Habermann, E. B., Naessens, J. M., Larson, D. W., & Liu, H. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj Digital Medicine*, 2, Article 43. <https://doi.org/10.1038/s41746-019-0122-0>
- Chen, J., Saha, S., & Bansal, M. (2024). ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7066–7085. <https://doi.org/10.18653/v1/2024.acl-long.381>
- Chen, X. (2023). Online health communities influence people’s health behaviors in the context of COVID-19. *PLOS ONE*, 18(4), e0282368. <https://doi.org/10.1371/journal.pone.0282368>
- Chen, Y., Wang, Z., Xing, X., Zheng, H., Xu, Z., Fang, K., Li, S., Wang, J., Wu, J., Liu, Q., & Xu, X. (2023). *BianQue: Balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT*. arXiv. <https://doi.org/10.48550/arXiv.2310.15896>
- China Internet Network Information Center. (2025). *The 55th statistical report on China’s Internet development*. CNNIC.

- Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., & Zettlemoyer, L. (2018). QuAC: Question answering in context. In *Proceedings of EMNLP 2018* (pp. 2174-2184). <https://doi.org/10.18653/v1/D18-1241>
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., . . . Wei, J. (2022). *Scaling Instruction-Finetuned Language Models*. arXiv. <https://doi.org/10.48550/arxiv.2210.11416>
- Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of EMNLP 2016* (pp. 2256-2262). <https://doi.org/10.18653/v1/D16-1245>
- Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 615–621). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-2097>
- Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education (7th ed.)*. Routledge. <https://doi.org/10.4324/9780203720967>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., . . . Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (pp. 8440-8451). <https://doi.org/10.18653/v1/2020.acl-main.747>
- Costello, R. E., Anand, A., Jameson Evans, M., & Dixon, W. G. (2019). Associations between engagement with an online health community and changes in patient activation and health care utilization: Longitudinal web-based survey. *Journal of Medical Internet Research*, 21(8), e13477. <https://doi.org/10.2196/13477>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.)*. Thousand Oaks, CA: Sage.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- de Lima, E., & Osório, F. de L. (2021). Empathy and mental health chatbots: A systematic review. *Journal of Medical Internet Research*, 23(11), e27239. <https://doi.org/10.2196/27239>
- De Raedt, L., Dumančić, S., Manhaeve, R., & Marra, G. (2020). From statistical relational to neuro-symbolic artificial intelligence. In C. Bessiere (Ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)* (pp. 4943–4950). International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2020/688>
- Deardorff, A., Masterton, K., Roberts, K., Kilicoglu, H., & Demner-Fushman, D. (2017). A protocol-driven approach to automatically finding authoritative answers to consumer health questions in online resources. *Journal of the Association for Information Science and Technology*, 68(7), 1724–1736. <https://doi.org/10.1002/asi.23806>
- DeepSeek AI. (2025a). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. <https://doi.org/10.48550/arXiv.2501.12948>
- DeepSeek AI. (2025b). *DeepSeek-V3 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2412.19437>
- Delin, J., Bateman, J. A., & Allen, P. (2002). A model of discourse for information retrieval and text summarisation. *Information Processing & Management*, 38(6), 799–820.
- Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1), 63-103. <https://doi.org/10.1162/coli.2007.33.1.63>

- Demner-Fushman, D., Abacha, A. B., Mrabet, Y., Benkovits, C., Roberts, K., & Shooshan, S. E. (2019). Consumer health information and question answering: Helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2), 194–201. <https://doi.org/10.1093/jamia/ocz196>
- Demner-Fushman, D., Kilicoglu, H., Roberts, K., Masterton, K., & Deardorff, A. (2015, September). *Consumer health question answering to automatically support NLM customer services* [Technical report]. U.S. National Library of Medicine, Lister Hill National Center for Biomedical Communications. <https://lhncbc.nlm.nih.gov/LHC-publications/pubs/ConsumerHealthQuestionAnsweringtoAutomaticallySupportNLMCustomerServices.html>
- Demner-Fushman, D., Mrabet, Y., & Ben Abacha, A. (2020). Consumer health information and question answering: Helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27(2), 194–201. <https://doi.org/10.1093/jamia/ocz152>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186). <https://doi.org/10.48550/arXiv.1810.04805>
- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., ... Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1286-1305. <https://doi.org/10.18653/v1/2021.emnlp-main.98>
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multi-agent debate*. arXiv. <https://doi.org/10.48550/arXiv.2305.14325>
- d'Avila Garcez, A. S., & Lamb, L. C. (2020). *Neurosymbolic AI: The 3rd wave*. arXiv. <https://arxiv.org/abs/2012.05876>
- Elwyn, G., Durand, M. A., Song, J., Aarts, J., Barr, P. J., Berger, Z., Cochran, N., Frosch, D., Galasiński, D., Gulbrandsen, P., Han, P. K. J., Härter, M., Kinnersley, P., Lloyd, A., Mishra, M., Perestelo-Perez, L., Scholl, I., Tomori, K., Trevena, L., . . . Van Der Weijden, T. (2017). A three-talk model for shared decision making: multistage consultation process. *BMJ*, 359, j4891. <https://doi.org/10.1136/bmj.j4891>
- Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., Cording, E., Tomson, D., Dodd, C., Rollnick, S., Edwards, A., & Barry, M. (2012). Shared decision making: a model for clinical practice. *Journal of General Internal Medicine*, 27(10), 1361-1367. <https://doi.org/10.1007/s11606-012-2077-6>
- Emanuel, E. J., & Emanuel, L. L. (1992). Four models of the physician-patient relationship. *JAMA*, 267(16), 2221-2226. <https://doi.org/10.1001/jama.1992.03480160079038>
- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, 196(4286), 129-136. <https://doi.org/10.1126/science.847460>
- Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E. J., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4, Article 5. <https://doi.org/10.1038/s41746-020-00376-2>
- Eysenbach, G. (2000). Consumer health informatics. *BMJ*, 320(7251), 1713-1716. <https://doi.org/10.1136/bmj.320.7251.1713>
- Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1-39. <http://jmlr.org/papers/v23/21-0998.html>
- Feng, V. W., & Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers* (pp. 60–68). Association for Computational Linguistics. <https://doi.org/10.5555/2390524.2390534>

- Feng, V. W., & Hirst, G. (2014). The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014* (pp. 940-949). <https://aclanthology.org/C14-1089/>
- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefter, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3), 59–79. <https://doi.org/10.1609/aimag.v31i3.2303>
- Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., & Song, M. (2019). Mining of textual health information from Reddit: Analysis of chronic diseases with extracted entities and their relations. *Journal of Medical Internet Research*, 21(6), e12876. <https://doi.org/10.2196/12876>
- Franklin, S., & Graesser, A. (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents. In *Lecture notes in computer science* (pp. 21–35). <https://doi.org/10.1007/bfb0013570>
- Friedman, C. P., & Wyatt, J. C. (2010). *Evaluation methods in biomedical informatics* (2nd ed.). Springer.
- Friedman, C., Alderson, P. O., Austin, J. H. M., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2), 161-174. <https://doi.org/10.1136/jamia.1994.95236146>
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gamer, M., Lemon, J., & Singh, I. F. (2019). *irr: various coefficients of interrater reliability and agreement (Version 0.84.1) [R package]*. CRAN. Retrieved from <https://cran.r-project.org/web/packages/irr/index.html>
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang J., He H., Thite A., Nabeshima N., Presser S., Leahy, C. (2020). *The Pile: An 800GB dataset of diverse text for language modeling*. arXiv. <https://arxiv.org/pdf/2101.00027>
- Garcez, A. d'Avila, Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). *Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning*. arXiv. <https://doi.org/10.48550/arXiv.1905.06088>
- Garg, R., Rebić, N., & De Vera, M. A. (2020). Information needs about cancer treatment, fertility, and pregnancy: Qualitative descriptive study of Reddit threads. *JMIR Cancer*, 6(2), e17771. <https://doi.org/10.2196/17771>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>
- Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., . . . Vinyals, O. (2024, March 8). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. arXiv.org. <https://arxiv.org/abs/2403.05530>
- Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. Association for Computational Linguistics. <https://aclanthology.org/2023.emnlp-main.751>
- Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). EMMA: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-7). IEEE. <https://doi.org/10.1109/ACII.2019.8925455>
- Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*

- (pp. 634-642). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/P17-1059>
- Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., & Zuccon, G. (2015). Overview of the CLEF eHealth evaluation lab 2015. In J. Mothe et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (pp. 429-443). Springer. https://doi.org/10.1007/978-3-319-24027-5_44
- Grand View Research. (2025). *Digital health market size & share report, 2030*. Grand View Research. <https://www.grandviewresearch.com/industry-analysis/digital-health-market>
- Greene, S. M., Tuzzio, L., & Cherkin, D. (2012). A framework for making patient-centered care front and center. *The Permanente Journal*, 16(3), 49-53.
<https://doi.org/10.7812/TPP/12-025>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337-355.
<https://doi.org/10.25300/MISQ/2013/37.2.01>
- Gu, G., Zhang, X., Zhu, X., Jian, Z., Chen, K., Wen, D., Gao, L., Zhang, S., Wang, F., Ma, H., & Lei, J. (2019). Development of a consumer health vocabulary by mining health forum texts based on word embedding: Semiautomatic approach. *JMIR Medical Informatics*, 7(2), e12704. <https://doi.org/10.2196/12704>
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-Specific Language Model Pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1-23. <https://doi.org/10.1145/3458754>
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105-117). Thousand Oaks, CA: Sage.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342-8360). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.acl-main.740>
- Hall, E. T. (1976). *Beyond culture*. Anchor Press.
- Hall, J. A., Roter, D. L., & Katz, N. R. (1987). Task versus socioemotional behaviors in physicians. *Medical Care*, 25(5), 399-412. <https://doi.org/10.1097/00005650-198705000-00004>
- Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Figueroa, A., ... Bressemer, K. K. (2023). *MedAlpaca – An open-source collection of medical conversational AI models and training data*. arXiv.
<https://doi.org/10.48550/arXiv.2304.08247>
- Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89-100. <https://doi.org/10.1093/jcmc/zmz022>
- Haskard Zolnierrek, K. B., & DiMatteo, M. R. (2009). Physician communication and patient adherence to treatment: A meta-analysis. *Medical Care*, 47(8), 826-834.
<https://doi.org/10.1097/MLR.0b013e31819a5acc>
- He, X., Chen, S., Ju, Z., Dong, X., Fang, H., Wang, S., Yang, Y., Zeng, J., Zhang, R., Zhang, R., Zhou, M., Zhu, P., & Xie, P. (2020). *MedDialog: Two large-scale medical dialogue datasets*. arXiv. <https://doi.org/10.48550/arXiv.2004.03329>
- He, Y., Zhu, Z., Zhang, Y., Chen, Q., & Caverlee, J. (2020). *Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition*. arXiv. <https://arxiv.org/abs/2010.03746>
- He, J., Fu, M., & Tu, M. (2019). Applying deep matching networks to Chinese medical question answering: a study and a dataset. *BMC Medical Informatics and Decision Making*, 19(S2), 52. <https://doi.org/10.1186/s12911-019-0761-8>

- Hernandez, E., Li, B. Z., & Andreas, J. (2023). *Inspecting and editing knowledge representations in language models*. arXiv. <https://arxiv.org/abs/2304.00740>
- Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, 19(2), 87–92.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hilario, M. (1997). An overview of strategies for neurosymbolic integration. In R. Sun & F. Alexandre (Eds.), *Connectionist-symbolic integration: From unified to hybrid approaches* (pp. 13–35). Psychology Press.
- Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (pp. 1-12). Lawrence Erlbaum. <https://escholarship.org/uc/item/79w838g1>
- Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: The view from here. *Natural Language Engineering*, 7(4), 275–300. <https://doi.org/10.1017/S1351324901002807>
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Irving, G. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016-30030. <https://arxiv.org/abs/2203.15556>
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Sage.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*. arXiv. <https://doi.org/10.48550/arXiv.1902.00751>
- Hovy, E., & Lavid, J. (2010). Towards a “science” of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1), 13–36.
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018* (pp. 328-339). <https://doi.org/10.48550/arXiv.1801.06146>
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LORA: Low-Rank adaptation of Large Language Models. arXiv. <https://doi.org/10.48550/arxiv.2106.09685>
- Hu, Z., Li, W., & Yang, H. (2025). Named entity recognition in online medical consultation using deep learning. *Applied Sciences*, 15(6), 3033. <https://doi.org/10.3390/app15063033>
- Huang, X., Zhang, J., Xu, Z., Ou, L., & Tong, J. (2021). A knowledge graph based question answering method for medical domain. *PeerJ Computer Science*, 7, e667. <https://doi.org/10.7717/peerj-cs.667>
- Huh, J., Marmor, R., & Jiang, X. (2016). Lessons learned for online health community moderator roles: A mixed-methods study of moderators resigning from WebMD communities. *Journal of Medical Internet Research*, 18(9), e247. <https://doi.org/10.2196/jmir.6331>
- Huo, P., Yang, Y., Zhou, J., Chen, C., & He, L. (2020). TERG: Topic-aware emotional response generation for chatbot. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. <https://doi.org/10.1109/IJCNN48605.2020.9206719>
- Hölldobler, S., & Kalinke, Y. (1994). Toward a massively parallel computational model for logic programming. In *Proceedings of the Workshop on Combining Symbolic and Connectionist Processing* (pp. 1–18).
- International Business Machines Corporation. (2023). *What are large language models (LLMs)?* IBM Think Blog. <https://www.ibm.com/think/topics/large-language-models>
- Jiang, F., Fan, Y., Chu, X., Li, P., Zhu, Q., & Kong, F. (2021). Hierarchical macro discourse parsing based on topic segmentation. *Proceedings of the AAAI*

- Conference on Artificial Intelligence*, 35(14), 13,152–13,160.
<https://doi.org/10.1609/aaai.v35i14.17554>
- Jiang, F., Xu, S., Chu, X., Li, P., Zhu, Q., & Zhou, G. (2018). MCDTB: A macro-level Chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3493–3504). Association for Computational Linguistics.
- Jin, D., Mehri, S., Hazarika, D., Padmakumar, A., Lee, S., Liu, Y., & Namazifar, M. (2023). *Data-Efficient Alignment of Large Language Models with Human Feedback Through Natural Language*. arXiv. <https://doi.org/10.48550/arXiv.2311.14543>
- Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 6421.
<https://doi.org/10.3390/app11146421>
- Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., & Yu, S. (2022). Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55(2), 1–36. <https://doi.org/10.1145/3490238>
- Jin, X., Yuan, Y., Chang, C., Wu, X., Tan, X., & Liu, Z. (2025). Telemedicine in China: Effective indicators of telemedicine platforms for promoting health and well-being among healthcare consumers. *Digital Health*, 11, 20552076251341163.
<https://doi.org/10.1177/20552076251341163>
- Joty, S., Carenini, G., & Ng, R. (2015). Codra: A novel discriminative framework for rhetorical analysis. *Computational linguistics*, 41(3), 385-435.
<https://aclanthology.org/J15-3002.pdf>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly*, 12(4), 571–586.
<https://doi.org/10.2307/249133>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020, January 23). *Scaling laws for neural language models*. arXiv.org. <https://arxiv.org/abs/2001.08361>
- Kautz, H. A. (2022). The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, 43(1), 45–58. <https://doi.org/10.1002/aaai.12036>
- Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., & Liu, B. (2023). Continual pre-training of language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=7fHij5G49H>
- Keselman, A., Tse, T., Crowell, J. A., Browne, A., Ngo, L., & Zeng, Q. T. (2008). Consumer health concepts that do not map to the UMLS: Where do they fit? *Journal of the American Medical Informatics Association*, 15(4), 496-506.
<https://doi.org/10.1197/jamia.M2612>
- Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2007). Assessing consumer health vocabulary familiarity: An exploratory study. *Journal of Medical Internet Research*, 9(1), e5. <https://doi.org/10.2196/jmir.9.1.e5>
- Kilicoglu, H., Abacha, A. B., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., & Demner-Fushman, D. (2018). Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1), 34. <https://doi.org/10.1186/s12859-018-2045-1>
- Kim, S., & Lee, J. (2024). Persona-based prompting for empathetic health counseling with large language models. *JMIR Medical Informatics*, 12(3), e60412.
<https://doi.org/10.2196/60412>
- Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., & Park, H. W. (2024). MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, 37. <https://doi.org/10.52202/079017-2522>
- Knowles, E. A., & Danzi, B. A. (2025). The role of online social support in mental health: Comparing rural and urban youth. *Children*, 12(2), 113.
<https://doi.org/10.3390/children12020113>

- Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., Kariippanon, K., & Coiera, E. (2019). The personalization of conversational agents in health care: Systematic review. *Journal of Medical Internet Research*, 21(11), e15360. <https://doi.org/10.2196/15360>
- Krešević, S., Giuffrè, M., Ajčević, M., Accardo, A., Crocè, L. S., & Shung, D. L., et al. (2024). Optimization of hepatological clinical guidelines interpretation by large language models: A retrieval augmented generation-based framework. *npj Digital Medicine*, 7, 102. <https://doi.org/10.1038/s41746-024-01091-y>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology (4th ed.)*. SAGE. <https://doi.org/10.4135/9781071878781>
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K., Jones, L., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7, 452–466. https://doi.org/10.1162/tacl_a_00276
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>
- Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of EMNLP 2017 (pp. 188-197)*. <https://doi.org/10.18653/v1/D17-1018>
- Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. arXiv. <https://doi.org/10.48550/arXiv.2104.08691>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://doi.org/10.48550/arXiv.2005.11401>
- Li, J., Chen, X., & Zhou, Y. (2023). Prompt-based summarization for biomedical question understanding. *Journal of Biomedical Informatics*, 149, Article 104582. <https://doi.org/10.1016/j.jbi.2023.104582>
- Li, R., Wang, H., Deng, Z., & Gao, B. (2022). Doctors' provision of online health consultation service and patient review valence: Evidence from a quasi-experiment. *Information & Management*, 59(5), 103360. <https://doi.org/10.1016/j.im.2020.103360>
- Li, S., Yavuz, S., Chen, W., & Yan, X. (2021). Task-adaptive pre-training and self-training are complementary for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 1006-1015)*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.findings-emnlp.86>
- Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL 2021*. arXiv. <https://doi.org/10.48550/arXiv.2101.00190>
- Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinearth*, 1(1), Article 9. <https://doi.org/10.1007/s44336-024-00009-2>
- Li, Y., Li, K., Ning, H., Xia, X., Guo, Y., Wei, C., Cui, J., & Wang, B. (2021). Towards an online empathetic chatbot with emotion causes. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1301–1310)*. Association for Computing Machinery. <https://doi.org/10.1145/3404835.3463042>
- Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(7), e40895. <https://doi.org/10.7759/cureus.40895>

- Liang, B., Wang, Y., & Tong, C. (2025). AI reasoning in deep learning era: From symbolic AI to neural-symbolic AI. *Mathematics*, 13(11), 1707. <https://doi.org/10.3390/math13111707>
- Linstone, H. A., & Turoff, M. (Eds.). (2002). *The Delphi method: Techniques and applications*. Newark, NJ: New Jersey Institute of Technology.
- Liu, C., Wang, W. Y. C., & Khan, G. (2023). Online health consultation: A systematic review for future research directions. In *Proceedings of the 7th International Conference on Medical and Health Informatics (ICMHI 2023)* (pp. 230–236). ACM. <https://doi.org/10.1145/3608298.3608340>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
- Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3469–3483). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.269>
- Liu, X., Ji, K., Fu, Y., Du, Z., Yang, Z., Tang, J., & Li, X. (2021). *P-Tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks*. arXiv. <https://doi.org/10.48550/arXiv.2110.07602>
- Liu, Y.-L., Yan, W., Hu, B., Li, Z., & Lai, Y. (2022). Effects of personalization and source expertise on users' health beliefs and usage intention toward health chatbots: Evidence from an online experiment. *Digital Health*, 8, 1–18. <https://doi.org/10.1177/20552076221129718>
- Lu, Y., Bartolo, M., Riedel, S., Stenetorp, P., & Schick, T. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. <https://doi.org/10.18653/v1/2022.acl-long.556>
- Luo, N., et al. (2025). Self-Training Large Language Models for Tool-Use Without Demonstrations. In *Findings of the NAACL 2025*. <https://doi.org/10.18653/v1/2025.findings-naacl.69>
- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), bbac409. <https://doi.org/10.1093/bib/bbac409>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B. P., Gupta, S., Yazdanbakhsh, A., & Clark, P. (2023). *Self-Refine: Iterative Refinement with Self-Feedback*. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2303.17651>
- Maharjan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M., Barnes, G., Thapa, R., Mao, Q., & Das, R. (2024). OpenMedLM: prompt engineering can outperform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1), 14156. <https://doi.org/10.1038/s41598-024-64827-6>
- Malhotra, N. K. (2010). *Marketing research: An applied orientation (6th ed.)*. Pearson.
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266. [https://doi.org/10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2)
- Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan. <https://doi.org/10.1057/9780230511910>
- McCarthy, J. (1980). Circumscription—A form of non-monotonic reasoning. *Artificial Intelligence*, 13(1–2), 27–39. [https://doi.org/10.1016/0004-3702\(80\)90011-9](https://doi.org/10.1016/0004-3702(80)90011-9)
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>

- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 35 (NeurIPS 2022).
https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf
- Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). *Augmented Language Models: a Survey*. arXiv.
<https://doi.org/10.48550/arxiv.2302.07842>
- Miao, Y., Liu, H., Lei, W., Chen, N., & Kan, M.-Y. (2024). Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6277-6295). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.339>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations (ICLR 2013)*, Workshop Track.
<https://doi.org/10.48550/arXiv.1301.3781>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook (3rd ed.)*. Sage Publications.
- Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Holt, Rinehart and Winston.
- Mills, F., Drury, J., Symons, C., Weston, D., Amlôt, R., & Carter, H. (2024). Online support groups, social identity, and the health and wellbeing of adults with Long Covid: An interview study. *Journal of Community & Applied Social Psychology*, 34(5), e2849. <https://doi.org/10.1002/casp.2849>
- Mills, F., et al. (2025). A mixed studies systematic review on the health and wellbeing effects of online support groups for chronic conditions. *npj Mental Health Research*, 4, Article 17. <https://doi.org/10.1038/s44271-025-00217-6>
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–265. <https://doi.org/10.1038/s41586-023-05881-4>
- Mordor Intelligence. (2025). Digital health market – Growth, trends, COVID-19 impact, and forecasts (2025–2030). *Mordor Intelligence*.
<https://www.mordorintelligence.com/industry-reports/digital-health-market>
- Myers, M. D. (1997). Qualitative research in information systems. *MIS Quarterly*, 21(2), 241–242. <https://doi.org/10.2307/249422>
- National Health Commission of China. (2020). *Guidelines for the Diagnosis and Treatment of Mental Disorders (2020 Edition)*. Retrieved from
<https://www.nhc.gov.cn/yzygj/c100068/202012/b4305ace9e14440792eb76d29602c88a.shtml>
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Nguyen, V., Karimi, S., Rybinski, M., & Xing, Z. (2023). MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, & A. A. Krisnadh (Eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 629–648). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2023.ijcnlp-main.42>
- Oh, S., Zhang, Y., & Park, M. S. (2016). Cancer information seeking in social question and answer services: Identifying health-related topics in cancer questions on Yahoo! Answers. *Information Research*, 21(3), paper 718. Retrieved from
<https://informationr.net/ir/21-3/paper718.html>

- Okesina, M. (2020). A critical review of the relationship between paradigm, methodology, design and method in research. *IOSR Journal of Research & Method in Education*, 10(3), 57–68. <https://doi.org/10.9790/7388-1003015768>
- Ong, L. M. L., de Haes, J. C. J. M., Hoos, A. M., & Lammes, F. B. (1995). Doctor-patient communication: A review of the literature. *Social Science & Medicine*, 40(7), 903–918. [https://doi.org/10.1016/0277-9536\(94\)00155-M](https://doi.org/10.1016/0277-9536(94)00155-M)
- OpenAI. (2025). Fine-Tuning GPT-4o for Customization. Retrieved from <https://openai.com/index/gpt-4o-fine-tuning>
- OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., . . . Malkov, Y. (2024, October 25). *GPT-4O System Card*. arXiv.org. <https://arxiv.org/abs/2410.21276>
- Orlikowski, W. J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, 3(3), 398–427. <https://doi.org/10.1287/orsc.3.3.398>
- Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information Systems Research*, 2(1), 1–28. <https://doi.org/10.1287/isre.2.1.1>
- Ormandy, P. (2011). Defining information need in health-Assimilating complex theories derived from information science. *Health Expectations*, 14(1), 92–104. <https://doi.org/10.1111/j.1369-7625.2010.00598.x>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. <https://doi.org/10.48550/arxiv.2203.02155>
- Park, M. S., He, Z., Chen, Z., Oh, S., & Bian, J. (2016). Consumers' use of UMLS concepts on social media: Diabetes-related textual data analysis in blog and social Q&A sites. *JMIR Medical Informatics*, 4(4), e41. <https://doi.org/10.2196/medinform.5748>
- Patrick, T. B., Monga, H. K., Sievert, M. E., Houston Hall, J. M., Longo, D. R., & Mitchell, J. A. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *Journal of Medical Internet Research*, 3(3), e24. <https://doi.org/10.2196/jmir.3.3.e24>
- Peirce, C. S. (1998). Abduction. In N. Houser, C. J. W. Kloesel, & the Peirce Edition Project (Eds.), *The essential Peirce: Selected philosophical writings, Volume 2 (1893–1913) (pp. 85–97)*. Indiana University Press.
- Peirce, C. S., & Houser, N. (Ed.). (1998). *The essential Peirce: Selected philosophical writings (Vol. 2, 1893–1913)*. Indiana University Press.
- Peng, C., Yang, X., Chen, A., Smith, K. E., Pournajatian, N. M., Costa, A. B., Lipori, G., Mitchell, D. A., Ospina, N. S., Ahmed, M. M., Hogan, W. R., Shenkman, E. A., Guo, Y., Bian, J., & Wu, Y. (2023). A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6, 210. <https://doi.org/10.1038/s41746-023-00958-w>
- Pennebaker, J. W. (2003). *The social, linguistic, and health consequences of emotional disclosure*. In J. Suls & K. A. Wallston (Eds.), *Social psychological foundations of health and illness (pp. 288–313)*. Blackwell. <https://doi.org/10.1002/9780470753552.ch11>
- Pervan, G. P. (1994). Information systems management: An Australian view of the key issues. *Australasian Journal of Information Systems*, 1(2), 32–44.
- Petrič, G., Cugmas, M., Petrič, R., & Atanasova, S. (2023). The quality of informational social support in online health communities: A content analysis of cancer-related discussions. *Digital Health*, 9, 20552076231155681. <https://doi.org/10.1177/20552076231155681>

- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 2463-2473). Association for Computational Linguistics. <https://aclanthology.org/D19-1250>
- Ping An Healthcare and Technology Co., Ltd. (2023). *Annual report 2022 (Ping An Good Doctor)*. <https://group.pingan.com/resource/pingan/IR-Docs/2023/pingan-ar22-presentation.pdf>
- Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, 46(1–2), 77–105. [https://doi.org/10.1016/0004-3702\(90\)90005-K](https://doi.org/10.1016/0004-3702(90)90005-K)
- Pounds, G. (2011). Empathy as “appraisal”: Developing a new language-based approach to the exploration of clinical empathy. *Journal of Applied Linguistics and Professional Practice*, 7(2), 145–168. <https://doi.org/10.1558/japl.v7i2.145>
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008* (pp. 2961-2968). Retrieved from <https://aclanthology.org/L08-1093/>
- Pugachev, A., Artemova, E., Bondarenko, A., & Braslavski, P. (2023). Consumer health question answering using Off-the-Shelf components. In *Lecture notes in computer science* (pp. 571-579). https://doi.org/10.1007/978-3-031-28238-6_48
- Qwen Team. (2024). *Qwen2.5 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2412.15115>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Rafailov, R., Sharma, Y., Mitchell, E., & Ermon, S. (2023). *Direct Preference Optimization: Your language model is secretly a reward model*. arXiv. <https://doi.org/10.48550/arXiv.2305.18290>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67. <https://jmlr.org/papers/v21/20-074.html>
- Rai, A. (2017). Editor’s comments: Diversity of design science research. *MIS Quarterly*, 41(1), iii–xviii. <https://www.jstor.org/stable/26629633>
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1359. <https://doi.org/10.1056/NEJMra1814259>
- Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of EMNLP 2016*, 2383-2392. <https://doi.org/10.18653/v1/D16-1264>
- Reddy, S., Chen, D., & Manning, C. D. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7, 249-266. https://doi.org/10.1162/tacl_a_00266
- Rice, R. E., & Love, G. (1987). Electronic emotion: Socioemotional content in a computer-mediated communication network. *Communication Research*, 14(1), 85–108. <https://doi.org/10.1177/009365087014001005>
- Roberts, A., Raffel, C., & Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)* (pp. 5418-5426). Association for Computational Linguistics. <https://aclanthology.org/2020.emnlp-main.437>
- Roberts, K., Demner-Fushman, D., Topping, J. M., et al. (2014a). Overview of the GARD question corpus and its applications for consumer health QA. *BMC Medical Informatics and Decision Making*, 15(Suppl 2), S3.

- Roberts, K., Kilicoglu, H., Fiszman, M., & Demner-Fushman, D. (2014b). Automatically classifying question types for consumer health questions. *Journal of Biomedical Informatics*, 51, 239-248. <https://doi.org/10.1016/j.jbi.2014.05.004>
- Roberts, K., Masterton, K., Fiszman, M., & Demner-Fushman, D. (2014c). Annotating question decomposition on complex medical questions. *Proceedings of the 2014 Workshop on Biomedical Natural Language Processing*, 39-47. <https://aclanthology.org/W14-4605/>
- Rocktäschel, T., & Riedel, S. (2017). *End-to-end differentiable proving*. arXiv. <https://doi.org/10.48550/arXiv.1705.11040>
- Roter, D. L., & Hall, J. A. (1989). Studies of doctor-patient interaction. *Annual Review of Public Health*, 10, 163-180. <https://doi.org/10.1146/annurev.pu.10.050189.001115>
- Roter, D. L., & Hall, J. A. (2006). *Doctors talking with patients, patients talking with doctors: Improving communication in medical visits (2nd ed.)*. Praeger.
- Roter, D., & Larson, S. (2002). The Roter Interaction Analysis System (RIAS): Utility and flexibility for analysis of medical interactions. *Patient Education and Counseling*, 46(4), 243-251. [https://doi.org/10.1016/S0738-3991\(02\)00012-5](https://doi.org/10.1016/S0738-3991(02)00012-5)
- Rowlands, H., Aylwin-Foster, E., Hotopf, M., Rayner, L., Simpson, A., Lavelle, G., & Lawrence, V. (2023). A qualitative exploration of the psychosocial needs of people living with long-term conditions and their perspectives on online peer support. *Health Expectations*, 26(5), 2075-2088. <https://doi.org/10.1111/hex.13814>
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1)*. MIT Press. <https://doi.org/10.7551/mitpress/5236.001.0001>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. <https://doi.org/10.1038/323533a0>
- Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1988). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. MIT Press. <https://mitpress.mit.edu/9780262680530/parallel-distributed-processing-volume-1>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735. <https://doi.org/10.2307/412243>
- Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. V., . . . Rush, A. M. (2022). *Multitask prompted training enables zero-shot generalization*. In International Conference on Learning Representations (ICLR 2022). <https://doi.org/10.48550/arXiv.2110.08207>
- Savery, M., Ben Abacha, A., Gayen, S., & Demner-Fushman, D. (2020). Question-driven summarization of answers to consumer health questions (MEDIQA-AnS). *Scientific Data*, 7, 219. <https://doi.org/10.1038/s41597-020-00667-z>
- Schnock, K. O., Snyder, J. E., Fuller, T. E., Duckworth, M., Grant, M., Yoon, C., Lipsitz, S., Dalal, A. K., Bates, D. W., & Dykes, P. C. (2019). Acute care patient portal intervention: Portal use and patient activation. *Journal of Medical Internet Research*, 21(7), e13336. <https://doi.org/10.2196/13336>
- Schuhmann, C., Beaumont, R., Vencu, R., Cade Gordon, Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Robert Kaczmarczyk, Jenia Jitsev. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35, 25278-25294. <https://doi.org/10.5555/3600270.3602103>
- Seedorff, M., Peterson, K., Nelsen, L., Cocos, C., McCormick, J., Chute, C. G., & Pathak, J. (2013). Incorporating expert terminology and disease risk factors into consumer

- health vocabularies. In *Biocomputing 2013* (pp. 421-432). World Scientific.
https://doi.org/10.1142/9789814447973_0041
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2020.emnlp-main.425>
- Shastri, L., & Ajanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16(3), 417-494.
<https://doi.org/10.1017/S0140525X00030910>
- Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020). *AutoPrompt: Eliciting knowledge from language models with automatically generated prompts*. arXiv. <https://doi.org/10.48550/arXiv.2010.15980>
- Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, 8(4), 303-320. [https://doi.org/10.1016/0010-4809\(75\)90009-9](https://doi.org/10.1016/0010-4809(75)90009-9)
- Shrout, T. P., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.
- Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467-482. <https://www.jstor.org/stable/985254>
- Simon, H. A. (1996). *The sciences of the artificial (3rd ed.)*. MIT Press.
<https://doi.org/10.7551/mitpress/12107.001.0001>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172-180. <https://doi.org/10.1038/s41586-023-06291-2>
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaeckermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D. R., Corrado, G. S., Matias, Y., Azizi, S., & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3), 943-950. <https://doi.org/10.1038/s41591-024-03423-7>
- Sorin, V., Brin, D., Barash, Y., Klang, E., & co-authors. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26, e52597.
<https://doi.org/10.2196/52597>
- Sternberg, R. J. (2011). *Cognitive psychology (6th ed.)*. Wadsworth Cengage Learning.
- Street, R. L., Jr., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Education and Counseling*, 74(3), 295-301.
<https://doi.org/10.1016/j.pec.2008.11.015>
- Suchman, A. L., Markakis, K., Beckman, H. B., & Frankel, R. (1997). A model of empathic communication in the medical interview. *JAMA*, 277(8), 678-682.
<https://doi.org/10.1001/jama.1997.03540320082047>
- Sukhbaatar, S., Szlam, A., Weston, J., & Fergus, R. (2015). End-to-end memory networks. *Advances in Neural Information Processing Systems*, 28, 2440-2448.
<https://doi.org/10.48550/arXiv.1503.08895>
- Sun, M., & Chai, J. Y. (2007). Discourse processing for context question answering based on linguistic knowledge. *Knowledge-Based Systems*, 20(6), 511-526.
<https://doi.org/10.1016/j.knosys.2007.04.005>

- Takeda, H., Veerkamp, P., Tomiyama, T., & Yoshikawa, H. (1990). Modeling design processes. *AI Magazine*, 11(4), 37–48. <https://doi.org/10.1609/aimag.v11i4.855>
- Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7, 258. <https://doi.org/10.1038/s41746-024-01258-7>
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(9), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
- Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023, July 18). *Llama 2: Open foundation and Fine-Tuned chat models*. arXiv.org. <https://arxiv.org/abs/2307.09288>
- Toyhom. (2019). *Chinese Medical Dialogue Data* [Data set]. GitHub. <https://github.com/Toyhom/Chinese-medical-dialogue-data>
- Tran, J., Hao, Z., Xie, Z., Yiu, K., & Neves, A. L. (2025). Nationwide trends of online consultations in China: A 10-year sequential cross-sectional study of 65,305 clinicians. *BMJ Public Health*, 3(2), e002296. <https://doi.org/10.1136/bmjph-2024-002296>
- Tseng, Y.-M., Huang, Y.-C., Hsiao, T.-Y., Chen, W.-L., Huang, C.-W., Meng, Y., & Chen, Y.-N. (2024). Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 16612–16631). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.969>
- Vaishnavi, V., & Kuechler, W. (2015). *Design science research in information systems*. CRC Press. <https://doi.org/10.1201/b18448>
- Varshney, N., Dolin, P., Seth, A., & Baral, C. (2024). The art of defending: A systematic evaluation and analysis of LLM defense strategies on safety and over-defensiveness. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 13111–13128). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.776>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30 (NeurIPS 2017)* (pp. 5998–6008). Curran Associates. <https://doi.org/10.5555/3295222.3295349>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Verberne, S., Boves, L., Coppen, P. A., & Oostdijk, N. (2006). Discourse-based answering of why-questions. In *Traitement Automatique des Langues, Volume 47, Numéro 2: Discours et document: traitements automatiques [Computational Approaches to Discourse and Document Processing]* (pp. 21–41). <https://aclanthology.org/2006.tal-2.2/>
- Verberne, S. (2007, July). Paragraph retrieval for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 922–922).
- vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems*, 37, 205–224. <https://doi.org/10.17705/1CAIS.03709>

- Voorhees, E. M., & Hersh, W. (2012). Overview of the TREC 2012 Medical Records Track. In *Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012)*.
- Voorhees, E. M., & Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 200–207). Association for Computing Machinery. <https://doi.org/10.1145/345508.345577>
- Vu, T., Lester, B., Constant, N., Al-Rfou, R., Cer, D., & Narang, S. (2022). *SPoT: Better frozen model adaptation through soft prompt transfer*. arXiv. <https://doi.org/10.48550/arXiv.2110.07904>
- Vydiswaran, V. G. V., Mei, Q., Hanauer, D. A., & Zheng, K. (2014). *Mining consumer health vocabulary from community-generated text*. AMIA Annual Symposium Proceedings, 2014, 1150-1159.
- Wadhwa, S., Weng, W.-H., Luo, Y., & Wallace, B. C. (2023). *RedHOT: A corpus of Reddit health online talk for modelling patient-centred outcomes*. arXiv. <https://doi.org/10.48550/arXiv.2210.06331>
- Walsham, G. (1993). *Interpreting information systems in organizations*. Chichester, UK: Wiley.
- Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *Information Systems Journal*, 4(1), 74–81. <https://doi.org/10.1111/j.1365-2575.1995.tb00082.x>
- Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1), 3–43. <https://doi.org/10.1177/009365096023001001>
- Wang, G., Yang, G., Du, Z., Fan, L., & Li, X. (2023). *ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation*. arXiv. <https://doi.org/10.48550/arXiv.2306.09968>
- Wang, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., Yuan, W., & Topol, E. J. (2025). *A survey of LLM-based agents in medicine*. arXiv. <https://doi.org/10.48550/arXiv.2502.11211>
- Wang, J., Li, W., Lin, P., & Mu, F. (2021). Empathetic response generation through graph-based multi-hop reasoning on emotional causality. *Knowledge-Based Systems*, 233, 107547. <https://doi.org/10.1016/j.knosys.2021.107547>
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Li, X., Ge, B., Zhu, D., Yuan, Y., Shen, D., Liu, T., & Zhang, S. (2023). *Prompt Engineering for Healthcare: Methodologies and applications*. arXiv. <https://doi.org/10.48550/arxiv.2304.14670>
- Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7, Article 41. <https://doi.org/10.1038/s41746-024-01029-4>
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>
- Wang, W., Yang, Y., & Wu, F. (2025). Towards data- and knowledge-driven AI: A survey on neuro-symbolic computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2), 878–899. <https://doi.org/10.1109/TPAMI.2024.3483273>
- Wang, X., Li, J., Chen, S., et al. (2025). Huatuo-26M, a large-scale Chinese medical QA dataset. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 3828–3848). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-naacl.211>
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models*. In International Conference on Learning Representations (ICLR 2023). <https://doi.org/10.48550/arXiv.2203.11171>

- Wang, Y., et al. (2022). *Aligning language models with self-generated instructions*. arXiv. <https://doi.org/10.48550/arXiv.2212.10560>
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii. <https://doi.org/10.2307/4132319>
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*. <https://arxiv.org/abs/2109.01652>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. <https://doi.org/10.48550/arXiv.2201.11903>
- Welivita, A., & Pu, P. (2023). A survey of consumer health question answering systems. *AI Magazine*, 44(4), 482–507. <https://doi.org/10.1002/aaai.12140>
- Weston, J., Chopra, S., & Bordes, A. (2014). *Memory networks*. arXiv. <https://doi.org/10.48550/arXiv.1410.3916>
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. arXiv. <https://doi.org/10.48550/arxiv.2302.11382>
- Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., Bradley, R., & Heywood, J. (2010). Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2), e19. <https://doi.org/10.2196/jmir.1549>
- Wiese, G., Weissenborn, D., & Neves, M. (2017). Neural domain adaptation for biomedical question answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 281–289). Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-1029>
- World Health Organization. (2021). *Global strategy on digital health 2020–2025*. World Health Organization. <https://iris.who.int/handle/10665/344249>
- Wu, S., Hsu, W., & Lee, M. L. (2024). EHChat: A knowledge-grounded, empathy-enhanced language model for healthcare interactions. In J. Hale, K. Chawla, & M. Garg (Eds.), *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)* (pp. 141–151). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.sicon-1.10>
- Wu, T., et al. (2024). *Large Language Models Alignment via Comparative RL*. University of California, Berkeley. <https://doi.org/10.48550/arXiv.2403.06754>
- Xu, J., Gan, Z., Cheng, Y., & Liu, J. (2020). Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5021–5031). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.451>
- Yadav, S., Cobeli, S., & Caragea, C. (2023). Towards understanding consumer healthcare questions on the web with semantically enhanced contrastive learning. *Proceedings of the ACM Web Conference 2023*, 1773–1783. <https://doi.org/10.1145/3543507.3583449>
- Yadav, S., Gupta, D., & Demner-Fushman, D. (2022b). CHQ-Summ: A dataset for consumer healthcare question summarization. *Proceedings of the 29th International Conference on Computational Linguistics*, 7020–7030. <https://doi.org/10.48550/arXiv.2206.06581>
- Yadav, S., Gupta, D., Ben Abacha, A., & Demner-Fushman, D. (2022a). Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128, 104040. <https://doi.org/10.1016/j.jbi.2022.104040>
- Yang, A., Li, A., & Liu, X. (2024). *Qwen2 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2407.10671>
- Yang, A., Zhou, S., Li, B., Zhao, Y., & Qwen Team. (2025). *Qwen3 technical report: Advancing multilingual and instruction-aligned large language models*. arXiv. <https://doi.org/10.48550/arXiv.2505.09388>

- Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, Y., & Zan, H. (2024). Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 19332–19340. <https://doi.org/10.1609/aaai.v38i17.29907>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv. <https://doi.org/10.48550/arxiv.2305.10601>
- Yao, S., Yu, D., Zhao, J., Shafiei, M., Lin, B. Y., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing reasoning and acting in language models*. arXiv. <https://doi.org/10.48550/arXiv.2210.03629>
- Yin, R. K. (2018). *Case study research and applications: Design and methods (6th ed.)*. Sage Publications.
- Ying, L., Collins, K. M., Wei, M., Zhang, C. E., Tan, Z.-X., Weller, A., Tenenbaum, J. B., & Wong, L. (2023). *The Neuro-Symbolic Inverse Planning Engine (NIPE): Modeling probabilistic social inferences from linguistic inputs*. arXiv. <https://doi.org/10.48550/arXiv.2306.14325>
- Yoo, Y., Henfridsson, O., & Lyytinen, K. (2010). The new organizing logic of digital innovation: An agenda for information systems research. *Information Systems Research*, 21(4), 724–735. <https://doi.org/10.1287/isre.1100.0322>
- Yu, H., Lee, M., Kaufman, D., Ely, J., Osheroff, J. A., Hripcsak, G., & Cimino, J. (2007). Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *Journal of Biomedical Informatics*, 40(3), 236-251. <https://doi.org/10.1016/j.jbi.2007.03.002>
- Yuan, H. (2025). Agentic large language models for healthcare: Current progress and future opportunities. *Medicine Advances*, 3(1), 37–41. <https://doi.org/10.1002/med4.70000>
- Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J. L., Moor, M., Lundberg, S. M., Major, V., Ashley, E. A., Harrington, R. A., Boyd, J., Nelson, J., Langlotz, C. P., & Hiesinger, W. (2024). Almanac: Retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2), Article A1oa2300068. <https://doi.org/10.1056/A1oa2300068>
- Zeng, Q. T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1), 24-29. <https://doi.org/10.1197/jamia.M1761>
- Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E. M., & Greenes, R. A. (2004). Positive attitudes and failed queries: An exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics*, 73(1), 45–55. <https://doi.org/10.1016/j.ijmedinf.2003.12.015>
- Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A. C., Goryachev, S., & Ngo, L. (2007). Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, 9(1), e4. <https://doi.org/10.2196/jmir.9.1.e4>
- Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 179–214. <https://doi.org/10.1145/984321.984322>
- Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). Self-attention and positional encoding. In *Dive into deep learning*. https://d2l.ai/chapter_attention-mechanisms-and-transformers/self-attention-and-positional-encoding.html
- Zhang, M., Dou, S., Wang, Z., & Wu, Y. (2022). Focus-driven contrastive learning for medical question summarization. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.48550/arXiv.2209.00484>
- Zhang, S., Zhang, X., Wang, H., Cheng, J., Li, P., & Ding, Z. (2017). Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Applied Sciences*, 7(8), 767. <https://doi.org/10.3390/app7080767>

- Zhang, X., Peng, B., Tian, Y., Zhou, J., Zhang, Y., Mi, H., & Meng, H. (2024). *Self-Tuning: Instructing LLMs to Effectively Acquire New Knowledge through Self-Teaching*. arXiv. <https://doi.org/10.48550/arxiv.2406.06326>
- Zhang, X., Zhou, X., Liu, Y., & Yang, Q. (2025). A survey of multi-AI agent collaboration: Theories, frameworks, and applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. <https://doi.org/10.1145/3745238.3745531>
- Zhang, Y. P., Zhang, Y., Liu, W. H., Yan, Y. T., & Wei, H. H. (2018). Comprehensive unmet needs and correlations with quality of life in Chinese cancer patients. *European Journal of Cancer Care*, 27(2), e12813. <https://doi.org/10.1111/ecc.12813>
- Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec: Improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6, 52. <https://doi.org/10.1038/s41597-019-0055-0>
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2023). *Calibrate before use: Improving few-shot performance of language models*. ICML. <https://doi.org/10.48550/arXiv.2102.09690>
- Zheng, C., Sabour, S., Wen, J., Zhang, Z., & Huang, M. (2023). AugESC: Dialogue augmentation with large language models for emotional support conversation. *Findings of the Association for Computational Linguistics: ACL 2023*, 1552–1568. <https://doi.org/10.18653/v1/2023.findings-acl.99>
- Zheng, G., et al. (2025). *Self-Training Large Language Models with Confident Reasoning Paths*. arXiv. <https://doi.org/10.48550/arXiv.2505.17454>
- Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)* (pp. 730-739). AAAI Press. <https://doi.org/10.1609/aaai.v32i1.11325>
- Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9), 2337-2348. <https://doi.org/10.1007/s11263-022-01653-1>
- Zhou, Y., Schick, T., Schütze, H., & Lewis, M. (2023). *Large language models are human-level prompt engineers*. arXiv. <https://doi.org/10.48550/arXiv.2211.01910>
- Zhu, M., Ahuja, A., Juan, D.-C., Wei, W., & Reddy, C. K. (2020). Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3840–3849). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.342>
- Zhu, S., Zhou, J., & Xu, N. (2025). The impact of perceived quality on patients' adoption and usage of online health consultations: An empirical study based on trust theory. *Healthcare*, 13(14), 1753. <https://doi.org/10.3390/healthcare13141753>
- Ziebland, S., & Wyke, S. (2012). Health and illness in a connected world: How might sharing experiences on the internet affect people's health? *The Milbank Quarterly*, 90(2), 219-249. <https://doi.org/10.1111/j.1468-0009.2012.00662.x>
- Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, 36(4-5), 326-333. <https://doi.org/10.1016/j.jbi.2003.09.015>
- Zimmermann, C., Del Piccolo, L., Bensing, J., Bergvik, S., De Haes, H., Eide, H., Fletcher, I., Goss, C., Heaven, C., Humphris, G., Kim, Y., Langewitz, W., Meeuwesen, L., Nuebling, M., Rimondini, M., Salmon, P., Van Dulmen, S., Wissow, L., Zandbelt, L., & Finset, A. (2010). Coding patient emotional cues and concerns in medical consultations: The Verona coding definitions of emotional sequences (VR-CoDES). *Patient Education and Counseling*, 82(2), 141-148. <https://doi.org/10.1016/j.pec.2010.03.017>

Appendix

Appendix 1. XML Configuration used to Define the Annotation Interface in Label Studio. Note: Original text translated from Chinese into English for clarity.

```
<View>
  <!-- Display user query text -->
  <Text name="merged_input" value="$__merged__"/>

  <!-- Disease complexity selection and disease list -->
  <Header value="Disease Complexity" />
  <Choices name="disease_complexity" toName="merged_input" choice="single" required="true">
    <Choice value="SIN" />
    <Choice value="MUL" />
  </Choices>
  <TextArea name="disease_list" toName="merged_input" perRegion="false" rows="3"
  placeholder="Please list specific diseases, such as depression, diabetes, etc." />

  <!-- Asker-patient relationship -->
  <Header value="Ask-patient Relationship" />
  <Choices name="asker_patient_relation" toName="merged_input" choice="single" required="true">
    <Choice value="SELF" />
    <Choice value="OTHER" />
    <Choice value="NON" />
  </Choices>

  <!-- Annotate user medical needs -->
  <Labels name="medical_need" toName="merged_input" choice="multiple" showInline="true">
    <Label value="M-N-GMI"/>
    <Label value="M-N-ETI"/>
    <Label value="M-N-DIA"/>
    <Label value="M-N-TREAT"/>
    <Label value="M-N-PROG"/>
  </Labels>

  <!-- Annotate medical needs related context -->
  <Labels name="medical_context" toName="merged_input" choice="multiple" showInline="true">
    <Label value="M-C-ELA"/>
    <Label value="M-C-BACK"/>
    <Label value="M-C-CON"/>
  </Labels>

  <!-- Annotate user emotional needs -->
  <Labels name="emotional_need" toName="merged_input" choice="multiple" showInline="true">
    <Label value="E-N-FEEL"/>
    <Label value="E-N-VIEW"/>
  </Labels>

  <!-- Annotate emotional needs related context -->
  <Labels name="emotional_context" toName="merged_input" choice="multiple" showInline="true">
    <Label value="E-C-CAUSE"/>
  </Labels>

  <!-- Rewrite/explain annotated span -->
  <TextArea name="span_notes" toName="merged_input" perRegion="true" placeholder="Provide
  additional explanation or rewrite for the annotated content..." rows="3"/>

  <!-- Link each context span to its corresponding need span -->
  <Relations>
    <Relation value="related_contexts"/>
  </Relations>
</View>
```

Appendix 2. An Example of Multi-level Annotation Interface in Label Studio.

【身高体重】: 174cm, 80.1kg
 【疾病】: 阿片类药物成瘾及规范使用
 【患病时长】: 大于半年 已就诊医院科室: 江苏省人民医院 疼痛科 南京鼓楼医院 疼痛科 淮安市第三人民医院 用药情况: 抗焦虑抑郁药, 安眠药, 镇痛药 (曲马多、羟考酮) 奥士康定, 氯硝西泮
 【怀孕情况】:
 【过敏史】: 无(2021-10-26填写)
 【既往病史】: 手术: 痔疮手术(2021-06-18填写) 手术: 一年多前摔伤臀部, 当时无疼痛, 具体不详。2021.06月行肛周脓肿手术(2021-12-18填写) 手术: 尾骨切除史(2021-12-27填写) 1年前患者做车上掉下来臀部着地: 楼梯上滚下来臀部着地, 当时有疼痛, 休息后恢复。(2022-01-24填写) 手术: 做了胸腰部手术, 抽了一块髓突, 目前疼痛, 麻木 (2024-12-07填写) 重大疾病: 无(2021-06-18填写) 放疗: 无(2021-06-18填写) 慢性疾病: 无(2021-06-18填写) 达到痊愈: [1-15填写]
 【疾病描述】: 抗抑郁药焦虑药服用三年了, 不想吃了, 不想吃药了, 目前停药三天, 全身乏力, 躯体疼痛, 头晕头痛, 希望医生帮我怎么缓解。
 【希望获得的帮助】: 希望医生帮我怎么缓解。

M-N-GMI 6 | M-N-ETI 7 | M-N-DIA 8 | M-N-TREAT 9 | M-N-PROG 0

M-C-ELA q | M-C-BACK w | M-C-CON e

E-N-FEEL t | E-N-VIEW a

E-C-CAUSE s

Appendix 3. An Example in CHQA-MedEmo Dataset. Note: Original text translated from Chinese into English for clarity.

```
{
  "query_id": "12",
  "disease_description": "Doctor, hello! Previously I visited the gastroenterology department because of gastrointestinal function disorder, and the doctor prescribed Deanxit for me. While taking it the effect was very good. After one year of use I stopped the medication, because the doctor at that time did not instruct a dose-reduction plan, so I just stopped directly from taking one pill a day. But ten days after stopping I experienced very severe withdrawal reactions. At first I endured for two months by willpower, but later I could not hold out and went to see a psychiatrist. Then the psychiatrist kept prescribing substitute medications to attempt relieving symptoms, but after many types of drugs none worked, and the side-effects were large. Basically for almost every drug I took it for one week and then switched to another. The drugs I tried include Lexapro (Escitalopram), Duloxetine, Mirtazapine, Milnacipran, Tansospirone. It has been eight months since I stopped Deanxit, but now the withdrawal reaction still remains, and recurs, with fluctuating severity—sometimes better, sometimes worse—with symptoms of fatigue, muscle ache, headache and head pressure, bloated stomach, nausea, palpitations, etc. May I ask how long this withdrawal reaction will last, and how it should be treated? I met several fellow patients online withdrawing from Deanxit; they all recommended that I restart Deanxit and then taper gradually over about one year. They all say that if Deanxit is stopped suddenly without tapering, recovery will not happen, no other medication can substitute, one must taper slowly to recover. But Deanxit has too strong dependency potential; I do not dare to restart it myself now, for fear I might not be able to stop later! Therefore I hope the doctor will give some valuable advice: should I take Deanxit again and then reduce gradually to withdraw, or what should I do? Thank you!",
  "height_weight": "181 cm, 53 kg",
  "disease": "Deanxit withdrawal reaction",
  "duration_of_illness": "More than half a year",
  "pregnancy_status": "",
  "allergy_history": "None (filled in 2021-09-04) Allergic to Amoxicillin (filled in 2022-04-30) None (filled in 2023-06-26)",
  "past_medical_history": "Surgery: none (2022-04-30) Major disease: hepatitis B (2022-04-30) Radiotherapy/chemotherapy: none (2022-04-30) Chronic disease: hepatitis B, has been taking antiviral medication since 2009 and is currently taking telbivudine daily (2021-09-04) Chronic disease: hepatitis B (2022-04-30)",
  "desired_help": "I hope the doctor can give some precious advice: Should I restart taking Deanxit and then taper off again, or what should I do? Thank you!"
}
```

```
"query_id": "12",
"medical_needs": [
  {
    "need_id": "M1",
    "need_type": "M-N-PROG",
    "need_text": "May I ask how long this withdrawal reaction will last?",
    "related_contexts": [
      {
        "relation_type": "M-C-ELA",
        "context_text": "Deanxit withdrawal reaction, recurring, sometimes better, sometimes worse; symptoms of fatigue, muscle soreness, headache and head pressure, stomach bloating, nausea, heart palpitations, etc."
      },
      {
        "relation_type": "M-C-BACK",
        "context_text": "It has been eight months since I stopped Deanxit, yet the withdrawal reaction still persists."
      }
    ]
  },
  {
    "need_id": "M2",
    "need_type": "M-N-TREAT",
    "need_text": "Should I restart Deanxit and then taper off again, or what should I do?",
    "related_contexts": [
      {
        "relation_type": "M-C-ELA",
        "context_text": "It has been eight months since I stopped Deanxit, yet the withdrawal reaction still persists, and recurs with fluctuating severity—fatigue, muscle soreness, headache and head pressure, stomach bloating, nausea, heart palpitations, etc."
      },
      {
        "relation_type": "M-C-BACK",
        "context_text": "Previously I visited gastroenterology for gastrointestinal functional disorder, then doctors prescribed Deanxit for me. After one year's use I stopped the medication because the doctor did not instruct tapering. I abruptly stopped one pill daily. But ten days later I experienced severe withdrawal reactions, initially enduring for two months by willpower, later heading to psychiatry. The psychiatrist kept prescribing substitute medications to try to relieve symptoms, but after many types none worked and side-effects were large."
      }
    ]
  },
  {
    "relation_type": "M-C-CON",
    "context_text": "Fellow patients suggested I restart Deanxit and then taper off; but I feel Deanxit has too strong dependence potential, I dare not take it again, for fear I will be unable to quit later."
  }
]
},
"emotional_needs": [
  {
    "need_id": "E1",
    "need_type": "E-N-VIEW",
    "need_text": "Deanxit dependence is too strong; I dare not restart it, I'm afraid I will fail to quit later.",
    "related_contexts": [
      {
        "relation_type": "E-C-CAUSE",
        "context_text": "Previously I visited gastroenterology for gastrointestinal functional disorder and was prescribed Deanxit; after one year I stopped without tapering. Ten days later I had very severe withdrawal reactions, endured two months by willpower, but later could not cope and went to psychiatry, where substitute medications were tried but with no effect and large side-effects."
      }
    ]
  },
  {
    "need_id": "E2",
    "need_type": "E-N-FEEL",
    "need_text": "Suffering",
    "related_contexts": [
      {
        "relation_type": "E-C-CAUSE",
        "context_text": "Deanxit withdrawal is severe and lasting long, and recurs with fluctuating severity."
      }
    ]
  }
]
}
```

Answer(Benchmark): I understand that you have been experiencing withdrawal reactions for eight months since stopping Deanxit, accompanied by recurring physical and psychological symptoms, and that you feel pain and helplessness. This distress arises not only from fatigue, headaches, gastrointestinal discomfort and other somatic symptoms, but also from the defeat you've felt after multiple attempts to substitute medications that failed and caused significant side-effects. Patients in similar situations often share these feelings, so your experience is entirely understandable (Response to E-N-FEEL). You mentioned concerns about Deanxit dependence, and that worry is perfectly reasonable. Because you experienced severe withdrawal after abrupt discontinuation of Deanxit and subsequent substitute medications failed to relieve you, your fear of restarting is well justified. In treatment decision-making, this concern must be fully acknowledged and respected (Response to E-N-VIEW). Clinically, when withdrawal reactions occur following abrupt cessation of Deanxit without tapering, they can indeed last for several months or even longer than a year, and the symptoms often fluctuate (waxing and waning), but in most patients they gradually diminish over time and with physiological adaptation (Response to M-N-PROG). From the perspective of overall disease course and symptom management, emphasis should be placed on a systematic treatment of anxiety-related disorder rather than focusing solely on the withdrawal reaction itself. The suggestion from fellow patients that you "restart Deanxit, then taper slowly over about a year" is used in some cases and may help alleviate persistent withdrawal symptoms in the short term, but there is also risk of re-establishing dependence and then facing difficulty in stopping again. Your concern about Deanxit dependence is therefore reasonable. However, given your history of multiple medications with poor effect, if after thorough evaluation of liver function and overall health status conventional antidepressant or anxiolytic medications still fail to improve symptoms, then under close monitoring it may be considered to cautiously restart Deanxit at a low dose and taper gradually, personalised to you, to reduce risk. Meanwhile, it is essential to emphasise the importance of comprehensive treatment including psychotherapy (such as cognitive behavioural therapy, mindfulness intervention), lifestyle management (regular routine, moderate exercise, adequate nutrition) and continued follow-up. These measures complement pharmacotherapy, improve overall mind-body status, and increase the possibility of long-term recovery (Response to M-N-TREAT).

Appendix 4. Prompt for Question Understanding. Note: Original text translated from Chinese into English for clarity.

Role

You are a professional clinical assistant specializing in extracting medical and emotional needs and related contextual information in consumer health queries.

Task

Given a user_query, generate a structured JSON response by following the steps below:

1. Identify all user emotional needs (emotional_need):

- Emotional state (emotional_need_feel): Expressions of the user's emotional feelings regarding their current situation, illness, or treatment, intended to seek understanding, empathy, or emotional support. Example: "I really don't know what to do, I feel like I'm falling apart."
- Subjective view (emotional_need_view): Expressions of the user's personal opinions or attitudes toward their situation, illness, or treatment, intended to convey concerns, beliefs, or judgments. Example: "The side effects of this medication are too severe, I don't want to take it anymore."

2. For each identified emotional need (emotional_need), extract the associated background information (context) from the user_query.

- Causal explanation (context_cause): The reason or trigger underlying the emotional need, explaining why the user feels or believes a certain way. Example: "I have tried many medications, none of them worked, and all had serious side effects."

3. Identify all user medical needs (medical_need), with each entry containing only one specific question. Categories are as follows:

- General medical information (medical_need_general): Requests for broad or non-specific medical knowledge, such as general information about a disease, medication, or treatment procedure. Example: "Can you tell me how this drug works?"
- Etiology inquiry (medical_need_etiology): Requests for the cause or origin of a symptom or disease. Example: "What causes my headache?"
- Diagnostic clarification (medical_need_diagnosis): Requests for possible diagnoses or clarification of a condition based on specific symptoms. Example: "What exactly is wrong with me?"
- Treatment suggestion (medical_need_treatment): Requests for concrete treatment advice, medication recommendations, or therapeutic options. Example: "What is the best way to treat this condition?"
- Prognosis inquiry (medical_need_prognosis): Requests about the likely course, outcome, or recovery timeline of a disease. Example: "Will this illness resolve on its own?"

4. For each identified medical need (medical_need), extract the associated background information (context) from the user_query.

- Elaboration (context_elaboration): Additional explanations or expansions of the need, such as symptom lists, details, or supplementary descriptions. Example: "I am experiencing dizziness, nausea, and fatigue."
- Background (context_background): Essential contextual information for understanding the need, such as prior treatments or medication history. Example: "I had been taking Deanxit for a year."
- Condition (context_condition): Constraints or preferences related to treatment suggestions (medical_need_treatment), such as subjective willingness or external advice. Example: "I don't want to take any more medication."

Output

You are expected to return a structured JSON object, including all detected needs and related context. A sample user query and expected output are shown below.

user_query:

```
{
  "query_id": "12",
  "symptom_description": "Hello! I previously consulted the gastroenterology department due to gastrointestinal dysfunction, and the doctor prescribed Dailixin for me. When I was taking it, the effect was excellent; however, after taking it for a year, I stopped the medication. Since the doctor did not instruct me to taper off, I simply stopped taking one tablet per day. But ten days after discontinuing, I experienced severe withdrawal symptoms. I managed to endure them for two months with sheer willpower, but eventually, unable to bear it, I sought help from a psychiatrist. The psychiatrist kept prescribing substitute medications to try to alleviate my symptoms, but after trying many different drugs, none proved effective, and all had significant side effects—in fact, I ended up switching medications about every week. The medications I have taken include Laishipu, duloxetine, mirtazapine, minaprine, and tandospirone. It has now been eight months since I stopped taking Dailixin, yet the withdrawal symptoms persist, fluctuating—sometimes better, sometimes worse—with manifestations such as fatigue, muscle soreness, headache and head pressure, abdominal bloating, nausea, and palpitations. May I ask how long these withdrawal symptoms will last? I have met several fellow patients online who are also undergoing Dailixin withdrawal. They all suggest that I resume taking Dailixin and then gradually taper off over about a year. They say that abruptly stopping Dailixin without tapering will not lead to recovery, that no substitute medication can replace it, and that only a slow taper will bring about recovery. However, Dailixin is so addictive that I am now afraid to resume it, fearing I won't be able to quit in the future! Therefore, I hope you can offer some advice on how to treat this. Thank you!",
```

```

"height_weight": "181cm, 53kg",
"diagnosis": "Deanxit withdrawal symptoms",
"duration": "More than 6 months",
"pregnancy": "",
"allergy_history": "None (2021-09-04), allergic to amoxicillin (2022-04-30), None (2023-06-26)",
"medical_history": "No surgeries (2022-04-30)... Chronic disease: Hepatitis B, on antiviral medication since 2009...",
"requested_help": "I would like the doctor's advice: should I restart Deanxit and taper off gradually, or what else should I do?"
}

output:
{
  "query_id": "12",
  "medical_needs": [
    {
      "medical_need_prognosis": "How long will the withdrawal symptoms last?",
      "context_elaboration": "Deanxit withdrawal symptoms keep recurring, with fatigue, muscle pain, headache, nausea, palpitations...",
      "context_background": "It has been eight months since I stopped taking Deanxit, and I'm still experiencing symptoms."
    },
    {
      "medical_need_treatment": "Should I restart Deanxit and taper off gradually, or what else should I do?",
      "context_elaboration": "Still suffering from withdrawal symptoms after eight months...",
      "context_background": "Initially diagnosed with gastrointestinal dysfunction and prescribed Deanxit... After sudden discontinuation, severe withdrawal appeared...",
      "context_condition": "Deanxit is highly addictive and I'm afraid of taking it again, fearing I might not be able to stop next time."
    }
  ],
  "emotional_needs": [
    {
      "emotional_need_view": "Deanxit is too addictive and I'm afraid to take it again",
      "context_cause": "Originally prescribed Deanxit for gastrointestinal issues, stopped abruptly after a year, then experienced severe withdrawal..."
    },
    {
      "emotional_need_feel": "Distress",
      "context_cause": "Severe, recurring, long-lasting withdrawal symptoms from Deanxit."
    }
  ]
}

User query: {user_query}

```

Appendix 5. Prompt for Answer Generation. Note: Original text translated from Chinese into English for clarity.

Role

You are a clinical decision-support assistant. You are skilled at generating structured health consultation responses by following clinical reasoning instead of generic advice.

Task

Given user_query, follow the steps below to generate a coherent, comprehensible, and professional natural language paragraph response:

1. For each emotional state identified in the user_query, generate a response as follows:

- Understanding: Explicitly demonstrate understanding of the user's specific emotional state and, based on the context, show understanding of the reasons behind that emotion.
- Support: Convey signals of companionship, encouragement, or assistance, and emphasize that solutions or coping options exist.

2. For each viewpoint or preference expressed in the user_query, generate a response as follows:

- Understanding: Explicitly demonstrate understanding of the user's specific viewpoint or preference, and based on the context, show understanding of the reasons or experiences that led to it.
- Acceptance: Express non-judgmental respect and supportive acknowledgment of the user's perspective.

3. For each treatment-related request in the user_query, generate a response as follows:

- Condition Appraisal: Evaluate the preferences, opinions, or external suggestions mentioned in the condition field, explaining whether they are reasonable and providing a concise rationale based on clinical guidelines.
- Contextual Integration: Integrate relevant background information and reasonable preferences to provide a guideline-consistent and individualized recommendation.

4. For all other medical informational needs in the user_query, integrate relevant background information and provide individualized, guideline-consistent recommendations.

Note: Guideline refers to the Diagnostic and Treatment Protocols for Mental Disorders (2020 Edition) issued by China's National Health Commission.

Output

You are expected to return a natural text response in fluent, coherent language. A sample user query and corresponding expected output are shown below:

```
user_query:
{
  "query_id": "152",
  "disease_description": "I had a consultation last November for insomnia and anxiety. I took Deanxit as prescribed and it worked very quickly, but the insomnia did not improve. I stopped Deanxit in April this year and tried stopping estazolam for more than ten days, during which I slept relatively well. Later, due to stress from certain events, I restarted estazolam—going from one tablet every two days to one tablet daily—and also went back on Deanxit for a period. I saw several TCM doctors during this time without benefit; the last two said I was depressed due to 'liver qi stagnation'. The medicines they prescribed didn't help; instead I felt uncomfortable, with poor appetite, low mood, and poor sleep. One TCM doctor even prescribed sertraline and lorazepam. I've taken sertraline for about 20 days; I took lorazepam for seven days and then switched back to estazolam. Now one tablet of estazolam only yields three to four hours of sleep, and one tablet of alprazolam doesn't work either. I don't know what to do—what medicine should I take and how should I take it? I have both anxiety and depression. Sometimes my legs and arms feel—well, not numb exactly—more like trembly without actually shaking; sometimes I suddenly feel hot and sweaty. Please help me with treatment; it's been too hard! Also, I tapered off Deanxit slowly without much reaction and I'm thinking of taking it again. I feel sertraline might be worse than it; I'm not sure whether it's effective for depression. Are the two together too much? For the insomnia, which one should I switch to or should I increase the dose?",
  "height_weight": "162 cm, 55.4 kg",
  "conditions": "Insomnia, anxiety, depression",
  "illness_duration": "More than half a year",
  "pregnancy": "",
  "allergies": "None (reported on 2022-09-25)",
  "past_medical_history": "",
  "desired_help": "How can I control my condition?"
}
```

output: "I completely understand that you feel exhausted and helpless after long-term insomnia and repeated medication adjustments. Many patients in similar situations also feel anxious and uneasy—we can work together to explore approaches that may help you more.

You mentioned that Deanxit worked quickly for you in the past, whereas sertraline seems less helpful. This preference based on your own experience is entirely understandable. Individual responses to medication do vary, and clinicians do adjust treatment accordingly. Your feedback is very important in treatment decisions.

From clinical-guideline perspectives, sertraline is a first-line antidepressant and generally needs continuous use for at least 4–6 weeks to take effect gradually. Your current course may be insufficient, so it may be premature to judge it ineffective. Deanxit, although fast-acting, can carry risks of dependence and tolerance with long-term use, so it is usually not the first choice for maintenance treatment. The two drugs are generally not used together long term, to avoid increased side-effect risk.

As for sleep aids, such as estazolam and alprazolam, they can help in the short term, but long-term use tends to lead to tolerance and dependence—the reduced effect you're seeing is typical. International and domestic guidelines recommend time-limited, low-dose use of these medications with gradual tapering. At the same time, non-pharmacological interventions such as sleep hygiene and cognitive behavioural therapy for insomnia are advised to improve sleep quality.

In your case, a more suitable Direction is: continue sertraline in a standardized way and complete an adequate treatment course, then reassess at follow-up whether to switch or adjust the dose; gradually taper the sleep medications to avoid long-term dependence; and combine this with regular routines, relaxation training, and psychological support. With this comprehensive management, sleep and mood usually improve step by step."

User query: {user_query}

AI Usage Disclosure

This thesis was prepared with limited assistance from artificial intelligence (AI) tools solely for language and presentation refinement. AI-based writing tools were used to support grammar correction, formatting adjustments, and minor improvements in clarity and readability.

All intellectual contributions, including the development of research questions, theoretical framing, study design, data collection, analysis, interpretation of results, and conclusions, were undertaken independently by the author. AI tools were not used to generate substantive academic content or research findings.

The author assumes full responsibility for the integrity, originality, and accuracy of all work presented in this thesis.