POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Prompt Engineering's Influence on Result Reliability in Large Language Models

Laurea Magistrale in Computer Science Engineering - Ingegneria Informatica

Author: Nicola della Volpe

Advisor: Prof. Cinzia Cappiello

Academic year: 2022-2023

1. Introduction

In the ever-evolving landscape of computer engineering, the advent of Large Language Models (LLMs) marks the beginning of a significant transformation. These models have become integral components in various applications, ranging from natural language understanding to content generation. The exponential expansion of their utility brings potential risks, including the generation of inappropriate, harmful, or misleading content, as well as concerns regarding privacy and data security. These considerations highlight the need to check how reliable Large Language Models (LLMs) are. Understanding that the accuracy of their responses depends on the questions asked, we delve into prompt engineering as a key factor affecting response quality.[1]

1.1. Objectives

The thesis aims to achieve the following objectives:

1. Assess the reliability of the Large Language Models (LLMs) by defining metrics like correctness, coherence, relevance, and completeness. This sets the criteria for analyzing the quality of LLM responses.
2. Understand how prompt engineering strategies affect the quality of LLM responses. It involves examining various prompt structures and techniques across diverse datasets and domains to determine their effectiveness or drawbacks.

1.2. Unveiling the Large Language Models

At the heart of Language Models (LMs) lies a meticulous training process aimed at predicting the probability distribution over sequences of tokens within a text. Through an autoregressive approach, these models compute the likelihood of the next token based on the context provided by preceding ones[1].

$$P(w) = P(w_1) \cdot P(w_2|w_1) \cdot \ldots \cdot P(w_T|w_1, \ldots, w_{T-1}) \quad (1)$$

Here, $w := w_1 \ldots w_T$ represents a sequence of $T = |w|$ tokens, and $P(w_t|w_1, \ldots, w_{t-1})$ with $t = 1, \ldots, T$ is the probability the LM predicts for the token w_t given the previous $t-1$ tokens.

Moving beyond conventional LMs, a Large Language Model (LLM) distinguishes itself by its substantial size, reaching into the magnitude of

[1] Repository with experiment notebooks: `https://github.com/NickdV99/Polimi-MasterThesis-2023`

tens of millions to billions of model parameters and training data. These models exhibit emergent abilities not observed in their smaller counterparts, pushing the boundaries of what is achievable in NLP.

A pivotal advancement in language modeling emerged with the introduction of Transformers, central to their efficacy is the self-attention mechanism, wherein each token evaluates the significance of neighboring tokens in the context of a given sequence. This mechanism plays a pivotal role in disambiguating references and ensuring coherent text generation. By gauging the relevance of nearby words, LLMs can craft responses that resonate with human understanding.

1.3. Evaluating Response Quality in Large Language Models

Despite the impressive capabilities of Large Language Models (LLMs), evaluating the quality of their responses presents significant challenges, including alignment with user intent and addressing security risks such as privacy concerns and misinformation propagation.

Alignment techniques, such as those outlined by [3], have made notable strides in mitigating misbehaviors exhibited by LLMs, with platforms like ChatGPT demonstrating success in aligning models with human values.

However, challenges persist in evaluating alignment extent and designing appropriate evaluation tasks. Considerations such as the "HHH" (Helpful, Honest, and Harmless) principle guide alignment evaluation.

Researchers and practitioners employ diverse methodologies and metrics to assess response quality, including *human evaluation, automated metrics, task-specific metrics, adversarial testing, ethical and bias evaluation, user studies, and explainability metrics.*

Human evaluation provides nuanced feedback on aspects such as relevance, coherence, and overall quality beyond automated metrics. Automated metrics like BLEU, ROUGE, METEOR offer efficiency but may not always align perfectly with human judgments.

Task-specific metrics tailor evaluation criteria to specific applications or domains, providing a focused assessment of LLM capabilities within particular contexts.

Adversarial testing exposes weaknesses or biases in LLMs by challenging model robustness.

Ethical and bias evaluation aims to ensure equitable and responsible deployment of LLMs, while user studies collect feedback to evaluate usability and user experience.

Explainability metrics address the increasing emphasis on model explainability.

1.4. Open Challenges

As we navigate the complexities of evaluating Large Language Models, certain open challenges emerge, demanding further exploration and attention.

Designing metrics that comprehensively assess LLM capabilities across various tasks, i.e., the generalization across them, remains a challenge. Balancing task diversity with benchmark simplicity is crucial for gauging the true extent of LLM capabilities.

In addition, questions arise regarding the validity of existing metrics in accurately evaluating LLM performance, particularly in professional fields where precision is paramount. Refining evaluation metrics to align with professional tasks is essential for assessing LLMs' real-world applicability.

Addressing these challenges is essential for advancing our understanding of LLM capabilities and enhancing their utility across diverse domains. The following focus on methodologies explicitly designed to tackle these challenges, with a primary emphasis on *assessing* and *improving* the fundamental aspect of response reliability in Large Language Models.

2. Methodology

Our approach involves the following steps:

1. **Dataset Selection and Preparation:** Diverse datasets relevant to the research goals are selected, including Multiple-Choice Question Answer and Open-Ended Answer datasets from various domains. These datasets are appropriately formatted for analysis.
2. **Definition of Key Metrics:** Key metrics such as correctness, coherence, relevance, and completeness are introduced and defined to assess response quality.
3. **Question Extraction:** A subset of 50 questions is extracted from each dataset.

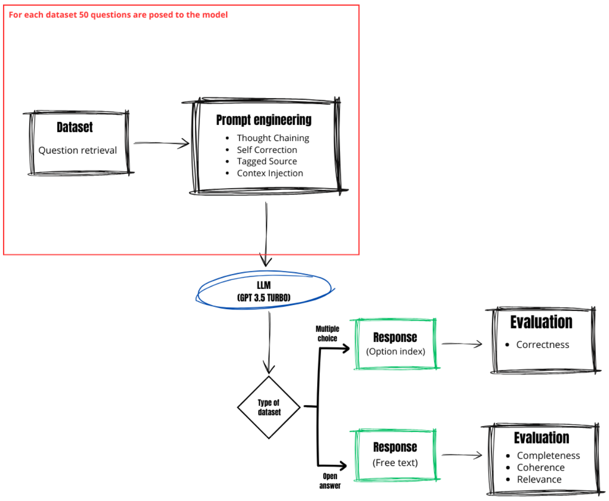

Figure 1: Methodology Flowchart

4. **Prompt Engineering:** Various prompt engineering strategies like Thought Chaining, Self Correction, Context Injection, and Tagged Source are systematically applied to the specific question. The goal is to query the LLM to generate tailored responses aligned with the unique characteristics of the prompt.
5. **LLM Evaluation:** The original and engineered questions are posed to the LLM (GPT-3.5 Turbo), and responses are collected for each question.
6. **Evaluation Criteria:** For Multiple Choice Question Answering (MCQA) datasets, correctness is evaluated by comparing the model's response (an option index) with the correct option in the ground truth dataset. For Open-Ended Answer datasets, completeness, coherence, and relevance of the free-text responses are evaluated.

2.1. Assessment Metrics

We defined different evaluation functions in order to assess the diverse aspect of quality response:

- **Correctness** assesses how well the model-generated responses match the correct answers present in the reference dataset(ground truth). The correctness percentage and a list of correctness scores (1 for correct, 0 for incorrect) for each evaluated response are calculated.
- **Coherence** measures the similarity between the model-generated response and the correct response provided in the dataset (ground truth). It utilizes cosine similarity between embedding vectors of generated and correct responses to determine coherence scores.
- **Completeness** assesses how much of the correct response is covered or included in the generated response. It calculates the ratio of common words between the generated response and the correct response to the total number of words in the correct response.
- **Relevance** evaluates how relevant the generated response is to the original question. It utilizes the pre-trained Sentence-BERT model to calculate the similarity between embedding vectors of responses and questions.

2.2. Dataset Choice

Here we highlight the critical role of dataset choice in evaluating LLM responses. We base our selection on:

- **Type of Dataset:** Our chosen datasets comprise *open-answer datasets* and *multiple-choice question-answer (MCQA) datasets*. Open-answer datasets are designed to assess coherence, completeness, and relevance, while MCQA datasets simplify assessment with strictly categorized choices of correct or incorrect.
- **Dataset Domain:** To ensure a comprehensive evaluation across various domains, we selected datasets spanning multiple subjects, and domain-specific ones, particularly in the fields of Medicine and Mathematics.

2.3. Strategic Embrace of Prompt Engineering

In our exploration of methodologies to enhance Large Language Models (LLMs), namely *Prompt Engineering*, *Retrieval-Augmented Generation (RAG)*, and *Parameter Efficient Fine-Tuning (PEFT)*, a definitive decision has been made to strategically focus on the art and science of *Prompt Engineering*. This deliberate choice stems from the recognition that Prompt Engineering stands out as the most cost-effective and efficient avenue for refining and enhancing the reliability of LLM responses.

Prompt Engineering involves a meticulous approach to crafting queries for LLMs. It plays a pivotal role in shaping the quality of model outputs by emphasizing precision, completeness, and the provision of illustrative examples. This method serves as a potent means to "tune" the model without necessitating updates, facilitating swift iterations for continuous improvement. However, it is crucial to acknowledge the inherent limitations of Prompt Engineering, specifically, it cannot impart knowledge to the LLM beyond its training data.

In the realm of *Prompt Engineering*, numerous techniques have been explored to enhance the model's cognitive processes[2]. Notable approaches include:

- **Chain of Thoughts (CoT):** The CoT approach advocates for a systematic and methodical chain of reasoning before generating responses. This methodology emphasizes breaking down complex tasks into smaller, more manageable steps, akin to a structured thought process. The prompt, *"Think step by step before answering"*, serves as a directive aimed at promoting a systematic approach to response generation.
- **Chain of Thoughts with Self-Correction (CoT-SC):** The CoT-SC approach refines the Chain of Thought (CoT) methodology by introducing self-consistency as a guiding principle for response generation. This approach prompts the model to provide multiple answers to a given prompt, thereby fostering diversity and enabling evaluation based on different perspectives. The prompt, *"Think step by step before answering and give three answers: if a domain expert were to answer, if a supervisor were to answer, and your answer"*, embodies a multi-faceted approach to response generation, encompassing perspectives from both domain-specific expertise and supervisory guidance.
- **Tagged Source Prompts:** Tagged Context Prompts represent an advancement in prompt design aimed at enhancing prompt effectiveness by incorporating additional context and verification mechanisms. These tagged prompts may include metadata or annotations that provide further insights into the context or intended meaning of the prompt, facilitating more precise and targeted response generation. The prompt, *"Provide details and include sources in the answer"*, underscores the importance of thoroughness and citation in response generation.
- **Self-Correct:** This approach aligns with the principles of self-refinement and accuracy improvement, offering a systematic method to mitigate errors and enhance the overall reliability of model outputs. By engaging in this self-refinement process, language models can enhance their ability to produce reliable, contextually appropriate, and accurate outputs. Moreover, by incorporating principles of critical thinking and self-assessment, models can contribute to more trustworthy and credible interactions in various domains and applications. The prompt *"Choose the most likely answer from the list. Then carefully double-check your answer. Think about whether this is the right answer, would others agree with it? Improve your answer as needed."* encourages a critical self-assessment process for the model, promoting reflection and refinement of generated responses.
- **Context Injection:** Context Injection prompts serve as a guiding mechanism to provide specific information or scenarios tailored to the model's training objectives and target domain. By injecting context into the prompt, the model is equipped with relevant background information or situational cues, which aids in refining its understanding and generating contextually appropriate responses. The effectiveness and specificity of Context Injection prompts vary depending on the datasets utilized.

3. Experiments

The experimental methodology employed in this study centers around the iterative application of prompt strategies using OpenAI's GPT-3.5 Turbo model. The experiments aim to comprehensively evaluate the model's performance across various datasets and prompt conditions, shedding light on its capabilities and limitations. The experiments are conducted using a Python codebase developed to interact with the

GPT-3.5 Turbo API and execute prompt-driven queries. This codebase facilitates the systematic application of prompt strategies across different datasets, ensuring a nuanced analysis of the model's behavior.

3.1. Datasets

A variety of datasets are utilized in the experiments, ranging from multiple-choice question-answer datasets to open-answer datasets. These datasets cover domains such as medical, mathematics, and general knowledge, providing a comprehensive evaluation of the model's performance.

3.1.1 medMCQA Dataset

The medMCQA dataset focuses on evaluating the model's proficiency in answering multiple-choice medical questions. It encompasses a diverse set of medical topics, requiring the model to select the correct answer option for each question.

3.1.2 mathQA Dataset

The mathQA dataset evaluates the model's mathematical problem-solving capabilities through multiple-choice questions spanning various mathematical concepts and difficulty levels.

3.1.3 RACE Dataset

The RACE dataset assesses the model's reading comprehension skills by presenting passages followed by multiple-choice questions. It covers diverse topics and difficulty levels to evaluate overall comprehension.

3.1.4 Multi-Context Long Answer Dataset

The Multi-Context Long Answer dataset challenges the model to generate detailed and comprehensive open-ended responses across a range of general knowledge questions.

3.1.5 MedQuad-MedicalQnA Dataset

The MedQuad-MedicalQnA dataset focuses on open-answer questions in the medical domain, testing the model's ability to generate detailed and accurate medical responses.

3.2. Iterative Application

The experimental process involves the iterative application of prompt strategies to each dataset. For every distinct dataset, 50 questions are posed for each prompt type, ensuring a comprehensive evaluation of the model's performance. As the experiments progress, the responses generated by the model are systematically collected and stored for subsequent analysis. This meticulous data collection process spans across different datasets and prompt conditions, creating a diverse repository of model-generated content.

4. Result Analysis

The analysis of the experiment results offers valuable insights into the effectiveness of different prompt strategies across various datasets, shedding light on trends and observations that provide meaningful implications for enhancing language model performance.

4.1. Correctness in General Domains

When assessing the RACE Middle and RACE High datasets, which represent general domains, several noteworthy trends emerge. In the RACE Middle dataset, both the original question and the Chain of Thoughts (CoT) prompt achieved commendable performance with an 80% correctness rate. Conversely, the CoT with Self Consistency (CoTSC) prompt yielded notably lower scores of around 40%, indicating the adverse impact of excessive constraints on the model's responsiveness. In the RACE High dataset Tagged Source and Self Correct prompts excelled at 80%. These findings suggest that simplicity and structured guidance can positively influence the model's performance in general domains.

4.2. Correctness in Specific Domains

In specific domains represented by the medMCQA and mathQA datasets, lower overall correctness percentages were observed. The behavior of CoTSC prompts remained anomalous across experiments, consistently scoring lower than other prompts. Notably, prompting engineering failed to confer benefits in certain scenarios, with the model demonstrating a propensity to provide correct answers when already within its knowledge base. However, employing a chain of thought prompting strategy yielded significantly higher results in the mathQA dataset, highlighting the efficacy of guiding the model through step-by-step reasoning, akin to human-like cognitive approaches.

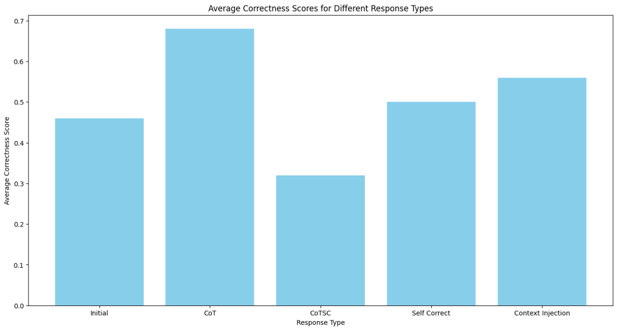

Figure 2: MathQA Average Correctness Score

4.3. Coherence, Completeness and Relevance

Analyzing the results from the MedQuad-MedicalQnA and MultiContextLongAnswer datasets reveals distinct trends across various prompts. Guided reasoning prompts, such as the Chain of Thoughts (CoT), consistently exhibit superior coherence and completeness compared to other prompts. However, the effectiveness of self-correction mechanisms, as seen in the Chain of Thoughts with Self Correction (CoTSC) prompts, is less pronounced, with mixed results indicating potential challenges in maintaining coherence and completeness. Interestingly, prompts incorporating source material, like the Tagged Source prompts, consistently outperform others in coherence, completeness, and relevance. In the end, the impact of context injection prompt varies, showing slight improvements in coherence and completeness in some cases but depending on the specific domain or task at hand.

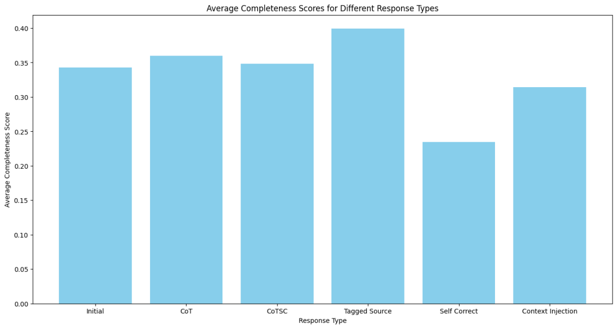

Figure 3: MedQUAD Average Completeness Score

5. Conclusions

To conclude, in this thesis we conducted an extensive exploration of prompt engineering and its implications for response quality in language models.

Important takeaways have emerged. Firstly, we found that guided reasoning prompts, such as Chain of Thoughts (CoT) and strategies incorporating source retrieving like Tagged Source, significantly enhance response quality. Conversely, prompt strategies involving self-correction mechanisms pose challenges that may outweigh their benefits. Additionally, the effect of inject contextual cues can vary.

Looking ahead, there are several avenues for future exploration and research. This includes comparative analyses across different models such as GPT-4, Llama 2 or GEMINI, and the expansion of prompt engineering pipelines to integrate techniques like Retrieval-Augmented Generation (RAG) and Fine-Tuning.

By pursuing these avenues of exploration, researchers can deepen our comprehension of prompt engineering and response quality in language models.

References

[1] Google Developers. Introduction to Large Language Models, 2023-08-08. Accessed: February 10, 2024.

[2] OpenAI. Techniques to improve reliability, 12-09-2022.

[3] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023.