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.

2. 3. Proposed solutions for data generation in educational question-answering systems

2.3.1. Educational question-answering systems and characteristics of educational domain QA

A question-answering (QA) system is a crucial area in natural language processing (NLP), where the system is required to answer user questions based on available or retrieved information. The core of a QA system is its ability to understand the question, search and extract information from documents or existing data, and provide accurate, comprehensive, and contextually relevant answers.

When applied to the educational domain, the task becomes more complex due to the nature of the data in the educational environment, which mainly includes policy texts, training regulations, study guides, and documents related to student admissions and academic processes. The distinctive feature of educational QA is that questions often involve legal regulations, requiring high precision and, in some cases, reasoning across multiple sources of information.

For example, in educational QA related to training regulations, questions may concern scholarship review processes, graduation requirements, or admissions procedures. To answer accurately, the system must connect and thoroughly understand the content of each clause in the regulations to provide persuasive and logical responses. Thus, solutions proposed for educational QA systems must not only focus on handling simple questions but also emphasize the ability to reason logically and synthesize information from multiple sources.

2.3.2. The role of data in training and fine-tuning models for QA systems

Data plays a central role in training and fine-tuning LLMs to meet the needs of QA systems. The quality, structure, and completeness of the data directly affect the model's accuracy and generalizability. For educational QA systems, data must be clearly structured, accurate, and diverse, including policy documents, regulations, guidelines, and policies related to training, admissions, and scholarships.

Training data must not only reflect all aspects of the educational process but also be structured so that the model can handle a wide range of questions, from simple extraction questions to complex reasoning questions that require connections between different parts of the documents. This means that the model needs to be trained on datasets with a balanced distribution of question complexity to ensure the model can understand and answer questions of varying complexity.

2.3.3. Key QA Datasets

Surveying existing QA datasets is crucial in developing LLMs for question-answering systems. These datasets provide a strong foundation for training and fine-tuning models, helping them become effective at extracting information and answering questions. Popular QA datasets such as SQuAD 1.1, SQuAD 2.0, and TREC-8 QA have provided a strong foundation for training and fine-tuning QA models. These datasets are designed with clear question and answer structures, enabling the model to develop accurate information extraction capabilities.

- SQuAD 1.1: Provides questions and answers based on a fixed passage of text, with the goal of finding and extracting the exact passage containing the answer from the document. This dataset mainly focuses on questions where answers can be found directly in the text.
- SQuAD 2.0: An improved version of SQuAD 1.1 that adds questions without answers, requiring the model not only to extract information but also to identify when an answer does not exist in the text. This helps the model develop the ability to recognize cases where questions cannot be answered based on available data.
- TREC-8 QA: Focuses on information retrieval and extraction tasks from a large database, requiring the model to retrieve information from multiple sources of documents to provide detailed and accurate answers.

In the Vietnamese context, datasets like ViQuAD and ViLLM-Eval have been developed to support question-answering systems in the Vietnamese language. Notably, ViQuAD is a SQuAD-style dataset customized for the Vietnamese context, helping improve the model's ability to process Vietnamese language and answer questions based on Vietnamese text.

2.3.4. Solutions for data generation in educational QA systems using prompt techniques

To address educational QA systems, applying prompt engineering techniques is essential to guide large language models in generating accurate and contextually relevant data for the educational environment. Techniques like Zero-Shot, Few-Shot, Chain of Thought (CoT), and Retrieval-Augmented Generation (RAG) can be used depending on the nature of the question and the requirements for accuracy and reasoning from the input data.

- Zero-Shot Prompting: This can be used in situations where there are no specific examples and is suitable for simple questions that require direct answers from regulatory documents. For example, when answering questions like "What are the admission requirements?", the system can rely on fixed regulations to provide an accurate answer without complex context.
- Few-Shot Prompting: This is particularly useful when providing the model with a few examples before performing the task. This technique enhances the model's ability to understand the context and provide more accurate answers to moderately complex questions. This is ideal for questions like “How is a scholarship review conducted?” where the model needs to understand and apply examples from previous scholarship processes.
- Chain of Thought (CoT): CoT should be applied when questions require multi-step reasoning or analyzing related elements in regulatory documents. For example, when answering a question about the scholarship review process, CoT helps the model break the task into smaller steps, from eligibility conditions to the review process to the final decision, thus providing a systematic and accurate answer.
- Retrieval-Augmented Generation (RAG): RAG is suitable when information needs to be retrieved from external sources or multiple parts of a document to provide the most up-to-date and comprehensive answer for the user. This is especially important in the educational environment, where regulations and policies frequently change, and it is necessary to retrieve information from multiple documents to provide the most accurate and current answer.

Thus, selecting the appropriate prompt technique depends on the nature of the question, the complexity of the information to be retrieved, and the requirements for accuracy in the response. Combining various techniques can help optimize the data generation process, ensuring that the QA system can provide accurate, valuable, and contextually appropriate answers for the educational environment.

2.4. Results and Discussion

2.4.1. Zero-Shot Prompting

Zero-Shot Prompting shows effectiveness when applied to simple tasks, such as generating short and direct answers from texts. However, when faced with questions that require deep reasoning or contain multiple pieces of connected information, Zero-Shot models tend to produce inaccurate or incomplete responses.

For example, when tasked with generating questions and answers from admissions regulations, Zero-Shot tends to extract only simple questions like "When does the admissions period start?" without delving into other important details in the text. These results indicate that Zero-Shot is not suitable for tasks requiring multi-step reasoning or processing complex information.

2.4.2. Few-Shot Prompting

Few-Shot Prompting shows a clear improvement over Zero-Shot in handling more complex tasks. By providing a few examples, the model can better understand the context and generate more accurate questions and answers. However, the results also show that when tasks require multi-step reasoning or gathering information from multiple sources, Few-Shot still does not achieve optimal effectiveness.

For example, when tasked with generating questions about clauses in educational regulations related to the number of credits required to complete a program, Few-Shot is more capable of identifying and generating suitable questions like "How many mandatory credits must students

complete in the training program?" However, for more complex tasks, such as questions about differences between clauses, Few-Shot still struggles to generate detailed and comprehensive answers.

2.4.3. Chain of Thought (CoT)

The CoT technique produces outstanding results when applied to tasks requiring multi-step reasoning. The answers generated by the model are not only accurate but also follow clear logical reasoning, making it easier for the reader to understand and follow the reasoning process.

In an experiment with complex questions like "How can the admissions process be conducted fairly for all students?" CoT broke down the question into logical steps and solved each step systematically, resulting in detailed and coherent answers.

2.4.4. Retrieval-Augmented Generation (RAG)

The RAG technique allows the model to retrieve information from external sources, which increases the accuracy and relevance of the answers. In the experiment, RAG demonstrated its distinct advantage when the model was asked to answer questions requiring new or previously untrained data.

For example, when asked to provide the latest information on university admissions regulations at Hanoi Metropolitan University, RAG retrieved data from reliable sources and generated detailed and accurate responses. However, one limitation of this technique is the longer processing time and greater computational resource requirements due to the need to retrieve information from multiple sources.

TABLE 1 COMPARISON OF PROMPTING TECHNIQUES WITH DISCUSSION

Technique	Strengths	Weaknesses	Best Use Cases
Zero-Shot Prompting	Effective for simple tasks, fast processing	Struggles with complex reasoning, incomplete answers	Simple Q&A, direct text extraction
Few-Shot	Better contextual understanding with examples, improved accuracy	Limited performance on multi-step reasoning	Generating questions/answers from structured documents
Chain of Thought	Excels in multi-step reasoning, provides logically structured answers	Higher computational cost, requires structured	Complex problem-solving, logical reasoning
Retrieval-Augmented Generation	Retrieves external data for up-to-date, detailed responses	Longer processing time, higher computational resource demand	Updating responses with the latest information, handling dynamic data

The table 1 outlines the strengths and weaknesses of different prompting techniques in educational question-answering systems. Zero-Shot Prompting is highly efficient for simple tasks but struggles with deep reasoning, making it unsuitable for complex questions. Few-Shot Prompting improves upon Zero-Shot by using a few examples to enhance contextual understanding, yet it still faces challenges with multi-step reasoning. Chain of Thought (CoT) excels in logical reasoning by breaking down problems into structured steps, producing more accurate and comprehensible answers, though it requires structured prompt design and additional computational resources. Retrieval-Augmented Generation (RAG) delivers the most accurate and up-to-date responses by retrieving external information, but it demands more processing time and computational resources. The choice of technique depends on task requirements—Zero-Shot and Few-Shot are ideal for efficiency, while CoT and RAG are better suited for reasoning-intensive and information-dense tasks.

3. CONCLUSION

This research experimented with and evaluated various prompt techniques for generating question-answering data from educational policies. The results show that techniques like Chain of Thought (CoT) and Retrieval-Augmented Generation (RAG) are particularly effective in handling complex tasks, while Zero-Shot and Few-Shot are better suited for

simpler tasks.

The study provides a comprehensive evaluation of prompt techniques, contributing to the development of automatic question-answering systems in education. The experimental results indicate that applying advanced techniques such as CoT and RAG can improve the reasoning ability and accuracy of the system, especially in contexts requiring multi-step reasoning or information retrieval from multiple sources.

One limitation of this research is that techniques like CoT and RAG require large computational resources and longer processing times compared to simpler techniques like Zero-Shot and Few-Shot. In the future, we propose expanding the scope of research by experimenting with larger and more diverse datasets, including data from social media, websites, and other multimedia sources. Optimizing language models to reduce computational costs in data generation is also a potential research direction.

REFERENCES

1. T. B. M. B. R. N. e. a. Brown (2020), "Language models are few-shot learners," *Advances in Neural Information Processing Systems,* pp. 33, 1877-1901.
2. J. W. X. S. D. e. a. Wei (2022), "Chain of thought prompting elicits reasoning in large language models.," *arXiv preprint arXiv:2201.11903*.
3. P. P. E. P. A. e. a. Lewis (2020), "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems,* pp. 33, 9459-9474.
4. A. W. J. C. R. e. a. Radford (2019), "Language models are unsupervised multitask learners.," OpenAI blog.
5. C. S. N. R. A. e. a. Raffel (2020), " Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research,* pp. 21(1), 5485-5551.

NGHIÊN CỨU MỘT SỐ KỸ THUẬT TẠO LỜI NHẮC VÀ GIẢI PHÁP SINH DỮ LIỆU CHO HỎI ĐÁP MIỀN ĐÀO TẠO

Tóm tắt: *Nghiên cứu này xem xét các kỹ thuật tạo lời nhắc bao gồm: Zero-Shot (lời nhắc không ví dụ), Few-Shot (lời nhắc kèm ví dụ), Chain of Thought (Chuỗi tư duy-CoT) và Retrieval-Augmented Generation (Tạo sinh tăng cường truy suất- RAG) trong hệ thống hỏi đáp giáo dục. Nghiên cứu đánh giá hiệu quả của các kỹ thuật này trong việc xử lý các câu hỏi phức tạp, lập luận qua nhiều bước và tạo ra các câu trả lời chính xác. Kết quả cho thấy CoT và RAG vượt trội trong lập luận logic và tổng hợp thông tin từ nhiều nguồn, trong khi Zero-Shot và Few-Shot hoạt động hiệu quả hơn đối với các nhiệm vụ đơn giản với chi phí tính toán thấp hơn. Bài báo nhấn mạnh vai trò quan trọng của thiết kế gợi ý trong việc tạo bộ dữ liệu và tối ưu hóa mô hình, đồng thời đề xuất các chiến lược như tối ưu hóa gợi ý, cập nhật dựa trên truy xuất và cân bằng tài nguyên để cải thiện hệ thống hỏi đáp trong giáo dục.*

Từ khóa*: Hệ thống hỏi đáp miền giáo dục, kỹ thuật tạo lời nhắc, mô hình ngôn ngữ lớn, sinh dữ liệu, tập dữ liệu.*