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 Prompting Techniques in solving Brazilian University Entrance Exams

By: Rafaela Rolim Santana

Student Number: 2210257114

Supervisor 1: Patrick Link, BSc.

Vienna, 27.03.2025

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

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.

Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.

Aenean nec lorem. In porttitor. Donec laoreet nonummy augue.

Suspendisse dui purus, scelerisque at, vulputate vitae, pretium mattis, nunc. Mauris eget neque at sem venenatis eleifend. Ut nonummy.

Fusce aliquet pede non pede. Suspendisse dapibus lorem pellentesque magna. Integer nulla.

Donec blandit feugiat ligula. Donec hendrerit, felis et imperdiet euismod, purus ipsum pretium metus, in lacinia nulla nisl eget sapien. Donec ut est in lectus consequat consequat.

Acknowledgements

**Keywords:** Keyword1, Keyword2, Keyword3, Keyword4, Keyword5

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.

Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.

Aenean nec lorem. In porttitor. Donec laoreet nonummy augue.

Table of Contents

[1 This is the heading of the first chapter 5](#_Toc179533587)

[1.1 Heading level 2 5](#_Toc179533588)

[1.1.1 Heading level 3 5](#_Toc179533589)

[1.1.1.1 Heading level 4 5](#_Toc179533590)

[2 This is the heading of the second chapter 5](#_Toc179533591)

[2.1 Heading level 2 5](#_Toc179533592)

[2.1.1 Heading level 3 6](#_Toc179533593)

[2.1.1.1 Heading level 4 6](#_Toc179533594)

[Bibliography 7](#_Toc179533595)

[List of Figures 8](#_Toc179533596)

[List of Tables 9](#_Toc179533597)

[List of Abbreviations 10](#_Toc179533598)

[Documentation table of AI-based tools 11](#_Toc179533599)

[A: Heading of Appendix A 12](#_Toc179533600)

[B: Heading of Appendix B 13](#_Toc179533601)

# Introduction

## Motivation

Large Language Models (LLMs) are increasingly being used in educational contexts — from assisting with content generation to acting as intelligent tutoring systems. Their potential to provide immediate and personalized support makes them attractive tools for learning environments. However, for these models to be truly beneficial in educational settings, their outputs must be **reliable, accurate, and consistent**, especially when answering exam-style questions where factual correctness is crucial.

Students often rely on such models to help them **understand concepts, solve exercises, and prepare for exams**. Although LLMs can generate coherent text, their accuracy on exam-style questions varies widely. In this context, selecting an appropriate prompting technique can significantly influence the quality and correctness of the generated responses. Yet, there is currently limited guidance on which prompting methods are best suited for specific subjects or tasks.

Given the wide range of disciplines present in university entrance exams — such as Mathematics, Natural Sciences, Human Sciences, and Languages — it becomes essential to evaluate how different prompting strategies perform across domains. A systematic and subject-sensitive evaluation could help students and educators understand how to best leverage LLMs for educational use.

## ****Goals****

The goal of this thesis is to design and evaluate a framework for assessing the performance of different prompting techniques in solving standardized high-school level questions. More specifically, the objectives are:

* To investigate and compare three prompting techniques in terms of accuracy in solving Brazilian university entrance exams.
* To determine which subject areas benefit most or least from each technique.
* To provide guidelines for optimizing prompts to improve educational outcomes.

## Research Questions

 **RQ1**: How does the application of Chain-of-Thought [1], Chain-of-Verification [2] and Self-Refine [3] prompting techniques affect LLM response accuracy on Brazilian University Entrance Exams?

 **RQ2**: How does the effectiveness of Chain-of-Thought [1], Chain-of-Verification [2] and Self-Refine [3] prompting techniques vary across different subject areas?

## ****Expected Outcome****

The expected outcome of this work is a **framework** for evaluating and comparing prompting strategies using standardized questions and objective performance metrics.

**Moreover, this work aims to make recommendations** for students and educators on how to select and use prompting techniques effectively.

# Methodology

## Experimental Setup

* Data Source
  + **Exam Questions:** This study uses a dataset comprising questions from the 2024 ENEM (Brazilian University Entrance Exam). It consists of 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.
  + **Subject Areas:** Questions are categorized into major subject domains (e.g., Mathematics, Language, Sciences, and Humanities) to allow analysis of performance variations (RQ2).
* Model and Environment
  + **Language Model:** Experiments are conducted using GPT-3.5
  + **Implementation:** The prototype is implemented in a Jupyter Notebook, where the experimental pipeline is organized into data ingestion, prompt generation, response evaluation, and logging of results.

## Prompting Techniques

### Chain-of-Thought (CoT)

 **Concept:** This technique involves including a few-shot set of examples that contain not only input–output pairs but also explicit intermediate reasoning steps (the “chain of thought”).

 **Implementation:** For each ENEM question, the prompt provides exemplars where the model “thinks aloud” before arriving at a final answer.

### Chain-of-Verification (CoVe)

 **Concept:** CoVe is designed to reduce factual hallucinations by having the model verify its own responses.

 **Implementation:** The process is divided into four steps:

* **Baseline Generation:** The model generates an initial answer.
* **Planning Verification:** The model produces a series of verification questions aimed at fact-checking the baseline response.
* **Execution of Verifications:** Each verification question is independently answered, ensuring that the process does not simply echo the original output.
* **Final Response:** The model generates a final, revised answer that integrates the results of the verifications.

### Self-Refine

 **Concept:** Self-Refine is an iterative method in which the model first generates an initial answer, then produces feedback on its own response, and subsequently refines its answer based on this feedback.

 **Implementation:** The process involves alternating between generating a candidate answer and a self-generated critique until a stopping criterion (e.g., a maximum number of iterations or a satisfactory confidence level) is met.

## Implementation

 **Prompt Construction:** For each technique, tailored few-shot prompts are designed.

* In Chain-of-Thought prompting, each response should include a detailed reasoning sequence.
* CoVe prompts are structured to first generate an initial answer, then generate verification questions for that answer, execute the verifications, and finally revise the initial answer taking the verifications into account.
* Self-Refine prompts include instructions for self-feedback and iterative correction.

 **Data Logging:** The notebook stores, for every question, the generated prompts, the LLM’s responses at each iteration (or step), and the correctness of the final answer based on the official ENEM answer key.

 **Iteration Control:** For Self-Refine, a stopping condition is defined so that the feedback loop is exited if the last answer is the same as the previous one or if a maximum of 3 iterations is achieved.

## Evaluation Metrics

# Solution

# Discussion

Bibliography

|  |  |
| --- | --- |
| [1] | H. Balzert, Lehrbuch der Objektmodellierung - Analyse und Entwurf mit der UML 2, 2. Ausg., Elsevier GmbH, München 2005., München: Elsevier GmbH, 2005. |
| [2] | K. W. Wagner, Performance Excellence. Der Praxisleitfaden zum effektiven Prozessmanagement, München: Hanser Fachbuch, 2007. |

List of Figures

[Figure 1: Example of name and year printed on spine. 6](#_Toc330300567)

List of Tables

[Table 1: Schedule for “Applied Mathematics”. 6](#_Toc330300577)

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