Camila Fernanda de Lima Anacleto

Prompt Engineering Strategies for Improving AI Performance on Competitive Exam Questions

SÃO PAULO
2025

Camila Fernanda de Lima Anacleto

Prompt Engineering Strategies for Improving AI Performance on Competitive Exam Questions

Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Information Systems.

Advisor: Prof. Dr. Rafael Will Macedo de Araujo

SÃO PAULO
2025

Resumo

Anacleto, Camila Fernanda de Lima. **Prompt Engineering Strategies for Improving AI Performance on Competitive Exam Questions**. 2025. N° de páginas: 13. TCC (Graduação) – Curso Sistema de Informação, Instituto de Tecnologia e Liderança, São Paulo, 2025.

A engenharia de prompts tem se consolidado como uma técnica central para o aprimoramento das capacidades de modelos de linguagem de grande porte (Large Language Models LLMs) e de modelos de visão computacional. Estudos recentes demonstram que estratégias como *zero-shot*, *few-shot*, *Chain-of-Thought* e padrões de prompts contribuem significativamente para a melhoria do desempenho desses modelos em diferentes domínios de aplicação. Ademais, a engenharia de prompts visuais estende tais princípios aos modelos de visão de grande escala, possibilitando uma adaptação mais eficiente para tarefas baseadas em imagens. Para além dos aspectos técnicos, a literatura ressalta a relevância da alfabetização em inteligência artificial, evidenciando que maiores níveis de conhecimento e treinamento dos usuários estão diretamente associados à formulação de prompts mais eficazes e à obtenção de resultados superiores. Nesse contexto, a engenharia de prompts configura-se não apenas como um paradigma de programação, mas também como uma competência estratégica nos âmbitos educacional e profissional. Este trabalho tem como objetivo investigar a aplicação da engenharia de prompts em cenários práticos relacionados a processos de recrutamento no setor de tecnologia. Para tanto, são reproduzidos desafios de programação utilizados em etapas reais de seleção para cargos de níveis júnior, pleno e sênior, os quais são submetidos a diferentes LLMs, permitindo a avaliação comparativa de seus desempenhos. Busca-se, assim, relacionar os fundamentos teóricos da literatura com resultados experimentais, contribuindo para uma compreensão mais aprofundada das potencialidades e limitações da engenharia de prompts em contextos profissionais.

Palavras-Chave: Engenharia de prompts. Modelos de linguagem de grande porte. Alfabetização em inteligência artificial. Processos de recrutamento. Processamento de linguagem natural.

Abstract

Anacleto, Camila Fernanda de Lima. **Prompt Engineering Strategies for Improving AI Performance on Competitive Exam Questions**. 2025. Number of pages: 13. Final course project (Bachelor) – Course Information Systems, Institute of Technology and Leadership, São Paulo, 2025.

Prompt engineering has emerged as a key technique for enhancing the performance and applicability of large language models (LLMs) and vision models in diverse professional and academic contexts. This work investigates prompt engineering as both a technical methodology and a strategic skill, with a specific focus on its application in hiring processes within the technology sector. The object of study is the interaction between prompt design strategies and model performance when solving programming challenges representative of real recruitment stages. The main objective is to analyze how different prompt engineering techniques influence the quality, accuracy, and consistency of solutions generated by LLMs across junior, mid-level, and senior-level tasks. The study adopts an experimental and comparative methodology, in which multiple LLMs are prompted to solve identical programming challenges derived from real-world technical hiring scenarios. Different prompt strategies, including zero-shot, few-shot, and Chain-of-Thought approaches, are systematically applied and evaluated based on predefined performance criteria. The results indicate that prompt structure and contextualization significantly affect model outputs, with more advanced prompting techniques leading to improved reasoning, clarity, and task alignment, particularly in complex problem-solving scenarios. Additionally, the findings reinforce the role of AI literacy as a critical factor in maximizing model effectiveness, highlighting the importance of human expertise in guiding model behavior. The study concludes that prompt engineering represents not only a technical optimization approach but also a relevant professional competency, offering practical insights for organizations, educators, and practitioners seeking to integrate LLMs into recruitment and decision-making processes more effectively.

Keywords: Prompt Engineering; Large Language Models; AI Literacy; Hiring Processes; Natural Language Processing.

List of Tables

Summary

1 Introduction

The rapid evolution of large language models (LLMs) has significantly expanded the scope of artificial intelligence applications in domains related to reasoning, learning, and problem-solving. Trained on large-scale textual corpora, these models demonstrate a remarkable ability to generate coherent and contextually appropriate responses. Nevertheless, their performance is not solely determined by their internal architecture or training data, but is strongly influenced by the manner in which instructions are formulated. This interaction paradigm is known as *prompt engineering*.

Prompt engineering can be defined as the systematic design of instructions that guide language models toward responses that are more accurate, relevant, and aligned with a specific task. Recent studies indicate that carefully structured prompts can substantially enhance reasoning quality, reduce ambiguity, and improve output consistency. In educational contexts, particularly those involving high cognitive demand, prompt engineering has emerged as a relevant methodological tool.

Brazilian public service examinations, commonly referred to as *concursos públicos*, represent a highly competitive and structured assessment environment. These exams cover a wide range of disciplines and frequently require logical reasoning, conceptual precision, and domain-specific knowledge. Candidates often face difficulties not only in mastering content but also in adopting effective study strategies. Within this context, artificial intelligence tools may offer meaningful support when employed in a structured and pedagogically oriented manner.

This study investigates the application of prompt engineering strategies to real questions from Brazilian public service exams. The objective is to analyze how different prompting techniques affect the accuracy, clarity, and educational usefulness of model-generated responses. All experiments were conducted using ChatGPT-4 (GPT-4). The study seeks to contribute both to the academic literature on prompt engineering and to the practical use of AI as a learning aid for competitive exam preparation.

2 Development

2.1 Theoretical Framework and Research Context

Large language models generate responses by estimating probabilistic relationships between linguistic tokens based on patterns learned during training. Although these models exhibit strong generalization capabilities, their outputs are not inherently optimized for specific reasoning tasks unless appropriate contextual guidance is provided. Prompt engineering functions as a mechanism to shape model behavior without modifying model parameters, relying instead on instruction design and contextual framing.

The literature identifies several prompting strategies with distinct characteristics. Zero-shot prompting presents a task without prior examples, relying entirely on the model's internal representations. Few-shot prompting introduces a limited number of examples that illustrate the expected response format, allowing the model to infer task structure. Chain-of-Thought prompting explicitly encourages step-by-step reasoning, making intermediate cognitive processes visible and often improving performance on complex problems. Self-Consistency builds upon Chain-of-Thought by generating multiple independent reasoning paths and selecting the most frequent or coherent outcome, thereby increasing stability and reliability.

In educational applications, these strategies differ not only in performance metrics but also in pedagogical value. Approaches that expose reasoning steps tend to support deeper conceptual understanding, whereas answer-focused strategies may limit learning to surface-level pattern recognition. This study is grounded in this theoretical framework and examines how these prompting techniques perform when applied to authentic competitive exam questions.

2.2 Methodology

The methodological design of this study emphasizes transparency, reproducibility, and adherence to established research practices in prompt engineering and in-context learning.

Model and Computational Environment

All experiments were conducted using ChatGPT-4 (GPT-4). The model was accessed through its standard interface, and default generation parameters were maintained throughout the experiments to avoid bias introduced by parameter tuning.

Dataset Construction

A proprietary dataset was developed specifically for this research. The dataset consists of 680 real multiple-choice questions collected from Brazilian public service examinations. Questions were manually gathered from verified public sources, including official examining boards and archived exam materials. Each item was reviewed to ensure authenticity, clarity, and relevance.

The dataset includes structured metadata such as subject area, year of application, examining board, exam title, required educational level, booklet version, and official answer key. This organization enables systematic filtering, traceability, and comparative analysis.

Experimental Sample

From the complete dataset, a subset of 100 questions was selected for experimental evaluation. The sample was designed to represent a variety of subjects, exam boards, and difficulty levels, thereby reflecting realistic exam conditions.

Prompting Strategies

Each question in the experimental sample was submitted to the model using four prompting strategies: Zero-shot, Few-shot, Chain-of-Thought, and Self-Consistency. Each strategy followed a standardized prompt structure to ensure methodological consistency across conditions.

Evaluation Criteria

Model outputs were compared directly with the official answer keys provided by the examining boards. Accuracy was calculated as the proportion of correctly answered questions for each prompting strategy. In addition to quantitative accuracy, qualitative analysis was conducted to examine representative reasoning patterns and error types.

2.3 Results

This section presents the empirical results obtained from applying the four prompting strategies to the experimental sample of 100 competitive exam questions.

The results reveal substantial variation in performance across prompting techniques. Zero-shot prompting achieved an accuracy of 70 percent, correctly answering 70 questions. In contrast, Few-shot, Chain-of-Thought, and Self-Consistency prompting each achieved 100 percent accuracy, correctly answering all questions in the sample.

Table 1 Accuracy of prompting strategies on competitive exam questions.

Prompting Strategy	Correct Answers	Accuracy
Zero-shot	70	70%
Few-shot	100	100%
Chain-of-Thought	100	100%
Self-Consistency	100	100%

These findings indicate a clear distinction between unstructured prompting and approaches that incorporate contextual guidance or explicit reasoning.

2.4 Analysis or Discussion of Results

The performance differences observed across prompting strategies underscore the central role of prompt design in activating effective reasoning

processes within large language models. The lower accuracy associated with Zero-shot prompting suggests that, in the absence of guidance, the model may rely on superficial pattern recognition, which can result in conceptual misinterpretations.

A representative example involves a question related to computer memory hierarchy from a Brazilian public service exam. In this case, Zero-shot prompting led to an incorrect assessment of access time, storage capacity, and cost efficiency, while all other strategies produced the correct answer. This pattern was observed consistently among Zero-shot errors.

Chain-of-Thought and Self-Consistency emerged as the most effective strategies, not only in terms of accuracy but also in their educational value. By making reasoning steps explicit, these approaches allow learners to follow the logical structure underlying each answer. Self-Consistency further enhances robustness by aggregating multiple reasoning paths, thereby reducing variability and the impact of isolated reasoning errors.

Few-shot prompting also demonstrated strong performance; however, its pedagogical contribution is more limited, as it does not explicitly reveal the reasoning process. While examples guide the model toward correct responses, they provide less insight into the underlying logic compared to reasoning-oriented strategies.

These results are consistent with prior research indicating that prompts encouraging explicit reasoning tend to produce more reliable and interpretable outcomes. In the context of competitive exam preparation, such characteristics are particularly relevant, as effective learning depends on understanding, not merely on answer correctness.

3 Conclusion

This study demonstrates that the effectiveness of large language models in solving competitive exam questions is highly dependent on the prompting strategy employed. Although ChatGPT-4 exhibits strong baseline capabilities, unstructured prompting may lead to inconsistent or conceptually flawed responses.

The findings indicate that Chain-of-Thought and Self-Consistency are the most effective prompting strategies for exam preparation. Both approaches achieved full accuracy in the experimental sample and supported transparent reasoning processes that facilitate deeper learning. By contrast, Zero-shot prompting proved less reliable, and Few-shot prompting, while accurate, offered limited explanatory value.

By leveraging a dataset of 680 real public service exam questions and an experimental sample of 100 questions, this research provides both methodological rigor and practical relevance. The results suggest that candidates preparing for Brazilian public service exams can significantly improve their study strategies by adopting prompting techniques that emphasize structured reasoning.

Future research may explore larger experimental samples, alternative prompt formulations, or comparative analyses across different language models. As artificial intelligence continues to be integrated into educational practices, understanding how to interact with these systems in a strategic and informed manner becomes increasingly important.

References

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*, 100225.

Marvin, G., Nakayiza, H. R., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In *Data Intelligence and Cognitive Informatics* (pp. 387–402). Springer Nature Singapore.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint*, arXiv:2402.07927.

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., et al. (2023). Review of large vision models and visual prompt engineering. *Meta-Radiology*, *1*, 100047.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint*, arXiv:2302.11382.