INSTITUTE OF TECHNOLOGY AND LEADERSHIP

INFORMATION SYSTEMS

PROMPT ENGINEERING FOR CIVIL SERVICE EXAMS: EVALUATING LARGE LANGUAGE MODELS IN MULTIPLE-CHOICE QUESTION SOLVING

An experimental study on AI prompt optimization for public exam preparation in Brazil

CAMILA FERNANDA DE LIMA ANACLETO

São Paulo

2025

CAMILA FERNANDA DE LIMA ANACLETO

PROMPT ENGINEERING FOR CIVIL SERVICE EXAMS: EVALUATING LARGE LANGUAGE MODELS IN MULTIPLE-CHOICE QUESTION SOLVING

An experimental study on AI prompt optimization for public exam preparation in Brazil

Course conclusion paper submitted to the Institute of Technology and Leadership (Inteli), as a partial requirement for obtaining the degree of Bachelor in Information Systems.

Advisor: Prof. Dr. Rafael Will Macedo de Araujo

São Paulo
2025

ABSTRACT

Prompt engineering has emerged as a central technique to enhance the reasoning and problem-solving capabilities of large language models (LLMs). Recent surveys and systematic studies demonstrate how strategies such as zero-shot, few-shot, Chain-of-Thought, and prompt patterns can substantially improve model performance across diverse domains (Sahoo et al., 2024; White et al., 2023; Marvin et al., 2024). Furthermore, the growing field of visual prompt engineering extends these advances to large vision models, enabling efficient adaptation for multimodal tasks (Wang et al., 2023).

Beyond technical methodologies, research emphasizes the relevance of AI literacy, showing that user expertise and intentional prompt design are directly correlated with improved outcomes (Knoth et al., 2024). This intersection between algorithmic performance and human interaction establishes prompt engineering as both a computational and cognitive skill with potential applications in education and professional training.

This work aims to investigate the role of prompt engineering in improving preparation for civil service exams in Brazil. The study focuses on multiple-choice questions with official answer keys, enabling objective evaluation of model accuracy and prompt effectiveness. By experimenting with various prompting techniques across different question categories and difficulty levels, this research seeks to identify strategies that maximize LLM performance and provide practical insights for optimizing study methods for public examination candidates.

Keywords: Prompt Engineering, Large Language Models, AI Literacy, Civil Service Exams, Artificial Intelligence in Education

SUMMARY

1 Introduction

The rapid advancement and adoption of large language models (LLMs) have reshaped how humans interact with artificial intelligence. These models, trained on massive and diverse datasets, demonstrate remarkable abilities in generating coherent text, solving reasoning problems, and providing contextualized explanations. However, their effectiveness is highly dependent on the quality and structure of the input they receive, a concept increasingly consolidated under the term *prompt engineering* (Marvin et al., 2024).

Prompt engineering refers to the practice of designing task-specific instructions, or prompts, that guide a model's behavior without modifying its internal parameters (Sahoo et al., 2024). From foundational strategies such as zero-shot and few-shot prompting to advanced methods like Chain-of-Thought and self-consistency, this discipline has evolved rapidly, proving essential for optimizing model performance across various applications. Structured methodologies, such as prompt pattern catalogs, further enhance prompt consistency, reuse, and effectiveness in applied contexts (White et al., 2023).

In addition to technical methods, research highlights the importance of AI literacy as a determinant of successful human–AI collaboration. Individuals who understand the capabilities and limitations of LLMs are more capable of crafting precise prompts and interpreting outputs effectively (Knoth et al., 2024). This connection positions prompt engineering not only as a technical methodology but also as an emerging cognitive and educational skill.

In the Brazilian context, civil service exams (*concursos públicos*) play a central role in the recruitment and qualification of professionals for government institutions. These exams are highly competitive and represent one of the most stable and prestigious career paths in the country, attracting millions of candidates each year. Most exams rely on multiple-choice questions with predefined answer keys, which makes them ideal for objective evaluation of both human and artificial intelligence performance.

The combination of prompt engineering and civil service exams presents an opportunity to explore how LLMs can support candidates in their preparation process. By leveraging effective prompting strategies, these models can help simulate realistic test

conditions, generate explanations, and provide personalized study support. Furthermore, the structured and repetitive nature of multiple-choice questions offers a controlled environment for evaluating model accuracy, reasoning, and prompt sensitivity.

This study aims to bridge theoretical insights from the field of prompt engineering with practical experimentation using civil service exam datasets. By evaluating different prompting techniques and comparing model outputs to official answer keys, this research seeks to identify the most effective strategies for enhancing learning outcomes and supporting students preparing for public examinations. Ultimately, the work contributes to understanding how prompt engineering can democratize access to quality study tools, fostering greater inclusion, efficiency, and innovation in education through artificial intelligence.

2 Background and Related Work

The growing field of prompt engineering has attracted considerable attention as researchers and practitioners seek to optimize the interaction between humans and large language models (LLMs). Recent studies emphasize that model performance is not determined solely by the underlying architecture, but also by how tasks are formulated through natural language instructions.

A systematic survey by Sahoo et al. (2024) maps the landscape of prompt engineering, highlighting techniques such as zero-shot, few-shot, chain-of-thought, and automatic prompt optimization. Their results show how these approaches address different challenges in reasoning, text generation, and domain adaptation. Complementing this work, White et al. (2023) propose a catalog of *prompt patterns*, offering reusable templates that allow practitioners to design more consistent and effective interactions with LLMs.

From a broader perspective, Marvin et al. (2024) discuss prompt engineering not only as a technical method, but as a paradigm that reframes how humans interact with intelligent systems. They emphasize the dual role of prompt engineering: guiding models while simultaneously reflecting human creativity and problem-solving strategies. This view aligns with Knoth et al. (2024), who introduce the concept of AI literacy and its implications for prompt engineering. Their findings indicate that user expertise and training are crucial factors, with more knowledgeable users able to design prompts that reduce ambiguity and mitigate risks of bias or hallucination.

While much of the literature focuses on text-based models, visual prompt engineering has also emerged as an important research area. Wang et al. (2023) review large vision models and demonstrate how prompt-based techniques can be extended beyond natural language to multimodal tasks, enabling efficient adaptation of models to image-based applications. This signals a trend toward expanding prompt engineering into diverse modalities and contexts.

Taken together, these studies demonstrate the multifaceted nature of prompt engineering. Effective prompting combines methodological rigor, user literacy, and cross-domain adaptability, making it a promising research avenue not only for advancing technical performance but also for exploring practical applications in education, recruit-

ment, software engineering, and other professional contexts.

3 Dataset Development

To conduct the experiments proposed in this study, a custom database will be created containing multiple-choice questions from various Brazilian civil service exams. The objective is to build a structured and diverse dataset that reflects the range of difficulty, topics, and reasoning styles typically found in national examinations.

The dataset will include questions extracted from public and reputable sources, such as past exams from institutions like CESPE/Cebraspe, FGV, Vunesp, and FCC, among others. These institutions are known for their methodological rigor and represent different patterns of question formulation, allowing for a broad evaluation of the models' adaptability to distinct linguistic and logical styles.

Data will be collected manually and, when possible, through semi-automated web scraping tools designed to extract question text and answer keys from public repositories and official exam archives. After extraction, all entries will be reviewed to ensure accuracy, completeness, and proper formatting before integration into the main dataset.

The resulting database will serve as the foundation for prompt engineering experiments, where different LLMs will be tasked with answering the same set of questions under controlled prompting conditions (e.g., zero-shot, few-shot, and Chain-of-Thought). Each experiment will record model responses, reasoning explanations, and confidence levels when available. This structure will enable systematic comparison across models and prompt strategies, supporting quantitative analysis of accuracy and qualitative analysis of reasoning coherence.

Ultimately, the creation of this dataset contributes not only to the experiments of this research but also to the broader academic community by offering a reproducible and extensible benchmark for studying LLM performance in educational and assessment contexts in Brazil.

REFERENCES

Knoth, N., Tolzin, A., Janson, A., and 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., and Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In *Data Intelligence and Cognitive Informatics*, pages 387–402. Springer Nature Singapore.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and 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., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.