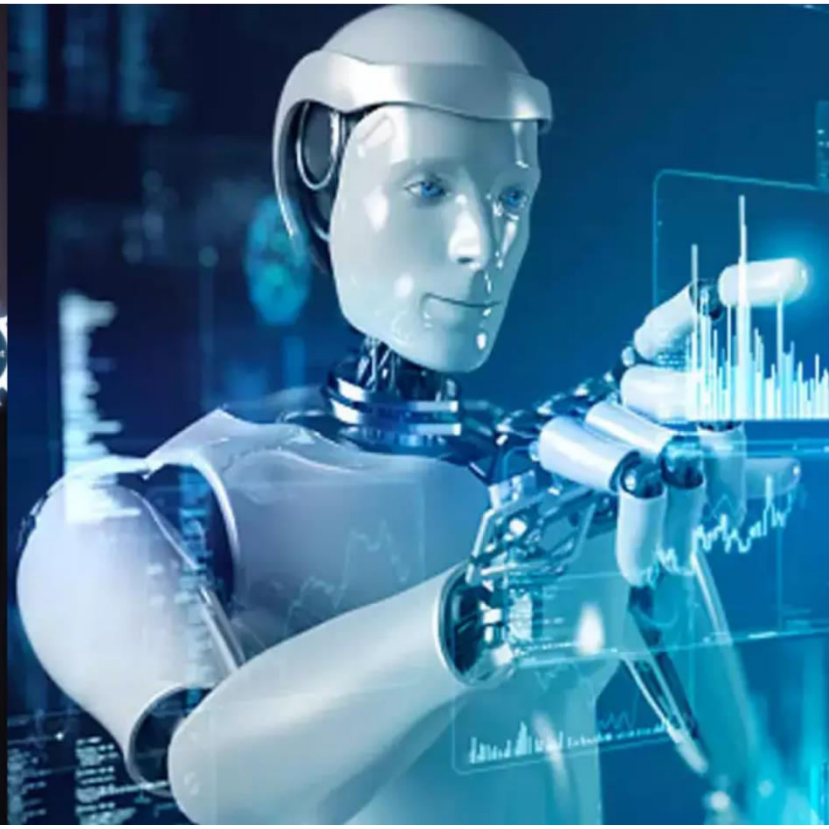
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.524

Volume 14, Issue 1, January 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 1, January 2025

|DOI: 10.15680/IJIRSET.2025.1401115|

A Comparative Study of Prompt Engineering Techniques for Large Language Models

Dr. Mital Vora

Assistant Professor, Department of Computer Science, T. N. Rao College, Rajkot, Gujarat, India

ABSTRACT: As artificial intelligence continues to evolve, it is important to understand large language models and how they function so that users can effectively utilize their capabilities. LLM generates human-like text and performs reasoning tasks across multiple domains. However, their performance strongly depends a lot on how user prompts are set up. Prompt engineering has thus become an effective method to train model behavior without modifying model parameters. This paper presents a comparative study of five commonly used prompt engineering techniques. An experiment is performed in which the GPT-based model and reasoning task are evaluated for different prompt strategies. The performance of the results was evaluated using reasoning accuracy and a human evaluation. The experimental results indicate that chain-of-thought and few-shot prompting enhance reasoning accuracy and increase output quality.

KEYWORDS: Prompt Engineering, Large Language Models, Generative AI, Chain-of-Thought, Comparative Study

I. INTRODUCTION

With the advance of AI, it has become more and more valuable to understand LLMs and how to use them well. LLMs are a significant step forward in AI systems, as they allow machines to produce human-like text, answer questions, and make reasoning based on input from a vast variety of domains.

Language modeling in large language models involves guessing the most likely intended sequence of words given input. Whereas traditional AI systems are informed by explicit task-specific programming, LLMs are trained on large amounts of data and can subsequently generalize to new tasks with the same underlying model. This versatility has revolutionized the way in which AI systems communicate with users—permitting natural, intuitive interactions.

In LLM-based setups, the prompt acts as a primary channel between users and their intentions. Slight modification in a prompt's wording, format, or setting can cause dramatic differences in the quality and relevancy of generated responses. Therefore, prompt engineering has become a critical field of research on how to design good prompts that can instruct language models to produce more accurate, relevant, and coherent responses.

Several prompt engineering techniques have been suggested and widely used. This paper aims to provide the comparative analysis by evaluating multiple prompt strategies on the same task using transparent evaluation criteria.

II. RELATED WORK

Prompt engineering has become an important way to make large language models (LLMs) work better without changing their internal settings. Prompt engineering focuses on carefully designing input instructions without retrain or fine-tune the models to generate accurate and useful output.

Wei *et al.* [1] presented Chain-of-Thought (CoT) prompting, which encourages language models to generate intermediate reasoning steps before producing a final answer. Their study demonstrated that reasoning-based prompts significantly improve performance on arithmetic, logical reasoning, and multi-step problem-solving tasks.

Reynolds and McDonell [2] presented the concept of prompt programming, where prompts are treated as a lightweight programming interface for controlling model behavior. They demonstrated that structured prompts, including role definitions and step-by-step instructions, allow users to more effectively guide model outputs across various tasks such

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 1, January 2025

|DOI: 10.15680/IJIRSET.2025.1401115|

as text transformation, summarization, and reasoning. Their work highlighted how prompts can be used in many different ways as a control tool.

Liu et al. [3] conducted an extensive survey of prompting methodologies in natural language processing. They put prompting methods into four groups: zero-shot, few-shot, instruction-based, and reasoning-based. Their analysis showed that prompt engineering lowers the cost of computation and makes it easier to adapt to different tasks. This makes it especially useful for real-world and large-scale applications.

Mishra et al. [4] investigated instruction-based prompting utilizing the Natural Instructions benchmark. Their research demonstrated that offering clear, natural language task descriptions facilitates models' ability to generalize to novel tasks. But they also said that instructions that are too complicated might make the model less flexible, which shows that prompt design needs to be balanced.

Brown et al. [5] showed that large language models can do a lot of different things with few-shot prompting, which means that examples are built right into the prompt. Their research showed that model scale and good prompt design can lead to strong performance without training on specific tasks. This is the basis for modern prompt engineering research.

Schick and Schütze [6] demonstrated that efficient prompting methods can also enhance smaller language models. Their results showed that the design of prompts is very important for getting task knowledge, and that it can sometimes make up for a smaller model size. This work showed that prompt engineering isn't just for big models.

Zhou et al. [7] examined automatic prompt engineering, in which models autonomously generate and refine prompts. Their findings indicated potential enhancements in performance; nevertheless, the authors acknowledged difficulties concerning transparency and assessment. This study shows a move toward automation, but it also makes people worry about how easy it is to understand.

Most existing studies focus on individual prompt engineering techniques and evaluate them under different experimental conditions. There is limited research that compares multiple prompting strategies using the same tasks and evaluation criteria. This gap highlights the need for a comparative study of prompt engineering methods for large language models.

III. EXAMPLES OF PROMPT ENGINEERING TECHNIQUES

This research analyses five prevalent prompt engineering methodologies. A straightforward reasoning problem demonstrates how each technique directs the model in distinct ways.

- **Prompting with No Shots**
  Zero-shot prompting just gives you the task description. For example, "How fast is a car that goes 120 km in 2 hours?" This method is easy, but it often doesn't work well for reasoning tasks.
- **Prompting with Few Shots**
  Few-shot prompting has only a few examples. For example, a bike goes 60 kilometres in 2 hours. 30 km/h is the speed. Now figure out how fast a car goes when it travels 120 km in 2 hours. This helps the model figure out what kind of reasoning pattern to expect.
- **Prompting Based on Instructions**
  Instruction-based prompting gives clear steps. For example, "Use the formula Speed = Distance ÷ Time." The distance is 120 km and the time is 2 hours. Find out how fast it is. This makes things less confusing and more consistent.
- **Prompting Based on Role**
  Role-based prompting gives a character. For example, "You teach physics." Tell me how to figure out how fast a car goes when it goes 120 km in 2 hours. This often makes the quality of the explanation better.
- **Prompting the Chain of Thought**
  Chain-of-thought prompting helps people think through things step by step. For example, "Solve the problem step by step and tell me why you did it." This method makes reasoning tasks more clear and accurate.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 1, January 2025

|DOI: 10.15680/IJIRSET.2025.1401115|

IV. EXPERIMENT EVALUATION METHODOLOGY

All experiments utilized a GPT-based large language model accessed via a publicly available conversational interface. To make sure it was fair, the model was used without any fine-tuning or changing of parameters.

A task based on reasoning was chosen because it lets you objectively check for correctness. To account for differences in responses, each type of prompt was used five times. The main question stayed the same for all of the prompt strategies, but each one added some extra help.

I used both accuracy-based assessment and human evaluation to collect and evaluate the model outputs.

Accuracy Calculation
To find out how accurate the model was, I compared its final answer to the known correct answer. One point was given for each correct answer, and zero points were given for each wrong answer. Accuracy is the percentage of correct answers over several runs.

Human Evaluation
Three independent evaluators used a five-point scale to rate each response based on how relevant, clear, correct, and useful it was. The final human score is the average of all the evaluators' and runs' scores.

V. RESULTS AND ANALYSIS

Reasoning Accuracy
Table 1 presents the reasoning accuracy obtained for each prompting technique. Accuracy was calculated based on five independent runs per technique.

Table 1. Reasoning Accuracy Across Prompt Engineering Techniques

Prompting Technique	Number of Runs	Correct Responses	Accuracy (%)
Zero-shot	5	3	60
Few-shot	5	4	80
Instruction-based	5	3	60
Role-based	5	3	60
Chain-of-Thought	5	4	80

The results show that few-shot and chain-of-thought prompting achieved higher accuracy compared to zero-shot prompting.

Human Evaluation Results
Human evaluation results are summarized in **Table 2**. Scores represent average ratings provided by three evaluators across multiple runs.

Table 2. Average Human Evaluation Scores

Prompting Technique	Relevance	Clarity	Correctness	Usefulness	Avg Score
Zero-shot	3.0	3.1	3.0	3.2	3.1
Few-shot	4.0	4.1	4.0	3.9	4.0
Instruction-based	3.8	3.9	3.7	3.8	3.8
Role-based	3.6	3.7	3.5	3.6	3.6
Chain-of-Thought	4.2	4.3	4.1	4.2	4.2

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 1, January 2025

|DOI: 10.15680/IJIRSET.2025.1401115|

The human evaluation results align closely with accuracy findings, indicating that structured prompting techniques not only improve correctness but also enhance perceived response quality.

Analysis

The comparative analysis shows that prompt engineering considerably influences model performances. Zero-shot prompting is easy to use, but it doesn't always lead to good reasoning. Prompting based on instructions and roles makes things clearer, but it only slightly improves accuracy.

Few-shot prompts take advantage of the guidance from examples, whereas the chain-of-thought approach is the most powerful by guiding the reasoning process explicitly. These findings suggest that prompting strategies that encourage structured reasoning are particularly effective for logic-based tasks.

Limitations

This comparative study focuses on a single model and a small set of reasoning tasks. Human evaluation involves subjectivity, although multiple evaluators were used to reduce bias. Future studies may explore additional tasks, models, and automated evaluation techniques.

VI. CONCLUSION

This paper provided a comparative analysis of prompt engineering techniques for large language models within a logical and transparent experimental framework. The results show that structured prompting methods, especially chain-of-thought and few-shot prompting, greatly improve the quality of reasoning and output. The results show that prompt engineering is a good way to make LLM work better without having to retrain it.

REFERENCES

[1] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proc. ACL*, 2022.
[2] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," *arXiv preprint*, 2021.
[3] P. Liu, W. Yuan, J. Fu, *et al.*, "Pre-train, Prompt, and Predict: A Survey of Prompting Methods in Natural Language Processing," *ACM Comput. Surv.*, 2023.
[4] S. Mishra, S. Khashabi, D. Jurafsky, *et al.*, "Natural Instructions: Benchmarking Generalization to New Tasks from Natural Language Instructions," in *Proc. ACL*, 2022.
[5] T. Brown, B. Mann, N. Ryder, *et al.*, "Language Models Are Few-Shot Learners," in *Proc. NeurIPS*, 2020.
[6] T. Schick and H. Schütze, "It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners," in *Proc. NAACL*, 2021.
[7] Y. Zhou, Y. Liu, S. Cheng, *et al.*, "Large Language Models Are Human-Level Prompt Engineers," *arXiv preprint*, 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details