# Exploring Prompt Engineering for Grapevine Pruning Decision Making with Limited Data

Wanrui He[1], Ben McGuinness[1], Geoffrey Holmes[1], Dale Fletcher[1], Mike Duke[1], Chi Kit Au[1], Henry Williams[2], David Smith[2], Shen Hin Lim[1]
wh237@students.waikato.ac.nz

**Abstract**

To address the challenges posed by the shortage of agricultural labour, this paper explores the use of Prompt Engineering (PE) to guide Large Language Models (LLMs) for grapevine pruning decision-making. Based on real grapevine images and structured data, five prompt templates were designed and tested on four mainstream LLMs: Gemma-3, Llama-4, Gemini 2.5 Pro, and GPT-o3. Experimental results show that some models excel in consistency—GPT-o3 achieved full consistency (1.0) under Prompt 3, and Gemma-3 performed well under Prompts 2 and 3 (0.98 and 0.96). However, inter-model agreement remained low (Fleiss' Kappa $< 0.07$) across all conditions. These findings confirm the feasibility of using PE to simulate expert pruning in data-scarce environments and offer insights for building scalable, interpretable decision-support systems in agriculture.

## 1 Introduction

As the problem of structural shortage of global agricultural labour is becoming more and more prominent, the automation and intelligence of agriculture production has become a key research direction to guarantee food security and sustainable development of the industry. Among many agricultural operations, grapevine pruning has become a key challenge for intelligent agricultural technology due to its high dependency on expert experience and complex decision-making dimensions. This challenge is particularly prominent in the main wine producing countries such as New Zealand, where seasonal labour constraints and low pruning efficiency have severely restricted the improvement of industrial efficiency. Traditional approaches to grapevine pruning automation have primarily relied on computer vision and robotic systems, where techniques such as 3D modelling and RGB-D perception have been used to identify vine structures and assist robotic pruning. While these methods demonstrate the feasibility of perception driven automation, they remain limited by structural complexity and the lack of flexible decision-making [Botterill *et al.*, 2017][Silwal *et al.*, 2021]. In recent years, large language models (LLMs) have demonstrated breakthrough capabilities in complex task planning and cross domain reasoning, providing a new paradigm for the development of intelligent decision-making systems in agriculture. However, the specificities of agricultural scenarios including the high heterogeneity of plant structures, the strong dependency on visual modalities, and the scarcity of expert annotated data have led to significant challenges for the implementation of LLMs in specific farming operations. Especially in the typical scenario of grapevine pruning, how to make generic LLMs accurately simulate the dual decision-making mechanism of "cut judgment" and "pruning strategy" of agricultural experts in the absence of large-scale labelled data is still a key scientific issue that has not been systematically solved in current research.

Unlike rule-based systems, pruning strategies vary among viticulturists due to differing experience and production goals, making universal rules impractical. LLMs, with their adaptive reasoning and generalization, can emulate multiple expert styles and generate context-appropriate pruning decisions, especially valuable when expert consensus is lacking or vines differ structurally.

This paper proposes a Prompt Engineering based approach to explore how to simulate the pruning decision-making process of agricultural experts by guiding a generalized Large Language Model (LLM) with different types of prompts without the need for fine-tuning the model or constructing complex model structures. We constructed a variety of pruning task based on real grapevine images and their structural data, designed 5 different kinds of prompt templates, and conducted experiments with mainstream LLMs (Gemma-3, Llama-4, Gemini 2.5 Pro, and GPT-o3), to evaluate their consistency, robustness and inter-model agreement in agricultural tasks. These were applied to mainstream LLMs and their consistency, robustness and inter-model agree-

*[1]STEM Division, University of Waikato, Hamilton, New Zealand
†[2]CARES, University of Auckland, Auckland, New Zealand

ment were evaluated. The following sections will sequentially introduce research on grape pruning robots and prompt engineering (Section 2), experimental design and prompt construction methods (Section 3), evaluation metrics and experimental results (Section 4), as well as discussion, limitations, and future directions of the study (Section 5).

This paper contributes to exploring the stability and sensitivity of prompt design for LLMs in grapevine pruning tasks when expert-annotated data is scarce. The research focuses is not on directly replacing expert judgment, but rather on first examining whether LLMs can produce stable and reproducible outputs under different prompt conditions. This work lays the groundwork for future systematic comparisons between LLM decisions and the actual results of expert annotations.

## 2 Related Work

Large Language model (LLM) have made breakthroughs in natural language processing, generalized reasoning, and code generation in recent years, and the attempts in the field of agriculture have gradually increased. For example, one study evaluated the performance of GPT-4 and other models in generating crop pest and disease management recommendations, and used GPT-4 as an evaluator to measure its coherence, accuracy, and utility, and the results showed that GPT-4 could achieve an accuracy of about 72%, indicating the feasibility of LLMs in agriculture decision-making. Assistance [Yang *et al.*, 2024]. In robot control, the Ahn et al.'s framework uses language models for transforming high level command into executable action plans to select the most feasible and helpful manipulation skills, demonstrating the potential of linguistic reasoning in robot task planning [Ahn *et al.*, 2022]. However, most of the current research focuses on Q&A and generation suggestions tasks, and there is still a gap in the systematic validation of LLM in structural decision-making and complex agricultural operations, while data scarcity and complexity in agricultural scenarios also limit the generalization ability of the model.

Prompt engineering, as a lightweight method for guiding LLMs to accomplish specific tasks, has been widely used in code generation, data analysis, educational assessment, and other fields. In recent years, prompt design has been systematically categorized as zero-shot, few-shot and chain-of-thought etc. strategies [Sahoo *et al.*, 2024]. In medical, legal and other professional fields, some studies have systematically investigated different prompt types and turning methods, and carried out in-depth evaluations in terms of robustness and stability [Liu *et al.*, 2023][Zhu *et al.*, 2023]. In the field of agriculture, there is a lack of prompt engineering research for complex structural decisions. In this paper, it introduce the prompt engineering system into grapevine pruning task for the first time, systematically design the prompt type from the perspectives of task format, reasoning style, rule-based etc., and evaluate its output consistency, robustness and inter-model agreement, so as to provide a new direction for the introduction of "language driven reasoning" in the field of agriculture.

## 3 Methodology

In this study, a LLM based decision-making system for grapevine pruning was constructed, and its general architecture is shown in Figure 1. The grapevine data used in the experiment was provided by MaaraTech [Williams *et al.*, 2023], which contains structured descriptive text that has been processed, covering key attributes such as branch diameter, internode length, number of nodes and location (left or right). The system incorporates a variety of designed prompts to guide the LLM's reasoning on the input data and outputs predictions for retain and remove of each branch. To objectively evaluate the model performance, an evaluation system was also designed to evaluate the prediction results. The overall process consists of four steps: Input prompt engineering, LLM inference, pruning prediction and evaluation. It can effectively measure the performance of different models and prompt designs to ensure the scientific validity and comparability of the experimental results.

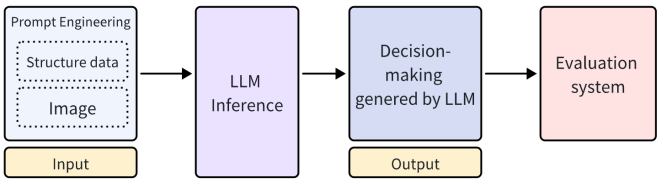


Figure 1: Overall Architecture of the LLM-based Grapevine Pruning Decision System

To enhance transparency and reproducibility, Figure 2 and Table 1 provide representative examples of the structured grapevine data used in this study. Figure 2 shows a grapevine sample from a commercial vineyard in New Zealand, where each branch is colour labelled to facilitate consistent identification and annotation. There labels correspond to the structured attributes summarized in Table 1, including diameter, internode length, node count, and spatial position. Diameter 1 and 2 are the diameters (in mm) of a branch at two random different measurement points are used to assess the branch's thickness. Internode 1 and 2 are the length (in mm) of the internode from first node to second node position, reflecting the vigor of the branch and the spacing of the bud. Position indicates the spatial distribution of the branches in the overall structure of the vine (left/right),

which is used to maintain the balance between the left and right sides of the vine. Node indicates the number of available buds on a branch, which directly affects the potential fruiting capacity and pruning strategy for the following season. During actual grapevine pruning, experts typically weigh multiple factors when deciding which shoots to "retain" or "remove". Typical decision-making logic includes: prioritizing the removal of old canes due to their diminished fruiting capacity; in terms of diameter, branches that are too thin or too thick are unsuitable as fruiting canes; internode length reflects branch vigor and bud spacing, with excessively short or long internodes affecting the following season's fruit set; the number of nodes determines the branch's potential fruiting capacity; simultaneously, maintaining lateral balance in branch placement is crucial to prevent vine structural imbalance. These expert principles are guided into the LLM's reasoning process through prompting, enabling it to simulate expert decision-making to a certain degree. It should be noted, however, that this study intentionally avoided incorporating excessive restrictive rules in prompt design to ensure it accurately reflects the LLM's stability and sensitivity when confronted with diverse prompts. For instance, only informed the model of factors experts typically consider during decision-making (such as diameter, internode length, number of nodes, and the age of the cane), but refrained from using absolute statements in the prompt, such as "old canes must be removed". instead, the prompts maintained a degree of openness to observe variations in the model's autonomous reasoning under different prompt conditions.


Figure 2: Example of grapevine with colour labelled canes for pruning decision-making experiments

The data used in the experiment was provided directly by MaaraTech [Williams *et al.*, 2023] to ensure the authenticity of the study. The data source covered a commercial vineyard in New Zealand. All data were rigorously manually labelled to ensure accuracy and consistency of branch structure information. Key information such as branch diameter, internode length, number of nodes, and spatial location were covered in the data of each vine, providing data to support the model's pruning decisions. Unlike many studies that rely on the image-to-text conversion procession, this study does not require additional image preprocessing and feature extraction, but instead directly uses the structured data, thus avoiding potential information loss and ensuring that the experimental results truly reflect the growth status and pruning needs of grapevines.

To systematically explore the potential and performance of prompt engineering in complex decision-making tasks in agriculture, this paper designed and applied five representative types of prompts, which are categorized according to the systematic definitions in related studies [Sahoo *et al.*, 2024] and matched with the actual characteristics of the grapevine pruning task. Specifically, Prompt 1 belongs to the Chain-of-Thought (CoT) reasoning type, which requires the model to reason step-by-step based on the attributes of multiple branches and ultimately give the results of retaining and removing them, reflecting the characteristics of step-by-step logical reasoning. Prompt 2 belongs to Structured Chain-of-Thought (SCoT), which guides the model to rank and interpret the branches under the quantitative scoring system of 1-5, and ensure the rationality and comparability of the output results. Prompt 3 is Chain-of-Knowledge (CoK), which introduces explicit pruning rules and prioritization weights to give the model an advantage in stability and hallucination avoidance. Prompt 4 is Chain-of-Symbol (CoS) symbolic reasoning, which strengthens the image visual to decision mapping capability by using colour marked branches as the main input. Prompt 5 is Program-of-Thought (PoT), which requires the model to synchronize images and structured data to output multiple types of pruning decisions, including remaining and removing branches, thus reflecting the ability of procedural and multilevel reasoning. These five types of prompts complement each other in terms of reasoning style, knowledge dependency, image input and procedural generation, which lays the experimental foundation for the subsequent comparison of the decision-making performance of LLM. A representative prompt example is presented in Appendix A to illustrate the template design, and Appendix B further demonstrates an output example under Prompt example generated by Llama-4, highlighting how the model applies the designed prompts in practice.

So as to comprehensively evaluate the applicability of LLM in the grapevine decision-making task, this study is based on the latest ranking of the Artificial Analysis platform [Artificial Analysis, 2025] for the model selec-

Table 1: Example of structured grapevine branch data used as input for the decision-making experiments

| **Feature** | **Blue** | **Light Blue** | **Red** | **Green** | **Orange** | **White** | **Yellow** | **Purple** | **Black** | **Pink** |
|---|---|---|---|---|---|---|---|---|---|---|
| Diameter 1 | 12.29 | 15.34 | 12.76 | 13.97 | 11.45 | 16.70 | 10.61 | 7.14 | 13.53 | 9.55 |
| Diameter 2 | 11.54 | 13.84 | 11.14 | 12.43 | 10.75 | 15.09 | 10.22 | 7.67 | 13.00 | 9.11 |
| Internode 1 | 58.14 | 32.80 | 19.95 | 55.41 | 16.84 | 49.31 | 40.32 | 75.05 | 56.39 | 28.89 |
| Internode 2 | 80.68 | 47.79 | 51.51 | 72.78 | 34.00 | 36.41 | 71.42 | 86.85 | 69.30 | 38.59 |
| Position | Left | Right | Left | Left | Right | Left | Right | Left | Left | Right |
| Node Count | 14 | 13 | 13 | 8 | 13 | 14 | 15 | 12 | 14 | 18 |
| Length Type | Yes | Old Cane | Yes | Old Cane | Yes | Old Cane | Yes | Yes | Yes | Yes |

Table 2: Overview of five prompt templates designed for LLM pruning decision-making

| **Prompt** | **Prompt Type** | **Abbreviation** | **Reasoning Style** | **Input Modality** | **Decision Output** |
|---|---|---|---|---|---|
| 1 | Chain-of-Thought | CoT | Step-by-step logical reasoning | Structured cane attributes | Retain/Remove list |
| 2 | Structured CoT | SCoT | Scoring-based structured reasoning | Structured cane attributes | Score (1–5) + Retain/Remove |
| 3 | Chain-of-Knowledge | CoK | Rule-based + prioritization | Structured data + pruning rules | Retain/Remove decision |
| 4 | Chain-of-Symbol | CoS | Symbolic / colour-based mapping | Colour labelled images | Retain/Remove decision |
| 5 | Program-of-Thought | PoT | Procedural & multi-level reasoning | Image + structured data | Retain/Remove + explanation |

tion, and combines the research task's multimodal input requirements of the research task for multimodal inputs, two open-source LLM models and two closed-source LLM models were selected. The core criteria for selection include: 1) multimodal support capability to handle both textual and image data; 2) performance and stability of the models in the comprehensive ranking; 3) accessibility and interface compatibility; 4) representativeness in academia and industry. During the experiments, all models are run under a unified framework, receiving the same input data and prompt types to ensure comparable results.

For the open-source models, Gemma-3 and Llama-4 were selected, where Gemma-3 has a stable performance in the ranking and good support for multimodal reasoning, which is able to handle structured inputs and incorporate logic chaining reasoning, and LLaMa-4, the latest open source model released by Meta, is ranked high in language comprehension and reasoning tasks, and both open-source models can be used to provide flexibility through the API from Together.ai [Together AI, 2025]. Together.ai API provides flexible parameter regulation functions, such as temperature setting and batch running, to facilitate the controllability of comparison experiments.

For closed-source models, Gemini 2.5 Pro and GPT-o3 were chosen. Gemini 2.5 Pro is the flagship multimodal model released by Google DeepMind, which is ranked high in the list and attracts much attention for its performance in complex reasoning and task stability; while GPT-o3 is a model optimized for real-world applications by OpenAI. GPT-o3 is OpenAI's next-generation model optimized for real-world applications, which maintains a leading position in natural language understanding and generation, and is invoked through the ChatGPT interface, providing strong inference consistency and contextual adaptation. The introduction of these two closed-source models ensures a fair comparison with open-source models in the one hand, and provided diverse architectures and training paradigms for research on the other.

Table 3 summarizes the four models selected for this study, including their type, release company, context window size, image input support, and release data. As shown in the table, Gemma-3 supports a context window of up to 128k tokens with image input capability; Llama-4 provides a significantly larger context window of 10000k tokens. Among the closed-source models, Gemini 2.5 Pro has a 1000k tokens context window and supports multimodal input, while GPT-3o supports a 200k tokens context window and image input capability. This comprehensive overview highlights the complementary strengths of the selected models in the comparative evaluation.

To systematically evaluate the performance of different LLMs in grapevine pruning decision-making, three types of core metrics are designed in this study: consistency, prompt robustness and inter-model agreement. These indicators construct the evaluation system from the three dimensions of internal stability of the model, robustness to input design, and cross model consensus to ensure the scientific validity and comparability of the experimental results.

1) Consistency

Consistency is used to measure the stability of the output of the same model under fixed prompt conditions for multiple inference, whether the model is able to give consistent pruning decisions is repeated experiments. In the experiment, each model is run 10 times in the same prompt, $M$ is the number of repetitions(0 to 10), and the

Table 3: Summary of selected open-source and closed-source LLMs used in this study [Artificial Analysis, 2025]

| LLM Model | Type | Creator | Context Window | Image Input Support | Release Date |
|---|---|---|---|---|---|
| Gemma-3 | Open-source | Google | 128k tokens | Yes | March, 2025 |
| Llama-4 | Open-source | Meta | 10,000k tokens | Yes | April, 2025 |
| Gemini 2.5 Pro | Closed-source | Google | 1,000k tokens | Yes | June, 2025 |
| GPT-o3 | Closed-source | OpenAI | 200k tokens | Yes | April, 2025 |

frequency of each branch being decided as removed, $f_{rm}$, and retained, $f_{rt}$ are counted respectively. The overall consistency rate is then calculated by majority voting, and Cohen's Kappa coefficient is used to correct for the effect of random consistency. The formula is as follows [Elazar *et al.*, 2021][Achiam *et al.*, 2023] [Hackl *et al.*, 2023] [Vieira *et al.*, 2010] :

$$Consistency = \frac{1}{N}\sum_{i=1}^{N}\frac{\max\left(f_{rm,i},\ f_{rt,i}\right)}{M} \tag{1}$$

where *i=1* to $N$ denotes the *i-th* number of total $N$ branches.

2) Prompt Robustness

Prompt Robustness is used to assess the consistency of the model's decisions under different prompts, reflecting its robustness to the design of input prompts. To evaluate the stability of a model under different prompts, first should define a prompt robustness metric [Zhu *et al.*, 2023][Razavi *et al.*, 2025][Zhuo *et al.*, 2024].

Set the model be $M_{PR}$, the set of prompts be:

$$P = \{p_1, p_2, \ldots, p_k\}$$

And the set of branches be:

$$C = \{c_1, c_2, \ldots, c_N\}$$

For a branch $c_i$, under a prompt $p_j$ , it's the average over multiple runs is defined as:

$$r_{i,j} = \frac{1}{M}\sum_{t=1}^{M}\mathbf{1}\left\{d_{i,j}^{(t)} = Retain\right\} \tag{2}$$

where 1 is the Iverson bracket function $t$ is the index of the run from 1 to $M$.

The average deviation of branch $c_i$ under prompt $p_j$ compared with other prompts is, where $k$ is total number of prompts:

$$\Delta_{i,j} = \frac{1}{k-1}\sum_{k=1k\neq j}^{k}|r_{i,j} - r_{i,k}| \tag{3}$$

Thus, the robustness of branch $c_i$ under prompt $p_j$ is defined as:

$$R_{i,j} = 1 - \Delta_{i,j} \tag{4}$$

With values ranging from [0,1]. A larger value indicates that the model's outputs are less affected by prompt variations, representing higher robustness.

The average robustness of model $M_{PR}$ under prompt $P_j$ is defined as follows:

$$R_{M_{PR},P_j} = \frac{1}{N}\sum_{i=1}^{N}R_{i,j} \tag{5}$$

where $N$ is the total number of branches, and $R_{i,j}$ denotes the robustness score of the $i$-th branch under prompt $P_j$.

3) Inter-model Consistency

Inter-model consistency is used to measure the consistency and consensus of pruning decisions of different models under the same prompt. The specific method is: for each branch, the remove/retain judgment results of multiple models are aggregated, and Fleiss' Kappa is used to calculate the overall consistency index. At the same time, the consistency matrix between model pairs is constructed to reveal the similarity of different model combinations is specific decisions. The formula is as follows: [Fleiss, 1971][Amiri-Margavi *et al.*, 2025][Badshah and Sajjad, 2024]:

$$Consistency = \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{K}p_{ij}^2 - \frac{1}{K}\right)}{1 - \frac{1}{K}} \tag{6}$$

where $K$ represents the number of models and $N$ represents the total number of grapevine branches of each vine.

Thus, $i$ is the index of the *i-th* branch in the range $1 \leq i \leq N$, $j$ is the index of the *j-th* model in the range $1 \leq j \leq K$, and $p_{i,j}$denotes the proportion of the *i-th* branch selected under the *j-th* model.

## 4 Results

This study verifies the feasibility and potential of guiding LLMs for grapevine pruning decision-making through PE in an agricultural data scarce scenario. As can be seen from the heat map in Figure 3, there are significant differences in the performance of the different models in terms of consistency (left) and robustness (right), reflecting their ability to adapt to different prompt types.

In terms of consistency, the average consistency of all four models under the five types of prompts is maintained at a high level (0.82-1). Among them, GPT-o3 reached full consistency (1) under prompt 3, showing very high output stability; Gemma-3 performed well under prompts 2 and 3 (0.98 and 0.96 respectively), suggesting its strong stability under structured rubric categorization and rule-based prompts; and Lllama-4 reached a peak consistency under prompt 4 (0.98), indicating that it responds very consistently to symbolic reasoning prompts; whereas Gemini 2.5 Pro performs consistently overall (range 0.84-0.98), but performs slightly lower at prompt 4 (0.84), indicating that it still fluctuates in processing certain types of prompts. However, these results also reflect that the reasoning paths of different models on the same prompt may differ significantly, and their decision logics may be inconsistent despite a high eventual consistency, which can pose a coordination challenge when fusing multiple models.

In terms of robustness, the overall distribution range from 0.68 to 0.82, indicating that all models have a certain output uncertainty under prompt changes; Gemma-3 performs best under prompt 2 and 3 (both 0.82), indicating that it is more resistant to fluctuations in scoring and rule-based prompt; GPT-o3 has a more balanced distribution of robustness (0.7-0.8), reflecting its adaptability to different prompt; Gemini 2.5 Pro also maintains high robustness under most of the cues, but slightly decreases under prompt 3 and 4 (0.72 and 0.76); and Llama-4 has relatively low robustness under prompt 1 and 4 (0.69 and 0.74), indicating that it may still be affected by prompt perturbation under generic and symbolic prompts. This further suggests that some of the models have a significant dependency on the prompt type and are prone to performance fluctuations when the input lacks sufficient detail or structured features.

Taken together, the models perform well in term of consistency, especially in terms of highly stable output ability under specific prompts. However, there is still room for optimization in terms of robustness. The sensitivity difference of the models under different prompt also suggests that need to combine the characteristics of the models to choose a more compatible and generalizable type of prompt when designing the prompt for pruning decision-making. And consider prompt adaptive mechanisms in actual deployments to minimize performance fluctuations.

As can be seen from Figure 4, the overall level of inter-model consistency under different prompt is low, and none of the Fleiss' Kappa values exceed 0.1, indicating that there are significant differences in the pruning decisions of different models under the same prompt conditions. Specifically, the highest level of inter-model consistency was found for prompt 3 (0.056), suggesting

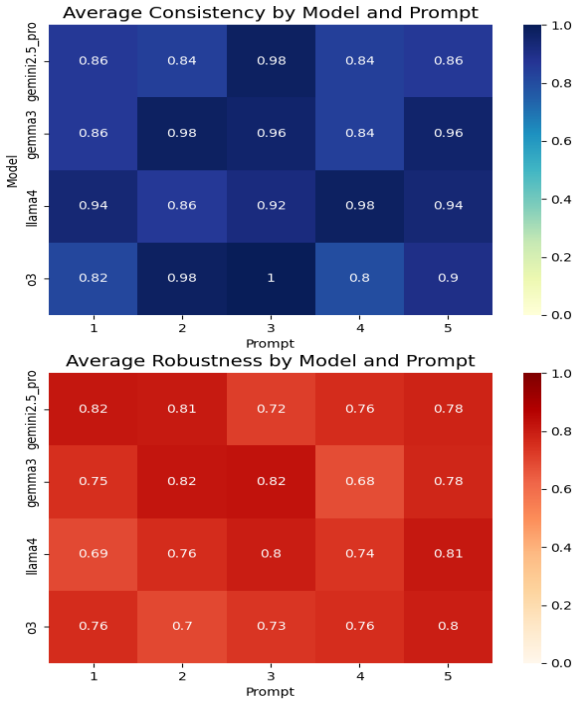


Figure 3: Average Consistency and Robustness of Models Across Different Prompts

that this type of prompt narrowed that inference differences between models to a certain extent, but still did not reach a high level of agreement. Prompt 2 (0.035) and prompt 5 (0.037) had the next highest results, suggesting that they helped to some extent to reach relatively close judgments across models. The lowest results were observed for prompts 1 and 4 (0.019 and 0.020), suggesting that when the prompts provided only general requirements without detailed guidance, the models performed less effectively. This is also a side note that the limited size and diversity of the current dataset may not be sufficient to cover the more complex changes in vine structure in actual production, and especially lacks validation of the data across seasons and varieties, thus limiting the ability of the model to generalize across multiple scenarios.

Overall, while individual models perform well on the consistency metrics, but cross-model consistency is still at a low level. This suggests that LLMs with different architectures and training paradigms have significant differences in reasoning logic when responding to the same pruning task, making it difficult to form a unified expert-level decision.

In addition to the above quantitative evaluation results, this study encountered a practical bottleneck during the experimental process. Specifically, there is a greater challenge in accurately extracting pruning decisions (remove or retain) from the output responses of

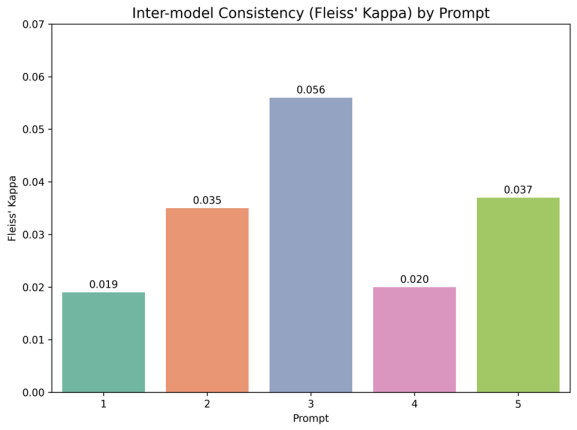


Figure 4: Inter-model Consistency (Fleiss' Kappa) Across Prompts

LLMs. In the initial stage, tried to adopt a rule-based parser method, but due to the significant differences in word usage, terminology and text structure among different models, it is difficult to design a set of generalized rules that can cover all cases. As a result, the parser could only be done manually, which not only significantly increases the workload, but also exposed the limitations of the current direct application of LLMs' outputs to the automated process of grapevine pruning. This issue suggests that future research needs to explore mechanisms for standardizing the output format or incorporation auxiliary automated extraction models to reduce manual intervention and increase automation.

## 5 Conclusion

This study validates the feasibility and potential of guiding LLMs for grapevine pruning decision-making through prompt engineering in an agricultural data scarce environment. And systematically evaluated the performance of open-source and closed-source models in terms of consistency, robustness, and inter-model consistency based on five types of representative prompts. The results show that some of the models (e.g. Gemma-3, Llama-4 and GPT-o3) perform well in terms of consistency, but the overall robustness is still in the middle of the range, and the degree of consensus among different models is low. This suggests that although LLM is able to consistently model pruning decisions under specific prompt, there are differences in their reasoning logics, making it difficult to form a unified expert level judgment.

At the application level, the prompt engineering framework proposed in this study effectively reduces the dependence on large scale labelled data, demonstrating its scalability and adaptability as a prototype viticulture decision support system. Meanwhile, the results also highlight the importance of combining multimodal data with expert rules to enhance model interpretability and stability.

It should be noted that some may question the value of consistency analysis, arguing that if models consistently generates erroneous results, the stability or alignment of its ouput become meaningless, accuracy should instead directly compared against expert ground truth. While this concern is valid, consistency is in fact a prerequisite for any meaningful accuracy assessments. If a model cannot maintain stable outputs under identical prompt conditions, its measured accuracy becomes unreliable, as evaluation results may fluctuate randomly across different runs. Therefore, this study first focuses on testing whether LLMs can sustain stable reasoning behavior across different prompts conditions in data-scarce scenarios. While this study primarily focuses on stability and robustness rather than correctness, we recognise that validation against expert viticulturists is an essential next step. Even a small subset of expert-labelled vines would enable quantitative evaluation of pruning correctness and further connect the LLM reasoning process to practical viticultural decision-making. We acknowledge that only one expert has provided initial qualitative feedback on the model outputs so far. A more extensive validation against multiple expert annotations is currently being prepared as part of ongoing work. We are engaging experts and collecting a larger set of annotated vine data to quantitatively validate model correctness and agronomic soundness in subsequent studies.

Future research will focus on 1) domain knowledge based prompt optimization to enhance robustness and inference reliability; 2) multimodal fusion of image and structured data to improve decision accuracy and interpretability; 3) introduction of expert manual evaluation to systematically compare the accuracy of LLM decision-making with expert judgments to validate its utility in real agriculture scenarios; 4) exploration the relationship between prompt engineering and traditional machine learning models (e.g. CNN and RNN), and design "prompt + neural network" experiments to take advantage of the inference ability of LLM and the structured learning advantage of deep learning to further improve the stability and accuracy of pruning decisions.

Overall, the starting point of this study is not to directly measure the correctness of LLM pruning results, but rather to first examine its stability and sensitivity under different prompting conditions. What we aim to replicate in LLMs is the decision-making process experts follow during actual pruning, selecting the most suitable fruiting cane based on metrics such as branch diameter, internode length, node count, spatial position, and new/old status. this process-oriented evaluation lays the groundwork for subsequent integration of expert anno-

tation data to further validate LLM accuracy.

## References


[Achiam *et al.*, 2023] Josh Achiam, Sarah Adler, Sharan Agarwal, Leila Ahmad, Ilge Akkaya, Francisco L. Aleman, and Bob McGrew. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Benjamin David, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[Amiri-Margavi *et al.*, 2025] Amir Amiri-Margavi, Iman Jebellat, Ehsan Jebellat, and Seyed P. M. Davoudi. Enhancing answer reliability through inter-model consensus of large language models. In *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 299–316, Cham, June 2025. Springer Nature Switzerland.

[Artificial Analysis, 2025] Artificial Analysis. Ai leaderboards. https://artificialanalysis.ai/leaderboards/models, 2025. Accessed: 2025-08-24.

[Badshah and Sajjad, 2024] Saad Badshah and Hassan Sajjad. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235*, 2024.

[Botterill *et al.*, 2017] Tim Botterill, Simon Paulin, Richard Green, Steven Williams, Jiamin Lin, Vaughan Saxton, and Simon Corbett-Davies. A robot system for pruning grape vines. *Journal of Field Robotics*, 34(6):1100–1122, 2017.

[Elazar *et al.*, 2021] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[Fleiss, 1971] Joseph L. Fleiss. Measuring nominal scale consistency among many raters. *Psychological Bulletin*, 76(5):378, 1971.

[Hackl *et al.*, 2023] Veronika Hackl, Anna E. Müller, Markus Granitzer, and Michael Sailer. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. *Frontiers in Education*, 8:1272229, 2023.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Razavi *et al.*, 2025] Amir Razavi, Mehrdad Soltangheis, Nima Arabzadeh, Shima Salamat, Mehdi Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *Proceedings of the European Conference on Information Retrieval*, pages 303–313, Cham, April 2025. Springer Nature Switzerland.

[Sahoo *et al.*, 2024] Prateek Sahoo, A. K. Singh, Souvik Saha, Vikas Jain, Subhankar Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[Silwal *et al.*, 2021] Ashwin Silwal, Felipe Yandun, Aditya Nellithimaru, Terry Bates, and George Kantor. Bumblebee: A path towards fully autonomous robotic vine pruning. *arXiv preprint arXiv:2112.00291*, 2021.

[Together AI, 2025] Together AI. Together ai models. https://api.together.ai/models, 2025. Accessed: 2025-08-24.

[Vieira *et al.*, 2010] Susana M. Vieira, Uzay Kaymak, and João M. Sousa. Cohen's kappa coefficient as a performance measure for feature selection. In *Proceedings of the International Conference on Fuzzy Systems*, pages 1–8. IEEE, July 2010.

[Williams *et al.*, 2023] Hamish Williams, Daniel Smith, Javad Shahabi, Thomas Gee, Mohammad Nejati, Brendan McGuinness, and Bruce A. MacDonald. Modelling wine grapevines for autonomous robotic cane pruning. *Biosystems Engineering*, 235:31–49, 2023.

[Yang *et al.*, 2024] Shu Yang, Ziyu Yuan, Shuo Li, Rui Peng, Kai Liu, and Peng Yang. Gpt-4 as evaluator: Evaluating large language models on pest management in agriculture. *arXiv preprint arXiv:2403.11858*, 2024.

[Zhu *et al.*, 2023] Kaiyu Zhu, Jian Wang, Jun Zhou, Zizheng Wang, Hong Chen, Yiqing Wang, and Xiaodong Xie. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306*, 2023.

[Zhuo *et al.*, 2024] Jialiang Zhuo, Shuo Zhang, Xiaofei Fang, Hongbo Duan, Dawei Lin, and Kai Chen. Prosa: Assessing and understanding the prompt sensitivity of llms. *arXiv preprint arXiv:2410.12405*, 2024.

## Appendix A. Prompt Template Example

**Prompt Example**

You are an expert viticulturist specializing in grapevine cane pruning. Your task is to evaluate a given vine based on specific attributes and decide which canes should be removed entirely to optimize fruit yield. For this type of grape, only two canes should be retained for fruit production. Evaluate the following attributes of the vine:
- Cane Colors: blue, light blue, red, green, orange, white, yellow, purple, black, pink
- Cane Attributes:
- Diameter 1 (mm)
- Diameter 2 (mm)
- Internode 1 (mm)
- Internode 2 (mm)
- Position (Left/Right)
- Node Count
- Old Cane Status (Yes/No)

The data for the canes is as follows:
Data Input:

| | blue | light blue | red | green | orange | white | yellow | purple | black | pink |
|---|---|---|---|---|---|---|---|---|---|---|
| diameter 1 | 12.29 | 15.34 | 12.76 | 13.97 | 11.45 | 16.70 | 10.61 | 7.14 | 13.53 | 9.55 |
| diameter 2 | 11.54 | 13.84 | 11.14 | 12.43 | 10.75 | 15.09 | 10.22 | 7.67 | 13.00 | 9.11 |
| internode 1 | 58.14 | 32.80 | 19.95 | 55.41 | 16.84 | 49.31 | 40.32 | 75.05 | 56.39 | 28.89 |
| internode 2 | 80.68 | 47.79 | 51.51 | 72.78 | 34.00 | 36.41 | 71.42 | 86.85 | 69.30 | 38.59 |
| position | Left | Right | Left | Left | Right | Left | Right | Left | Left | Right |
| node count | 14 | 13 | 13 | 8 | 13 | 14 | 15 | 12 | 14 | 18 |
| old cane | Yes | Old Cane | Yes | Old Cane | Yes | Old Cane | Yes | Yes | Yes | Yes |

Make your decision based on typical viticulture practices aimed at achieving a balanced yield. Consider factors such as cane diameter, internode length, position, and the age of the cane in your evaluation.

**Final Answer Format:**
Decision: [canes to be removed]
Reason: [your explanation based on viticulture practices]

## Appendix B. Prompt Output Example

**Prompt Output Example from Llama-4**

**Step 1: Understanding the Goal of Grapevine Cane Pruning**
The goal of grapevine cane pruning is to optimize fruit yield by selecting the healthiest, most productive canes. Typically, only two canes are retained for fruit production. The selection criteria include cane diameter, internode length, position, node count, and the age of the cane.

**Step 2: Evaluating Cane Diameter**
Canes with diameters between 10–15 mm are considered optimal. Too thin (¡10 mm) or too thick (¿15 mm) canes may not produce well.

**Step 3: Assessing Internode Length**
Balanced internode length indicates vigor. Very long internodes may mean excess vigor (more leaves than fruit), while very short ones may indicate low productivity.

**Step 4: Considering Cane Position**
Even distribution (left/right) improves sunlight exposure and air circulation, important for yield and disease prevention.

**Step 5: Evaluating Node Count**
Higher node count suggests more fruiting sites, but must be balanced with cane health.

**Step 6: Assessing Old Cane Status**
Old canes often have reduced productivity. Given all canes are old in this case, selection must prioritize the best physical attributes.

**Step 7: Analyzing Given Data**
- Light blue: diameter 15.34/13.84 mm, internode 32.8 mm, node 13, Right
- White: diameter 16.7/15.09 mm, internode 49.31 mm, node 14, Left
- Other canes show less optimal traits (too thin, too long internodes, or fewer nodes).

**Step 8: Applying Selection Criteria**
- Light blue: promising diameter and node count.
- White: slightly thick, but strong diameter and good node count.
- Others: less optimal due to thinness, low node counts, or excessive internode length.

**Step 9: Final Decision**
**Decision:** Remove blue, red, green, orange, yellow, purple, black, pink
**Retain:** light blue, white
**Reason:** Light blue and white canes provide the best balance of diameter, internode length, node count, and position, making them the strongest candidates for retention. Other canes are removed due to less favorable attributes.