RESEARCH ON PROMPT ENGINEERING AND PROPOSED SOLUTIONS FOR DATA GENERATION IN EDUCATIONAL QUESTION-ANSWERING SYSTEMS

Trần Thị Thu Phương, Nguyễn Quốc Tuấn, Lê Thị Hằng
Hanoi Metropolitan University
Bùi Đức Trung
CNTT D2021A - Hanoi Metropolitan University

Abstract: *This research explores prompt engineering techniques—Zero-Shot, Few-Shot, Chain of Thought (CoT), and Retrieval-Augmented Generation (RAG)—to assess their effectiveness in educational question-answering systems. It evaluates their ability to handle complex queries, reason through multi-step problems, and generate accurate, contextually relevant responses. Results indicate that CoT and RAG are particularly effective for tasks requiring logical reasoning and multi-source information synthesis, while Zero-Shot and Few-Shot methods are more efficient for straightforward questions with lower computational demands. The research highlights the crucial role of prompt engineering in enhancing model performance and generating high-quality datasets for educational applications. Practical solutions are proposed, including optimizing prompts for different question types, leveraging retrieval-based methods to dynamically update responses, and balancing efficiency with computational costs. These findings contribute to advancing educational question-answering systems, enabling Large Language Models (LLMs) to deliver precise, well-contextualized, and reliable responses in academic settings.*

Keywords*: Educational question-answering system, data generation, dataset, large language models, prompt engineering.*

Nhận bài ngày 10.01.2024; gửi phản biện, chỉnh sửa, duyệt đăng ngày 20.02.2025
Liên hệ tác giả: Trần Thị Thu Phương; email: tttphuong2@daihocthudo.edu.vn

1. INTRODUCTION

Since the development of large language models like Generative Pre-trained Transformer-GPT-3 and GPT-4, prompt-based text generation techniques have become an important tool across various applications. In the educational context, automatic question-answering systems can support students and educators in information retrieval and answering queries regarding academic regulations, admissions guidelines, or specific rules related to learning and exams. When properly optimized, these systems can save time and enhance learning efficiency, especially in environments with high demands for accuracy and detailed information.

One of the biggest challenges in implementing these question-answering systems is how language models process and respond to complex questions, particularly in contexts requiring multi-step reasoning or information from multiple sources. Techniques such as Zero-Shot, Few-Shot, and Chain of Thought (CoT) have been developed to improve model capabilities in understanding and handling complex tasks. However, the effectiveness of each technique in specific tasks has not been thoroughly evaluated in the educational context.

Several studies have explored the impact of prompt engineering on language models. Brown et al introduced the concept of few-shot learning in GPT-3, demonstrating its ability to generalize across multiple tasks with minimal examples [1]. Wei et al later proposed Chain of Thought (CoT) prompting, which significantly improved multi-step reasoning tasks in LLMs [2]. Meanwhile,

Lewis et al. introduced Retrieval-Augmented Generation (RAG) as an approach that enhances model responses by integrating external knowledge sources, improving factual accuracy in question-answering systems [3]. These works highlight the evolving strategies for improving language model performance in different domains.

In the Vietnamese context, research on large language models for question-answering remains limited. Trang et al examined Vietnamese question-answering systems using BERT-based models but found challenges in handling complex queries and maintaining accuracy. Sang T. Truong et al explored prompt tuning for Vietnamese LLMs, but applications in education remain underdeveloped.

The goal of this study is to experiment with and evaluate prompt techniques for data generation in question-answering systems and to compare the effectiveness of each method in tasks of varying difficulty. Through several experiments, the study proposes methods to use prompts for data generation in educational question-answering systems.

This research contributes to the field of prompt engineering and question-answering systems by providing solutions for data generation in educational question-answering systems. The results will provide important insights for building datasets for other question-answering systems using LLMs, particularly in the Vietnamese language.

2. CONTENT

2.1 Prompt engineering

Prompt engineering is a crucial method in using LLMs to optimize output results. A prompt is a set of instructions or queries provided by the user to trigger the model to process and generate a corresponding response. A typical prompt includes key components such as instructions, context, input data, and output indicators to help the model clearly understand the task's requirements and context. Prompt techniques play a key role in guiding the model to generate suitable data while ensuring accuracy and handling diverse tasks.

This research focuses on four popular prompt techniques: Zero-Shot Prompting, Few-Shot Prompting, Chain of Thought (CoT), and Retrieval-Augmented Generation (RAG). Each technique has its characteristics and applications suitable for different tasks.

2.1.1 Zero-Shot prompting

Zero-Shot Prompting involves asking the model to solve a task without any prior examples, solely based on the initial prompt. This approach is well-suited for simple tasks that require clear and concise information [1] [4]. For example, if a model is asked to classify text based on sentiment, such as neutral, negative, or positive, Zero-Shot Prompting can quickly provide a response if the content is not too complex.

Example:

Prompt: "Classify the following text as neutral, negative, or positive."

Text: ""I think this vacation was okay"

Result: "Neutral."

However, the limitation of this technique becomes apparent when facing more complex tasks that require deep reasoning or processing of structured data. In such cases, the results are often inaccurate and incomplete. Zero-Shot Prompting is widely used due to its simplicity and minimal resource requirements, but it is less effective when dealing with tasks that require multi-level reasoning or analysis from various sources.

2.1.2. Few-shot prompting

The Few-Shot technique introduces a small number of examples to the model before it tackles a task. This approach enables the model to grasp the context of the task more effectively and enhances the accuracy of its responses, especially for tasks with greater complexity [1]. It is widely appreciated for improving the model's ability to learn from provided examples, significantly boosting response precision when dealing with tasks of moderate difficulty [5].

Example:

Few-Shot Prompt: "Here are a few examples of text classification. Classify the new text based on these examples."

- Example 1: "I am very satisfied with the service."
  - Result: "Positive."
- Example 2: "The service was not as expected."
  - Result: "Negative."
- New text: "This vacation was not bad."
  - Generated result: "Neutral."

Few-Shot Prompting significantly improves accuracy compared to Zero-Shot, especially when tasks require a certain level of reasoning. However, it is still not strong enough to handle tasks that demand multi-step reasoning or when information needs to be gathered from multiple sources. The responses may lack comprehensiveness and depth in such situations.

2.1.3. Chain of thought – CoT

Chain of Thought (CoT) is a technique that breaks down a task into sequential reasoning steps, allowing the model to analyze each step logically before reaching a final conclusion. CoT is particularly useful for tasks requiring multi-step reasoning, such as scientific problems or tasks requiring complex analysis [2].

Example:

Prompt: "Explain the process of climate change and its impact on ecosystems."

Step-by-step reasoning:

1. Greenhouse gas emissions increase due to human activities.
2. This raises the Earth's temperature, causing global warming.
3. Global warming leads to melting ice caps, rising sea levels, and extreme weather events.
4. The ultimate result is the destruction of ecological environments.

By using CoT, the model can reason step by step in a logical and coherent manner, thereby increasing the accuracy and reasonableness of the response. However, this technique requires more computational resources and longer processing time compared to simpler techniques like Zero-Shot and Few-Shot. CoT is especially useful in complex tasks that require continuous reasoning.

2.1.4. Retrieval-augmented generation - RAG

Retrieval-Augmented Generation (RAG) is a technique that combines retrieving information from external sources with generating data from a large language model. This technique allows the model to retrieve new information from various sources and then synthesize and generate the most accurate and up-to-date response [3].

Example:

- Prompt: "Provide the latest information on university admissions regulations at Hanoi Metropolitan University."
- RAG process: The model retrieves information from the latest regulatory documents and then generates a response based on the retrieved information.

RAG is an effective technique when the task requires new information or needs verification from multiple sources. However, its effectiveness depends largely on the quality of the data sources retrieved, and managing computational resources must be optimized to ensure processing time is not extended [3].

In general, each prompt technique has its own advantages and disadvantages, suited to different types of tasks. Zero-Shot and Few-Shot are suitable for simple tasks, requiring fewer resources and offering faster processing times. Meanwhile, CoT and RAG are better choices for more complex problems that require multi-step reasoning or the retrieval of information from multiple sources. The choice of method depends on the complexity of the task, as well as the requirements for processing time and resources.