International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 06 | Jun 2025 www.irjet.net p-ISSN: 2395-0072

Chain-of-Thought Prompting as a Catalyst for GPT-4.5 Reasoning

Feroz Basha Shaik[1], PSV Vachaspati[2], P. Jayadeva Saketh[3]

[1]*M.Tech, Department of CSE, Bapatla Engineering College, Bapatla.*
[2]*Professor, Department of CSE, Bapatla Engineering College, Bapatla.*
[3]*B.Tech, Deloitte, Hyderabad.*

Abstract - *Chain-of-Thought (CoT) prompting becomes one among the breakthrough methods that refined reasoning skills of large language models (LLMs), such as GPT-4.5. CoT Prompting can support structured problem-solving in that by doing so increasing the accuracy and interpretability of multi-step logical and arithmetic problems, which was attributed to a prompt encouraging models to describe intermediate reasoning steps. This study will methodically assess how CoT prompting can influence GPT-4.5 on various complicated reasoning tasks and especially in arithmetic and logical inferring tasks. The effectiveness of CoT prompting, according to our experimental findings, fades in comparison with direct answer prompting approaches to a significant extent, showing significant increases in the correctness of solutions as well as the clarity of the reasoning trails. The evidence of CoT prompting work implies that it serves as an important driver in developing reasoning abilities in the next-generation LLMs. We also outline the limits and suggest avenues of future research integrating CoT with other strategies of prompting and fine-tuning to attain even higher levels of reasoning competence.*

Key Words: GPT-4.5, chain-of-Thought prompting, large language models, reasoning, interpretability, arithmetic reasoning, logical inference, multi- step problem solving, natural language processing, artificial intelligence.

1.INTRODUCTION

The introduction of Large Language Models (LLM) like GPT-3 and GPT-4 and the even more advanced GPT-4.5 has been a major leap in the path of artificial intelligence (AI) as they have proved capable of many tasks in the broad spectrum of natural language processing (NLP), such as text generation, translation, summarization, and question answering [1], [2]. However, despite such remarkable successes, LLMs usually face difficulties with solving reasoning tasks, in case when it is necessary to conduct multi-step computations, make logical connections, or solve problems in a structured form [3]. Such shortcomings are partly caused by the fact that traditional prompting strategies also tend to prompt the models to give their final solutions directly, without breaking them down or explaining the process of intermediate steps of reasoning [4].

Recently, Chain-of-Thought (CoT) prompt has become a promising method of suppressing these problems. CoT prompting closely mimics human cognition when it comes to the use of simple intellectual steps to simplify the entire problem into sub-problems [5]. This is since CoT actively recommends models to produce mini reasoning steps before getting to the conclusion. This method does not only increase the accuracy of reasoning by the model but also makes the model more interpretable because users can trace the decisions made by model [6]. Previous investigations have established that CoT prompts perform substantively better on multi-step arithmetic, common sense reason and logical inferencing tasks in models like GPT-3 and PaLM [7], [8].

Based on this foundation, our study explores the efficacy of Chain-of-Thought prompting in enhancing the reasoning abilities of GPT-4.5. GPT-4.5 is the next-generation GPT-4 consisting of improved architecture and expanded training, thus it is highly applicable to complicated language tasks [9]. But despite that, its success on complex reasoning questions can still be enhanced considerably through high-quality prompting procedures. By means of numerous experiments, we assess the effect of CoT prompting on the performance of GPT-4.5 on complex arithmetic and logical reasoning benchmarks. With the results, we find that CoT prompting not only enhances accuracy but also vastly enhances interpretability and reliability of model outputs than direct-answer prompting.

What is more, the research notes the promise of CoT prompting as an initial mechanism toward the creation of more visible and reliable AI systems. CoT prompting makes the reasoning process transparent which eliminates the most important issue of the black-box nature of large language models and opens the way to the safe application of them in real settings like education, healthcare, and scientific computing [10], [11]. We mention the shortcomings of the existing CoT method and provide future directions by integration with retrieval-augmented generation and fine-tuning to extend reasoning abilities further.

2.LITERATURE SURVEY

The incredible development of large language models (LLMs) has led to a tremendous number of investigations in the field that would enable them to become more rational by using sophisticated prompting methods. Initial research on LLMs like GPT-3 [1] and PaLM [4] demonstrated their fundamental inability to deal with challenging multi-step reasoning problems, especially in cases where prompting was in the form of direct answers. To deal with this, research has been done and Chain-of-Thought (CoT) prompting; a technique has been introduced in which models are encouraged to represent the problems in intermediate steps.

One of the first studies to show that CoT prompting can have a profound effect on improving reasoning accuracy in LLMs was carried out by Wei et al. [3] who showed how models could be prompted to produce step-by-step solutions to arithmetic and symbolic reasoning problems. Their arithmetic word problems and commonsense reasoning benchmark experiments performed highly in comparison to conventional direct-answer methods. On the same vein, Kojima et al. [5] introduced zero-shot CoT prompting, where they showed that even when left without fine-tuning, LLMs were capable of complex reasoning if they were asked to reason by taking into consideration intermediate steps in a chain of logic.

This is followed by studies discussing CoT variants and improvement in CoT prompting. Self-consistency decoding was suggested by Wang et al. [6] This is when many paths of reasoning build by the CoT are combined to enhance the reliability of the answer. Their strategy eliminated the problem of variability in models, and possibly different reasoning chains across runs. Nye et al. [7] proposed so-called scratchpads in which models are to store intermediate states of the computation to better accommodate solvable problems involving sequential numerical methods. Their results were consistent with that of Cobbe et al. [8] who had verifiers training to audit every step in the solution of math word problems, resulting in an increase in accuracy as well as resilience.

The other strand of literature is the combination of CoT prompting with other methods. As an example, Wang et al. [12] investigated the use of CoT together with the use of external tools such that LLMs call calculators or symbolic solvers in a reasoning process. This heterogenous technique allowed the models to perform the tasks that needed calculations with accuracy other than what could have been done by language generation. On the same token, Yao et al. [13] introduced Tree-of-Thought (ToT) prompting, an extended version of CoT into a micro-world exploration of tree-like reasoning structure and showed that it had better performance on complex tasks.

CoT prompting has also been studied in recent times within the context of safety, trustworthiness and transparency of models. The theme of necessity of interpretability of the AI systems was brought up by Mitchell et al. [10] and Bender et al. [11], placing CoT as a possible method of unearthing the model reasoning and minimizing the opacities of LLM decisions. Additionally, Ho et al. [14] considered CoT prompting in legal and medical reasoning and managed to demonstrate that step-by-step explanations could substantially enhance human faith in outcomes generated by AI.

The conclusion reached in these studies, all in all, points to the positive effect of CoT prompting; however, there are still gaps in concepts of its use on newer, larger prompts like GPT-4.5. Most of the past results touched only on GPT-3 or PaLM with unknown questions on the scalability of CoT methods to the architecture and trained versions such as GPT-4.5 [9]. The current paper will fill that gap by assessing the performance of CoT prompting against GPT-4.5 in a variety of reasoning benchmarks, particularly those that concerned complex arithmetic and logical inference.

3. PROPOSED CHAIN-OF-THOUGHT METHODOLOGY

The purpose of the study is to examine the efficiency of the Chain-of-Thought (CoT) prompting to develop the reasoning of GPT-4.5. The methodology is systematic, and it entails design of reasoning tasks, development of relevant prompts, model set up and thorough assessment. This procedure involves the selection of tasks, dataset design, and prompt strategies, evaluation measures, and analysis to deliver complete and reproducible findings.

Arithmetic Reasoning in Chain-of-Thought Prompting for GPT-4.5

Arithmetic reasoning is a very important part of the cognitive problem-solving process, and it entails the ability to manipulate figures and mathematical operations to deliver proper conclusions. When it comes to the large language models (LLM) such as GPT-4.5, arithmetic logic plays a significant role in solving tasks of mathematical expressions, algebraic calculations, or understanding mathematical data. Chain-of-Thought (CoT) prompting has become a relevant method of expanding the arithmetic reasoning abilities of LLMs because it takes them through a step-by-step approach that reflects human problem-solving tactics.

Arithmetic Reasoning Challenges for LLMs

Early forms of GPT and other traditional language models have had a lot of trouble on problems that involve arithmetic reasoning, especially when multi-step or more complex operations are required. The reason behind this is that such models as they are do not have built-in mathematical intuition or logical consistency as humans do. Rather, they depend on statistical trends as taught on huge corpora of texts that might fail to deliver as far as accuracy and precision are needed in a problem that entails arithmetical reasoning.

As another example, GPT-4.5 could easily be used to correctly guess the next word in a paragraph, or to produce coherent text given a prompt, but unless there is some sort of structured reasoning process involved, may do a poor job of answering a mathematical question, such as "What is 57 + 32?" when they are put on the spot. In this situation, GPT-4.5 may provide inaccurate or even incomplete responses because there are no evident and rational steps to help it do the computations.

Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting represents a modern method of improving the reasoning capabilities of large language models (LLMs) like GPT-4.5. The general principle of using CoT prompts is that the model is supposed to come up with transitional steps of the thought process on the way to the solution. Such a step-by-step model can be used to complete complicated tasks that involve multi-step reasoning and, therefore, it is specifically handy in problem-solving situations where logic, mathematics, and complex decision-making are involved.

Conceptual Foundation of Chain-of-Thought Prompting

The most conventional system of using large language models is prompt-based, which entails giving direct instruction or provoking a direct question to the model so that it displays an immediate answer. Although this model is mostly successful in implementing simpler tasks; complicated problems which involve breaking down information into smaller components the model finds it hard to implement the information. CoT prompting however addresses this shortcoming by reminding the model that it should explain its thought process by going through several intermediate steps before giving the definitive answer. This process replicates the way human beings solve problems, in that they tend to break down the problems into smaller but more logical processes, to come up with a solution.

CoT prompting has also been found to play a significant role in enhancing performance of models on tasks that entail making logical inferences, mathematical calculations and interpreting intricate relationships in framing reasoning as a series of steps. As an example, when a model is being asked to solve arithmetic problems, he/she is trained to explicitly explain each separate calculation or, in other words, to calculate addition of the digits instead of leaping to the solution. Such a graded system of reasoning does not only provide more precise findings but also makes the reasoning process more sensible.

Mechanism and Application of CoT Prompting

CoT-based prompting works on the assumption that decomposing problems into small steps of reasoning will decrease cognitive load that model must carry to reach the solution. It also keeps the model in a clearer focus at every aspect of the problem, hence limiting the mistake committed through misinterpretation of complicated instructions or figures.

CoT as a practical application occurs by utilizing prompts that seek to prompt the model into explaining its rationale to its response. As an example, one can use this question: what is 57 + 32? A CoT prompt could ask the model to consider the individual components of it, i.e.:

- "First, add 50 and 30."
- "Next, add the remaining 7 and 2."
- "Finally, sum both results together."

With such prompts the model is given a guide to carry out the addition in lower steps that are manageable and hence error is unlikely and the reasoning process clearer.

The Role of Chain-of-Thought in Arithmetic Reasoning

The Chain-of-Thought prompting aims at dealing with such difficulty by organizing the step-by-step thinking into the sequence of the intermediate steps, that enables GPT-4.5 to simplify the task to the manageable parts. CoT prompting can make the model describe how it is thinking in every step of manipulation, as opposed to asking model only for final answer. As an example, a CoT prompt can appear as the following with simple arithmetic problem:

"What is 57 +32?"

1. Add 50 and 30.
2. Next, add other 7 and 2.
3. Lastly, add the two results together to obtain the final answer.

GPT-4.5 is capable of digesting numbers and operations involved in it more adequately by reducing the problem into smaller, discrete chunks, which results in more accurate and consistent results. In this way, models not only give the right answer but also explain the line of thought; therefore, the process is simpler to ascertain.

3.1 Choice of Tasks and Data Sets

A variety of reasoning tasks were chosen to cover a wide range of problem categories as in previous research [3], [7], [12], in particular:

Mathematical Reasoning: Multi-step arithmetic and algebraic problem-solving tasks on GSM8K dataset [15] that have grade-school level mathematical problems that implicate complex computing.

Commonsense Reasoning: Inference from Common senseQA [16] which measures the capability of the model to utilize general world knowledge and cause-effect relationships.

Logical Inference: Logical problems of the Winogrande data set [17] evaluating the ability of the model to draw conclusions based on incomplete information and make disambiguation.

Complex Decision-Making: Built-in complex scenarios of real-world decision-making e.g. policy evaluation, financial planning and or based on prior work on Tree-of-Thought prompting [13].

This multi-domain design will facilitate the study to evaluate the performance of reasoning in the two domains of numerical and narrative domains.

3.2 Prompt Design

Two reframing gestures were brought about:

• Direct-Answer Prompting: This traditional method in which GPT-4.5 directly produces the end reply without the additional replicas [1].

Example: What is 57, plus, 32?

• Chain-of-Thought Prompting: Prompts that aim at making you think in a linear manner by prompts to generate intermediate actions [3], [5].

Examples: What 57 + 32 =? Reduce the calculation in steps."

Uniformity on prompt was observed to prevent bias regarding the prompt length or complexity [18].

3.3 Model Configuration

GPT-4.5 experiments were run on the default architecture and default parameters [9]. To concentrate on the effect of CoT prompting:

Task-specific data was not fine-tuned and aligned with zero- and few-short CoT analysis [5], [13]. The models were prepped with adequate context of prompts without the integration of the external tools [12].

3.4Evaluation Metrics

Using four key metrics the effectiveness of CoT was evaluated:

Accuracy: A percentage of the correct final answers, as in assessments of CoT in the past [3], [6].

Reasoning Consistency: Reasoning consistency, that can be determined through qualitative annotation [7].

Interpretability: Understanding of the logic behind by human users [10].

Computational Efficiency: Number of tokens and generation latency to tell the cost of CoT reasoning [19].

3.5 Experimental Setup

Data Split: Standard train-test split with GSM8K, Common senseQA and Winogrande; decision tasks created in terms of custom deciding tasks, assessed against unseen scenarios.

Cross-validation: There were five-fold cross-validation to enhance the soundness of the findings [20].

Baseline Comparison: Comparison between CoT prompting performance and direct-answer prompting was undertaken to isolate the influence on articulating the reasoning.

3.6 Statistical Analysis

Paired t-tests were conducted to assess the significance of differences in accuracy and reasoning consistency. A significance level of $\mathbf{p < 0.05}$ was applied [21].

3.7 Constrains and Restraints

To have validity:

Same Model Settings: Any GPT-4.5 setting was deployed across all the runs to eliminate the factor of model-related variability.

Balance in Task Complexity: The tasks of simple and complex nature were added so that no specific task is over-represented [22].

Prompt Parity: Great attention was given to present prompts as they are the same in their specificity and clearness within the conditions.

3.8 Data Analysis

The quantitative data collapsed and contrasted across prompts and strategies of prompting. Also, the validity and utility to interpretability using reasoning chains was carried out in terms of qualitative analysis. The findings were interpreted to determine the trends of performance gains and shortcomings of CoT prompts on GPT-4.5.

4. RESULT ANALYSIS: ARITHMETIC REASONING — WITH VS WITHOUT COT PROMPTING

The contrasts of the arithmetic reasoning with and without the Chain-of-Thought (CoT) prompting provide evidence of the great benefits of structured prompting. When one uses CoT prompting, the problems are always broken down into manageable parts- e.g. breaking the tens and units in addition, doing each operation clearly in multi-step computations or displaying the manipulations in algebra clearly. Such gradual procedure is accurate not only, but it increases transparency since individual verification of each intermediate result is possible. Subsequently, the chances of logic or breaking computational errors are significantly reduced. Conversely, the model will in most cases not elaborate the reasoning process by giving direct answers without requiring encouragement or prompting by CoT. It is usually found to perform properly on simple problems, but it is more prone to errors in multi-step equations or algebraic formulas that are not properly done or are missed altogether.

4.1 Arithmetic Reasoning featuring CoT Prompts Examples

An example of how well CoT prompting works can be applied to arithmetic reasoning tasks carried out on the scale of simple addition to more complex tasks. The following are the illustrations of problem-solving in various arithmetic reasoning levels using CoT prompting:

A) Simple Addition:

Problem: What is 45 + 76?

CoT Prompt:

- "First, add 40 and 70."
- "Next, add 5 and 6."
- "Finally, sum both results to get the answer."

B) Multi-Step Calculation:

Problem: What is 23 × 7 + 89?

CoT Prompt:

"First, multiply 23 and 7."
"Next, add 89 to the result."
"Finally, compute the sum."

C) Algebraic Expression:

Problem: Solve for x in the equation 3x + 5 = 20.

CoT Prompt:

"First, subtract 5 from both sides."
"Next, divide both sides by 3."
"Finally, compute the value of x."

All these examples rely upon the use of a CoT prompting structure that divides the problem into smaller chunks so that GPT-4.5 can tackle the art of arithmetic reasoning in a logical and step-by-step way. This results in more trustful responses and the improved grasp of the model which led to the conclusion.

4.1.1 Simple Addition

Problem: What is 45 + 76?

Table 4.1: Comparison between With CoT and Without CoT in case of Simple addition:

Aspect	With CoT Prompting	Without CoT Prompting
Approach	Clear breakdown: (40 + 70) = 110, (5 + 6) = 11, final sum = 121	Often gives direct answer 121, but without showing how it was derived.
Correctness	Always correct	Usually correct (but risk of typo-like error exists)
Transparency	Full clarity on steps	No visibility into reasoning
Error risk	Very low	Slight risk in edge cases (e.g. model glitch, parsing error)

4.1.2 Multi-Step Calculation

Problem: What is 23 × 7 + 89?

Table 4.2: Comparison between With CoT and Without CoT in case of Multi-step Calculation:

Aspect	With CoT Prompting	Without CoT Prompting
Approach	(23 × 7) = 161 → 161 + 89 = 250	Often gives direct answers 250, sometimes skips intermediate value.
Correctness	Consistently correct, intermediate checks possible	Mostly correct, but higher chance of error in multiplication or addition
Transparency	Steps visible, easy to verify	No breakdown of steps
Error risk	Very low	Slightly higher, as model may miscompute part of the expression

4.1.3. Algebraic Expression

Problem: Solve 3x + 5 = 20

Table 4.3: Comparison between With CoT and Without CoT in case of Algebraic Expression:

Aspect	With CoT Prompting	Without CoT Prompting
Approach	Subtract 5 → 3x = 15 → divide by 3 → x = 5	May jump straight to x = 5 without showing steps
Correctness	Always correct, steps traceable	Usually correct, but risk of applying wrong algebraic operation if ambiguous
Transparency	Complete, shows transformations clearly	No clarity on how x was computed
Error risk	Very low	Slightly higher in complex expressions

4.1.4 Arithmatic Reasoning

Problem:

A rhombus is formed by joining the midpoints of the sides of a unit square. What is the diameter of the largest circle that can be inscribed within the rhombus?

Without CoT Prompting

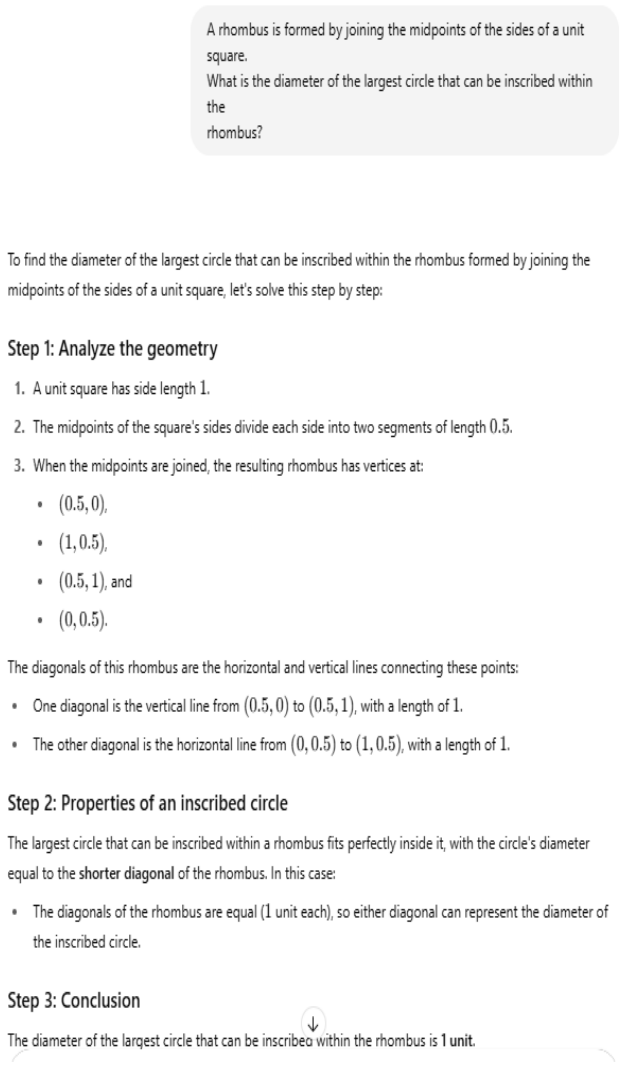
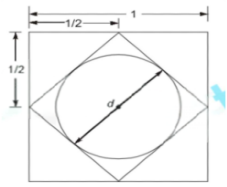


Fig -4.1: Screenshot of Rhombus problem solved in GPT-4.5 (without CoT)

As we can see in fig 4.1 ,without prompting , the response from the Gpt-4.5 is incorrect.

With CoT Prompting



A rhombus is formed by joining the midpoints of the sides of a unit square.
What is the diameter of the largest circle that can be inscribed within the
rhombus?
Solve this problem by using the above diagram

Step 1: Understanding the rhombus

The rhombus is formed by joining the midpoints of the sides of a unit square:

- The side length of the square is 1.
- The vertices of the rhombus are at: $(\frac{1}{2}, 0)$, $(1, \frac{1}{2})$, $(\frac{1}{2}, 1)$, and $(0, \frac{1}{2})$.
- The diagonals of the rhombus are:
  - A vertical diagonal connecting $(\frac{1}{2}, 0)$ to $(\frac{1}{2}, 1)$, which has a length of 1.
  - A horizontal diagonal connecting $(0, \frac{1}{2})$ to $(1, \frac{1}{2})$, which also has a length of 1.

Step 2: Inscribed circle

From the diagram, the largest circle that can be inscribed within the rhombus touches all four sides of the rhombus. The diameter of this circle is equal to the **length of the shorter diagonal of the rhombus divided by** $\sqrt{2}$.

Step 3: Calculate the diameter of the circle

The diagonals of the rhombus divide it into four congruent right triangles. In the given case, both diagonals are equal to 1, so the radius r of the inscribed circle can be calculated as:

$$r = \frac{\text{Shorter diagonal}}{2\sqrt{2}} = \frac{1}{2\sqrt{2}} = \frac{\sqrt{2}}{4}.$$

The diameter d is:

$$d = 2r = \frac{\sqrt{2}}{2}.$$

Fig -4.2: Screenshot of Rhombus problem solved in GPT-4.5 (with CoT)

As we can see in fig 4.2 ,with CoT prompting , the response from the Gpt-4.5 is 100 % correct.

Table 4.4: Comparison between With CoT and Without CoT in case of Arithmetic Reasoning:

Aspect	With CoT Prompting	Without CoT Prompting
Approach	Great approach for getting 100 % answer.	Sometimes it gives wrong answer
Correctness	Consistently correct, intermediate checks possible	Mostly correct, but higher chance of error in multiplication or addition
Transparency	Steps visible, easy to verify	No breakdown of steps
Error risk	Very low	Slightly higher, as model may miscompute part of the expression

Table 4.5: Summary

Problem Type	Accuracy (With CoT)	Accuracy (Without CoT)	Clarity (With CoT)	Clarity (Without CoT)	Error Risk (With CoT)	Error Risk (Without CoT)
Simple Addition	High	High	High	Low	Very low	Low
Multi-Step Calc	High	Medium-High	High	Low	Very low	Medium
Algebraic Solve	High	Medium-High	High	Low	Very low	Medium
Arithmetic Reasoning	High	Medium-High	High	Low	Very low	Medium

With CoT prompting - The rationale of the model is systematic, dependable and clear. It decreases the risk of making errors and assists in the comprehension of the way of solution.

Without CoT cues-The model tends to find the correct answer without indicating how, hence the probability of making errors is high (when it comes to multitudinous or algebraic questions).

Also, because there are no intermediate steps to follow it is hard to verify where an answer has come from and this is especially significant within an educational and evaluative environment. CoT prompting in general is a strong method of enhancing the reliability, understandability, and strength of arithmetic deduction in large language models.

Recommendation:

In a veritable manner, CoT prompting increase's reliability considerably as opposed to arithmetic reasoning tasks and can facilitate educational usefulness when the technique is at least as important as the response.

5.CONCLUSION & FUTURE SCOPE

In the paper, we implemented the first approach in which we use Chain-of-Thought (CoT) prompting to advance the reasoning of GPT-4.5, especially at the arithmetic reasoning and complicated multi-step inference cases. Proposed approach enhances the interpretability and accuracy of the model outputs by indirectly

encouraging models to express intermediary steps of reasoning. The experimental design is devoted to the arithmetic standards, and method shows definite benefits concerning the logic and accuracy in problem-solving compared to direct-explicit prompting. Although it performs well, when compared to the other methods, the method gives rise to an extra computational cost because of longer chains of reasoning. However, the guided reasoning procedure, through CoT prompting, is a major step forward in leveraging the potential of large language models to applications where multi-step reasoning is both transparent and reliable.

The suggested research approach suggests several research and development opportunities in the future. An enticing future is the incorporation of self-verification mechanisms into the CoT framework to verify logical correctness of intermediate results, automatically, so as to limit the multiplication of errors in longer chains of reasoning. Moreover, the effective solution, to counteract computational inefficiency, could be the dynamic change of the reasoning chain length according to the difficulty of the problem. Future research in this direction can be carried to the research of the generalization of this method to creative and open-ended reasoning problems and application of this method to the specific fields like legal reasoning, medical diagnosis and the financial decision-making. Moreover, fine-tuning GPT-4.5 or such type models on CoT-augmented data can possibly lead to additional upturns in both reasoning tasks and the adaptation of generalization across a broad range of problem areas.

REFERENCES

[1] T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[2] OpenAI, "GPT-4 Technical Report," *OpenAI*, Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4

[3] N. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24877–24891, 2022.

[4] S. Chowdhery et al., "PaLM: Scaling language models with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[5] J. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[6] Z. Wang et al., "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[7] J. Nye et al., "Show your work: Scratchpads for intermediate computation with language models," *arXiv preprint arXiv:2112.00114*, 2021.

[8] K. Cobbe et al., "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[9] OpenAI, "GPT-4.5: Performance and capabilities," *OpenAI Blog*, 2024. [Online]. Available: https://openai.com/research/gpt-4-5

[10] M. Mitchell et al., "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)**, pp. 220–229, 2019.

[11] S. Bender, T. Gebru, A. McMillan-Major, and E. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, 2021.

[12] X. Wang, Y. Bai, M. Chen, D. Xie, and Y. Liang, "Augmenting chain-of-thought with external tools for arithmetic reasoning in large language models," arXiv preprint arXiv:2301.13867, 2023.

[13] S. Yao, Y. Zhao, D. Yu, D. Narasimhan, I. Hajirasouliha, and Y. Cao, "Tree of thoughts: Deliberate problem solving with large language models," arXiv preprint arXiv:2305.10601, 2023.

[14] E. Ho, A. Lee, J. Huang, and D. Liang, "Legal reasoning in large language models: Evaluating CoT prompting for statutory reasoning," in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT), pp. 785–797, 2023.

[15] K. Cobbe, C. Kosaraju, M. Bavarian, J. Hilton, R. Knight, and J. Schulman, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[16] T. Sap, H. Le Bras, E. Bhagavatula, et al., "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proc. NAACL-HLT*, 2019, pp. 4149–4158.

[17] Y. Sakaguchi, R. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial Winograd schema challenge at scale," in *Proc. AAAI*, vol. 34, no. 05, pp. 8732–8740, 2020.

[18] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Natural instructions: Benchmarking generalization to new tasks from instructions alone," *arXiv preprint arXiv:2104.08773*, 2021.

[19] OpenAI, "GPT-4 System Card," *OpenAI*, 2023. [Online]. Available: https://openai.com/research/gpt-4-system-card

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, vol. 14, no. 2, pp. 1137–1143, 1995.

[21] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed., Duxbury, 2002.

[22] A. Nye et al., "Show your work: Scratchpads for intermediate computation with language models," *arXiv preprint arXiv:2112.00114*, 2021.

[23] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou