Proposal

Degree Program Computer Science Bachelor Dual

**Optimizing LLM Performance in Brazilian University Entrance Exams: A Comparative Study of CoT, CoVe and Self-Refine Prompting**

By: Rafaela Rolim Santana  
Student number: 2210257114

Advisor: BSc. Patrick Link

Vienna, 24.02.2025

# Introduction

## Background and Context

Large Language Models (LLMs) are becoming integral to educational settings, assisting students in learning, problem-solving, and exam preparation. Their ability to generate responses based on vast amounts of data makes them promising tools for improving learning outcomes. However, their effectiveness in structured academic assessments, such as the annual Brazilian University Entrance Exam (ENEM), remains underexplored. These exams cover diverse subjects, including mathematics, natural sciences, humanities, and languages, each requiring different reasoning styles.

Prompting techniques such as **Chain-of-Thought (CoT)** [1]**, Chain-of-Verification (CoV)** [2]**,** and **Self-Refine** [3] have been developed to enhance LLM reasoning, accuracy, and consistency. While these techniques have demonstrated improvements in various domains, there is limited research assessing their comparative effectiveness in solving multiple-choice questions across different subject areas.

## Problem

Despite their potential, LLMs are prone to errors, hallucinations, and inconsistencies in their responses [4]. Students often rely on LLMs to learn and need reliable and accurate responses. The challenge is identifying which prompting techniques yield the most precise answers in a structured testing environment and whether their effectiveness varies across different academic subjects.

This study aims to address two key research questions:

* How do different prompting techniques affect LLM response accuracy on Brazilian University Entrance Exams?
* How does the effectiveness of prompting techniques vary across different subject areas (objective vs. subjective subjects)?

By systematically analyzing LLM performance across multiple subjects, this research seeks to determine the best strategies for improving answer accuracy and consistency in an educational context.

## Relevance to the Field

As LLMs continue to shape the future of education, understanding how to optimize their responses is crucial for their responsible and effective implementation. This research contributes to the growing field of **AI in education** by providing empirical insights into LLM prompting strategies, bridging the gap between **machine learning advancements and practical educational applications.**

By evaluating multiple prompting techniques and their impact on different subject areas, this study will help educators and students refine how LLMs are used for learning and assessment. The findings will also contribute to a **framework for evaluating prompting techniques** and **guidelines for their application in standardized testing environments**, ensuring that AI-driven educational tools are both effective and trustworthy.

## Current State of Research

Several advanced prompting techniques have been developed to improve the **accuracy, reasoning, and consistency** of **Large Language Models (LLMs)** in complex problem-solving tasks. These approaches address different challenges, including **hallucination reduction, step-by-step reasoning, and iterative self-improvement,** making them particularly relevant in educational applications such as standardized exams.

### ****Chain-of-Thought (CoT) Prompting****

**Chain-of-Thought (CoT) prompting** encourages LLMs to generate **intermediate reasoning steps** before arriving at a final answer. This structured reasoning process enhances **problem-solving accuracy** across tasks requiring logical deductions, including **math, commonsense, and symbolic reasoning** [1]. **Their research** demonstrated that CoT significantly improves LLM performance compared to standard direct-answer prompting.

### ****Self-Consistency****

**Self-Consistency** builds on CoT by generating **multiple reasoning paths** for the same query and selecting the **most consistent** response among them. This method reduces variance in model outputs and increases reliability in **complex reasoning tasks. Their research** showed that self-consistency leads to higher accuracy than single-path CoT by filtering out **inconsistent or erroneous** reasoning chains [5].

### ****Chain-of-Verification (CoVe)****

**Chain-of-Verification (CoVe)** aims to **reduce hallucinations** by introducing a **verification phase** after the initial response. The model generates verification questions, answers them **independently**, and refines its initial response based on the consistency of its verification steps. This **fact-checking mechanism** ensures that responses are **more reliable and evidence-based** and that itsignificantly improves factual accuracy across various knowledge-based tasks [2].

### ****Self-Refine****

**Self-Refine** enables LLMs to iteratively **analyze and improve their own outputs** by generating **self-feedback and refinements** in multiple cycles. Unlike CoT, which relies on a **single-step reasoning process**, Self-Refine allows models to **correct mistakes, enhance clarity, and optimize responses** dynamically. **Research** shows that this approach leads to **better-structured and more accurate responses**, particularly in long-form text generation [3].

### ****Self-Reflection****

**Self-Reflection** further enhances model reasoning by enabling **introspection**—the model systematically evaluates its own responses, identifies errors, and **adjusts its reasoning process**. This iterative method improves performance in **logic-based** tasks and helps the model refine its understanding **without external input** [6]. **Research** found that self-reflection significantly **reduces logical inconsistencies** and enhances model confidence in its answers [6].

# Research Question

## Goals

This research aims to systematically evaluate the effectiveness of different prompting techniques when applied to Large Language Models (LLMs) in educational assessment contexts. By analyzing performance across various academic subjects using Brazilian university entrance exams as a benchmark, the research seeks to develop evidence-based guidelines for optimizing LLM interactions in educational applications. The findings will contribute to understanding how to effectively leverage LLMs for educational purposes while maintaining accuracy and reliability.

## Primary Research Question

How do different prompting techniques affect the accuracy and consistency of LLM responses when solving Brazilian university entrance exam questions across different knowledge areas?

## Secondary Research Question

How does the effectiveness of prompting techniques vary across different subject areas?

## Hypotheses

### Primary Hypothesis

**H1**: Structured prompting techniques that incorporate verification steps, such as Chain-of-Verification and Self-Refine, will yield higher accuracy rates compared to simple Chain-of-Thought prompting across all subject areas.

### Secondary Hypotheses

**H2a**: Verification-based prompting techniques will show greater improvement in objective subjects (Maths and Natural Sciences) than in subjective subjects (Languages and Human Sciences) when compared to Chain-of-Thought prompting.

**H2b**: The consistency of LLM responses (measured by repeated trials) will be higher with verification-based prompting techniques compared to Chain-of-Thought prompting across all subject areas.

# Methods

This study employs a quantitative experimental design to evaluate the performance of different prompting techniques on LLM responses to university entrance exam questions. The experiment will compare three prompting techniques: Chain-of-Thought (CoT) [1], Chain-of-Verification (CoVe) [2], and Self-Refine [3] across different subject areas.

## Dataset

The study will use the 2023 and 2024 ENEM dataset for the experiment, each containing 180 questions on Languages, Human Sciences, Natural Sciences, and Mathematics. This data was not available during the LLMs' training period, contributing for the integrity of the experiment results.

## Implementation of Prompting Techniques

### Chain-of-Thought (CoT)

* Implementation of standard CoT prompting [1] with step-by-step reasoning
* Few-shot examples provided for each subject area
* Format:
  + Question → Reasoning Steps → Final answer

### Chain-of-Verification (CoVe)

* Implementation of Chain-of-Verification [2] methodology
* Four-step process:
* Draft initial response
* Plan verification questions to fact-check the draft
* Answer verification questions independently
* Generate final verified response
* Format:
  + Question → Baseline Answer → Plan Verification Questions → Execute Verifications → Final Answer

### Self-Refine

* Implementation of Self-Refine [3] methodology
* Three-step process:
  + Generate initial output
  + Receive feedback on initial output
  + Refine output according to feedback
* Format:
  + Question → Baseline Answer → Self-Feedback → Final Answer

## Evaluation Metrics

### Accuracy

* Answer accuracy: Percentage of correct answers compared to ground truth
* Subject-wise accuracy: Performance breakdown by subject area
* Comparative accuracy: Improvement over baseline (CoT)

### Consistency

* Response consistency score: Percentage of identical answers across multiple runs
* Standard deviation of responses for each question

## Experimental Procedure

### Prepare Data

1. Question categorization by subject area
2. Removal of questions requiring image comprehension
3. Standardization of question format
4. Development of few-shot examples for each prompting technique

### Test Protocol

1. Each question will be tested with all three prompting techniques
2. Multiple runs (n=5) per question to assess consistency
3. Use of controlled temperature and sampling parameters

# Expected Results

## Primary Results

Verification-based prompting techniques (Chain-of-Verification and Self-Refine) will outperform basic Chain-of-Thought prompting across all subjects, with estimated:

* 10-15% improvement in overall accuracy
* 85% consistency in responses across multiple runs

## Subject-Specific Performance

* Mathematics/Natural Sciences:
  + 15-20% accuracy improvement
  + 90% consistency across runs
  + More reliable self-verification
* Languages/Human Sciences:
  + 5-10% accuracy improvement
  + 70-80% consistency across runs
  + Higher variation in verifications

## Practical Outcomes

* Framework for evaluating prompting techniques in education
* Subject-specific prompting guidelines
* Identification of LLM limitations in educational applications
* Better understanding of verification strategies' effectiveness

# Technical Requirements

* LLM: GPT-4 and MaritacaAI via API
* Programming Language: Python
* Key Libraries: OpenAI, Pandas, NumPy, SciPy
* Version Control: Git

# Bibliography

|  |  |
| --- | --- |
| [1] | J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le und D. Zhou, „Chain-of-thought prompting elicits reasoning in large language models,“ *Advances in neural information processing systems,* 2022. |
| [2] | S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz und J. Weston, „Chain-of-verification reduces hallucination in large language models,“ arXiv preprint arXiv:2309.11495, 2023. |
| [3] | A. Madaan, N. T. P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck und A. Yazdanbakhsh, „Self-refine: Iterative refinement with self-feedback,“ *Advances in Neural Information Processing Systems 36,* 2024. |
| [4] | L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin und T. Liu, „A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,“ November 2023. [Online]. Available: https://arxiv.org/abs/2311.05232v2. [Zugriff am 1 December 2024]. |
| [5] | Wang, Xuezhi, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery und D. Zhou, „Self-consistency improves chain of thought reasoning in language models.,“ arXiv preprint arXiv:2203.11171, 2022. |
| [6] | M. Renze und E. Guven, „Self-Reflection in LLM Agents: Effects on Problem-Solving Performance,“ arXiv preprint arXiv:2405.06682, 2024. |