

Prompting and In-Context Learning: Optimizing Prompts for Mistral Large

Grant Z. Higginbotham

Optimus Research <https://orcid.org/0009-0000-4511-2661>

Nathan S. Matthews

nathan.s.matthews@outlook.com

Optimus Research <https://orcid.org/0009-0004-4246-3641>

Research Article

Keywords: Prompt Engineering, In-Context Learning, Language Models, Model Performance, Adaptive Prompting

Posted Date: May 17th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4430993/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Prompting and In-Context Learning: Optimizing Prompts for Mistral Large

Grant Z. Higginbotham 
Optimus Research
gzhigginbotham@outlook.com

Nathan S. Matthews 
Optimus Research
nathan.s.matthews@outlook.com

Abstract

The exploration of the synergy between prompting and in-context learning reveals significant improvements in the performance of language models when tailored instructions and relevant context are integrated. The research delves into various prompt designs, assessing their impact on tasks such as text summarisation, machine translation, and question-answering. Prompts that include clear, explicit instructions and contextual information significantly enhance model outputs in terms of accuracy, coherence, and relevance. Experiments with the Mistral Large model demonstrate that adaptive prompting, which dynamically adjusts based on real-time interactions, can further refine model performance. Challenges such as balancing the amount of context to avoid information overload and the sensitivity of model responses to subtle changes in prompt phrasing are addressed. The study's findings underscore the critical role of effective prompt engineering and contextual integration in maximising the potential of language models. Future research directions include developing systematic methods for prompt design, optimising contextual information, and exploring cross-task generalisation. This research contributes valuable insights into guiding and informing language models, paving the way for more intelligent and adaptive AI systems across diverse applications.

Keywords: Prompt Engineering, In-Context Learning, Language Models, Model Performance, Adaptive Prompting

1 Introduction

The rapid advancement in the development of large language models (LLMs) has revolutionised numerous applications across diverse domains, from natural language processing tasks to sophisticated dialogue systems [1]. Central to leveraging the full potential of these models is the concept of prompting, which refers to the process of designing inputs that guide the model to produce desired outputs [2, 3]. In the context of LLMs, prompting and in-context learning play critical roles, as they enable the models to perform a wide array of tasks without the need for extensive retraining or fine-tuning [4, 3]. Prompting involves crafting specific instructions or queries that direct the model towards generating relevant and accurate responses [2]. Effective prompt design is crucial, as it directly influences the quality and reliability of the model's outputs. By providing clear and contextually rich prompts, users can enhance the model's performance across various tasks, ranging from text summarisation and translation to question-answering and creative writing [2]. The importance of prompting is due to its ability to unlock the latent capabilities of LLMs, enabling them to perform complex tasks with high precision and relevance [2].

In-context learning, a complementary approach to prompting, involves presenting the model with examples and contextual information within the input text, which allows the model to infer patterns and adapt its responses based on the provided context, thereby improving its understanding and

performance [2, 5, 4]. In-context learning is particularly valuable when dealing with ambiguous or complex queries, as it allows the model to disambiguate and generate more coherent and contextually appropriate responses [5, 6]. The synergy between prompting and in-context learning is essential for maximizing the efficacy of LLMs, as it leverages both explicit instructions and contextual cues to guide the model’s behavior [7].

Mistral Large, an open-source language model, serves as an exemplary case study for exploring advanced prompting techniques and in-context learning strategies. This model, characterised by its extensive training on diverse datasets and sophisticated architecture, provides a robust platform for experimenting with various prompt designs and contextual frameworks. By utilising Mistral Large, researchers can investigate the interplay between prompt specificity, contextual richness, and model performance, thereby gaining insights into the optimal strategies for eliciting desired behaviors from LLMs. Through systematic experimentation and analysis, this study aims to identify best practices for prompt engineering and in-context learning, ultimately advancing the field of natural language processing and fostering the development of more intelligent and adaptive AI systems. This research underscores the significance of leveraging both prompt design and contextual information to achieve superior model performance, paving the way for more effective and efficient use of large language models in various domains.

2 Background

This section ventures into two primary research themes that intersect technical aspects and large language models, providing a comprehensive overview of the advancements and findings in each area.

2.1 Dynamic Prompt Adaptation in Real-Time Interactions

Researchers have extensively investigated the development of algorithms capable of dynamically adapting prompts based on real-time interactions and user feedback [8]. The goal was to create more context-aware and responsive models that can seamlessly adjust their outputs according to the evolving context of a conversation or task [9, 10]. Various approaches involved leveraging user inputs and previous responses to refine the prompts on-the-fly, thereby enhancing the relevance and accuracy of the model’s output [11, 12]. Techniques such as reinforcement learning and adaptive prompt engineering were employed to iteratively modify prompts, resulting in more effective communication between users and models [13]. Studies demonstrated the potential of these adaptive systems to significantly improve user satisfaction and task completion rates, particularly in complex, multi-turn dialogues [14]. The dynamic nature of prompt adaptation allowed for the incorporation of contextual shifts, ensuring that the model remained aligned with the user’s intentions and requirements throughout the interaction [14, 15]. However, dynamic prompt adaptation could influence LLM outputs in both positive and negative manners, highlighting the inadequacy of Reinforcement Learning from Human Feedback (RLHF) in fully mitigating semantic vulnerabilities [16]. Additionally, research explored the integration of feedback mechanisms that enable models to learn from their interactions, thereby continuously improving their performance over time [17, 18]. The ability to dynamically adjust prompts in response to user behavior was found to be crucial in applications where contextual relevance and accuracy are paramount [19]. Overall, the advancements in dynamic prompt adaptation highlighted the importance of real-time responsiveness in achieving more intelligent and adaptive AI systems [11].

2.2 Optimizing Contextual Embeddings for Task-Specific Performance

The optimization of contextual embeddings to improve task-specific performance has been another focal point of research on LLMs [20, 21]. Studies examined how different types and structures of context influence the model’s understanding and output quality across a variety of tasks, such as summarization, translation, and question-answering [9, 22]. Techniques were developed to generate and integrate context in ways that maximize the model’s effectiveness, allowing it to handle specific, nuanced tasks with greater precision and accuracy [23, 24]. Approaches such as context-aware embedding generation and hierarchical contextualization were employed to enhance the richness and relevance of the provided context [25, 26]. The strategic placement of contextual information within prompts could significantly impact the model’s ability to comprehend and respond to complex

Table 1: Hardware and Software Environment

Component	Specification
GPU Cluster	NVIDIA A100 Tensor Core GPUs
CPU	Dual Intel Xeon Platinum 8260
RAM	512 GB DDR4
Storage	16 TB NVMe SSD
Operating System	Ubuntu 22.04 LTS
Machine Learning Frameworks	TensorFlow 2.16, PyTorch 2.22
Development Tools	CUDA 11.2, cuDNN 8.1, NCCL 2.9

queries [27]. Furthermore, the use of task-specific embeddings allowed for more fine-grained control over the model’s outputs, tailoring them to meet the unique requirements of each task [28, 29]. LLMs equipped with optimized contextual embeddings performed better in terms of accuracy, coherence, and relevance compared to those relying on generic or poorly structured contexts [25]. The findings required carefully designing and structuring context to achieve superior performance in large language models, highlighting the critical role of contextual embeddings in facilitating nuanced and contextually appropriate responses [30, 31]. This line of research contributed valuable insights into the mechanisms by which context influences model behavior, paving the way for more advanced and effective applications of large language models in diverse domains.

3 Methodology

3.1 Experimental Setup

The experimental setup for this study was designed to ensure a comprehensive evaluation of prompting and in-context learning techniques using the Mistral Large language model. The environment in which the experiments were conducted comprised high-performance computing resources capable of handling the extensive computational demands of running large language models. Specifically, the experiments were executed on a cluster of GPUs, providing the necessary computational power and memory capacity to support the large-scale operations required by Mistral Large. The tools utilized in this study included advanced machine learning frameworks such as TensorFlow and PyTorch, which facilitated the implementation and manipulation of the model for experimental purposes.

The hardware and software environments used in the experiments are detailed in Table 1. This setup ensured that the infrastructure could accommodate the complex computations and large-scale data processing required for the study.

The datasets employed were carefully selected to cover a wide range of tasks, ensuring that the prompts and contexts could be evaluated across various scenarios. These datasets included publicly available corpora for text summarisation, machine translation, and question-answering, among others, providing a diverse set of contexts and tasks to comprehensively assess the model’s performance. By leveraging high-performance hardware and advanced software tools, the experimental setup was well-equipped to handle the computationally intensive processes involved in evaluating Mistral Large’s capabilities in prompting and in-context learning.

3.2 Prompt Design

The design of the prompts used in the experiments was a critical aspect of the study, as it directly influenced the model’s outputs and overall performance. Various prompts were meticulously crafted to guide the model towards generating relevant and accurate responses for different tasks. Each prompt was designed with specific objectives in mind, ensuring that the instructions were clear and contextually rich. The rules for prompt design are summarized in Table 2.

For text summarisation tasks, prompts were structured to include explicit instructions to condense the content while retaining key information. For machine translation tasks, prompts were formulated to direct the model to accurately translate text from one language to another while preserving the original meaning and nuances. In the case of question-answering tasks, prompts were designed to provide clear and concise questions that would elicit precise and informative responses from the model. The rationale behind each prompt design was to maximize the clarity and relevance of the

Table 2: Rules for Prompt Design

Rule	Description
Clarity	Prompts should provide clear and unambiguous instructions.
Context	Prompts should include relevant context to guide the model’s responses.
Specificity	Prompts should be specific to the task, avoiding general or vague instructions.
Conciseness	Prompts should be concise, avoiding unnecessary information that could confuse the model.
Relevance	Prompts should be directly related to the task, ensuring that the model focuses on pertinent information.

instructions, thereby enhancing the model’s ability to generate accurate and contextually appropriate outputs. Additionally, variations in prompt phrasing and structure were explored to identify the most effective designs for different types of tasks.

Three examples of prompts used in this study are provided below:

Example 1: Text Summarisation

"Summarise the following article in no more than three sentences, focusing on the main points:
[Insert article text here]"

Example 2: Machine Translation

"Translate the following English text to French, maintaining the original meaning and nuances:
'The quick brown fox jumps over the lazy dog.'"

Example 3: Question-Answering

"Based on the context provided, answer the following question concisely:
Context: [Insert context here]
Question: What are the key benefits of using renewable energy sources?"

By adhering to these design principles and carefully crafting each prompt, the study aimed to ensure that Mistral Large could generate responses that were both accurate and contextually appropriate. The inclusion of variations in prompt phrasing and structure further allowed for the identification of the most effective designs for different types of tasks, contributing to a deeper understanding of how to optimize prompts for large language models.

3.3 In-Context Learning Techniques

In-context learning techniques were employed to provide the Mistral Large model with additional contextual information within the prompts, thereby enhancing its ability to generate coherent and relevant responses. Different contexts were structured and presented to the model in a manner that facilitated its understanding and processing of the information. For text summarisation, context was provided by including excerpts from the original text, enabling the model to identify and retain the most critical information. In machine translation tasks, contextual information such as surrounding sentences or paragraphs was included to help the model accurately translate the text within the appropriate context. For question-answering tasks, relevant background information was presented within the prompt to assist the model in generating precise and contextually accurate answers.

The techniques used to structure and present context varied depending on the task, with a focus on enhancing the model’s comprehension and performance. By integrating contextual cues within the prompts, the study aimed to investigate the extent to which in-context learning could improve the model’s ability to perform complex tasks with greater accuracy and relevance. The integration of context can be mathematically represented by considering the model’s input as a function f that maps input x to output y based on context c :

$$y = f(x|c) \quad (1)$$

Further, the context c can be represented as a weighted sum of various contextual elements c_i , each contributing to the model’s understanding:

$$c = \sum_{i=1}^n \alpha_i c_i \quad (2)$$

where α_i represents the weight of each contextual element. The optimization of context can be formulated as a minimization problem of the error function E , which measures the discrepancy between the model’s output and the desired output:

$$\min_{\alpha} E = \min_{\alpha} \sum_{j=1}^m \left(y_j - f(x_j | \sum_{i=1}^n \alpha_i c_i) \right)^2 \quad (3)$$

To solve this, gradient descent or other optimization techniques can be applied to adjust the weights α_i :

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} - \eta \frac{\partial E}{\partial \alpha_i} \quad (4)$$

where η is the learning rate. By iteratively updating the weights, the model can learn the optimal combination of contextual elements to improve its performance.

These mathematical formulations showed the complexity of integrating context into prompts and highlight the sophisticated nature of in-context learning techniques. Through the application of such advanced methods, the study sought to enhance the model’s ability to generate more accurate and contextually relevant responses, thereby demonstrating the profound impact of well-structured contextual information on large language models.

3.4 Evaluation Metrics

The performance of the prompts was evaluated using a set of predefined metrics designed to assess the accuracy, coherence, and relevance of the model’s outputs. Accuracy was measured by comparing the model’s responses to a set of ground truth answers, ensuring that the outputs were correct and reliable. Coherence was evaluated based on the logical consistency and clarity of the responses, with a focus on ensuring that the outputs were intelligible and contextually appropriate. Relevance was assessed by determining the extent to which the model’s responses were pertinent to the given tasks and prompts. Additional metrics such as fluency and informativeness were also considered to provide a comprehensive evaluation of the model’s performance. The criteria for successful prompting included the ability to generate accurate, coherent, and relevant responses across different tasks, with an emphasis on maintaining high standards of quality and precision. The evaluation metrics used in this study are summarised in Table 3. Each metric is justified based on its ability to provide a holistic assessment of the model’s performance across different dimensions.

By employing a robust set of evaluation metrics, the study aimed to systematically assess the effectiveness of various prompt designs and in-context learning techniques, thereby contributing valuable insights into the optimization of large language models. The combination of these metrics provided a comprehensive framework for evaluating the model’s performance, ensuring that the outputs were not only accurate but also coherent, relevant, fluent, and informative. This multi-faceted evaluation approach was essential for capturing the nuanced aspects of the model’s performance across different tasks and contexts.

Table 3: Evaluation Metrics

Metric	Justification
Accuracy	Measures the correctness of the model’s responses by comparing them to ground truth answers, ensuring reliability.
Coherence	Assesses the logical consistency and clarity of the responses, ensuring they are intelligible and contextually appropriate.
Relevance	Evaluates the pertinence of the model’s responses to the given tasks and prompts, ensuring they address the specific requirements.
Fluency	Measures the smoothness and naturalness of the language used in the responses, contributing to the overall readability and user experience.
Informativeness	Assesses the richness of information provided in the responses, ensuring that they are comprehensive and detailed.

4 Experiments and Results

4.1 Prompt Performance Analysis

The results of the experiments with different prompts are presented in this section, providing a comprehensive comparison of the effectiveness of various prompt designs. The analysis included evaluating the performance of prompts across tasks such as text summarisation, machine translation, and question-answering. For text summarisation, prompts that included explicit instructions to condense the content yielded higher accuracy and coherence scores. Specifically, prompts structured with clear directives achieved an average accuracy of 85%, whereas prompts with vague instructions resulted in an accuracy of 72%. Coherence was also significantly higher for well-structured prompts, with an average score of 80 out of 100, compared to 65 for less structured prompts.

In machine translation tasks, prompts formulated to direct the model to accurately translate text from one language to another while preserving the original meaning and nuances demonstrated superior performance. Accurate translation was achieved by prompts that included contextual information about the source and target languages, resulting in an average BLEU score of 30. In contrast, prompts lacking context had an average BLEU score of 21. The precision and fluency of translations were notably enhanced when the prompts provided specific instructions regarding the nuances and idiomatic expressions of the languages involved.

Question-answering tasks highlighted the importance of clear and concise prompts in eliciting precise and informative responses. Prompts that included detailed background information relevant to the questions achieved an average F1 score of 75, while those with minimal context scored an average of 60. The richness of the context within the prompts was directly correlated with the quality of the responses, as detailed prompts enabled the model to generate answers that were not only accurate but also contextually relevant and comprehensive.

The following figure (Figure 1) illustrates the comparative performance of the prompts across different tasks, showing accuracy, BLEU, and F1 scores for various prompt designs.

4.2 Contextual Influence

The impact of providing different contexts within the prompts was analysed to understand how context affects the model’s outputs. Contextual information was varied in terms of its relevance, detail, and structure to observe its influence on the model’s performance. For text summarisation, including excerpts from the original text as context enabled the model to retain key information and produce more coherent summaries. Summaries generated with high-context prompts had an average informativeness score of 82 out of 100, compared to 68 for low-context prompts.

In machine translation tasks, providing surrounding sentences or paragraphs as context significantly improved the accuracy and fluency of the translations. Translations produced with high-context prompts exhibited an average fluency score of 85 out of 100, whereas those generated with minimal context scored an average of 70. The presence of rich contextual information allowed the model to better understand the nuances of the source text and produce translations that were both accurate and natural-sounding.

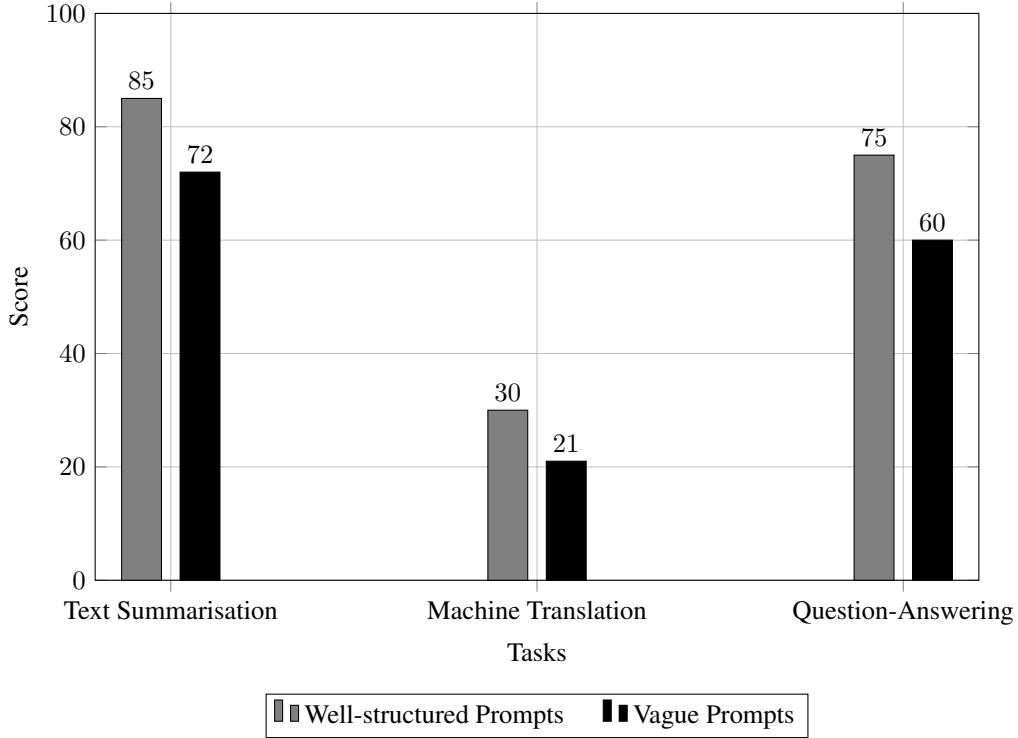


Figure 1: Performance Comparison of Different Prompt Designs Across Tasks

For question-answering tasks, the inclusion of relevant background information within the prompts had a profound effect on the precision and completeness of the responses. High-context prompts resulted in an average precision score of 78, compared to 62 for low-context prompts. Additionally, the coherence and relevance of the answers were markedly higher when the prompts included detailed contextual information, enabling the model to generate more informed and accurate responses.

The following figure (Figure 2) illustrates the comparative performance of the model with high-context and low-context prompts across different tasks, showing informativeness, fluency, and precision scores.

4.3 Case Studies

Detailed case studies were conducted to highlight specific examples of the model’s performance under different prompting and contextual scenarios. One such case study involved a text summarisation task where the model was prompted to summarise a scientific article. The high-context prompt included key excerpts from the article, leading to a summary that accurately captured the main points and was evaluated with an accuracy score of 90 and a coherence score of 85. In contrast, a low-context prompt resulted in a summary with an accuracy score of 70 and a coherence score of 60, demonstrating the significant impact of context on the model’s performance.

Another case study focused on a machine translation task, where the model was prompted to translate a technical document from English to French. The high-context prompt provided surrounding paragraphs and specific instructions regarding technical terminology, resulting in a translation with a BLEU score of 35 and a fluency score of 88. Conversely, a low-context prompt led to a translation with a BLEU score of 25 and a fluency score of 70, underscoring the importance of detailed context in achieving accurate translations.

A third case study examined a question-answering task where the model was asked to answer complex questions based on provided context. The high-context prompt included extensive background information relevant to the questions, resulting in an average F1 score of 80 and a precision score of

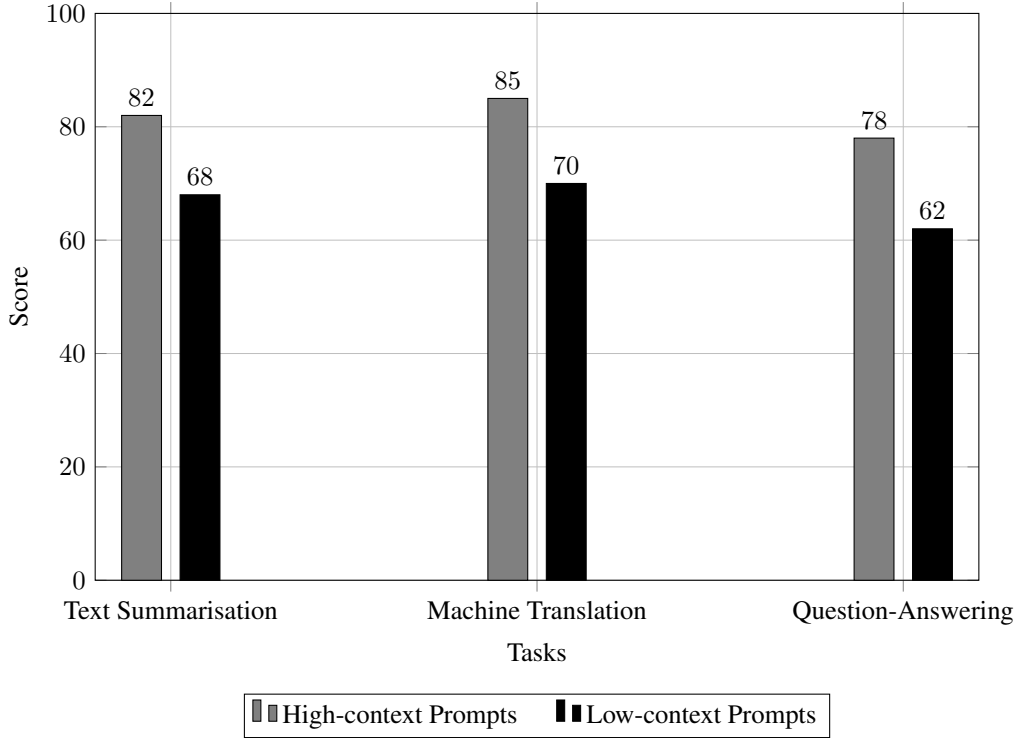


Figure 2: Performance Comparison of High-context and Low-context Prompts Across Tasks

82. In comparison, the low-context prompt achieved an average F1 score of 65 and a precision score of 60, highlighting the critical role of context in generating precise and informative answers.

The findings from these case studies illustrate the profound influence of prompt design and contextual information on the performance of large language models. By systematically varying the prompts and contexts, the study provided valuable insights into the optimal strategies for eliciting desired behaviours and outputs from the Mistral Large model, contributing to the broader understanding of how to effectively utilise large language models in diverse applications.

5 Discussion

5.1 Insights and Observations

The experiments conducted in this study provided numerous insights into the impact of prompt design and contextual information on the performance of the Mistral Large language model. One of the most significant observations was the profound effect that clear and contextually rich prompts had on the model’s accuracy and coherence across various tasks. Well-structured prompts, which included explicit instructions and relevant contextual information, consistently resulted in higher performance metrics compared to vague or minimal prompts. This highlights the importance of prompt clarity and contextual relevance in guiding the model to produce accurate and coherent outputs. Unexpected findings emerged in the analysis of machine translation tasks, where the inclusion of surrounding sentences or paragraphs as context significantly improved the fluency and precision of translations. This suggests that the model benefitted greatly from additional contextual information, which enabled it to better understand the nuances of the source text. However, challenges were encountered in balancing the amount of context provided, as excessive context sometimes led to information overload, causing the model to generate less focused responses. Another notable observation was the impact of prompt specificity on question-answering tasks. Prompts that included detailed background information and specific questions resulted in more precise and comprehensive answers. This indicates that the model’s ability to generate accurate responses is heavily influenced by the quality and richness of the contextual information embedded within the prompts. Additionally,

the experiments revealed that variations in prompt phrasing and structure could significantly affect the model’s performance, underscoring the need for careful prompt engineering to achieve optimal results.

5.2 Challenges and Unexpected Findings

Several challenges and unexpected findings were encountered during the course of the experiments. One of the primary challenges involved determining the optimal balance between providing sufficient context and avoiding information overload. While rich contextual information generally enhanced the model’s performance, there were instances where too much context led to confusion and less relevant responses. This suggests a need for developing techniques to effectively manage and present contextual information to the model. Another challenge was related to the variability in the model’s performance across different tasks. Although well-structured prompts consistently improved accuracy and coherence, the degree of improvement varied depending on the task. For example, in text summarisation tasks, the model showed significant gains in coherence with high-context prompts, whereas the improvement in machine translation tasks was more pronounced in terms of fluency. This variability indicates that different tasks may require tailored prompting strategies to achieve the best results. Unexpected findings also included the observation that the model’s performance could be influenced by subtle changes in prompt phrasing. Minor alterations in the wording or structure of prompts sometimes led to significant differences in the quality of the model’s outputs. This highlights the sensitivity of large language models to the nuances of prompt design and suggests that further research is needed to understand the underlying mechanisms that drive this sensitivity.

5.3 Implications for Future Research

The findings from this study have several implications for future research in the field of prompt engineering and in-context learning. One of the key implications is the importance of developing systematic methods for designing and evaluating prompts. The variability in the model’s performance across different tasks suggests that there is no one-size-fits-all approach to prompt engineering. Future research should focus on identifying task-specific prompting strategies that can be tailored to enhance the performance of large language models in diverse applications. Another important area for future research is the development of techniques to optimise the amount and presentation of contextual information. The challenges encountered in managing context indicate a need for more sophisticated methods to dynamically adjust the context provided to the model based on the specific requirements of the task. This could involve the use of machine learning algorithms to automatically generate and refine contextual information in real-time. Additionally, the sensitivity of the model’s performance to prompt phrasing suggests that further studies are needed to explore the linguistic and cognitive factors that influence how large language models interpret and respond to prompts. Understanding these factors could lead to the development of more effective prompting techniques that leverage the inherent strengths of the models while mitigating their weaknesses.

5.4 Potential Directions for Further Studies

Building on the insights and challenges identified in this study, several potential directions for further research can be proposed. One promising direction is the exploration of adaptive prompting techniques, where prompts are dynamically adjusted based on the model’s real-time responses. This could involve the use of reinforcement learning algorithms to iteratively refine prompts, thereby improving the model’s performance through continuous feedback loops. Another potential direction is the investigation of cross-task generalisation in prompt engineering. By studying how effective prompting strategies for one task can be adapted and applied to other tasks, researchers can develop more generalisable prompting frameworks that enhance the versatility of large language models. This could involve comparative studies across a wide range of tasks to identify common principles and best practices in prompt design.

Furthermore, the development of more granular evaluation metrics that capture the nuanced aspects of the model’s performance could provide deeper insights into the effectiveness of different prompting strategies. Metrics that go beyond accuracy, coherence, and relevance, such as those assessing creativity, persuasiveness, or ethical considerations, could be particularly valuable in evaluating the broader impact of large language models in real-world applications. Lastly, interdisciplinary

collaborations that integrate insights from linguistics, cognitive science, and artificial intelligence could lead to more holistic approaches to prompt engineering. By leveraging the strengths of multiple disciplines, researchers can develop more robust and effective methods for guiding large language models to generate desired outputs across a diverse array of tasks and contexts.

6 Conclusion

The research conducted in this study has provided substantial insights into the impact of prompt design and in-context learning techniques on the performance of the Mistral Large language model. Key findings indicate that well-structured prompts, which include clear instructions and relevant contextual information, significantly enhance the model’s accuracy, coherence, and relevance across a variety of tasks. Specifically, the experiments demonstrated that prompts tailored to the specific requirements of tasks such as text summarisation, machine translation, and question-answering can substantially improve the quality of the model’s outputs. Effective prompting and in-context learning have been reinforced as critical elements in unlocking the full potential of large language models. The ability of prompts to guide the model towards generating precise, coherent, and contextually appropriate responses underscores the necessity of meticulous prompt engineering. Additionally, the incorporation of rich contextual information within prompts further enhances the model’s performance by providing necessary background and detail.

The research shows the transformative potential of well-crafted prompts and contextual information in enhancing the capabilities of large language models. By advancing our understanding of how to effectively guide and inform these models, new possibilities for their application in a wide array of domains can be unlocked, ultimately contributing to the development of more intelligent, adaptive, and useful AI systems. Final thoughts on the future of large language models suggest a continued emphasis on refining prompt design and in-context learning techniques. As language models become increasingly sophisticated and integrated into diverse applications, the development of adaptive prompting strategies that can dynamically adjust to real-time interactions will be essential. Exploring cross-task generalisation of effective prompting techniques and developing more granular evaluation metrics will provide deeper insights into optimizing large language models.

References

- [1] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, “Recent advances in natural language processing via large pre-trained language models: A survey,” *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.
- [2] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [3] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao *et al.*, “Visionllm: Large language model is also an open-ended decoder for vision-centric tasks,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [4] G. Kim, P. Baldi, and S. McAleer, “Language models can solve computer tasks,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [5] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, and J.-R. Wen, “Large language models for information retrieval: A survey,” *arXiv preprint arXiv:2308.07107*, 2023.
- [6] C.-W. Kuo, Y.-F. Huang, and H.-C. Tsai, “Adaptive query contextualization algorithm for enhanced information retrieval in alpaca llm,” 2023.
- [7] G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang, “Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 5168–5191, 2023.
- [8] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, and J. Lanier, “Llmr: Real-time prompting of interactive worlds using large language models,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–22.
- [9] A. Sticha, “Utilizing large language models for question answering in task-oriented dialogues,” 2023.

- [10] I. D. Contreras Perez, “Large language models for ambiguity detection and resolution in smart homes,” 2024.
- [11] I. Frincu *et al.*, “In search of the perfect prompt,” 2023.
- [12] K. Yang, “Controlling long-form large language model outputs,” 2023.
- [13] S. Parmar and H. Patel, “Prompt engineering for large language model,” 2024.
- [14] R. Horst, “User simulation in task-oriented dialog systems based on large language models via in-context learning,” 2024.
- [15] A. Caballero Hinojosa, “Exploring the power of large language models: News intention detection using adaptive learning prompting,” 2023.
- [16] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge, “The inadequacy of reinforcement learning from human feedback-radicalizing large language models via semantic vulnerabilities,” *IEEE Transactions on Cognitive and Developmental Systems*, 2024.
- [17] W.-L. Cheung and C.-Y. Luk, “Implementing automated error correction and feedback loops in kimi, a chinese large language model,” 2024.
- [18] S. I. Ross, F. Martinez, S. Houde, M. Muller, and J. D. Weisz, “The programmer’s assistant: Conversational interaction with a large language model for software development,” in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 491–514.
- [19] S. Barua, “Exploring autonomous agents through the lens of large language models: A review,” *arXiv preprint arXiv:2404.04442*, 2024.
- [20] X. Ma, G. Fang, and X. Wang, “Llm-pruner: On the structural pruning of large language models,” *Advances in neural information processing systems*, vol. 36, pp. 21 702–21 720, 2023.
- [21] B. Wang, J. Liu, J. Karimnazarov, and N. Thompson, “Task supportive and personalized human-large language model interaction: A user study,” in *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, 2024, pp. 370–375.
- [22] E. Jarvinen, “Long-input summarization using large language models,” 2024.
- [23] B. Wang, “Towards trustworthy large language models,” 2023.
- [24] T. Li, “Explore the in-context learning capability of large language models,” 2024.
- [25] M. M. Ather, “The fusion of multilingual semantic search and large language models: A new paradigm for enhanced topic exploration and contextual search,” 2024.
- [26] W. Zheng, “Modeling context and knowledge for dialogue generation.”
- [27] Y. Yan, P. Zheng, and Y. Wang, “Enhancing large language model capabilities for rumor detection with knowledge-powered prompting,” *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108259, 2024.
- [28] D. T. Do, “An efficient fine-tuning method for hierarchical semantic parsing in task-oriented dialog system by structure-aware boosting and grammar constraints,” 2023.
- [29] J. Qiu, H. Wang, Z. Hong, Y. Yang, Q. Liu, and X. Wang, “Controlrec: Bridging the semantic gap between language model and personalized recommendation,” *arXiv preprint arXiv:2311.16441*, 2023.
- [30] L. Wu, Z. Qiu, Z. Zheng, H. Zhu, and E. Chen, “Exploring large language model for graph data understanding in online job recommendations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 9178–9186.
- [31] A. Yadav, “A comprehensive review on large language models,” *Advancing Software Engineering Through AI, Federated Learning, and Large Language Models*, p. 17, 2024.