Industrial Management & Data Systems

2205

Received 11 May 2025
Revised 29 July 2025
11 September 2025
Accepted 19 September 2025

Understanding medical information and emotional support needs in mental health questions with large language models

Chen Liu and William Yu Chung Wang
Waikato Management School, The University of Waikato, Hamilton, New Zealand, and
Gohar Khan
College of Technological Innovation, Zayed University, Dubai, United Arab Emirates

Abstract

Purpose – This study seeks to bridge the gap between users' multidimensional needs and the single-task capabilities of existing Mental Health Question Answering (MHQA) systems by tackling the underexplored challenge of jointly understanding medical informational needs and emotional support needs within complex consumer mental health inquiries.

Design/methodology/approach – Grounded in Rhetorical Structure Theory (RST), the proposed Multi-Needs and Context Recognition (MNCR) framework decomposes mental health question understanding task into four interrelated subtasks: Medical Needs Recognition (MNR), Medical Needs-related Context Extraction (MNCE), Emotional Needs Recognition (ENR) and Emotional Needs-related Context Extraction (ENCE). A new benchmark dataset, MHQ-MedEmo, was constructed through multi-layered semantic annotation of 703 clinical queries sourced from real-world online health consultation platforms. The performances of six base LLMs and two fine-tuned LLMs were evaluated across precision, recall, F1 score and latency metrics.

Findings – Dense, fine-tuned models strike the optimal balance between accuracy and latency for end-to-end MNCR tasks; subtask sensitivity varies markedly across different model architectures; fine-tuning consistently enhances overall performance; the joint-prompt strategy consistently improves both effectiveness and efficiency over the separate-prompt strategy and model architecture and scale significantly influence performance on MNCR subtasks.

Originality/value – This study introduces MNCR and MHQ-MedEmo, the first framework and benchmark for simultaneously understanding medical informational needs and emotional support needs in mental health questions. Comparative evaluation of eight LLMs reveals distinct model-specific strengths, guiding future architectures that balance accuracy and latency and offering concrete guidance for healthcare organizations seeking to deploy LLM-based MHQA solutions in practice.

Keywords Mental health, Rhetorical structure theory, Large language models

Paper type Research article

1. Introduction

Mental disorders afflict an estimated 970 million individuals worldwide, approximately one in eight people, thereby representing a leading contributor to the global burden of disease (World Health Organization, 2022). Despite the availability of evidence-based interventions, more than 75% of affected individuals in low- and middle-income countries receive no formal care, primarily due to persisting stigma, resource constraints, and systemic barriers (World Health Organization, 2022). This supply demand imbalance is further exacerbated by a global

Industrial Management & Data Systems
Vol. 126 No. 7, 2026
pp. 2205-2230
Emerald Publishing Limited
e-ISSN: 1758-5783
p-ISSN: 0263-5577
DOI 10.1108/IMDS-05-2025-0609

shortage of qualified mental health professionals, with psychiatrist-to-population ratios declining markedly in many regions (Weiner, 2022).

Mental health question answering (MHQA) systems, aiming to interpret users' psychological states, symptom descriptions, or treatment needs and to automatically deliver scientifically rigorous yet empathetic responses (Yao *et al*., 2021), have been demonstrated as an effective means of mitigating the chronic supply–demand imbalance in mental health care, expanding access and reducing wait times through automated, on-demand support (Rudd and Beidas, 2020).

However, most existing studies and applications of MHQA concentrate on discrete functionalities—automatic diagnosis, medical recommendation, or standalone psychotherapeutic dialogue—rather than delivering integrated, multi-dimensional support (Cruz-Gonzalez *et al*., 2025; Cilar Budler *et al*., 2023). This compartmentalized paradigm overlooks the complexity of real-world mental health consultations, where users often seek evidence-based medical advice and empathic emotional support in a single interaction. For example, a typical consumer query might request guidance on pharmacotherapy options while expressing feelings of hopelessness, necessitating a unified understanding and response strategy that integrates clinical accuracy with compassionate understanding (see Figure 1).

However, this disconnects between users' real-world needs and the capabilities of current MHQA systems exposes a significant research gap. Closing this gap demands solutions that can simultaneously identify and address both medical informational queries and emotional support needs within a single, cohesive framework. The advent of large language models (LLMs) presents a promising avenue: these models have shown strong potential for delivering evidence-based medical guidance (Roy *et al*., 2024; Singhal *et al*., 2025) alongside compassionate, empathetic support (Zhu *et al*., 2024; Zhang *et al*., 2024). Regrettably, transformer-based AI in MHQA has also predominately been evaluated in narrow domains such as diagnostic accuracy, or emotional support efficacy, without addressing seamless transitions between these components (Cruz-Gonzalez *et al*., 2025).

To this end, we propose a structured approach that decomposes multi-dimensional understanding into four complementary subtasks: Medical Needs Recognition (MNR), Medical Needs-related Context Extraction (MNCE), Emotional Needs Recognition (ENR), and Emotional Needs-related Context Extraction (ENCE). Grounded in Rhetorical Structure Theory (RST, Mann and Thompson, 1988), this decomposition captures both the central user

Question: 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. what medication should I take and how? 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! It's so difficult! 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. 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?

Figure 1. A representative example of a user-submitted mental health question in an online consultation context. Red-highlighted segments reflect expressed emotional support needs, whereas blue-highlighted segments capture the user's informational inquiries regarding medical treatment. Note: Original text translated from Chinese into English for clarity. Source: Authors' own work

intents (nuclei) and their supporting context (satellites), enabling more nuanced understanding of complex mental health questions.

To support research in this area, we introduce MHQ-MedEmo, a novel benchmark corpus consisting of 703 real-world clinical mental health questions annotated with both medical and emotional needs and their contextual spans. We further benchmark three state-of-the-art LLMs—GPT-4o, DeepSeek-V3, and Qwen2.5-Max—under a unified evaluation framework based on semantic similarity-driven partial matching. Our findings demonstrate the feasibility of automatically disentangling and extracting dual-mode support requirements from complex mental health questions, establishing robust baselines for future enhancements in empathetic, clinically informed MHQA systems.

The remainder of this paper is organized as follows: Section 2 reviews related literature; Section 3 presents the proposed framework and tasks; Section 4 describes the dataset construction; Section 5 details the experimental evaluation; and Section 6 discusses the findings and concludes.

2. Literature review

Mental-health question answering (MHQA) has advanced through four key paradigms—rule-based systems, traditional machine-learning classifiers, deep-learning and graph-based models, and large language models (LLMs)—each introducing new capabilities yet still falling short of jointly capturing users' medical information and emotional support needs. To chart these developments and uncover persistent gaps, this review is organized as follows: Section 2.1 traces the historical evolution of MHQA systems; Section 2.2 categorizes methods for question understanding within MHQA; Section 2.3 examines the rise of LLMs and their impact on both comprehension and response generation; and Section 2.4 evaluates existing semantically annotated MHQ datasets. At the conclusion of each section, we identify open challenges that motivate the design of our MNCR framework.

2.1 The evolution of MHQA systems

A typical MHQA system comprises three sequential modules: (1) Question understanding, responsible for question formulation, answer type detection, intent classification, slot/entity extraction, multi-intent detection, and discourse parsing; (2) Knowledge integration, which retrieves and merges evidence from medical databases, guidelines, or commonsense knowledge graphs; and (3) Response generation, which produces final answers via rule-based templates, retrieval mechanisms, or generative models augmented with safety filters and ethical guardrails (Siddals *et al*., 2024).

MHQA architectures have advanced through four epochs: (1) Rule-based Prototypes. Early MHQA system ELIZA employed decomposition-reassembly scripts to emulate a Rogerian therapist (Weizenbaum, 1966), and PARRY simulated paranoid reasoning through belief-based rules (Colby *et al*., 1972). While these systems underscored the necessity of intent representation, they failed to generalize beyond fixed patterns and lacked any true contextual or emotional comprehension. (2) Statistical NLP. The advent of CRF-based intent and slot tagging (Tang *et al*., 2015) and BiLSTM-CRF sequence models (Zhou *et al*., 2016) markedly improved entity recognition and question-type classification. However, these methods could not effectively handle multi-intent queries or capture discourse-level dependencies, limiting their applicability in complex MHQA scenarios. (3) Deep-learning and graph-based models. Neural architectures introduced richer semantic and causal reasoning (Peng *et al*., 2022; Garg *et al*., 2022; Tu *et al*., 2022). Despite their advances, these models typically address either medical or emotional dimensions in isolation, leaving the multi-need challenge unfulfilled. (4) Transformer and LLM Era. The transformer breakthrough (Vaswani *et al*., 2017) enabled pre-trained contextual encoders such as BERT (Devlin *et al*., 2019) and GPT-3 (Brown *et al*., 2020), which have been adapted to MHQA via domain-specific variants like ClinicalBERT for

medical QA and few-shot prompting in emotional support settings. These models excel at generating fluent, empathetic responses but remain largely "answer-first" with limited mechanisms for unified, fine-grained extraction of coexisting medical and emotional intents.

Each phase introduced critical innovations—rule-based clarity, statistical precision, neural semantic depth, and generative fluency—yet none have fully bridged medical-informational and emotional-support needs within a single, discourse-aware process. These enduring gaps form the basis for our MNCR framework, which explicitly integrates nucleus-satellite reasoning and emotional appraisal to achieve truly multi-dimensional MHQA understanding.

Beyond English-only settings, early non-English MHQA systems have begun to appear, for example the Spanish emotional-support agent Sólo Escúchame (Ramírez *et al.*, 2024), the Arabic AraHealthQA shared task built on the MentalQA corpus (Alhuzali *et al.*, 2025), and Chinese systems such as MeChat (Qiu *et al.*, 2023), SoulChat (Chen *et al.*, 2023a, b), and Psychat (Qiu *et al.*, 2024). However, recent reviews emphasize that non-English MHQA still lacks culturally grounded corpora and robust metrics for code-switching, idioms, and localized symptom expression, limiting external validity beyond English. This gap motivates multilingual datasets, modeling frameworks, and evaluation protocols tailored to cross-lingual and cultural phenomena (Guo *et al.*, 2024).

2.2 Question understanding in MHQA

Having surveyed MHQA's architectural evolution, we now focus on its Question Understanding module—a cornerstone for any robust MHQA system (Kilicoglu *et al.*, 2018). Effective MHQA understanding must go beyond simple intent labeling to include domain-specific entity extraction, multi-intent detection, and discourse-level reasoning. We categorize existing methods into three streams.

Rule and statistical-based approaches initially applied handcrafted templates, CRFs for intent detection, and BiLSTM-CRF taggers for slot extraction (Tang *et al.*, 2015; Zhou *et al.*, 2016), enabling reliable entity recognition but lacking sensitivity to context and discourse relations. Graph and knowledge-driven models such as GLHG (Peng *et al.*, 2022) and CAMS (Garg *et al.*, 2022) explicitly model hierarchical and causal links, while MADP (Chen and Liu, 2025) simulates CBT dialogue via multi-agent planning. These methods enrich semantic and causal inference but typically address either medical or emotional dimensions in isolation. Hybrid reasoning and commonsense-augmented frameworks like MISC (Tu *et al.*, 2022) integrate external engines (e.g. COMET) for fine-grained emotion inference, guiding strategy selection. Although these approaches enhance emotional understanding, they still neglect concurrent medical-information needs.

Despite these advances, two key limitations remain. First, most methods are trained on non-clinical corpora (e.g. ESConv, Reddit), which differ markedly from real-world clinical consultations. Second, by focusing on either emotional or causal facets, they overlook the multifaceted nature of consumer mental health queries, which often bundle medical, emotional, and social support needs. These gaps underscore the necessity of our MNCR framework, which jointly extracts both what users need and why they need it.

2.3 LLMs in MHQA

Building on the non-LLM methods surveyed above, recent research has increasingly leveraged large pre-trained language models to enhance MHQA—yet primarily for generating fluent, empathetic responses rather than for deeper query comprehension. Early LLM adaptations such as Psy-LLM (Lai *et al.*, 2023) and GPT-3 few-shot prompting demonstrated that minimal tuning could produce coherent counseling dialogue, but evaluations focused almost exclusively on usefulness, coherence, and empathy, leaving intent parsing unmeasured. Safety and quality-oriented platforms, for example, ChatCounselor (Liu *et al.*, 2023) and the stage-aware SuDoSys chatbot (Chen *et al.*, 2024), introduced WHO-aligned templates and

content filters to mitigate harmful outputs, yet success remained tied to human satisfaction ratings rather than systematic understanding metrics.

Emerging comprehension-focused methods include intent-aware prompting (Ma *et al.*, 2025), which classifies conversational goals before generation; and retrieval-augmented pipelines with slot-filling for symptoms and medications (Lahiri and Hu, 2024). Despite these advances, reviews still characterize MHQA LLM use as "answer-first" (Guo *et al.*, 2024), highlighting an urgent need for unified, fine-grained benchmarks and datasets that evaluate multi-dimensional intent parsing. These gaps underscore the importance of developing semantically annotated resources to measure and improve LLMs' question understanding capabilities.

2.4 Semantically annotated MHQ datasets

Semantically annotated MHQA datasets provide critical labels on question attributes—such as decomposition structures, focus entities, question categories, and emotional states—that are essential for developing and evaluating question understanding modules (e.g. type classifiers, focus recognizers, emotional intent detectors) (Kilicoglu *et al.*, 2018). We identify two main categories of such datasets that closely align with our research objectives.

The first category comprises MHQA-specific, semantically annotated corpora. In the medical information support subdomain, MentalQA consists of 500 Arabic patient–doctor exchanges annotated across six question categories (diagnosis, treatment, anatomy and physiology, epidemiology, healthy lifestyle, provider choice) and three answer strategies (information provision, direct guidance, emotional support) (Alhuzali *et al.*, 2024). MHQA offers a multiple-choice benchmark of 2,475 expert-verified gold instances and 56.1 K pseudo-labeled QA pairs extracted from PubMed abstracts, spanning anxiety, depression, trauma, and obsessive-compulsive domains across four question types (factoid, diagnostic, prognostic, preventive) (Racha *et al.*, 2025). In the emotional support domain, PsyQA includes 22K Chinese counseling questions paired with 56K long-form responses, each annotated for support strategies and emotional cues (Sun *et al.*, 2021). ESConv contains 1,300 multi-turn help-seeker/supporter dialogues with pre-chat emotion and situation labels, as well as turn-level annotations of eight emotional support strategies (Liu *et al.*, 2021). CAMS annotates 3,155 Reddit posts with causal interpretation spans and categories to capture underlying reasons for mental health issues (Garg *et al.*, 2022). ExTES leverages recursive LLM generation to produce 11,177 complete emotional support dialogues, each annotated by scenario and strategy (Zheng *et al.*, 2023). Finally, EHDChat introduces 33,303 doctor–patient conversations grounded in verified medical knowledge, annotated with both informational and empathetic strategies (Wu *et al.*, 2024).

The second category encompasses semantically annotated datasets for complex consumer health questions, which—although not exclusively focused on mental health—offer valuable methodologies for modeling and annotating intricate question structures. The GARD dataset pioneers question decomposition by segmenting 1,467 consumer queries into 2,937 subquestions, each labeled with one of 13 question types and associated with focus diseases (Roberts *et al.*, 2014). The CHQA-email corpus extends semantic annotation to 1,740 NLM customer-service emails, labeling named entities, question focus, category, type, and trigger terms within structured question frames (Kilicoglu *et al.*, 2018). CHQ-SocioEmo is the first dataset to introduce non-informational labels for consumer health questions, covering 1,500 community posts annotated for basic emotion categories and social support need types (Alasmari *et al.*, 2023).

This overview highlights that most existing MHQA dataset annotation efforts remain predominantly answer-centric; question labels are often too coarse to fully and accurately characterize user states and intents. In Table 1, we present a comparative analysis of our proposed MHQ-MedEmo dataset against existing annotated consumer health question (CHQ) datasets.

Table 1. Comparison between MHQ-MedEmo and other CHQ dataset in terms of data source and annotation objects

Dataset	Data source	Lang	Annotation objects				
			MN	MNC	EN	ENC	ANS
GARD (2014)	Health website	EN	✓	✓	✓		
CHQA-email (2018)	Health website	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					✓
MHQ-MedEmo(Ours)	*Online health consultation platform*	*CH*	✓	✓	✓	✓	

Note(s): 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)

Source(s): Authors' own work

3. Framework and task design for multi-dimensional MHQ understanding

Instead of merely listing isolated labels, our aim is to project the full communicative intent of a mental-health query onto a theory-driven discourse scaffold. We therefore ground the Multi-Needs and Context Recognition (MNCR) framework in two complementary theories.

3.1 Rhetorical Structure Theory (RST)

Introduced by Mann and Thompson (1988), RST explains textual coherence by modeling functional links between spans of discourse. Each link pairs a nucleus—the clause that realizes the writer's primary communicative goal—with one or more satellites that elaborate, justify, condition, or otherwise support that nucleus. Because RST relations are domain-agnostic yet semantically labeled (e.g. Cause, Condition, Background), they provide a principled way to represent why an utterance is made, what information is central, and which clauses are merely supporting. In clinical dialogue, the nucleus–satellite distinction maps naturally onto a patient's core question (e.g. "What treatment should I take?") and the contextual details that modulate the answer (symptom history, constraints, personal preferences). By leveraging RST, our framework treats medical and emotional requests as discourse nuclei and extracts their satellites to capture the reasoning context essential for safe guidance.

3.2 Appraisal theory of clinical empathy

Building on Systemic Functional Linguistics, Pounds (2011) extends appraisal theory to healthcare communication, arguing that clinicians must respond differently to Feelings (direct expressions of emotion such as fear, sadness, or anxiety) and to Evaluative Viewpoints (attitudinal stances such as criticism, skepticism, or hopelessness). Feelings call for acknowledgment and emotional validation, whereas viewpoints require cognitive reframing or informational reassurance. Distinguishing these two emotional meanings is therefore critical for generating context-appropriate empathic responses. By importing this dichotomy into ENR, we ensure that the model does not merely tag "emotion words" but recognizes the type of emotional need, enabling downstream systems to tailor counseling tone and strategy.

Together, RST provides the structural backbone (nucleus vs. satellite), while appraisal theory supplies the emotional taxonomy (feeling vs. viewpoint). Their integration allows

MNCR to capture both the informational and affective dimensions of mental-health queries within a single, theoretically principled representation. By aligning these two theories, we derive four inter-dependent subtasks that together capture both the medical and emotional nuclei of a question and their corresponding satellites (Figure 2).

3.3 Task 1: Medical Needs Recognition (MNR)

In RST terms, MNR identifies the nucleus of a mental-health query: the patient's explicit informational request. 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 clear understanding of the patient's evidence-based medical information needs.

3.4 Task 2: 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.

3.5 Task 3: Emotional Needs Recognition (ENR)

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

3.6 Task 4: Emotional Needs-related Context Extraction (ENCE)

For each emotional need, ENCE extracts causal satellites (cause) that describe the underlying triggers for the expressed emotions. Incorporating such causal information has been shown to enhance empathetic response generation (Li *et al.*, 2021; Wang *et al.*, 2021).

4. Benchmark dataset description

4.1 Data collection and preparation

The mental health questions utilized in this study were sourced from haodf.com, one of China's earliest and most reputable online healthcare platforms. To build our corpus, we initially identified 1,639 psychiatry-related user queries from MedDialog's publicly available records (Zeng *et al.*, 2020). To capture more recent consultation styles, we subsequently developed a Python-based web crawler to randomly extract an additional 500 queries from the platform's 2024 mental health consultation logs. Each query entry comprises ten fields: query

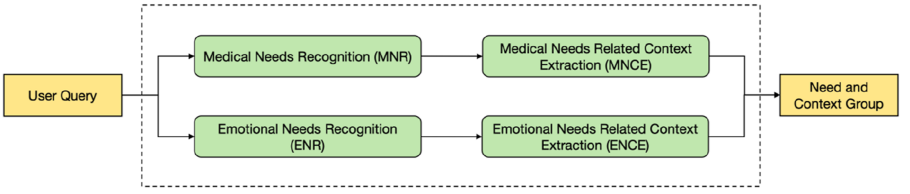

Figure 2. MNCR framework and tasks designed for mental health question understanding. Source: Authors' own work

ID, symptom description, height and weight, diagnosis, duration of illness, pregnancy status, pregnancy history, allergy history, past medical history, and the patient's stated help request.

From the combined total of 2,139 cases, we applied a semi-automated screening procedure—integrating manual review with Python scripts—to select 703 high-quality records that met all the following inclusion criteria: (1) Emotional content: the query expresses at least one explicit emotional cue (e.g. fear, despair, or uncertainty). (2) Text-only format: both question and answer are presented purely as structured text, with no multimedia elements. (3) Clarity and completeness: the patient's inquiry is self-contained and unambiguous. (4) Non-follow-up case: the query represents an initial consultation rather than a revisitation.

4.2 Data annotation schema

Building on the RST framework introduced in Section 3, we further elaborate on how RST principles inform the design of our annotation schema. In the context of consumer health questions, each user need is treated as a nucleus and is further categorized into two domains: medical information needs and emotional support needs.

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

(1) General medical information (*M-N*-GMI). Request 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?").

(2) Etiology (*M-N*-ETI). Request the cause or origin of a specific symptom or disease (e.g. "What causes my headache?").

(3) Diagnosis (*M-N*-DIA). Request diagnostic clarification refers to specific symptoms (e.g. "What the hell is wrong with me?").

(4) Treatment (*M-N*-TREAT). Request specific treatment advice, medication recommendations, or therapy options (e.g. "How can I treat this condition effectively?").

(5) Prognosis (*M-N*-PROG). Ask about the development trend of the disease, possible consequences, recovery cycle, etc. (e.g. "Will the disease heal on its own?").

4.2.2 Medical needs related context (M-C). For each recognized medical need, annotate zero, one or more related context segments as satellites. Each satellite is explicitly linked to the nucleus via a rhetorical relation label that captures its discourse function, such as:

(1) Elaboration (M-C-ELA), where further detail or explanation is added (e.g. "I am currently experiencing dizziness, nausea, and fatigue").

(2) Background (M-C-BACK), which provides necessary information for understanding the user's current concern (e.g. "I have been taking Duloxetine (Cymbalta) for one year").

(3) Condition (M-C-CON), which specifies the circumstances under which the need is applicable (e.g. "I no longer wish to continue medication").

4.2.3 Emotional support needs (E-N). Emotional need categories are derived from Pounds' appraisal framework for language-based clinical empathy:

(1) 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!").

(2) 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").

4.2.4 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:

(1) Cause (E-C-CAUSE), which offers a reason or trigger for this 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").

This approach enables us to represent the semantic dependency between a user's medical or emotional intent and its contextual foundation, thus moving beyond flat intent classification to a more structured representation of user concerns. It also supports the evaluation of LLMs not only on whether they can identify the user's explicit needs, but whether they correctly link those needs to the relevant supporting information—an essential capability for accurate and empathetic health question answering.

4.3 Data annotation and revision

We recruited four Chinese postgraduate students specializing in clinical psychiatry to participate in the annotation and revision of the corpus. Each user query was independently annotated by two annotators. In cases of inconsistency, a third annotator was assigned to adjudicate and finalize the labels.

To ensure consistency and high-quality annotation, we 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 five queries to familiarize themselves with the task specifications.

Annotation tasks were conducted using Label Studio, an open-source data-labeling platform that supports custom text-span and relation annotations, audit trails, and JSON format exports.

4.4 Inter-annotator agreement analysis

We adopted a semi-automated, manual adjudication protocol to quantify inter-annotator agreement (IAA) for our 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 1) their labels were identical and 2) 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 0.83, 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 need_type labels were identical and their need_text spans satisfied the 50% rule. At the context level. Within each matched need, supporting context_text spans were compared under the same rule, contingent on identical relation_ type labels.

For each of the 11 (label + span) units, we computed overall agreement. We then produced frequency-weighted macro averages across four dimensions, medical needs, medical needs related context, emotional needs, and emotional needs related context relations.

As shown in Table 2, for the annotations of medical needs and related context unites, the overall agreement is 80.17 and 80.95%, respectively. For the annotations of emotional needs and related context unites, the overall agreement is 78.00 and 85.45%. 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.

4.5 Dataset statistics and analysis

The benchmark dataset comprises 703 clinical mental health questions—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%).

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 3). 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 2. 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)	

Source(s): Authors' own work

Table 3. Statistics of MHQ-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

Source(s): Authors' own work

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, we 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, we 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, we 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 3). Fear and sadness emerged as the dominant emotional categories, underscoring the emotional burden frequently expressed in consumer mental health questions.

To better understand the subjective perspectives expressed in the E-N-VIEW subtype of emotional needs, we 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, we observed that GPT-4o frequently misclassified conceptually specific instances into the residual Other category. To mitigate this issue and improve thematic reliability, we performed focused reannotation of all entries initially assigned to Other. As a result, the corpus was organized into six distinct themes (see Figure 4): (1) Opinions on treatment approaches: Reflections on medical interventions, medications, or treatment plans (249 cases). (2) Opinions on disease etiology: Subjective interpretations regarding the causes or perceived triggers of the condition (23). (3) Opinions on diagnostic results: Judgments about the accuracy, credibility, or meaning of diagnostic outcomes (31). (4) Opinions on disease prognosis: Views related to expectations or concerns about the future course of the illness (9). (5) Cognitive expressions: Statements reflecting patients' beliefs, reasoning patterns, or interpretations about themselves, others, or the broader world (104). (6) Other: Expressions that could not be clearly assigned to the above categories (92).

5. Experiment

5.1 Task formalization

Let $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{iL_i}\}$ be a consumer mental-health question of length L_i. For each Q_i, our goal is to jointly understand both its medical informational needs and emotional support needs,

Figure 3. 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. Source: Authors' own work

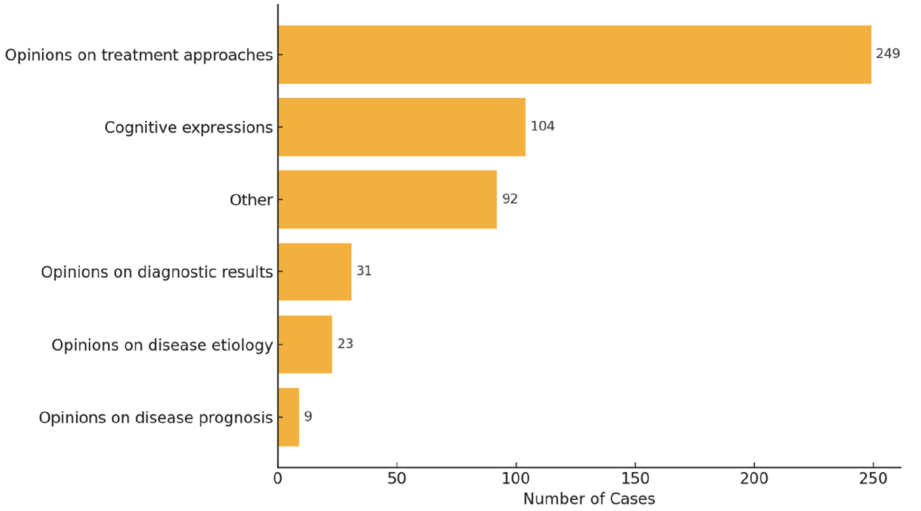

Figure 4. Thematic distribution of subjective perspectives in E-N-VIEW entries. Source: Authors' own work

along with their associated contextual information. We decompose this Multi-Needs and Context Recognition (MNCR) task into four subtasks:

5.1.1 Task 1: Medical Needs Recognition (MNR). Identify all medical-need spans $m_{ij} \subset Q_i$, and assign each span to one of five identified categories $c_{ij} \in$ {general medical information, etiology, diagnosis, treatment, prognosis}, according to their request medical information support category.

5.1.2 Task 2: Medical Needs related Context Extraction (MNCE). For each pair (m_{ij}, c_{ij}), extract every related context span $s_{ik} \subset Q_i$, and label it with a identified relation $r_{ik} \in$ {elaboration, background, condition}, where condition is applied only if c_{ij} = treatment.

5.1.3 Task 3: Emotional Needs Recognition (ENR). Identify all emotional-need span $e_{il} \subset Q_i$, and assign each to one of two categories $d_{il} \in$ {view, feeling}, corresponding to viewpoint expressions or feeling expressions.

5.1.4 Task 4: Emotional Needs related Context Extraction (ENCE). For each pair (e_{il}, d_{il}), extract every related context span $f_{ip} \subset Q_i$, and label it uniformly as cause.

5.2 Experimental settings

To ensure effective training and robust evaluation, we stratified the MHQ-MedEmo dataset by year and by task label, then split it into 60% training (422 samples), 20% validation (140 samples), and 20% testing (141 samples). This partition preserves temporal diversity and maintains balanced label distributions across all four subtasks (MNR, MNCE, ENR, ENCE). We then conducted a series of experiments on the MHQ-MedEmo test set to assess various LLMs' ability to jointly identify medical informational needs, emotional support needs, and their associated contextual information in mental-health queries.

5.2.1 Models. We selected six base LLMs not only for their superior performance in multi-task language understanding scenarios but also to ensure coverage of diverse architectures, parameter scales, availability tiers, and model categories (see Table 4). All models were accessed via their respective APIs, ensuring a fair comparison under consistent experimental conditions. The evaluated models are as follows:

GPT-4o (OpenAI, 2024a, b)*:* As OpenAI's flagship multimodal model, GPT-4o is capable of processing text, audio, and visual inputs in real-time. It achieved an impressive score of 88.7% on the Massive Multitask Language Understanding (MMLU) benchmark, reflecting robust general knowledge and reasoning capabilities across diverse domains.

DeepSeek-V3 (DeepSeek, 2025b)*:* 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, 2025a, b)*:* 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.

Table 4. Base LLMs adopted in this study

	Qwen2.5–72B	Qwen2.5-MAX	Qwen3-235B-A22 B	DeepSeek-V3	DeepSeek-R1	GPT-4o
# Architecture	Dense	MoE	MoE	MoE	Dense	Dense, Multimodal
# Total Params	72B	325 B	235B	671 B	260B	~1.8 T (estimated)
# Activated Params	72B	22 B	22B	37 B	260B	~1.8 T (estimated)
# Availability	Open Source	Open Source	Open Source	Open Source	Open Source	Closed Source
# Model Category	General	General	Reasoning	General	Reasoning	General

Note(s): MoE is short for mixture-of-expert

Source(s): Authors' own work

Qwen3-235B-A22 B (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 Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). It demonstrated leading performance on multiple benchmarks, including Arena-Hard, LiveBench, LiveCodeBench, and GPQA-Diamond.

Qwen2.5-72B (Yang *et al*., 2024)*:* Qwen2.5–72B is a dense transformer model with 72 billion parameters, trained on more than 15 trillion tokens. It is optimized for balanced performance and efficiency, supporting various downstream tasks such as code generation, instruction following, and multilingual QA. It serves as a strong open-source baseline in the Qwen2.5 family.

5.2.2 Prompt design. To move beyond treating each subtask in isolation, we embed RST's nucleus–satellite relations and appraisal theory into a single structured prompt (see Figure 5). 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 model performance, we incorporate two widely adopted prompting strategies:

 # 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. The illustration of the joint prompt design with CoT and few-shot prompt strategies. Source: Authors' own work

(1) *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: (1) 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 . . ."; (2) 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." (3) 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.

(2) *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 combination of CoT and few-shot prompting ensures token-level clarity, consistency across subtasks, and supports robust, end-to-end semantic parsing in complex mental health queries.

To demonstrate the benefit of the unified prompt design, we also conducted a comparative ablation experiment against a baseline that invokes four separate prompts for the four sub-tasks.

5.2.3 Joint fine-tuning. Although DeepSeek-V3 (685 B parameters) and Qwen2.5-Max (325 B parameters) are open-source and theoretically support parameter-efficient fine-tuning, their sheer scale renders direct adaptation impractical under typical academic or modest cloud budgets. To evaluate domain adaptation without compromising architectural consistency, we instead applied QLoRA-based fine-tuning to Qwen2.5–72B, and instruction-tuning to GPT-4o via OpenAI's Fine-tuning API.

5.2.3.1 GPT-4o fine-tuning. We converted 422 training and 140 validation examples from raw JSON into the JSONL format required by the Chat Completions schema. All hyperparameters—batch_size, learning_rate_multiplier, and num_epochs—were left at OpenAI's proprietary default settings ("Auto"), which are not publicly documented. The job completed successfully in approximately 51 min 26 s.

5.2.3.2 Qwen2.5-72B fine-tuning. The same MNCR training and validation splits were converted to JSONL per the ChatML schema. We froze the base model's weights and injected trainable low-rank adapter modules into each transformer layer following the QLoRA protocol. We set the LoRA rank to 8, α to 16, batch_size to 16, and learning_rate to 3e−4. Training used Alibaba Cloud Bairuan's High-Efficiency SFT profile and finished successfully in about 31 min 40 s.

5.3 Evaluation method

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, we adopt 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: (1) it belongs to the same query_id; (2) its parent need unit has been successfully matched; (3) it has the same relation_type label; and (4) 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*., 2023a, b).

In our implementation, we utilize OpenAI's text-embedding-3-large model to generate high-dimensional (up to 3,072 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, 2024a, b).

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, we adhere to standard information retrieval metrics as outlined by Manning *et al.* (2008). We define 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 = TP/(TP + FP)

Recall = TP/(TP + FN)

F1-score = 2 * (Precision * 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.

Beyond isolated evaluations, we further introduce a Joint Exact-Match (JEM) metric to assess holistic performance across all four sub-tasks. Specifically, for a given query, a prediction is counted as correct only if the system simultaneously produces a correct MNR, ENR, MNCE, and ENCE output. This stringent metric directly reflects the system's ability to deliver fully consistent and integrated multi-dimensional understanding. JEM is calculated as follows:

$$\mathrm{EM}_t(q) = \begin{cases} 1, \textit{if } \mathrm{FP}_t(q) = 0 \textit{ and } \mathrm{FN}_t(q) = 0 \\ 0, \textit{otherwise} \end{cases} \quad (t \in \{\mathrm{MNR}, \mathrm{ENR}, \mathrm{MNCE}, \mathrm{ENCE}\})$$

$$\mathrm{JEM}(q) = \prod_t \mathrm{EM}_t(q) \ (q \text{ is a single query instance})$$

$$\mathrm{JEM} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{JEM}(q) \ (Q \text{ is the set of all test queries; } |Q| \text{ is its size})$$

Considering the practical application, we 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's built-in timing hooks.

5.4 Experimental results

Figure 5 plots overall F1-score against average processing time per query for all eight model variants, while Tables 5 and 6 summarize precision (P), recall (R), and F1-scores for each subtask under both prompt-based and fine-tuning paradigms. Several clear patterns emerge:

5.4.1 Overall performance vs. latency trade-off. As shown in Figure 6, 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-A22 B ($F1 = 0.651$, 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.

5.4.2 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 behind (Qwen3-235B, 0.802; Qwen2.5-MAX, 0.771; DeepSeek-V3, 0.671). 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.

5.4.3 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.

5.4.4 Impact of prompt strategy. Across all six base models, the joint-prompt (JP) strategy consistently improves both effectiveness and efficiency over separate-prompt (SP) (see Table 7). These trends indicate that a single-pass joint prompt not only reduces inference cost, but also enhances cross-task consistency-most notably for emotional needs and their related contexts-thereby improving end-to-end, joint exact-match performance.

5.4.5 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

Table 5. Performance results of models (Qwen2.5–72B, Qwen2.5-72B-finetuning, Qwen2.5-max and Qwen3-235B)

	Models											
	Qwen2.5–72B base			Qwen2.5-72B-finetuning			Qwen2.5-max base			Qwen3-235B-A22 B base		
Tasks	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	0.589
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*

Source(s): Authors' own work

Table 6. Performance results of models (GPT-4o, GPT-4o-finetuning, DeepSeek-V3 and DeepSeek-R1)

	Models											
	GPT-4o Base			GPT-4o-finetuning			DeepSeek-V3 base			DeepSeek-R1 base		
Tasks	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	0.828	0.746	0.707	0.726
ETI	0.385	0.500	0.435	0.471	0.800	0.593	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	0.621	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	0.400
TREAT	0.876	0.698	0.777	0.833	0.785	0.808	0.918	0.779	0.843	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	0.474	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	0.726
CON	0.867	0.289	0.433	0.550	0.468	0.506	0.537	0.449	0.489	0.574	0.700	0.631
ELA	0.779	0.815	0.796	0.856	0.836	0.846	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*

Source(s): Authors' own work

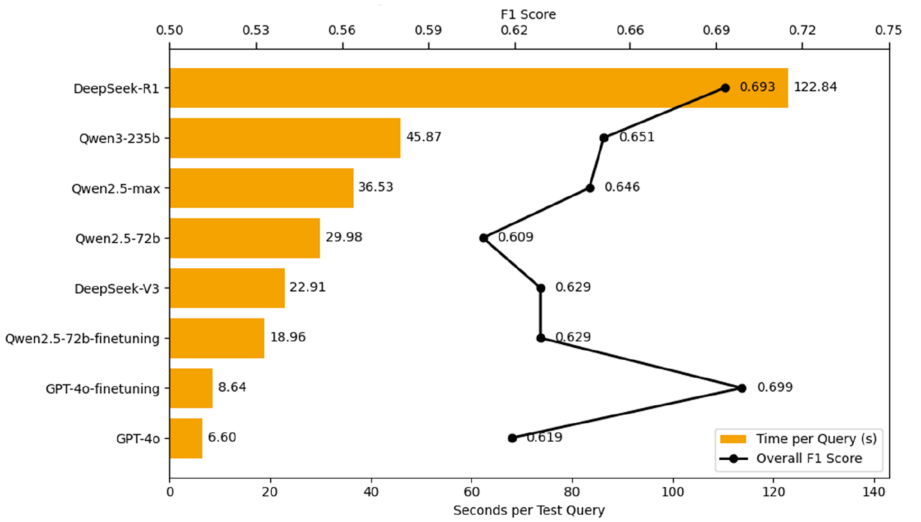

Figure 6. Overall F1 score and processing time per test query across 8 LLMs. Source: Authors' own work

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 stems 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.

6. Discussion and conclusion

6.1 Theoretical implications

Our findings illuminate several key theoretical contributions. First, grounded in Rhetorical Structure Theory (Mann and Thompson, 1988), we introduce the MNCR framework, which decomposes each consumer mental health query into four interrelated subtasks: MNR, MNCE, ENR, and ENCE. This systematic decomposition operationalizes both core user intents (nuclei) and their supporting context (satellites) within a unified theoretical model, offering a principled basis for future consumer health questions understanding architecture design.

Second, we developed MHQ-MedEmo, the first large-scale, multi-layer annotated corpus of 703 real-world clinical mental health questions collected from online consultation platforms. By annotating both medical information needs and emotional support needs along with their exact contextual spans, MHQ-MedEmo fills a critical resource gap and enables rigorous, reproducible evaluation of dual-mode need recognition. Moreover, this dataset has been shown to effectively adapt LLMs to the MNCR tasks through fine-tuning.

Finally, our extended analysis of experiment results reals that: (1) Observed variability, where MoE models excel on sparse inputs yet underperform on multi-intent queries, aligns with recent findings on routing inefficiencies in traditional MoE architectures, suggesting a theoretical imperative to refine expert selection mechanisms; (2) The substantial gains from

Table 7. Performance results of six base models with different prompt strategies

	Models											
	GPT-4o Base		Qwen2.5–72B base		Qwen2.5-max base		Qwen3-235B-A22 B base		DeepSeek-V3 base		DeepSeek-R1 base	
Performance	SP	JP	SP	JP	SP	JP	SP	JP	SP	JP	SP	JP
L	11.78	6.60↓	42.56	29.98↓	59.58	36.53↓	77.21	45.87↓	36.20	22.91↓	186.43	122.84↓
JEM	0.220	0.355↑	0.184	0.291↑	0.206	0.340↑	0.241	0.397↑	0.220	0.362↑	0.262	0.433↑
F1(MNR)	0.705	0.709↑	0.645	0.686↑	0.651	0.682↑	0.635	0.651↑	0.774	0.790↑	0.753	0.762↑
F1(ENR)	0.300	0.412↑	0.304	0.418↑	0.356	0.445↑	0.410	0.507↑	0.337	0.385↑	0.386	0.462↑
F1(MNCX)	0.582	0.654↑	0.598	0.646↑	0.643	0.710↑	0.676	0.690↑	0.622	0.656↑	0.701	0.737↑
F1(ENCX)	0.476	0.712↑	0.515	0.718↑	0.560	0.771↑	0.640	0.802↑	0.591	0.671↑	0.720	0.790↑

Note(s): L stands for latency (time per query, s). JEM stands for Joint Exact-Match. JP stands for Joint Prompt strategy. SP stands for Separate Prompt strategy

Source(s): Authors' own work

fine-tuning, especially on emotionally complex subtasks, reinforce the theoretical advantage of parameter-efficient adaptation in domain-specific contexts; (3) Our findings that smaller, well-aligned models (e.g. Qwen2.5-72B-finetuned) match or exceed larger open-source counterparts highlight a theoretical shift: alignment and prompt engineering may be more pivotal for generalization in multi-dimensional mental health question understanding than sheer parameter scale.

6.2 Managerial implications

From an implementation perspective, our findings offer concrete guidance for healthcare organizations seeking to deploy LLM-based MHQA solutions in practice. These organizations must carefully align model choice, data curation, and real-time monitoring to ensure both clinical accuracy and empathetic user engagement.

Model selection criteria. Dense, LoRA-fine-tuned variants—such as GPT-4o-finetuning and Qwen2.5-72B-finetuning—provide the most coherent balance of high accuracy and low latency, making them the preferred choice for real-time MHQA applications. In contrast, MoE models (e.g. DeepSeek-R1, Qwen2.5-MAX) excel in context-heavy subtasks like Medical Context Extraction (MNCX) but incur higher inference costs and routing-induced variability on multi-dimensional queries.

Data curation and targeted fine-tuning. Effective deployment also hinges on robust data curation and training: augmenting corpora with richly annotated emotional utterances—particularly in underrepresented domains like etiology and prognosis—and applying targeted fine-tuning have demonstrated marked improvements in model sensitivity and specificity.

Operational guardrails and governance. Healthcare organizations must establish guardrails, tracking performance drift, user safety metrics, and regulatory compliance thresholds, to detect degradation and manage risk. Transparent reporting dashboards, role-based access controls for retraining triggers, and routine audits of ensemble weights reinforce governance and stakeholder trust in AI-driven mental health support systems.

6.3 Limitations and future research

This study has several limitations that highlight promising directions for future research. First, although our baselines are encouraging, all evaluated models still underperform on emotional-demand recognition, underscoring the difficulty of capturing nuanced affective cues. Moreover, due to the relatively small number of samples in the fine-tuning and test datasets, performance on etiology and prognosis recognition is also suboptimal. Future work should therefore explore advanced affective-computing techniques, such as sentiment-aware pretraining or explicit empathy modeling, and construct dedicated fine-tuning datasets for etiology and prognosis to improve emotional-intent detection. Second, certain content categories (notably etiology and prognosis) remain challenging. Developing specialized annotation schemas or employing curriculum-learning strategies could help models better grasp the linguistic subtleties in these domains. Third, our evaluation is limited to a single Chinese benchmark (MHQ-MedEmo), six base and two fine-tuned LLMs. Extending validation to additional datasets, languages, and more emerging open-source architectures will be crucial for assessing generalizability. Fourth, although the fine-tuned GPT-4o variant delivers the best trade-off between accuracy and latency for our multi-dimensional consumer question understanding task, practical deployment must also consider both cost and access restrictions, such as the inability to connect to international LLM APIs from mainland China (OpenAI, 2025). Finally, future research should further align LLM-based MHQA system design and validation with emerging regulatory frameworks. For example, under the EU AI Act (2024), conversational assistants intended to inform diagnosis or treatment will typically be treated as high-risk AI, implying requirements for risk management, data governance, transparency, human oversight, logging, and post-market monitoring. In the U.S., the FDA's Software as a Medical Device (SaMD) approach (2021) is risk-based: chatbots that make

medical claims require clinical evaluation (including human-factors validation and real-world performance evidence), whereas general-wellness tools without diagnostic claims may sit outside device oversight.

Ethical considerations

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—we 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.

References

Act, E.A.I. (2024), *The Eu Artificial Intelligence Act*, European Union, Brussels.

Alasmari, A., Kudryashov, L., Yadav, S., Lee, H. and Demner-Fushman, D. (2023), "CHQ-SocioEmo: identifying social and emotional support needs in consumer-health questions", *Scientific Data*, Vol. 10 No. 1, p. 329, doi: 10.1038/s41597-023-02203-1.

Alhuzali, H., Alasmari, A. and Alsaleh, H. (2024), "MentalQA: an annotated Arabic corpus for questions and answers of mental healthcare", doi: 10.48550/arXiv.2405.12619.

Alhuzali, H., Shamout, F.E., Abdul-Mageed, M., Abouzahir, C., Abu Daoud, M., Alasmari, A., Al-Eisawi, W., Al-Monef, R., Alqahtani, A., Ayash, L., Habash, N. and Kharouf, L. (2025, November), "AraHealthQA 2025: The first shared task on Arabic health question answering", in *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 107-118.

Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999), "Beyond kappa: a review of interrater agreement measures", *Canadian Journal of Statistics*, Vol. 27 No. 1, pp. 3-23, doi: 10.2307/3315487.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... and Amodei, D. (2020), "Language models are few-shot learners", doi: 10.48550/arXiv.2005.14165.

Chen, Q. and Liu, D. (2025), "MADP: multi-agent deductive planning for enhanced cognitive-behavioral MHQA", doi: 10.48550/arXiv.2501.15826.

Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q. and Xu, X. (2023a), "Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations", Association for Computational Linguistics (ACL), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1170-1183.

Chen, Y., Lu, X., An, J. and Duan, H. (2023b), "Automatic ICD-10 coding: deep semantic matching based on analogical reasoning", *Scientific Reports*, Vol. 13, 38761.

Chen, Y., Zhang, X., Wang, J., Xie, X., Yan, N., Chen, H. and Wang, L. (2024, September), "Structured dialogue system for mental health: an LLM chatbot leveraging the PM+ guidelines", in *International Conference on Social Robotics*, pp. 262-271, Springer Nature Singapore, Singapore.

Cilar Budler, A., Budler, B. and Wang, X. (2023), "Conversational agents for depression screening: a systematic review", *Journal of Affective Disorders*, Vol. 325, pp. 1-10.

Colby, K.M., Weber, S., Hilf, F.D., Fleischman, G.M. and Barr, H.C. (1972), "PARRY—A computer program to simulate paranoid processes", *Proceedings of the Second International Joint Conference on Artificial Intelligence (IJCAI)*, p. 333.

Cruz-Gonzalez, P., He, A.W.J., Lam, E.P., Ng, I.M.C., Li, M.W., Hou, R., Chan, J.N.M., Sahni, Y., Vinas Guasch, N., Miller, T., Lau, B.W.M., Sánchez Vidaña, D.I. and Vidaña, D.I.S. (2025),

"Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications", *Psychological Medicine*, Vol. 55, e18, doi: 10.1017/s0033291724003295.

DeepSeek AI. (2025a), "DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [preprint]", *arXiv*, doi: 10.48550/arXiv.2501.12948.

DeepSeek AI. (2025b), "DeepSeek-V3 technical report", [Computer software], *Hugging Face*, available at: https://huggingface.co/deepseek-ai/DeepSeek-V3

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), "BERT: pre-training of deep bidirectional transformers for language understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Association for Computational Linguistics, pp. 4171-4186.

Garg, M., Saxena, C., Saha, S., Krishnan, V., Joshi, R. and Mago, V. (2022), "CAMS: an annotated corpus for causal analysis of mental health issues in social media posts", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, ELRA, pp. 6387-6396.

Guo, Z., Lai, A., Thygesen, J.H., Farrington, J., Keen, T. and Li, K. (2024), "Large language models for mental health applications: systematic review", *JMIR Mental Health*, Vol. 11 No. 1, e57400, doi: 10.2196/57400.

Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S.E., Rodriguez, L., Masterton, K. and Demner-Fushman, D. (2018), "Semantic annotation of consumer health questions", *BMC Bioinformatics*, Vol. 19 No. 1, p. 34, doi: 10.1186/s12859-018-2045-1.

Lahiri, A.K. and Hu, Q.V. (2024), "Alzheimerrag: Multimodal retrieval augmented generation for PubMed articles", *arXiv preprint, arXiv:2412.16701*.

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y. and Wang, Z. (2023), "Psy-LLM: scaling up global mental health psychological services with AI-based large language models", doi: 10.48550/arXiv.2307.11991.

Li, Y., Li, K., Ning, H., Xia, X., Guo, Y., Wei, C., Cui, J. and 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*, ACM, pp. 1301-1310.

Liu, J.M., Li, D., Cao, H., Ren, T., Liao, Z. and Wu, J. (2023), "Chatcounselor: a large language models for mental health support", *arXiv preprint, arXiv:2309.15461*, doi: 10.48550/arXiv.2309.15461.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y. and Huang, M. (2021), "Towards emotional support dialog systems", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 3608-3619.

Ma, J., Na, H., Wang, Z., Hua, Y., Liu, Y., Wang, W. and Chen, L. (2025, January), "Detecting conversational mental manipulation with intent-aware prompting", in *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9176-9183.

Mann, W.C. and Thompson, S.A. (1988), "Rhetorical structure theory: toward a functional theory of text organization", *Text and Talk*, Vol. 8 No. 3, pp. 243-281, doi: 10.1515/text.1.1988.8.3.243.

Manning, C.D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England.

OpenAI (2024a), "Hello GPT-4o", *OpenAI*, available at: https://openai.com/index/hello-gpt-4o/

OpenAI (2024b), "New embedding models and API updates", *OpenAI*, available at: https://openai.com/index/new-embedding-models-and-api-updates/

OpenAI (2025), "OpenAI API – supported countries and territories", *OpenAI Help Center*, available at: https://help.openai.com/en/articles/5347006-openai-api-supported-countries-and-territories (accessed 25 July 2025).

Peng, W., Hu, Y., Xing, L., Xie, Y., Sun, Y. and Li, Y. (2022), "Control globally, understand locally: a global-to-local hierarchical graph network for emotional support conversation", *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI, pp. 4324-4330.

Plutchik, R. (1980), "A general psychoevolutionary theory of emotion", in Plutchik, R. and Kellerman, H. (Eds), *Theories of Emotion*, Academic Press, New York, pp. 3-33.

Pounds, G. (2011), "Empathy as 'appraisal': a new critical framework for analysing empathy in healthcare discourse", *Journal of Applied Linguistics and Professional Practice*, Vol. 8 No. 2, pp. 179-202.

Qiu, H., He, S., Zhang, A., Li, Z. and Lan, Y. (2023), "Smile: single-turn to multi-turn inclusive language expansion via chatgpt for mental health support", arXiv preprint arXiv:2305.00450.

Qiu, H., Li, A., Ma, L. and Lan, Z. (2024, May), "Psychat: A client-centric dialogue system for mental health support", in *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, pp. 2979-2984.

Qwen Team (2024), "Qwen2.5 technical report", doi: 10.48550/arXiv.2412.15115.

Racha, S., Joshi, P., Raman, A., Jangid, N., Sharma, M., Ramakrishnan, G. and Punjabi, N. (2025), "MHQA: a diverse, knowledge-intensive mental health question answering challenge for language models", doi: 10.48550/arXiv.2502.15418.

Ramírez, B.G., Espejel, J.L., Díaz, M.D.C.S. and Linares, G.T.R. (2024), "Sólo Escúchame: Spanish emotional accompaniment chatbot", arXiv preprint arXiv:2408.01852.

Roberts, K., Masterton, K., Fiszman, M., Kilicoglu, H. and Demner-Fushman, D. (2014), "Annotating question decomposition on complex medical questions", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, ELRA, pp. 2598-2602.

Roy, S., Banerjee, P. and Das, A. (2024), "Clinical decision support for bipolar depression using large language models", *Journal of Affective Disorders*, Vol. 346, pp. 85-92.

Rudd, B.N. and Beidas, R.S. (2020), "Digital mental health: the answer to the global mental health crisis?", *JMIR Mental Health*, Vol. 7 No. 6, e18472, doi: 10.2196/18472.

Siddals, S., Torous, J. and Coxon, A. (2024), "'It happened to be the perfect thing': experiences of generative AI chatbots for mental health", *NPJ Mental Health Research*, Vol. 3 No. 1, 48, doi: 10.1038/s44184-024-00097-4.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., Neal, D., Rashid, Q.M., Schaekermann, M., Wang, A., Dash, D., Chen, J.H., Shah, N.H., Lachgar, S., Mansfield, P.A., Prakash, S., Green, B., Dominowska, E., Agüera y Arcas, B., Tomašev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S.S., Barral, J.K., Webster, D.R., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A. and Natarajan, V. (2025), "Toward expert-level medical question answering with large language models", *Nature Medicine*, Vol. 31 No. 1, pp. 1-8, doi: 10.1038/s41591-024-03423-7.

Sun, H., Lin, Z., Zheng, C., Liu, S. and Huang, M. (2021), "PsyQA: a Chinese dataset for generating long counseling text for mental health support", *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Stroudsburg, pp. 1489-1503.

Tang, D., Qin, B. and Liu, T. (2015), "Document modeling with gated recurrent neural network for sentiment classification", *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1422-1432.

Tu, Q., Li, Y., Cui, J., Wang, B., Wen, J. and Yan, R. (2022), "MISC: a mixed strategy-aware model integrating COMET for emotional support conversation", doi: 10.48550/arXiv.2203.13560.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... and Polosukhin, I. (2017), "Attention is all you need", in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998-6008.

Wang, J., Li, W., Lin, P. and Mu, F. (2021), "Empathetic response generation through graph-based multi-hop reasoning on emotional causality", *Knowledge-Based Systems*, Vol. 233, 107547, doi: 10.1016/j.knosys.2021.107547.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... and Zhou, D. (2022), "Chain-of-thought prompting elicits reasoning in large language models", doi: 10.48550/arXiv.2201.11903.

Weiner, S. (2022, August 9), "A growing psychiatrist shortage and an enormous demand for mental health services", AAMC News, available at: https://www.aamc.org/news-insights/growing-psychiatrist-shortage-enormous-demand-mental-health-services

Weizenbaum, J. (1966), "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 9 No. 1, pp. 36-45, doi: 10.1145/365153.365168.

WHO (2022), *Mental Health*, World Health Organization, available at: https://www.who.int/health-topics/mental-health

Wu, S., Hsu, W. and Lee, M.L. (2024), "EHDChat: a knowledge-grounded, empathy-enhanced language model for healthcare interactions", *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pp. 141-151, doi: 10.18653/v1/2024.sicon-1.10.

Yang, A., Li, A. and Liu, X. (2024), "Qwen2 technical report", *arXiv*, doi: 10.48550/arXiv.2407.10671.

Yang, A., Li, A. and Zhou, J. (2025), "Qwen3 technical report", *arXiv*, doi: 10.48550/arXiv.2505.09388.

Yao, Y., Zhang, X. and Li, P. (2021), "Mental health question and answering system based on BERT model and knowledge graph technology", *Proceedings of the 10th International Conference on Knowledge and Systems Engineering*, pp. 472-476, doi: 10.1145/3500931.3501011.

Zeng, G., Chen, H., Yang, H., Wu, X., Liang, Y., Liu, J., . . . and Tang, J. (2020), "MedDialog: a large-scale medical dialogue dataset", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhang, H., Chen, Y., Wang, M. and Feng, S. (2024), "FEEL: a framework for evaluating emotional support capability with large language models", *Proceedings of the International Conference on Intelligent Computing*, pp. 96-107, doi: 10.1007/978-981-97-5618-6_9.

Zheng, Z., Sabour, S., Wen, J., Zhang, Z. and Huang, M. (2023), "ExTES: an extensible emotional support dialogue dataset generated by large language models", *Findings of ACL-IJCNLP*, Vol. 2023, pp. 1552-1568.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H. and Xu, B. (2016), "Attention-based bidirectional long short-term memory networks for relation classification", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 207-212.

Zhu, J., Jiang, Z., Zhou, B., Su, J., Zhang, J. and Li, Z. (2024), "Empathizing before generation: a double-layered framework for emotional support LLM", *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 490-503, doi: 10.1007/978-981-97-8490-5_35.

Further reading

Alibaba Cloud (2025), "Tongyi Qianwen (Qwen) project", *Alibaba Cloud*, available at: https://www.alibabacloud.com/en/solutions/generative-ai/qwenBanerjee

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Çelikyilmaz, A. and Choi, Y. (2019), "Comet: commonsense transformers for automatic knowledge graph construction", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 4762-4779.

NCBI Staff Choi, G.H., Yun, J., Choi, J., Goh, M.J., Sinn, D.H., Jin, Y.J., Kim, M.A., Yu, S.J., Jang, S., Lee, S.K., Jang, J.W., Lee, J.S., Kim, D.Y., Cho, Y.Y., Kim, H.J., Kim, S., Kim, J.H., Kim, N. and Kim, K.M. (2024), "The opportunities and risks of large language models in mental health", *npj Digital Medicine*, Vol. 7 No. 1, p. 2, doi: 10.1038/s41746-023-00976-8.

Sarkar, S., Gaur, M. and Chen, L.K. (2023), "A review of the explainability and safety of conversational agents for mental health", *ACM Transactions on Computing for Healthcare*, Vol. 4 No. 1, Article 1.

U.S. Food and Drug Administration (2021), "Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD) action plan", available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Corresponding author
William Yu Chung Wang can be contacted at: william.wang@waikato.ac.nz