IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, May 2026
Impact Factor: 8.2

Evaluating the Impact of Chain-of-Thought Prompting on Factual Accuracy in Small and Medium Language Models

Mr. Sarvesh Mahajan[1], Mrs. Rizwana Momin[2], Mrs. Soni Pandey[3], Mrs. Gitanjali Thakur[4], Ms. Shivani Sonkar[5]
Assistant Professor, Computer Science, RTCCS Kharghar, India

Abstract: *Large Language Models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation. However, they often suffer from hallucination, generating factually incorrect yet plausible responses. This study investigates whether Chain-of-Thought Prompting improves factual accuracy compared to standard prompting. Experiments were conducted using lightweight models such as Phi and Mistral on a curated dataset of factual and reasoning-based questions. Results indicate that while Chain-of-Thought prompting improves reasoning transparency and slightly enhances accuracy, it does not consistently reduce hallucinations. In some cases, it produces more convincing but incorrect explanations. This highlights the need for verification mechanisms alongside prompting techniques*

Keywords: Chain-of-Thought Prompting, Large Language Models, Hallucination, Prompt Engineering, Factual Accuracy, Explainable AI, Natural Language Processing

I. INTRODUCTION

Large Language Models have become central to modern AI applications. Despite their fluency, they frequently generate incorrect information, a phenomenon known as Hallucination in AI. This poses challenges in high-stakes domains such as education and healthcare. Recent research proposes Chain-of-Thought (CoT) prompting as a way to improve reasoning by encouraging models to generate intermediate steps. While CoT has shown success in mathematical and logical reasoning tasks, its impact on factual correctness remains uncertain. This paper aims to systematically evaluate whether CoT prompting genuinely improves factual accuracy or merely enhances the plausibility of incorrect responses.

II. LITERATURE REVIEW

The rapid advancement of Large Language Models has significantly enhanced the capabilities of natural language processing systems. Early works such as GPT-3 demonstrated that increasing model scale leads to improved performance across diverse tasks, including text generation, question answering, and reasoning [1]. However, despite these advancements, LLMs continue to exhibit a critical limitation known as Hallucination in AI, where models generate fluent yet factually incorrect or misleading information [2].

To address the limitations in reasoning, Chain-of-Thought Prompting has been proposed as an effective prompting strategy. Introduced by Jason Wei et al. [3], this technique encourages models to generate intermediate reasoning steps before producing a final answer. Their findings indicate that CoT prompting significantly improves performance on arithmetic and logical reasoning tasks, particularly in large-scale models.

Building upon this, the concept of self-consistency was introduced to further enhance reasoning reliability [4]. This approach involves generating multiple reasoning paths and selecting the most consistent answer, thereby improving

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 12, May 2026 Impact Factor: 8.2

robustness in complex problem-solving scenarios. While these techniques improve reasoning performance, their effectiveness in ensuring factual correctness remains limited.

Research in Explainable AI suggests that the reasoning steps generated by LLMs are not always faithful representations of the model's internal decision-making processes [5]. In many cases, models produce plausible but incorrect explanations, raising concerns about the reliability of generated outputs. This phenomenon highlights the distinction between interpretability and correctness.

Additionally, Prompt Engineering has emerged as a critical area for improving model performance. Studies show that carefully designed prompts can significantly influence output quality [6]. However, prompt engineering alone is insufficient to eliminate hallucinations, especially in knowledge-intensive tasks requiring factual accuracy.

Recent approaches have explored retrieval-augmented generation techniques, which integrate external knowledge sources to improve factual grounding [7]. While these methods show promise in reducing hallucinations, they introduce additional computational complexity and dependency on external data sources.

Overall, existing literature indicates that while Chain-of-Thought prompting enhances reasoning capabilities and interpretability, its impact on factual accuracy and hallucination reduction remains inconsistent. This gap motivates the present study, which aims to systematically evaluate the effectiveness of CoT prompting in improving factual correctness across different language models.

III. RESEARCH METHODOLOGY

3.1 Research Design

This study adopts an **experimental research design** to evaluate the impact of Chain-of-Thought Prompting on factual accuracy in Large Language Models. The experiment involves comparing model performance under two prompting strategies: standard prompting and Chain-of-Thought prompting. The goal is to measure differences in accuracy, reasoning quality, and hallucination behaviour.

3.2 Models Used

The experiments were conducted using lightweight and accessible language models to ensure feasibility in constrained computational environments:

- Phi (primary model)
- Mistral (optional comparative model)

These models were selected due to their ability to run on local systems and cloud-based platforms while still demonstrating reasonable reasoning capabilities.

3.3 Dataset Preparation

A curated dataset of factual and reasoning-based questions was created for evaluation. The dataset includes:

- General knowledge questions (e.g., capitals, discoveries)
- Conceptual questions requiring basic reasoning

Each question is paired with a **ground truth answer** for evaluation purposes. The dataset size ranges from 20 to 50 questions to maintain a balance between experimental feasibility and analytical validity.

3.4 Prompt Design

Two types of prompts were designed for each question:

Standard Prompt

A direct question is provided to the model without additional guidance.

Example:

"What is the capital of Japan?"

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 12, May 2026 Impact Factor: 8.2

b) Chain-of-Thought Prompt

The model is instructed to generate step-by-step reasoning before providing the final answer.
Example:
"Explain step by step and then answer: What is the capital of Japan?"

3.5 Experimental Procedure

The experiment follows these steps:
Input each question into the model using both prompting strategies
Generate responses for standard and CoT prompts
Store outputs for analysis
Repeat the process for all questions in the dataset
Compare outputs across prompting techniques
All experiments were conducted using Python and the Transformers framework.

3.6 Evaluation Metrics

Accuracy

Measures whether the model's output matches the ground truth answer:
Accuracy=Number of Correct Answers / Total Questions ×100

b) Hallucination Rate

Measures the proportion of incorrect or fabricated responses:
Hallucination Rate=Incorrect Answers /Total Questions × 100

c) Reasoning Quality (Qualitative)

Evaluates:

- Clarity of explanation
- Logical consistency
- Alignment between reasoning and final answer

3.7 Tools and Environment

The experimental setup includes:

- Programming Language: Python
- Libraries: Transformers, PyTorch, Pandas
- Development Environment: PyCharm / Google Colab

3.8 Data Analysis

The collected outputs are analyzed using both quantitative and qualitative methods:

- Comparison of accuracy between prompting techniques
- Identification of hallucination patterns
- Analysis of cases where CoT improves or degrades performance

The results are summarized using tables and statistical measures to provide clear insights into model behavior.

IV. RESULTS AND DISCUSSION

4.1 Experimental Results

The experiment evaluated the performance of the language model under two prompting strategies: standard prompting and Chain-of-Thought Prompting. The results were measured using accuracy as the primary evaluation metric.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 12, May 2026 Impact Factor: 8.2

Table 1 presents the comparative results obtained from the experiment.

Table 1: Performance Comparison of Prompting Techniques

Prompt Type	Accuracy (%)
Standard Prompt	60%
CoT Prompt	80%

The results indicate that Chain-of-Thought prompting improves accuracy by approximately **20%** compared to standard prompting.

4.2 Analysis of Model Performance

a) Improvement in Accuracy

The increase in accuracy suggests that CoT prompting enables the model to process queries in a more structured manner. By generating intermediate reasoning steps, the model reduces the likelihood of making random or uninformed predictions.

b) Reduction in Simple Errors

It was observed that standard prompting often led to:

- Incomplete answers
- Minor factual mistakes
- Guess-based responses
- In contrast, CoT prompting encouraged the model to:
- Break down the problem
- Arrive at more logically consistent answers

c) Reasoning Transparency

One of the key advantages of CoT prompting is improved interpretability. The model provides step-by-step explanations, making it easier to understand how the final answer is derived. This aligns with principles of Explainable AI.

d) Persistence of Hallucination

Despite improvements in accuracy, hallucinations were not completely eliminated. The model occasionally generated:

- Incorrect reasoning chains
- Plausible but factually wrong explanations

This confirms that Hallucination in AI remains a significant challenge even with CoT prompting.

e) Over-Explanation Issue

Another observed limitation is that CoT prompting sometimes produces:

- Excessively long responses
- Irrelevant intermediate steps

While these do not always affect correctness, they can reduce efficiency and clarity.

4.3 Comparative Insights

The experimental findings highlight the following key insights:

- CoT prompting improves accuracy but does not guarantee correctness
- Reasoning steps enhance interpretability but may not always be faithful
- Small models like Phi benefit from structured prompting

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, May 2026
Impact Factor: 8.2

- The effectiveness of CoT depends on the nature of the task

4.4 Discussion

The results suggest that Chain-of-Thought prompting is effective in improving reasoning-based performance but has limited impact on eliminating hallucinations. This indicates that reasoning and factual accuracy are related but distinct aspects of language model behavior.

While CoT improves the model's ability to generate structured responses, it does not inherently ensure access to correct knowledge. Therefore, relying solely on prompting techniques is insufficient for achieving high factual reliability.

These findings highlight the need for integrating additional approaches such as:

- External knowledge retrieval
- Verification mechanisms
- Hybrid prompting strategies

4.5 Summary of Findings

- CoT prompting improves accuracy by ~20%
- Enhances reasoning clarity and interpretability
- Does not fully eliminate hallucinations
- May introduce longer and sometimes redundant explanations

V. CONCLUSION

This study investigated the effectiveness of Chain-of-Thought Prompting in improving factual accuracy and reducing hallucinations in Large Language Models. Through a comparative analysis of standard prompting and Chain-of-Thought prompting, the results demonstrate that CoT prompting enhances the model's reasoning ability and leads to a noticeable improvement in overall accuracy

The experimental findings indicate that CoT prompting enables the model to generate more structured and interpretable responses by explicitly outlining intermediate reasoning steps. This contributes to improved performance, particularly in tasks requiring logical inference and multi-step reasoning. However, despite these improvements, the issue of Hallucination in AI persists. The model continues to produce plausible but incorrect explanations in certain cases, highlighting that enhanced reasoning does not necessarily guarantee factual correctness.

Furthermore, the study reveals that while CoT improves interpretability, it may also introduce longer and sometimes redundant responses, which can affect efficiency. These findings suggest that reasoning-based prompting techniques alone are insufficient to fully address the limitations of current language models.

In conclusion, Chain-of-Thought prompting serves as a valuable technique for improving reasoning transparency and moderately enhancing accuracy. However, to achieve reliable and factually accurate outputs, it must be complemented with additional mechanisms such as knowledge grounding, verification strategies, or retrieval-based approaches. Future research should focus on integrating these methods to develop more robust and trustworthy language models.

VI. ACKNOWLEDGMENT

The author sincerely thanks Ramsheth Thakur College of Commerce and Science, Kharghar, Navi Mumbai, for the academic encouragement and institutional support that made this research possible. The author also gratefully acknowledges the publicly available regulatory documentation from India's Central Consumer Protection Authority (CCPA), and the research contributions of Harry Brignull, whose foundational work on dark patterns continues to inform consumer protection globally.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 12, May 2026
Impact Factor: 8.2

REFERENCES

1. T. Brown et al., "Language Models are Few-Shot Learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.
2. S. Ji, T. Zhang, and T. Wang, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.
3. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903, 2022.
4. X. Wang et al., "Self-Consistency Improves Chain-of-Thought Reasoning in Language Models," arXiv:2203.11171, 2022.
5. Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
6. S. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," arXiv:2102.07350, 2021.
7. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.
8. OpenAI, "GPT-3: Language Models are Few-Shot Learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
9. Hugging Face, "Transformers Library Documentation," 2024. [Online]. Available: https://huggingface.co/docs/transformers