Comparative Analysis of Prompting Techniques of Reducing Hallucinations in LLMs

Jasleen Kaur
Department of Computer Science and Engineering
Bennett University
Greater Noida, Uttar Pradesh, India

Tanvi Kad
Department of Computer Science and Engineering
Bennett University
Greater Noida, Uttar Pradesh, India

ABSTRACT
Hallucination is one of the prominent challenges faced by LLMs in generating responses, hence, reducing the reliability and trustworthiness of LLMs across various domains. Effectiveness of these models highly depend on how they are prompted. While prior work explores individual prompting techniques, this research shows a comparative analysis of multiple free tier LLMs using seven prompting techniques- Zero-shot, One-shot, Few-shot, Tree-of-Thought, Chain-of-Thought, Thread-of-Thread and Graph-of-Thought in a unified framework. The authors evaluate the performance of these combinations on three metrices- accuracy, hallucination rate and response time, offering practical insights and empirical analysis for prompting strategies in real-world applications.

Keywords
Large Language Models, Hallucination, Prompting, Retrieval-Augmented Generation (RAG)

1. INTRODUCTION
Large Language Models show remarkable performances for Natural Language processing-driven tasks like answering questions, dialogue generation or summarizing tasks. Despite of fast and fluent responses, they tend to have hallucinated or fabricated citations spreading ill information.

However, their performance is very sensitive to how they are prompted. Despite of progress in evolution of prompting techniques, existing literature lacks multi-model evaluation and joint analysis of accuracy, hallucination rate and response latency.

This paper addresses this gap by evaluating multiple prompting strategies across multiple LLMs on three benchmark datasets- GSM8K, TruthfulQA, and MMLU.

2. RELATED WORK
[1] This survey paper presents a comprehensive survey of hallucinations in LLMs based on two core metrices- Prompt Sensitivity and Model Variability, according to which the models are further classified into four categories: prompt dominant, model-dominant, mixed-origin and unclassified. The authors show that Chain-of-Thought prompting is the most effective prompting strategy to reduces hallucination rates from 38.3% to 18.1%, however it still remains insufficient if model's knowledge is fundamentally limited.

[2] The study shows a survey of over thirty-two hallucination mitigation techniques in LLMs, separating them on based of two broad categories: prompt engineering and model development. They use RAG along with prompting strategies like UPRISE and SynTra, although the paper highlights that no single mitigation technique to reduce hallucination and focuses on combining multiple strategies are needed for deployment in high-stake domains.

[10] This paper shows that small models can be trained to flag specific parts of an answer that is unsure using reinforcement learning. Another part of research is focused on increasing reasoning ability in LLMs using techniques like Chain-ofThought prompting and self-consistency decoding. Such techniques help models produce intermediate reasoning before arriving at a final answer, thus increasing logical consistency and possibly reducing factual errors in their responses.

3. PROMPTING TECHNIQUES
This section defined the seven techniques evaluated in this study.

A. Zero-Shot Prompting: Model must rely on its pre-trained knowledge entirely, no examples provided.

B. One-Shot Prompting:Single example is given to the model explaining the task.

C. Few-Shot Prompting:It includes two or more examples including pattern recognition for model.

D. Chain-of-Thought (CoT): It involves step-by-step reasoning for mathematical tasks.

E. Tree-of-Thought (ToT): It explores multiple reasoning paths to tackle decision-making tasks.

F. Thread-of-Thought (ThOT):These involve structured reasoning threads for more complex tasks.

G. Graph-of-Thought (GoT); Used for non-linear problems by reasoning as graphs.

4. METHODOLOGY
The authors evaluated four LLMs, accessing them via the free Groq api to ensure replicability at zero cost:

1) LLaMA 3.1 8B – Lightweight, open-weights model from Meta featuring 8 billion parameters and offers improved reasoning and is optimized for speed. It uses an autoregressive transformer architecture.

2) LLaMA 3.3 70B (llama-3.3-70b-versatile) – Multilingual open-source language model. It is a text-only 70 billion parameter instruction tuned model designed to increase performance for multilingual dialogue and complex reasoning tasks.

3) LLaMA 4 Scout 17B – Multimodal Mixture Of expert model that specializes in visual understanding and text generation with a 10M token context length.

4) Qwen3 32B – Multilingual LLM from Alibaba. It has a dual mode mode functionality that switches between deep thinking mode and direct conversational mode for strong performance.

4.1 Datasets
a) GSM8K [3] comprising of mathematic reasoning

problems out of which ten problems were sampled uniformly from the test split.

b) TruthfulQA [4] consisting of questions designed to detect hallucination, sampled from the generation split with ten questions.

c) MMLU [5] covering 57 academic subjects with ten questions sampled to test declarative knowledge.

4.2 Models

a) LLaMA 3.1 8B (llama-3.1-8b-instant) [6]
b) LLaMA 3.3 70B (llama-3.3-70b-versatile) [7]
c) LLaMA 4 Scout 17B (meta-llama/llama-4-scout17b-16e-instruct) [8]
d) Qwen3 32B (qwen/qwen3-32b) [9]

4.3 Evaluation Metrices

a) Accuracy (ACC): Binary indicator (1=correct, 0=incorrect) by comparing model's final answer against the truth-label.
b) Hallucination Rate (HR): Binary indicator (1= fabricated, 0= grounded response) by manual inspection.
c) Response Time (RT): Wall-clock latency in seconds from query submission to complete response.

The entire testing code is publicly available at - Authors-SurveryPaperGitLink

5. RESULTS AND ANALYSIS

Tables 1a – 1c give us the accuracy by dataset and show some important patterns.

Table 1a – Accuracy (%) on GSM8K

Techniques	Model			
	LLaMA 3.1	***LLaMA 3.3***	***LLaMA 4 Scout***	***QWEN3***
Zero Shot	30	50	60	90
One - Shot	0	60	80	90
Few - Shot	80	90	90	90
CoT	70	90	90	90
ToT	80	90	90	90
GoT	70	90	90	90
ThoT	70	90	90	90

GSM8K (Mathematical Reasoning): The dataset highlights the most compelling evidence in favor of reasoning techniques, but only large models exhibit such behavior. Both LLaMA 3.3 70B, LLaMA 4 Scout, and Qwen3 32B have an accuracy rate of 90% for CoT, ToT, GoT, and ThOT – a ceiling effect that renders any further distinction between advanced techniques impossible. LLaMA 3.1 8B is the only model in the group that has an accuracy rate of 30% for Zero-shot and 0% for One-shot, thus proving that small models need detailed instructions for performing arithmetic operations. Few-shot (80%) for small models far exceeds Zero-shot (30%), by 50 percentage points.

Table 1b – Accuracy (%) on MMLU

Techniques	Model			
	LLaMA 3.1	***LLaMA 3.3***	***LLaMA 4 Scout***	***QWEN 3***
Zero Shot	100	100	100	10
One - Shot	100	100	100	30
Few - Shot	100	100	70	10
CoT	10	10	10	0
ToT	40	30	20	20
GoT	40	30	20	10
ThoT	10	30	20	20

MMLU (General Knowledge): MMLU demonstrates the biggest turnaround in the data set. Zero-shot and One-shot attain perfect scores of 100% accuracy on all three LLaMA versions, whereas CoT drops to 10%, i.e., a difference of 90%. ToT, GoT, and ThOT partly recover in performance (20–40%), but remain well behind the basic methods. The result is very important, since it shows that for factual multiple-choice problems, giving instructions to reason actively damages performance.

Table 1c – Accuracy (%) on TruthfulQA

Techniques	Model			
	LLaMA 3.1	***LLaMA 3.3***	***LLaMA 4 Scout***	***QWEN3***
Zero Shot	70	80	70	40
One - Shot	70	70	80	50
Few - Shot	80	80	80	60
CoT	60	50	50	40
ToT	50	60	50	40
GoT	60	70	60	50
ThoT	80	70	50	50

TruthfulQA (Factual Accuracy): In this dataset, Few-shot has the highest performance consistency with all the models (80% with LLaMA models). ThOT has relatively high performance (70-80% with LLaMA models), showing that the structured threading method used by ThOT keeps the facts coherent while doing open-ended question answering. Qwen3 32B has low performance consistently with all methods used for TruthfulQA task (40-60%) due to training data and instruction following differences.

Table 2: Performance Comparison of Prompting Techniques

Techniques	Metrics		
	Accuracy	***Hallucination Rate***	***Avg. Response Time***
Zero Shot	0.67	0.14	0.42
One - Shot	0.69	0.11	0.36
Few - Shot	0.78	0.17	0.52
CoT	0.48	0.48	1.39
ToT	0.55	0.41	1.77
GoT	0.57	0.38	1.28
ThoT	0.56	0.39	1.11

Table 2 presents the evaluation of various prompting techniques across multiple LLMs and how it affects accuracy, hallucination rates and the average response time of the models.

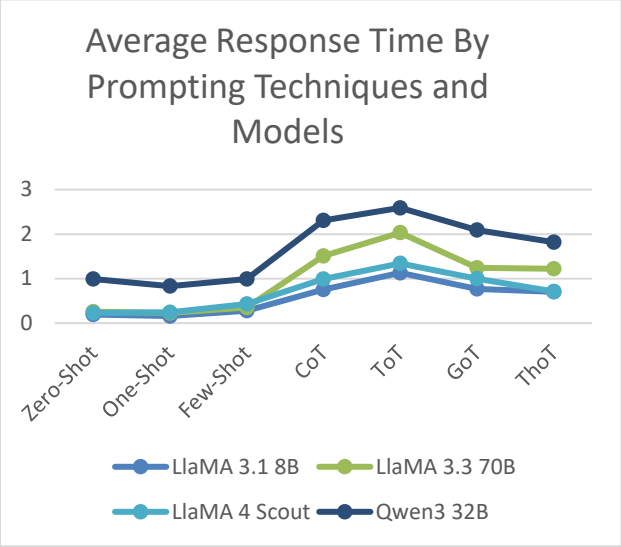

Figure 1

Figure 1 shows the considerable latency penalty incurred by advanced prompting approaches. One-Shot prompting offers the shortest average latency for all models (0.16-0.83s), followed closely by Zero-shot. Tree-of-Thought prompts incur the highest latency costs with Qwen3 32B having an average latency of 2.59s per query – almost 3.1× more than its Zero-shot latency (0.99s).

Specifically, for LLaMA 3.3 70B model, ToT incurs a penalty of 1.77s on top of Zero-shot (2.03s vs 0.26s) – a whopping 7.8× overhead – while failing to improve the accuracy on most datasets. In terms of practical implementation, this amounts to about 5.7 extra hours of latency per 10,000 queries in comparison to Zero-shot approach. Since GoT and ThOT also suffer from increased hallucinations, the cost-efficiency ratio of these two is unattractive in production use.

GoT and ThOT provide intermediate response latencies (0.70-2.09s) – lower than ToT but significantly higher than that of three simple baseline approaches. Of the reasoning approaches, ThOT is the best compromise in terms of speed and accuracy trade-off.

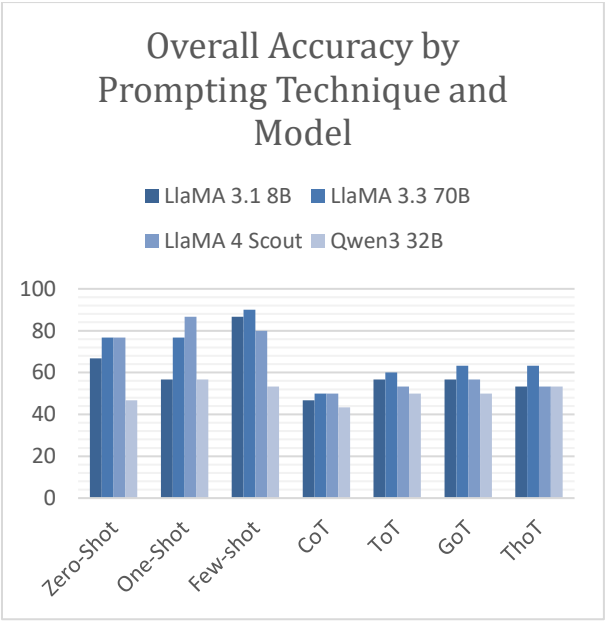

Figure 2

The aggregated results of all models and datasets are presented in Table 2 and Figure 1. Few-shot prompting demonstrates the highest accuracy (0.78) surpassing all other methods including reasoning frameworks. One-shot prompting has the second-highest accuracy (0.69), lowest hallucination rate (0.11) and fastest average response time (0.36s) making it an extremely efficient technique.

An important and counter-intuitive conclusion is made based on the results obtained: the lowest accuracy (0.48) is reached by Chain-of-Thought prompting while having the highest hallucination rate (0.48) among all seven techniques tested. This finding challenges the common idea of increased accuracy due to increased reasoning. An equality of the accuracy and hallucination rate of CoT (both 0.48) means that in case of failure of CoT, it is caused by hallucination rather than by mere ignorance.

The advanced reasoning techniques, ToT, GoT and ThOT are comparable to each other in terms of their accuracy (0.55-0.57) and hallucination rate (0.38-0.41) but significantly lag behind the three simpler baselines in both parameters.

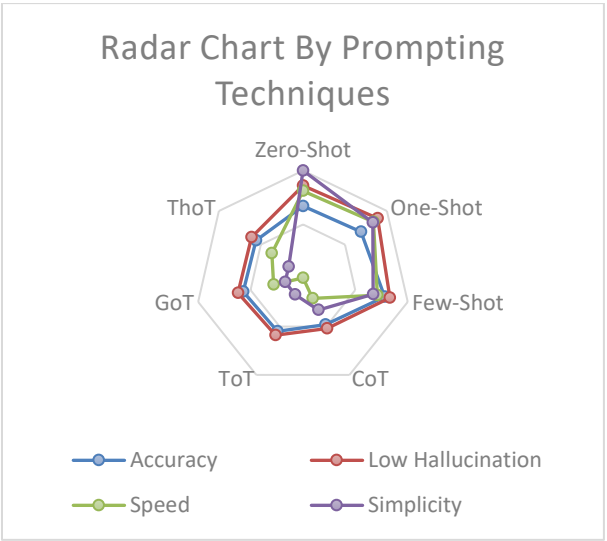

Figure 3

Figure 3 shows radar plots that visualize all four dimensions at once for each algorithm. The Few-shot approach shows the most balanced profile in terms of accuracy and low hallucination, whereas One-shot outperforms in speed and simplicity with comparable accuracy. CoT has the least balanced profile — its accuracy drops sharply in knowledge-based tasks along with increased hallucination and latency simultaneously. The Tree-of-Thought method shows the worst result in speed, which confirms its computational complexity in comparison with other algorithms.

6. KEY INSIGHTS

1) Simple Prompting Can Be More Effective

Few shot and one-shot prompting techniques outperformed more complex reasoning-based techniques in both accuracy and hallucination control suggesting that the assumption more reasoning gives better performance is not always true.

2) Reasoning Based Prompting Associated with Higher Hallucination Rates

Techniques like CoT, ToT and GoT demonstrated significantly higher hallucination rates (0.38-0.48) suggesting that models may generate incorrect reasonings and that errors can be caused by intermediate steps.

3) Trade-Off between Depth and Efficiency

Advanced reasoning methods showed to increase computational time however they do not guarantee factual correctness. Thus, the optimal prompting technique varies by task and no single technique is universally accurate.

4) Few Shot Prompting is the Most Reliable Technique

Few-Shot Prompting Technique provided the best balance between all the three of evaluation metrics, with the highest accuracy (0.78), a low hallucination rate (0.17), and reasonable response time (0.52s), making it a practical technique for real world applications.

5) Structured Reasoning Not Always Reduces hallucination

The study shows that structured reasoning techniques also do not guarantee factual correctness. Qwen3 32B(32B parameters) underperforms LLaMA 3.1 8B(8B parameters) consistently on accuracy tasks. Hence Hallucination is not just a reasoning problem, but also a knowledge and confidence evaluation problem.

6) Reasoning Behind Underperformance of Complex Techniques

The underperformance of CoT, ToT, ThOT, and GoT on MMLU and their high rates of hallucinations in all datasets are important consideration points. Two of the potential reasons behind this phenomenon are presented here. The first one is that the prompts involving complex reasoning would result in the generation of more tokens, which means the likelihood of having a mistake somewhere in the middle that would propagate to the end result becomes higher.

7) LLaMA 3.3 70B performs better than LLaMA 3.1 8B, which was the expected result due to scaling of the model. Nevertheless, the model Qwen3 32B underperforms LLaMA 3.1 8B despite the fifold increase in parameters. Therefore, it may be concluded that the number of parameters is not always indicative of the benchmark results; other factors have to be taken into consideration.

7. CONCLUSIONS

This paper presented a comparative study on the impact of various prompting techniques on hallucination in LLMs. Through the survey and evaluation of multiple datasets, the authors examined how different techniques influence the factual accuracy, hallucination rate and the average response time the LLMs took to answer. The results showed that prompting techniques play an important role in model reliability. Among all the techniques, Few-Shot prompting achieved the best accuracy while also maintaining a balance between hallucination rate, indicating the importance of contextual grounding in model responses. Meanwhile, the more complex and reasoning-based techniques like Chain-of-Thought, Tree-of-Thought did not improve performance and in some cases also introduced higher hallucination rates while also increasing the response time of the model. These findings counter the conventional belief that more reasoning will automatically lead to higher factual reliability. Rather, the report suggests that more straightforward and well-organized prompts can perform better than complex techniques, particularly with small to medium sized large language models. In addition, the discussion reveals a crucial trade-off between reliability, hallucination avoidance, and speed, where methods fostering cautiousness, including instruction-oriented prompts, can diminish hallucinations but raise the probability of abstentions, while methods oriented toward reasoning could enhance interpretability but not necessarily correctness. Overall, the survey signifies that hallucination is not limited to model architecture but is also intermittently dependent on prompt design and interaction strategy.

8. REFERENCES

[1] Anh-Hoang, D., Tran, V., and Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. Frontiers in Artificial Intelligence.

[2] Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models.

[3] K. Cobbe et al., “Training verifiers to solve math word problems,

[4] Lin,S, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in Proc. ACL, 2022.

[5] D. Hendrycks et al., “Measuring massive multitask language understanding,” in Proc. ICLR.

[6] Meta AI. (2024). LLaMA 3.1 8B Instruct. Meta AI. https://huggingface.co/meta-llama/Llama-3.1-8BInstruct

[7] Meta AI. (2024). LLaMA 3.3 70B Versatile. Meta AI. https://huggingface.co/meta-llama/Llama-3.3-70BInstruct

[8] Meta AI. (2025). LLaMA 4 Scout 17B Instruct. Meta AI. https://huggingface.co/meta-llama/Llama-4-Scout-17B16E-Instruct

[9] Qwen Team, Alibaba Cloud. (2025). Qwen3 32B. Alibaba Cloud. https://huggingface.co/Qwen/Qwen332B

[10] Wu, J., Liu, J., Zeng, Z., Zhan, T., Cai, T., & Huang, W. (2026). Mitigating LLM Hallucination via Behaviorally Calibrated Reinforcement LearningK. Cobbe et al., “Training verifiers to solve math word problems.