BACHELOR PAPER

Thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science in Engineering at the University of Applied Sciences Technikum Wien Degree Program Computer Science Dual

Comparing Chain-of-Thought, Chain-of-Verification and Self-Refine in solving Brazilian University Entrance Exams

By: Rafaela Rolim Santana

Student Number: 2210257114

Supervisor: Patrick Link, BSc.

Vienna, 02.04.2025

Declaration of Authenticity

“As author and creator of this work to hand, I confirm with my signature knowledge of the relevant copyright regulations governed by higher education acts (see Urheberrechtsgesetz/ Austrian copyright law as amended as well as the Statute on Studies Act Provisions / Examination Regulations of the UAS Technikum Wien as amended).

I hereby declare that I completed the present work independently and according to the rules currently applicable at the UAS Technikum Wien and that any ideas, whether written by others or by myself, have been fully sourced and referenced. I am aware of any consequences I may face on the part of the degree program director if there should be evidence of missing autonomy and independence or evidence of any intent to fraudulently achieve a pass mark for this work (see Statute on Studies Act Provisions / Examination Regulations of the UAS Technikum Wien as amended).

I further declare that up to this date I have not published the work to hand nor have I presented it to another examination board in the same or similar form. I affirm that the version submitted matches the version in the upload tool.”

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Place, Date |  | Digital Signature |

Abstract

Large Language Models (LLMs) are widely used by students to solve academic tasks. While these models can be helpful, they often produce incorrect or misleading answers, a problem known as hallucination. This study explores whether simple prompting strategies can reduce hallucinations and improve answer accuracy without requiring fine-tuning or external tools.

Three methods were evaluated: Chain-of-Thought (CoT), which guides the model to reason step by step; Chain-of-Verification (CoVe), which adds a verification phase to check the initial answer; and Self-Refine, which lets the model review and revise its own responses through feedback loops. Each method was applied to 180 questions from the Brazilian university entrance exam ENEM 2024, covering four subject areas. The test was run 40 times per question to measure not only accuracy but also consistency across repeated runs.

The results showed that Chain-of-Thought achieved the highest overall accuracy and consistency, especially in subjects like Mathematics and Human Sciences. Self-Refine and CoVe showed more variability, and sometimes their refined answer performed worse than the initial answer. These findings suggest that well-designed prompts can help reduce hallucinations in LLMs and improve their reliability in educational settings.

Acknowledgements

**Keywords:** Large Language Models, Prompt Engineering, Chain-of-Thought, Chain-of-Verification, Self-Refine, Hallucination, GPT-3.5 Turbo

I thank my parents, Márcio and Jane, and my sister, Fernanda, for believing in me and giving me the strength to face this journey; and for always being present, despite the ocean between us. This would not have been possible without their support.

I thank my friends Thais, Guilherme and Victor for being great companions, bringing joy into my life and becoming my chosen family in Vienna. I thank my friends from Colegagi for their friendship that goes beyond time and distance. I thank Roger, who inspired me to study Computer Science and stood by me through difficult times.

I thank the lecturers, employees and colleagues from the FH Technikum Wien, for the good time and great learnings. And finally, I thank my colleagues from A1 for the warm reception and for supporting my professional growth.

Table of Contents

[1 Introduction 5](#_Toc198037923)

[1.1 Motivation 5](#_Toc198037924)

[1.2 Tasks 5](#_Toc198037925)

[2 Methodology 6](#_Toc198037926)

[2.1 Ideal Solution 6](#_Toc198037927)

[2.2 Requirements 7](#_Toc198037928)

[2.3 Development Process 9](#_Toc198037929)

[2.3.1 Prepare Dataset 9](#_Toc198037930)

[2.3.2 Define Prompting Templates 9](#_Toc198037931)

[2.3.3 Test Prompting Techniques 10](#_Toc198037932)

[2.3.4 Evaluation and Data Collection 10](#_Toc198037933)

[2.4 Tests 10](#_Toc198037934)

[2.4.1 General Accuracy 10](#_Toc198037935)

[2.4.2 Accuracy by Subject 11](#_Toc198037936)

[2.4.3 Consistency 11](#_Toc198037937)

[2.5 Tools 11](#_Toc198037938)

[3 Solution 12](#_Toc198037939)

[3.1 Answering the Research Questions 12](#_Toc198037940)

[3.2 Prompting Techniques 13](#_Toc198037941)

[3.2.1 Chain-of-Thought (CoT) 13](#_Toc198037942)

[3.2.2 Chain-of-Verification (CoVe) 14](#_Toc198037943)

[3.2.3 Self-Refine 15](#_Toc198037944)

[3.3 Load QA dataset 16](#_Toc198037945)

[3.4 Run Tests 16](#_Toc198037946)

[3.5 Clean Data 16](#_Toc198037947)

[3.6 Analyse Performance 17](#_Toc198037948)

[3.6.1 General Accuracy 17](#_Toc198037949)

[3.6.2 Subject-Specific Accuracy 19](#_Toc198037950)

[3.6.3 Consistency 21](#_Toc198037951)

[3.6.4 Interesting Findings 23](#_Toc198037952)

[4 Discussion 24](#_Toc198037953)

[4.1 Potentials 24](#_Toc198037954)

[4.2 Limitations 25](#_Toc198037955)

[4.3 Possible Improvements 25](#_Toc198037956)

[4.4 Generalizability of the Solution 26](#_Toc198037957)

[4.5 Key Takeaways 26](#_Toc198037958)

[5 Bibliography 27](#_Toc198037959)

[6 Figures 28](#_Toc198037960)

[7 Tables 29](#_Toc198037961)

[List of Abbreviations 30](#_Toc198037962)

[Documentation table of AI-based tools 31](#_Toc198037963)

[A: Heading of Appendix A 32](#_Toc198037964)

[B: Heading of Appendix B 33](#_Toc198037965)

# Introduction

## Motivation

As a Computer Science student in the age of artificial intelligence, I have often turned to Large Language Models (LLMs) like ChatGPT to support my study routine. These tools act as accessible tutors, capable of answering questions, explaining concepts in simpler terms, and adapting to my individual learning pace. Especially attractive is the LLM’s ability to respond without judgment, enabling students like me to ask "dumb" questions, repeat topics as often as needed, and receive immediate, tailored feedback.

However, my positive experiences with LLMs were often met with a critical limitation: the **inconsistency and unreliability in their responses**. While LLMs often produce helpful answers, explanations were sometimes factually incorrect or misleading. In some cases, the error was obvious; in others, I would only recognize it after consulting external sources, raising the concern that I might be unintentionally **learning and reinforcing misconceptions.**

This observation reflects a broader problem: **LLMs can hallucinate**, producing outputs that sound plausible, but are factually incorrect (Ji, et al., 2023). This becomes a significant risk in educational settings, where students might not yet have the background knowledge to detect such inaccuracies. Students are left vulnerable to absorbing incorrect information with full confidence.

This concern is the central motivation of my thesis: **how can we reduce hallucinations in LLMs in a way that is simple and accessible to students?** While many solutions exist, such as fine-tuning models or incorporating retrieval systems, they are not easily accessible for a typical LLM chatbot user. Therefore, I focused on **prompt-based techniques** that require no special infrastructure, training, or external tools, and can be used with standard LLM interfaces like OpenAI's ChatGPT.

The use case I address is straightforward: students using LLMs to help them solve high-school level multiple-choice questions. If prompting techniques can reduce hallucinations and improve the reliability of answers in this setting, they would offer a **practical way to improve LLM-assisted learning** without requiring any technical expertise or system modifications from the user.

## Tasks

The primary goal of this thesis is to **evaluate the effectiveness of prompt-based techniques in reducing hallucinations in Large Language Models (LLMs)**, with a focus on scenarios where students use these models as learning aids. Instead of relying on model retraining or external tools, this work investigates whether **LLM response** reliability can be improved using prompt engineering alone.

To this end, the following research questions are investigated:

**RQ1**: *How does the application of Chain-of-Thought, Chain-of-Verification, and Self-Refine prompting techniques affect LLM response accuracy on Brazilian University Entrance Exams?*

**RQ2**: *How does the effectiveness of these prompting techniques vary across different subject areas?*

**RQ3**: *How consistent are the predictions of each method across repeated runs?*

These questions are explored through a **quantitative experiment** using a standardized high-school level exam (ENEM 2024) with multiple-choice questions across different subjects. This setup allows for automatic LLM response evaluation and comparison across methods.

The main outcome of this work is a **reproducible evaluation framework** implemented in Python Jupyter Notebooks. This framework loads the questions dataset and applies each prompting strategy using the OpenAI API. It then collects results across multiple test runs, and calculates key performance metrics such as accuracy, subject-wise performance, and consistency. All code, prompt templates, and evaluation logic were designed with modularity and transparency in mind, making it straightforward to extend the approach to other datasets or prompting techniques in future studies.

By providing empirical insights and a practical testing pipeline, this thesis aims to contribute to ongoing efforts in **making LLM-assisted learning more trustworthy.**

# Methodology

## Ideal Solution

An evaluation of prompting techniques for Large Language Models (LLMs) requires a controlled and repeatable experimental setup that allows for fair comparisons across methods. In an ideal scenario, the following conditions would be met:

**Fully deterministic LLM behaviour**

Each prompt would yield a predictable and reproducible response, eliminating randomness and isolating the effect of prompt structure alone.

**Access to model internals**

Evaluators would have control over the underlying LLM architecture, token-level probabilities, and intermediate reasoning states, allowing for in-depth analysis.

**Large-scale, domain-specific benchmark datasets**

Evaluation would be conducted on diverse, verified questions across multiple disciplines, ideally with additional information such as reasoning steps and difficulty levels.

**Automated multi-pass evaluation**

The system would support large-scale, automated prompting with batch processing, minimal API latency, and robust error handling.

**Statistical power and interpretability**

Sufficiently large numbers of prompt-response pairs would be generated per method to ensure statistical confidence, and the analysis would include meaningful performance metrics beyond accuracy, such as consistency and reasoning quality.

In practice, evaluating LLM behavior under these ideal conditions is limited by several factors, including API access constrains and computational cost. Therefore, this study adopts a **practical approximation** of the ideal solution by combining well-defined prompting templates, a standardized evaluation dataset, and repeated test runs. Response consistency was estimated with results from multiple test runs, and the accuracy difference between prompting techniques was validated using statistical tests.

The goal of this setup is not to eliminate uncertainty entirely, but to **systematically control for variability** while enabling reproducibility and transparent comparison across Chain-of-Thought, Chain-of-Verification, and Self-Refine prompting strategies.

## Requirements

To explore effective hallucination mitigation strategies, a literature review focused on **state-of-the-art techniques** was conducted. An overview of hallucination mitigation methods (see Figure 1) categorizes them into two main branches: Prompt Engineering and Developing Models. While model-based techniques like fine-tuning require infrastructure and model training expertise beyond the reach of typical users, prompt engineering methods stand out for their accessibility and low entry barrier. These methods can be applied directly using commercial APIs and chatbots, making them ideal for scenarios where students interact with LLMs as black-box systems.

A diagram of a company

AI-generated content may be incorrect.

Figure 1: Imama, “Beyond Traditional Fine-tuning: Exploring Advanced Techniques to Mitigate LLM Hallucinations,” Hugging Face Blog, Apr. 2024. [Online]. Available: <https://huggingface.co/blog/Imama/pr>. [Accessed: Apr. 10, 2025].

Within prompt engineering, three sub-categories are outlined: ***Retrieval-Augmented Generation (RAG)*, *Prompt Tuning***, and ***Self-Refinement through Feedback and Reasoning***. Based on feasibility, the first two were excluded: *RAG* relies on external knowledge bases or search modules (Lewis, et al., 2020), which a student may not have access to, and *Prompt Tuning* requires gradient-based optimization and training tokens (Lester, Al-Rfou, & Constant, 2021), which is impractical for everyday users.

This narrowed the focus to **self-refinement methods**: a subset of prompt engineering that enhances LLM performance through structured reasoning and self-improvement. Among the techniques surveyed, three suitable candidates were identified, being both effective and feasible:

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

Encourages models to produce intermediate reasoning steps, improving performance on complex multi-step problems (Wei, et al., 2022).

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

Adds a verification phase that critically assesses and revises the model’s initial answer, shown to reduce hallucinations (Dhuliawala, et al., 2023).

**SELF-REFINE**

Implements an iterative loop of feedback and refinement, helping models self-correct reasoning and improve factual accuracy (Madaan, et al., 2023).

These techniques were selected because they:

1. Require no model fine-tuning or external tools,
2. Operate entirely through prompting, and
3. Can be reproduced with standard access to the OpenAI API.

The design of the experiment was also shaped by the **evaluation needs of these techniques**. To assess factual correctness, the **ENEM 2024 university entrance exam** (Nunes, Primi, Pires, Lotufo, & Nogueira, 2023) **was selected. It contains 180 multiple-choice questions across four subject domains, along with the ground-truth answers**. This allowed for straightforward, automated scoring without the need to judge the reasoning quality, ideal for isolating the effects of prompt design alone.

Finally, the **evaluation metrics and prompt template structures** were based on the goals of the selected methods and the prior literature. Accuracy, subject-specific performance, and consistency were chosen to capture performance, domain sensitivity, and stability, respectively, aligning with how prior work evaluated CoT (Wei, et al., 2022), CoVe (Dhuliawala, et al., 2023), and Self-Refine (Madaan, et al., 2023). Structured prompt templates were created for each method for consistency across test runs while enabling method-specific logic such as verification or refinement loops.

Together, these requirements ensured that the study would not only provide insight into the effectiveness of self-refinement prompting methods, but also provide actionable results applicable in real-world student use cases.

## Development Process

This section describes the practical steps taken to implement the evaluation framework, from preparing the dataset to executing the prompts and collecting results.

### Prepare Dataset

The experiments were conducted using the publicly available ENEM 2024 exam dataset (Nunes, Primi, Pires, Lotufo, & Nogueira, 2023), which contains all multiple-choice questions from the 2024 Brazilian national university entrance exam (Exame Nacional do Ensino Médio). A total of **180 questions** were used, covering four official subject areas:

* Languages (Linguagens, Códigos e suas Tecnologias)
* Human Sciences (Ciências Humanas e suas Tecnologias)
* Natural Sciences (Ciências da Natureza e suas Tecnologias)
* Mathematics (Matemática e suas Tecnologias)

Each entry included the question, the multiple-choice alternatives and a ground-truth label.

### Define Prompting Templates

To ensure consistency and control over input formatting, **custom templates** were created for each of the three prompting methods under evaluation:

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

Used a 3-shot prompt with explicit step-by-step reasoning.

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

Extended the CoT format by generating and answering verification questions before revising them for the final answer.

**Self-Refine**

Combined CoT with an iterative feedback–refinement loop, where the model reviewed and improved its own answers.

Each template was implemented using functions that dynamically populated prompt slots with ENEM question data. These templates formed the backbone of the evaluation.

### Test Prompting Techniques

The execution pipeline was implemented in a **Jupyter Notebook** using the GPT-3.5 Turbo model via OpenAI API. For each method, each question was run **40 times** to capture variability and analyse consistency. Each response was recorded along with metadata such as timestamps, reasoning traces, and intermediate answers.

### Evaluation and Data Collection

For each prompt-response pair, the following data points were collected:

* The full model output, including reasoning and answer,
* The extracted answer letter (A–E),
* Whether the prediction was correct,
* The subject label,
* Additional metadata depending on the method:
  + **CoVe:** Verification questions and answers
  + **Self-Refine:** Iterative responses, feedback trace and intermediate answers

This structured dataset enabled an evaluation across multiple dimensions, including overall accuracy, subject-wise performance, and prediction consistency.

## Tests

### General Accuracy

General accuracy was calculated as the proportion of correct answers produced by each method across all ENEM questions and all runs. For each method, accuracy was computed per run (based on 180 questions), and then averaged across the 40 runs to obtain the mean accuracy and standard deviation. This approach captures both central performance and run-level variability, enabling fair comparison across methods.

This metric directly addresses **Research Question 1 (RQ1)**, which investigates whether certain prompting techniques lead to more accurate predictions overall. By aggregating performance across the full test set, general accuracy provides a high-level indicator of each method's effectiveness.

### Accuracy by Subject

To assess how performance varied by academic domain, accuracy was also computed separately for each of the four ENEM subjects. For every method and run, subject-specific accuracy was calculated as the mean correctness over questions in that subject area. These values were then aggregated to compute the mean and standard deviation of accuracy per subject and method.

This subject-wise breakdown was essential to answer **Research Question 2 (RQ2)**, which investigates whether prompting strategies differ in effectiveness depending on the domain of the question.

### Consistency

Consistency measures how reliably a prompting method produces the **same answer** to the **same question** across different runs. For each method, a question was marked as consistent if the same answer was predicted in all 40 runs. The overall consistency rate was computed as the proportion of questions classified as consistent.

To analyse domain-specific stability, consistency was also computed **per subject** by grouping questions by their subject label and calculating the proportion of consistently answered items within each group.

This metric was critical to answering **Research Question 3 (RQ3)**, which addresses the **reliability** of each prompting method under repeated evaluation.

## Tools

All experiments, data processing, and analyses were conducted using openly available tools and Python libraries. The following technologies were used throughout the development and evaluation process:

**Jupyter Notebook**  
Provided an interactive environment for developing, testing, and documenting the experimental pipeline. It facilitated reproducibility and allowed step-by-step analysis of intermediate outputs, errors, and statistical results.

**OpenAI API**  
Used to generate model responses via the GPT-3.5 Turbo language model. The API enabled programmatic submission of prompts and retrieval of outputs across multiple runs and prompting templates.

**pandas**  
Served as the primary tool for data manipulation, grouping, filtering, and cleaning of results. It was used extensively to structure run outputs, calculate performance metrics, and organize data for visualization.

**matplotlib** and **seaborn**  
These libraries were used for generating all plots and visualizations in the analysis. Bar plots, box plots, and significance markers were created to communicate results in a clear and informative way.

**numpy**  
Provided support for numerical operations, including array-based computations and statistical aggregation.

**scipy**  
Used for statistical testing, including Shapiro–Wilk tests for normality, Levene’s test for variance homogeneity, Kruskal–Wallis tests for group comparison, and McNemar’s tests for paired binary outcomes.

# Solution

## Answering the Research Questions

This research is guided by the following three questions:

**RQ1:** *How does the application of CoT, CoVe and Self-Refine affect LLM response accuracy on Brazilian University Entrance Exams?*

**RQ2:** *How does the effectiveness of CoT, CoVe and Self-Refine vary across different subject areas?*

**RQ3:** *How consistent are the predictions of Chain-of-Thought, Chain-of-Verification, and Self-Refine methods across repeated runs?*

Each prompting method was implemented as a reusable prompt template and applied to the same multiple-choice question set across multiple runs. The evaluation strategy is based on three core criteria:

**Accuracy**

The proportion of correct responses produced by each method across all questions.

**Subject-specific Performance**

The accuracy and consistency of each method when broken down by academic subject (Mathematics, Human Sciences, Languages, and Natural Sciences).

**Consistency**

The ability of a method to produce the same answer for the same question across repeated test runs.

Together, these analyses provide a comprehensive answer to the research questions. They help identify which technique produces the most accurate, stable, and subject-adaptable results, highlighting the most promising strategy for enhancing reliability in LLM-assisted educational tasks.

## Prompting Techniques

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

**Definition**

Chain-of-Thought (CoT) is a prompting technique designed to improve the reasoning capabilities of LLMs by encouraging them to produce step-by-step explanations before arriving at a final answer. Rather than directly outputting a single response, CoT prompts guide the model through intermediate reasoning steps that mimic human problem-solving behavior. As reported in *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (Wei, et al., 2022) , this structured form of output has shown to significantly improve performance on tasks requiring logical inference, arithmetic, and multi-step analysis.

**Key Steps in the Process**

In the CoT approach, the model follows a three-step process:

1. **Instructions**

The model is introduced to the task through a 3-shot prompt containing example ENEM questions and detailed explanations. It is then presented with the question it should solve by following the examples.

1. **Generate reasoning**

The model is expected to produce a chain of reasoning that explains the thought process leading to its answer.

1. **Explicit Final Answer**

After the reasoning steps, the model is prompted to select and declare the correct multiple-choice alternative.

**Implementation for ENEM Questions**

CoT was implemented using a Python prompt template tailored for ENEM-style multiple-choice questions. Each prompt included three examples taken from *Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams* (Nunes, Primi, Pires, Lotufo, & Nogueira, 2023), where the authors evaluated GPT-3.5 and GPT-4 models on ENEM exams. These examples were selected from the **ENEM 2022 exam** and span across three subject areas: **Languages, Human Sciences,** and **Mathematics**. Each example is formatted to reflect the CoT structure: the question is followed by the answer alternatives, a detailed explanation that analyzes the options, and a final line that explicitly states the correct alternative in a structure format. The explanations were derived from expert teacher discussions and public exam commentary resources, as also documented in the paper (Nunes, Primi, Pires, Lotufo, & Nogueira, 2023).

In the prompt template, the 3-shot block is followed by the question under evaluation, inserted dynamically during each test run. The model is instructed to produce a full reasoning chain and identify the final answer at the end. Each response was stored with metadata including the selected alternative and the full generated reasoning chain.

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

**Definition**

Chain-of-Verification (CoVe) is a prompting method developed to reduce hallucinations in LLMs by having the model **verify its own outputs**. Introduced in *Chain-of-Verification Reduces Hallucination in Large Language Models* (Dhuliawala, et al., 2023), CoVe operates on the assumption that LLMs can critically analyse and improve their responses when prompted appropriately. Rather than relying on a single pass, CoVe decomposes the generation process into separate verification steps.

**Key Steps in the Process**

The CoVe process involves **four sequential stages**:

1. **Generate a Baseline Response**  
   An initial draft is generated using the Chain-of-Thought template.
2. **Plan Verification Questions**  
   The model reflects on the baseline response and produces verification questions to check for factual correctness or reasoning steps.
3. **Execute Verifications**  
   Each verification question is answered in isolation, preventing the model from being biased by its own previous answers.
4. **Generate Final Verified Response**  
   Using the original question, the baseline response, and the results of the verifications, the model revises and outputs a final, refined answer.

**Implementation for ENEM Questions**

In this study, CoVe was implemented as a three-step pipeline. All steps are executed inside an iterative loop that repeats the CoVe pipeline over the ENEM questions.

The process begins by generating a **baseline response** using the 3-shot Chain-of-Thought template described in the CoT section. This response serves as the initial answer.

The second step uses the plan\_verification\_questions() function to generate verification questions. The prompt instructs the LLM to reflect on its previous answer and identify key factual or logical claims worth verifying.

Each verification question is then answered independently using the execute\_verifications() function. To ensure a robust verification process, each question is sent to the LLM in isolation, without providing the baseline response in the context: this emulates the *factored* verification setup, which was shown to prevent the model from repeating its hallucinations​ (Dhuliawala, et al., 2023).

Finally, the generate\_final\_verified\_answer() function constructs a full prompt that includes the original question, alternatives, the baseline answer, and the verification Q&A pairs. The LLM is then asked to synthesize this information into a revised, fully verified response. This final step mimics the ***Factor + Revise*** variant, which the original authors found to yield the **highest factual accuracy** (Dhuliawala, et al., 2023).

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

**Definition**

Self-Refine is an iterative prompting technique in which a language model improves its own output through cycles of self-critique and revision. Instead of producing a single-shot response, the model first generates an answer, then evaluates it by generating critical feedback, and finally uses that feedback to revise and refine its response. This method has shown to improve LLM performance across diverse tasks, including reasoning, coding, and factual QA (Madaan, et al., 2023).

**Key Steps in the Process**

The implementation consisted of three steps:

1. **Initial Answer**

The model answers the ENEM question using the Chain-of-Thought template, producing an initial answer.

1. **Feedback**

The model receives a second prompt asking it to critically analyze its own response, pointing out factual mistakes, logical inconsistencies, or poor reasoning.

1. **Refinement**

The model receives a third prompt to revise its answer taking the generated feedback into account. It is instructed to re-express the reasoning and conclude with a final answer in the specified format.

This feedback–refinement loop is repeated until the selected multiple-choice answer remains unchanged between iterations, or until a maximum of 10 iterations is reached.

**Implementation for ENEM Questions**

The implementation was carried as a Python pipeline. The function run\_self\_refine() loops through all ENEM questions, applying the Self-Refine logic defined in self\_refine\_enem(). For each question, the process begins by calling build\_cot\_prompt() to create an initial CoT response. Then, for up to 10 iterations:

* A feedback prompt is created using build\_feedback\_prompt(), asking the model to identify and critique weaknesses in its response.
* A refine prompt is created with build\_refine\_prompt(), instructing the model to revise its answer based on that feedback.
* If the final predicted answer remains the same as the one in the previous iteration, the loop stops.

Throughout this process, all iterations, responses, feedbacks, and selected answers are stored in a structured format, including metadata like subject and answer trace.

## Load QA dataset

The dataset used in this study was sourced from the open-access Maritaca AI ENEM dataset (Nunes, Primi, Pires, Lotufo, & Nogueira, 2023). It contains all multiple-choice questions from the 2024 Exame Nacional do Ensino Médio (ENEM), Brazil's national standardized exam used for university admissions.

To facilitate subject-specific accuracy analysis, each question was assigned to one of the four official ENEM subject areas based on its position in the exam. This mapping follows the fixed structure used in every ENEM booklet:

* Questions 1–45: Linguagens, Códigos e suas Tecnologias (Languages)
* Questions 46–90: Ciências Humanas e suas Tecnologias (Human Sciences)
* Questions 91–135: Ciências da Natureza e suas Tecnologias (Natural Sciences)
* Questions 136–180: Matemática e suas Tecnologias (Mathematics)

## Run Tests

A testing pipeline was developed so that it systematically applies each prompting technique to the full ENEM 2024 dataset. For each method there is a total of 40 responses per question.

Each test run executes a corresponding method-specific function and stores the results in structured CSV files for subsequent evaluation.

## Clean Data

After running and storing the raw results from all test iterations, a structured data cleaning pipeline was applied to extract the relevant fields for analysis. Each result file included the question ID, subject labels translated to English, ground truth and final answer, as well as the initial answer for CoVe and Self-Refine.

**Subject Translation**  
Since the subject labels in the original dataset were in Portuguese, they were mapped to their English equivalents to simplify analysis and visualization.

**Cleaning Process per Method**  
The process\_files() function iterated through all result CSV files for each method (CoT, CoVe, and Self-Refine). It detected new raw result files using a timestamped filename pattern and generated cleaned versions using method-specific cleaning functions.

## Analyse Performance

To compare the effectiveness of the prompting techniques, a quantitative analysis was performed based on three key metrics: overall accuracy, subject-wise accuracy, and consistency across runs. All analyses were conducted on the cleaned datasets generated for each of the 40 test runs.

### General Accuracy

A general comparison of accuracy distributions for each method was conducted to address the **Research Question 1**:

**RQ1**: *How does the application of CoT, CoVe and Self-Refine affect LLM response accuracy on Brazilian University Entrance Exams?*

#### **Computation**

To measure overall accuracy, the cleaned results from all runs of each method were combined into a single DataFrame. For each question, the correct flag was used as the basis for the analysis. Data was first grouped by method and run to compute per-run accuracy and then results were aggregated to calculate the mean and standard deviation.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 2: Overall Accuracy by Method

The following table shows the average accuracy across all ENEM questions for each technique, along with the standard deviation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Mean Accuracy** | **Standard Deviation** | **Runs** |
| CoT | 0.703750 | 0.189968 | 40 |
| CoVe | 0.672917 | 0.205645 | 40 |
| Self-Refine | 0.685694 | 0.193422 | 40 |

Table 1: General Mean Accuracy and Standard Deviation of CoT, CoVe and Self-Refine across test runs

#### **Statistical Analysis**

The next step was to assess whether the differences in overall accuracy between the three methods were statistically significant.

**Testing for Normality and Homogeneity**  
First it was tested whether the conditions for parametric analysis, such as ANOVA, were met:

* **Normality (Shapiro–Wilk test):**
  + CoT: W = 0.8983, p < 0.0001
  + CoVe: W = 0.9108, p < 0.0001
  + Self-Refine: W = 0.8929, p < 0.0001

All three distributions showed statistically significant deviation from normality (p < 0.05), indicating that the normality assumption was violated.

* **Homogeneity of Variances (Levene’s test):**
  + Statistic = 1.2022, p = 0.3014

The results of Levene’s test indicate that the variance across groups is not significantly different, satisfying the assumption of equal variances.

**Non-Parametric Comparison (Kruskal–Wallis test)**  
Given the non-normality of the data, the **Kruskal–Wallis H test** (Kruskal & Wallis, 1952) was applied to compare the distributions of accuracy across the three methods.

* **H₀ (Null Hypothesis):** The distributions of accuracy scores across the three prompting methods are the same.
* **H₁ (Alternative Hypothesis):** At least one method’s distribution differs significantly from the others.

**Interpretation**

* **H statistic**: 3.7021
* **p-value**: 0.1570

The Kruskal–Wallis test resulted in a p-value of 0.1570, which is above the conventional threshold of 0.05, so **it is not possible to reject the null hypothesis (H₀)**. The lack of statistical significance indicates that none of the methods consistently outperformed the others in a way that generalizes across all evaluation runs. Although Chain-of-Thought achieved the highest mean accuracy (70.3%), followed by Self-Refine and Chain-of-Verification (67.3%), these observed differences could be attributed to random variation.

**Conclusion for RQ1**

This finding brings an answer to **Research Question 1 (RQ1)**: while CoT exhibits higher performance, the statistical analysis does not support a definitive conclusion that it is more accurate than CoVe or Self-Refine.

### Subject-Specific Accuracy

A subject-wise comparison of accuracy distributions was conducted to address the **Research Question 2:**

**RQ2:** How does the effectiveness of Chain-of-Thought, Chain-of-Verification, and Self-Refine vary across different subject areas?

**Metric Aggregation**  
Accuracy scores were grouped by method, run, and subject, and the mean accuracy was calculated for each unique combination.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 3: Accuracy Mean and Standard Deviation per Method and Subject

**Hypotheses and Statistical Method**

**Conside**ring the violations of the normality assumption in the accuracy distributions, as established in the previous section, the **Kruskal–Wallis H test** was employed to determine whether the prompting methods differed significantly in accuracy for specific subjects.

* **H₀ (Null Hypothesis):** The distributions of accuracy scores across the three prompting methods are statistically indistinguishable for any subject.
* **H₁ (Alternative Hypothesis):** At least one method differs significantly in accuracy from the others for a given subject.

**Results**

Statistically significant differences were observed for **Mathematics** and **Human Sciences**, indicating that at least one method outperformed the others within these subjects. In both cases, Chain-of-Thought achieved the highest average accuracy. No significant differences were found for **Languages** and **Natural Sciences**, suggesting comparable performance across methods for these subjects.

Table 2: Average Accuracy by Subject and Kuskal-Wallis Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Subject** | **CoT Mean** | **CoVe Mean** | **Self-Refine Mean** | **Kruskal-Wallis H** | **p -value** | **p < 0.05** |
| Human Sciences | 0.955000 | 0.934444 | 0.942222 | 10.425922 | 5.445527e-03 | True |
| Languages | 0.803889 | 0.787778 | 0.785556 | 5.508679 | 6.365103e-02 | False |
| Mathematics | 0.495000 | 0.415000 | 0.475000 | 37.652330 | 6.666540e-09 | True |
| Natural Sciences | 0.561111 | 0.554444 | 0.540000 | 5.743476 | 5.660046e-02 | False |

**Interpretation**  
For **Mathematics** and **Human Sciences**, the p-values are both below the significance threshold of 0.05. Therefore, **the null hypothesis (H₀) is rejected** for Mathematics and Human Sciences, as there are statistically significant differences in accuracy between the prompting methods. In both of these subjects, **Chain-of-Thought achieved the highest mean accuracy** (Mathematics: 0.495, Human Sciences: 0.955), indicating it outperformed both Chain-of-Verification and Self-Refine.

For **Languages** (p = 0.06365103) and **Natural Sciences** (p = 0.05660046), the p-values are above the 0.05 threshold. Therefore, **it** **is not possible to reject the null hypothesis (H₀)** for these subjects, indicating that there is no statistically significant difference in accuracy between the three prompting methods.

**Conclusion for RQ2**

**These findings** bring an answer to Research Question 2 (RQ2): there was a significant difference between the methods’ performance in Mathematics and Human Sciences, with Chain-of-Thought offering the most benefit in both cases. In subjects like Languages and Natural Sciences, all methods perform similarly.

### Consistency

Large Language Models are stochastic, meaning the same prompt can lead to different outputs across runs. This section addresses **Research Question 3:**

**RQ3**: How consistent are the predictions of Chain-of-Thought, Chain-of-Verification, and Self-Refine methods across repeated runs?

A graph with different colored bars

AI-generated content may be incorrect.

Figure 4: Answer Consistency per Method and Subject

**Definition of Consistency**  
For this analysis, a prediction was marked as **consistent** if the model produced the same final answer for the same question across all 40 test runs.

**Overall Consistency**

|  |  |
| --- | --- |
| **Method** | **Consistency Rate** |
| CoT | 47.2% |
| CoVe | 29.4% |
| Self-Refine | 30.0% |

These results indicate that **Chain-of-Thought** produces the most stable predictions, followed by **Self-Refine** and **CoVe**.

To compare consistency between prompting methods, the McNemar’s test (McNemar, 1947) was selected. The test is designed for **paired binary data**, and in this study, each method's prediction was classified as either **consistent** or **inconsistent** across runs for a given question. This creates matched observations suitable for McNemar’s test.

The test assesses whether the **frequency of discordant outcomes** differs significantly between two methods, for example, whether one is more likely to produce consistent answers than the other.

**Hypotheses**

* **H₀ (Null Hypothesis):** The consistency proportions of the two methods being compared are equal (no difference in consistency).
* **H₁ (Alternative Hypothesis):** The consistency proportions differ between the two methods.

**McNemar’s Tests for Overall Consistency**

|  |  |  |
| --- | --- | --- |
| **Comparison** | **p -value** | **Significant** |
| CoT vs CoVe | < 0.0001 | True |
| CoT vs Self-Refine | < 0.0001 | True |
| CoVe vs Self-Refine | 1.0000 | False |

The results show that **CoT is significantly more consistent** than both CoVe and Self-Refine (p < 0.0001), while there is no statistically significant difference between CoVe and Self-Refine (p = 1.0000).

**McNemar’s Tests for Consistency by Subject**

|  |  |  |  |
| --- | --- | --- | --- |
| **Subject** | **Comparison** | **p -value** | **Significant** |
| Human Sciences | CoT vs CoVe | 0.0001 | True |
|  | CoT vs Self-Refine | 0.0020 | True |
|  | CoVe vs Self-Refine | 0.3438 | False |
| Languages | CoT vs CoVe | 0.0213 | True |
|  | CoT vs Self-Refine | 0.0004 | True |
|  | CoVe vs Self-Refine | 0.1460 | False |
| Mathematics | All comparisons | >0.06 | False |
| Natural Sciences | All comparisons | >0.3 | False |

**Interpretation**  
Based on the test results, **the null hypothesis** (**H₀) is rejected** for comparisons involving Chain-of-Thought (CoT) versus both Chain-of-Verification (CoVe) and Self-Refine, indicating that **CoT is significantly more consistent** overall (p < 0.0001). However, **the null hypothesis** (**H₀) is not rejected** for CoVe versus Self-Refine (p = 1.0000), suggesting no significant difference in their consistency.

At the subject level, CoT also shows significantly higher consistency than both CoVe and Self-Refine in **Human Sciences** and **Languages**, but no significant differences were observed in **Mathematics** and **Natural Sciences**. These results support the alternative hypothesis (H₁) for **Human Sciences** and **Languages domains.**

**Conclusion for RQ3**

**These results** answer Research Question 3 (RQ3): **Chain-of-Thought** produces significantly more consistent predictions than both **Chain-of-Verification** and **Self-Refine**, both overall and within specific subject areas like **Human Sciences** and **Languages.**

### Interesting Findings

A closer examination of the subject-wise accuracy gains reveals insights about the effectiveness of feedback-based techniques in this context. While Self-Refine and CoVe are designed to improve upon initial answers through iterative reasoning or verification, results show that this is not always the case.

A test was performed to assess what would happen if both methods had always stuck with their initial answer: the output produced using the Chain-of-Thought template, before any feedback or revision. Surprisingly, in several cases, this **baseline answer proved to be statistically more accurate than the final refined one**.

A graph with green and blue bars

AI-generated content may be incorrect.

Figure 5: Accuracy gain by sticking with the initial answer

Table 3: Statistical Analysis of Initial vs Final Answer Accuracy for CoVe and Self-Refine

A screenshot of a test

AI-generated content may be incorrect.

**CoVe**

For CoVe, the final refinement underperformed the initial CoT answer in **three out of four subjects** (Human Sciences, Languages, and Mathematics) with statistically significant drops in accuracy (p < 0.001 for Mathematics, for example). The only exception was in Natural Sciences, where the difference was not statistically significant. This pattern suggests that **CoVe’s verification mechanism may sometimes introduce errors or reinforce uncertainties** rather than resolving them, possibly due to flawed verification questions or overly cautious revisions.

**Self-Refine**

Self-Refine showed a consistent trend of **accuracy decrease after refinement** in most subject areas. In **Languages, Mathematics,** and **Natural Sciences**, the final answer generated after the feedback–refinement loop was **significantly worse** than the initial response generated using the Chain-of-Thought template (p < 0.001 for both Languages and Mathematics; p = 0.01 for Natural Sciences). These results suggest that, rather than improving the original answer, the **iterative feedback mechanism of Self-Refine often introduced noise or instability that led to poorer outcomes**.

**Conclusion**

Overall, these results highlight a **critical insight: answer refinement do not always lead to more accurate answers**. In fact, for certain subjects and methods, **the best result was already achieved in the first answer,** raising questions about when and how refinement should be applied. It suggests that a hybrid strategy, where refinement is selectively applied based on initial confidence or answer quality, may be a more effective approach in educational tasks.

# Discussion

## Potentials

A key strength of this study was the use of **multiple test runs** per method. Repeating the experiment 40 times helped reveal not only which methods were more accurate on average, but also which ones gave stable answers when faced with the same question multiple times. Measuring consistency is especially important when working with stochastic models like GPT-3.5 Turbo, where the same input can produce different outputs.

Another important part of the methodology was the use of **prompting templates**. Each template was tailored to follow the logic of the corresponding method. This preserved the unique reasoning structure of each method and ensured a fair comparison. Templates were applied consistently across all questions and test runs, reducing bias caused by formatting or instruction differences.

The **ENEM 2024 dataset** used in this study also provided a realistic and relevant test scenario. These are real multiple-choice questions from a national university entrance exam in Brazil, covering subjects like Math, Languages, and Sciences. The dataset was already labelled with correct answers and followed a fixed structure, which allowed for automatic scoring and subject-level analysis without manual grading.

Together, these methodological choices helped make the results reproducible, fair, and applicable to real-world student use cases.

## Limitations

This study has several limitations that should be considered when interpreting the results.

First, all experiments were conducted using **GPT-3.5 Turbo.** While this is a powerful model, newer or larger models like GPT-4 might behave differently. The results may not fully generalize to other LLMs with different training data or architectures.

Second, the evaluation was based on a **single dataset**, the ENEM 2024 Brazilian university entrance exam. Although it covers a variety of subjects, it is limited to **multiple-choice questions**, which have a fixed structure and predefined answer options. Therefore, the findings might not apply to open-ended questions, essay-style responses, or other formats.

Finally, **computational and API cost constraints** limited the scale of the experiment. Only 180 questions were tested, and each method was run 40 times. A larger sample of questions or additional runs could provide stronger statistical power and help detect smaller performance differences.

These limitations suggest that while the results are useful for understanding LLM performance in this specific setting, further research is needed to explore how these prompting techniques perform across other models, datasets, and question types.

## Possible Improvements

**Test with Multiple Models**  
Using other models, such as GPT-4, Claude, or LLaMA would help determine whether the findings hold across architectures and levels of capability. This could reveal whether certain prompting techniques are more model-dependent.

**Include More Diverse Datasets**  
Expanding the evaluation beyond ENEM 2024 to include other standardized exams or open-ended question formats would test how well the methods generalize.

**Analyze Reasoning Quality**  
While this study focused on accuracy, a qualitative analysis of the reasoning chains, feedback messages, or verification steps could offer deeper insight into how and why methods succeed or fail.

**Adaptive Prompting Strategies**  
One limitation of methods like Self-Refine is that they apply refinement to all questions, even when it is not needed. Future work could explore hybrid or adaptive strategies, where refinement is only triggered if the model’s confidence is low or if the reasoning shows signs of error.

**Larger-Scale Experiments**  
Running experiments on larger datasets or increasing the number of repeated runs per question could improve statistical power and help reduce uncertainty. This would allow for more robust conclusions and better handling of model variance.

## Generalizability of the Solution

The solution presented in this study is **modular and adaptable,** making it possible to apply it in other contexts beyond the ENEM 2024 dataset with small adjustments. Since all prompting techniques were implemented without relying on model fine-tuning or external tools, they can be reused with **any LLM accessible via API**.

The methodology can generalize well **to other multiple-choice exams**, especially those with a clear structure and labeled answers, such as SATs, or national exams from other countries. With small adjustments to the prompt templates (e.g., using domain-specific examples), the same approach could be applied to **different subject areas** or educational levels.

However, the current setup is limited when it comes to **open-ended tasks**, such as writing essay, programming, and answering open questions. In those cases, different evaluation methods would be required, including qualitative scoring or human feedback.

In summary, the solution is **general enough** for structured QA tasks across different exams and domains but would require adaptation for more complex applications.

## Key Takeaways

This study shows that it is possible to design an **evaluation framework for LLM prompting methods** using only publicly available tools and datasets. One key takeaway is the importance of **modular design**: by separating data loading, prompting, testing, and analysis into reusable components, the entire pipeline became easier to debug, extend, and scale.

The use of **multiple runs per method** was also essential. It revealed how inconsistent LLMs can be and why performance should not be measured with a single test. This is an important insight for researchers and developers working with LLMs: measuring only accuracy without considering variability can be misleading.

Finally, this study shows that **LLM experimentation can be practical and reproducible**, even for students or small teams. With clear goals, thoughtful design, and systematic testing, it is possible to generate meaningful insights without access to internal model weights or advanced infrastructure.

# Bibliography

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . . . Fung, P. (2023). *Survey of Hallucination in Natural Language Generation.* ACM computing surveys 55.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association, 47*(260), 583-621.

Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv preprint arXiv:2104.08691.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Zettlemoyer, L. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems (NeurIPS)*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., . . . al, e. (2023). SELF-REFINE: Iterative Refinement with Self-Feedback. *Advances in Neural Information Processing Systems 36*.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 153–157.

Nunes, D., Primi, R., Pires, R., Lotufo, R., & Nogueira, R. (2023). Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. arXiv2303.17003.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*.

# Figures

[Figure 1: Overview of Hallucination Mitigation Techniques in LLMs 8](#_Toc195175595)

[Figure 2: Overall Accuracy by Method 17](#_Toc195175596)

[Figure 3: Accuracy Mean and Standard Deviation per Method and Subject 19](#_Toc195175597)

[Figure 4: Answer Consistency per Method and Subject 21](#_Toc195175598)

# Tables

[Table 1: General Mean Accuracy and Standard Deviation of CoT, CoVe and Self-Refine across test runs 18](#_Toc195175846)

[Table 2: Average Accuracy by Subject and Kuskal-Wallis Results 20](#_Toc195175847)

# List of Abbreviations

|  |  |
| --- | --- |
| WWW | World Wide Web |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Documentation table of AI-based tools

|  |  |  |
| --- | --- | --- |
| **AI-based tools** | **Intended use** | **Prompt, source, page, paragraph...** |
| **DeepL Translate** | Translation of an article in English | Source (XXX), Chapter X on page X-X |
| **ChatGPT (4.0)** | Grammar and spelling | "Please list issues with spelling and grammar in the following text: ..." Entire document |
|  |  |  |

A: Heading of Appendix A

B: Heading of Appendix B