The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

Solving Linguistic Olympiad Problems with Tree-of-Thought Prompting

Zheng-Lin Lin
National Taiwan University
b09208026@ntu.edu.tw

Chiao-Han Yen
National Taiwan University
r11142008@ntu.edu.tw

Jia-Cheng Xu
National Taiwan University
r07142006@ntu.edu.tw

Deborah Watty
National Taiwan University
r11142012@ntu.edu.tw

Shu-Kai Hsieh
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

In this study, we delve into the efficacy of the Tree-of-Thought Prompting technique as a mechanism to address linguistic challenges and augment the reasoning capabilities of expansive language models. Specifically, we scrutinize the reasoning prowess of the Generative Pre-trained Transformer (GPT) model, which has garnered significant attention within the research and practitioner community. Utilizing the Tree-of-Thought Prompting methodology, we assess its utility in enhancing both the precision and response latency of the GPT model, especially for Linguistic Olympiad tasks demanding elevated reasoning competencies. Concurrently, we delineate inherent limitations within this approach and proffer avenues for future research to refine and optimize it. Code repo: https://github.com/chrizeroxtwo/ToT-LinguisticProblem

Keywords: Tree-of-Thought Prompting, Large Language Models, Machine Reasoning, Generative Pre-trained Transformer, Linguistic Olympiad

1 Introduction

Large language models (LLMs) have experienced significant evolution, showcasing their versatile abilities in tackling a wide range of natural language processing (NLP) tasks. The Generative Pre-trained Transformer (GPT) model stands out as one of the most extensively discussed and influential language models. By leveraging its foundation on large-scale text data pre-training, Liu et al. (2023) shows that GPT has given rise to numerous innovative applications across various domains.

Among these tasks, its exceptional reasoning ability has emerged as a subject of fascination among researchers and practitioners. The adeptness at proficient reasoning serves as a foundational element for various cognitive processes, shaping the intricate interplay between cognition and human capabilities. As such, understanding the underlying mechanisms of exceptional reasoning holds substantial implications for cognitive psychology and related disciplines. To investigate the capacity for reasoning, a common area of focus is complex problem-solving scenarios or logical reasoning tasks. Such subjects typically require individuals to analyze intricate information, discern patterns, and draw well-structured conclusions from the available evidence. The selected tasks may encompass both deductive reasoning puzzles and inductive reasoning challenges, enabling researchers to assess participants' cognitive abilities in various contexts.

The recently emerged research topic known as the "Rosetta Stone" problem addresses the aforementioned requirements effectively. This distinctive problem type involves the application of limited data to "solve" and establish correspondences between expressions in two distinct language systems (Bozhanov and Derzhanski, 2013).

The Rosetta Stone task combines linguistic problems to create a general task that can be tackled by individuals without specialized linguistic skills. It encompasses a genre of composition that presents linguistic facts and phenomena in an enigmatic form (Derzhanski and Payne, 2010). This eventually integrated into the Linguistic Olympiad (LO), akin to renowned competitions such as the United Kingdom Linguistics Olympiad (UKLO) [1] and the North American Computational Linguistics Open Competition (NACLO). [2]

The Linguistic Olympiad (LO) encompasses various types of problems, focusing on different linguistic aspects such as semantics, syntax, mor-

[1]https://www.uklo.org/
[2]https://nacloweb.org/

Table 1: NACLO(2022) - Seeing the Future

Lyo'awujwa'	English
"a'wen"	"I see you (sg.), I see him/her/them"
"a'weneɨ"	"I see you (pl.)"
"si'wen"	"you (sg.) see me, he/she/they see me"
"hi'wen"	"you (sg.) see him/her/them"
"kasi'wen"	"you (sg.) see us, he/she/they see us"
"in'wen"	"he/she/they see you (sg.)"
"in'weneɨ"	"he/she/they see you (pl.)"
...	...
""	"you (sg.) are going to see him/her/them"
""	"he/she/they are going to see you (sg.)"
""	"you (sg.) are going to see us"
""	"you (pl.) are going to see us"
""	"we are going to see you (pl.)"

phology, and phonology. These problems are presented in diverse question formats during the competition, including translation tasks, match-up exercises, multiple-choice questions, rule-induction challenges, as well as problems involving numbers and kinship terms. The integration of these linguistic problem types and formats aims to provide participants with a comprehensive and engaging platform to demonstrate their analytical skills. The demonstrated problem presented in Table 1.

Initiatives led by organizations such as OpenAI and Puzzling Machine [3] have undertaken investigative efforts within the domain of the Linguistic Olympiad (LO). These endeavors have primarily concentrated on addressing challenges that encompass numerical enigmas and translation exercises. A pivotal aspect of these initiatives has revolved around utilizing expansive language models, involving the conception of algorithmic structures and the creation of task prompts.

Reflecting upon the insights gained from these previous initiatives and acknowledging the continuous progress in the field of prompt engineering, we consider the viability of employing the cognitive framework outlined by Yao et al. (2023), commonly referred to as the "Tree-of-Thoughts (ToT)," to tackle the complexities presented by the Rosetta Stone challenge.

Following previous studies, including Puzzling Machine (Şahin et al., 2020a) and the OpenAI IMO (International Mathematics Olympiad) problem-solving experiment (Polu et al., 2022), we attempt to use ToT on Rosetta Stone questions to examine whether the approach works as well in this domain (for the structure of Tree-of-Thought see Figure 1). In short, we scored the output from GPT-3.5 and ToT Prompting compared to the open competition from Puzzling Machine. Furthermore, we compare the results with and without ToT Prompting. We use the data published on the Puzzling Machine website for testing. The paper is organized as follows: Section 2 summarizes related work and Section 3 discusses the LLM-based applications of prompt engineering. We then elaborate on the details of our experiment in Section 4 and provide discussion in Section 5, and finally, Section 6 concludes the paper.

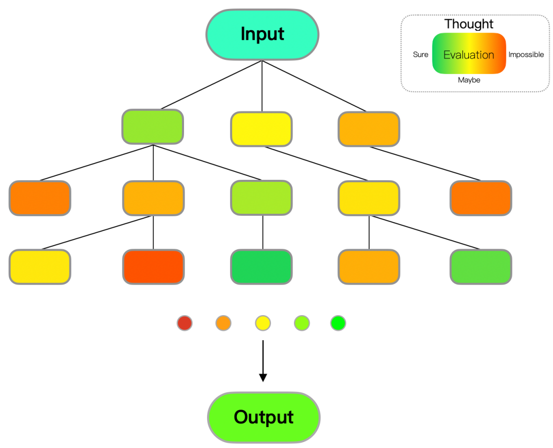

Figure 1: The structure of Tree-of-Thought Prompting.

[3] https://ukplab.github.io/PuzzLing-Machines/

2 Related Work

Previous work on solving the Rosetta Stone task can be referred to the `Puzzling Machine challenge` organized by Şahin et al. (2020a). This task focuses on the translation task type. 63% of the tasks in question ask the participant to translate from English to a foreign target language. The other 37% require translations from another language into English. They created an open competition before OpenAI published the GPT, and have been experimented with various of deep learning models. The best-performing model at the time in 2020 was the Phrase Based Statistical Machine Translation (PBSMT) by Koehn et al. (2007), which significantly surpassed other models employed as baselines such as Transformer (Vaswani et al., 2017) and FastAlign (Dyer et al., 2013). ChatGPT by OpenAI joined the competition in late 2022 with a test conducted by Jannis Vamvas. Remarkably, the performance exceeded that of PBSMT, achieving more than twice its score.

Another early work testing the ability to reason using Olympiad questions was done by OpenAI themselves (Polu et al., 2022). They tested the ability of ChatGPT to solve IMO problems with a mathematical focus, known as "statement curriculum learning". However, according to Liu et al. (2023), while the model is capable of non-trivial mathematical reasoning, its performance is still far below that of the best students in the competition.

In general, extracting information from language models such as GPT requires prompt engineering. One new method of designing a prompt proposed by Yao et al. (2023) is Tree-of-Thoughts (ToT), which was developed based on the Chain-of-Thought (CoT) prompting method (Wei et al., 2022) and can improve the output of an LLM for tasks requiring different types of reasoning including common sense, arithmetic and symbols. ToT uses Self-Consistency (Wei et al., 2022) to sample different reasoning paths and select the output with the highest possibility to increase accuracy. A rating system is used to evaluate candidate thoughts in each step after prompting. If the inference cannot reach the ideal output thought, it will turn to the sibling thoughts or backtrack in the case that no possible sibling thoughts exist. Yao et al. (2023) provide test results for three types of tasks: Game of 24, Creative Writing and Mini Crosswords. In the Game of 24 task, ToT far outperforms preceding methodologies such as CoT. Moreover, ToT has a pronounced capacity for adeptly addressing the cognitive demands of Mini Crosswords. While advancements in the domain of Creative Writing are perceptible, they did not attain commensurate prominence.

The results suggest that ToT might constitute a pivotal juncture in the realm of Prompt Engineering. Similarly to iterative reasoning, it allows different algorithms to enhance the thinking processes of the Language Model at the same time.

3 LLM-based Approaches

In the rapidly evolving landscape of NLP, the introduction of large language models (LLMs) represents a paradigm-shifting moment. These models, characterized by their enormous sizes, sometimes containing billions of parameters, have set unprecedented benchmarks in a myriad of NLP tasks, from translation to text generation. LLMs, such as GPT, leverage vast data to learn linguistic nuances, idiomatic expressions, and even factual knowledge. This enables them to generate human-like text and comprehend complex queries with remarkable accuracy.

The emergence of LLMs in NLP has paved the way for a new, important skillset: prompt engineering. As LLMs, such as GPT variants, are pre-trained on vast amounts of data and then fine-tuned for specific tasks, how questions or prompts are posed to these models becomes crucial in eliciting desired outputs. While LLMs have minimized the need for extensive task-specific architectures, they have introduced the challenge of designing effective prompts to guide the model's responses. Prompt engineering involves crafting, refining, or even chaining prompts to guide the model toward a specific type of answer or behavior. The art and science of prompt engineering are akin to "programming" these models, leveraging their vast knowledge in a controlled and predictable manner. White et al. (2023) has introduced a versatile framework for structuring prompts, providing specific rules and guidelines to engage LLMs effectively.

In essence, while LLMs have significantly reduced the complexities associated with traditional NLP model architectures, they have introduced an intricate dance of interaction, where prompt engineering emerges as a bridge between human intentions and model capabilities. The recent development of ChatGPT and GPT-4 is centered around the refinement of prompt engineering, a crucial aspect in improving interactions with these extensive language models (LLMs). Effective prompt engineering holds a pivotal role in advancing both ChatGPT and GPT-4. In our experiment, prompt engineering also plays a role, and we describe its application and limitations in the following sections.

4 Experiment

The Figure 2 shows the structure of the Tree-of-Thought we implemented.

4.1 Experimental Setup

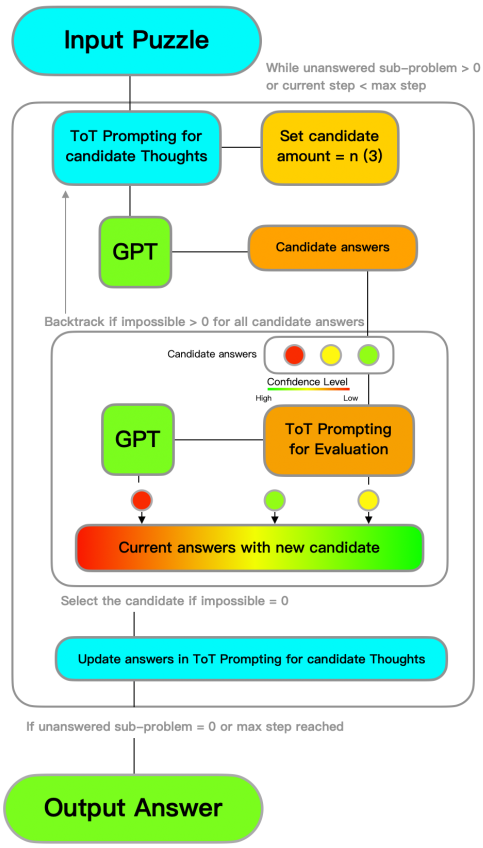

Figure 2: Tree-of-Thought implementation on solving language puzzles.

Benchmark. We employ the Puzzling Machine Benchmarks introduced by Şahin et al. (2020b) for our analysis. This benchmark comprises two main sections: Trial Data, containing 10 problems accompanied by answers, and Competition Data, containing 86 problems without provided solutions. Figure 3 shows an example of such a problem. All of these problems require iterative reasoning to solve. We carried out six rounds of experiments on the Competition Data using our prompting methods. Subsequently, we submitted our predictions to the Puzzling Machine 1.0 Officials for evaluation.

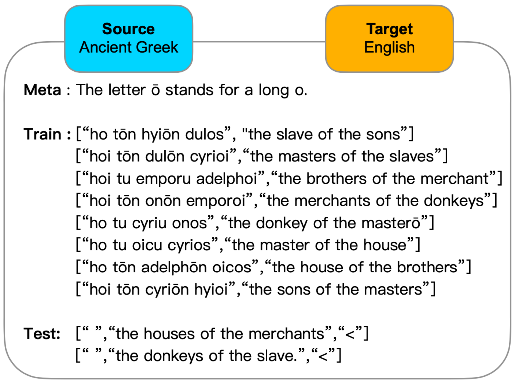

Figure 3: Example of a Puzzling Machine problem introduced by Şahin et al. (2020b). The symbols ’<’ and ’>’ in the Test part indicate the direction of the translation.

Baseline. We utilize Standard Input-Output Prompts accompanied by a few-shot exemplar approach, demonstrating the required output format for the language model (see Figure 4). The intended outcome is for the language model to provide answers addressing all sub-problems of each given linguistic problem at once.

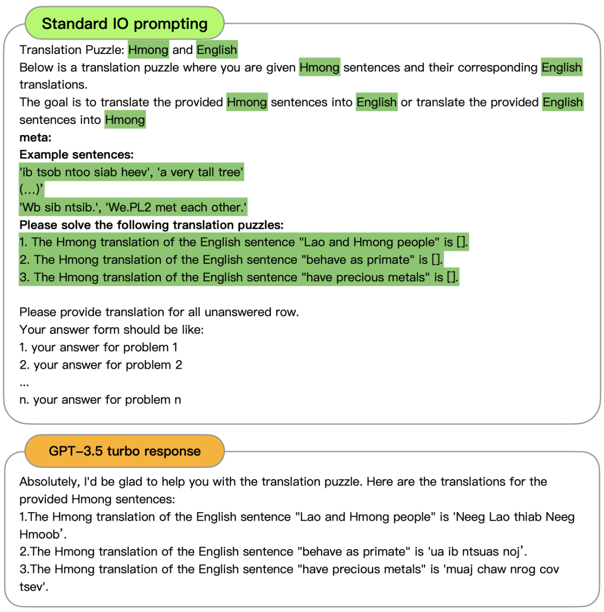

Figure 4: Standard Input-Output Prompting.

Tree-of-Thought Prompting. Considering the framework proposed by Yao et al. (2023) in their work on Tree-of-Thought Prompting, we adopt a systematic approach in this study. Our methodology involves instructing the language model to propose a set of candidate solutions sorted by their respective confidence levels to address one sub-problem at a time within each given linguistic problem (see Figure 5).

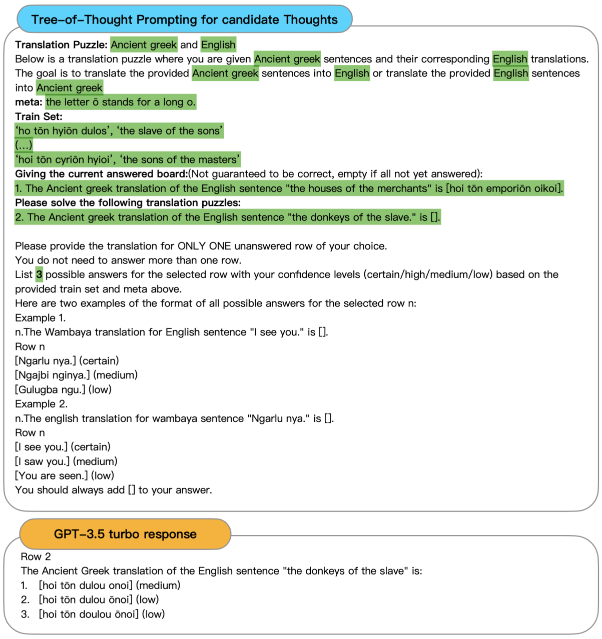

Figure 5: An instance of Tree of Thought Prompting. It proposes candidate solutions for one sub-problem. The highlighted components are adaptable, depending on the problem and its state.

A set is composed of three candidates. Subsequently, we task the language model to evaluate the current state of the chosen candidate solutions based on whether the adoption of a newly chosen candidate would introduce any contradictions among the answered sub-problems. The evaluation prompt example is shown in Figure 6. If the currently chosen candidate leads to a contradiction, an evaluation prompt containing the candidate with the second-highest confidence level would be provided to GPT to continue the evaluation process. In the case that contradictions occur within the whole set of new candidate solutions, a backtrack ensues. Ideally, this methodical progression facilitates a dynamic evaluation of the trajectory towards the correct resolution. Considering the cost of GPT output, a maximum thought-generating step can be established; the output will

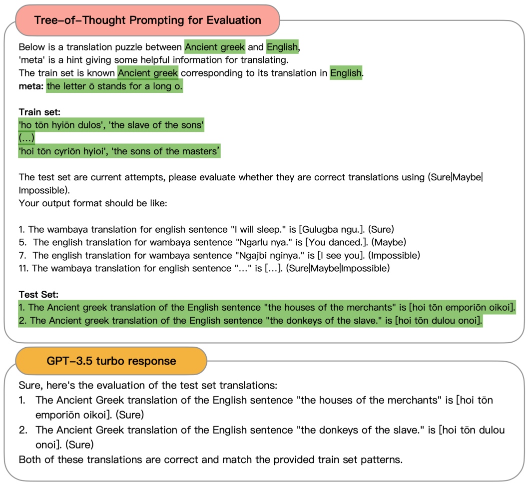

Figure 6: Illustration of an evaluation prompt within Tree-of-Thought Prompting. Following the introduction of a new candidate into the current given state, GPT-3.5 turbo is tasked with determining the presence of any contradictions. The highlighted components are adaptable, depending on the problem and its state.

be the deepest status with the most answers filled once this maximum step is reached. In our experiment, we conducted both unlimited steps and maximum step = 50.

Language Model. We opted to utilize the widely available GPT-3.5 Turbo, in contrast to the GPT-4 employed in original study of (Yao et al., 2023). We carried out ToT Prompting experiments employing two distinct temperature settings (0.5 and 0.7), in comparison to the study conducted by (Yao et al., 2023) in which the temperature of GPT-4 was set to 0.7. This allowed us to explore how variations in GPT output diversity and creativity could lead to better results.

4.2 Results.

As depicted in Figure 7, the combined average results for solving English into Foreign language and Foreign language into English puzzles reveal that within the context of the Puzzling Machine Benchmarks, the baseline method of Standard IO Prompts with both temperature = 0.5 and 0.7 outperforms the Tree-of-Thought Prompting approach with various temperature and step settings. The scores indicate that the baseline approach generates solutions that are slightly more accurate and consistent not only based on word-level metric BLEU-2 (Papineni et al., 2002) and character-level metrics chrF (Popović, 2015) and characTER

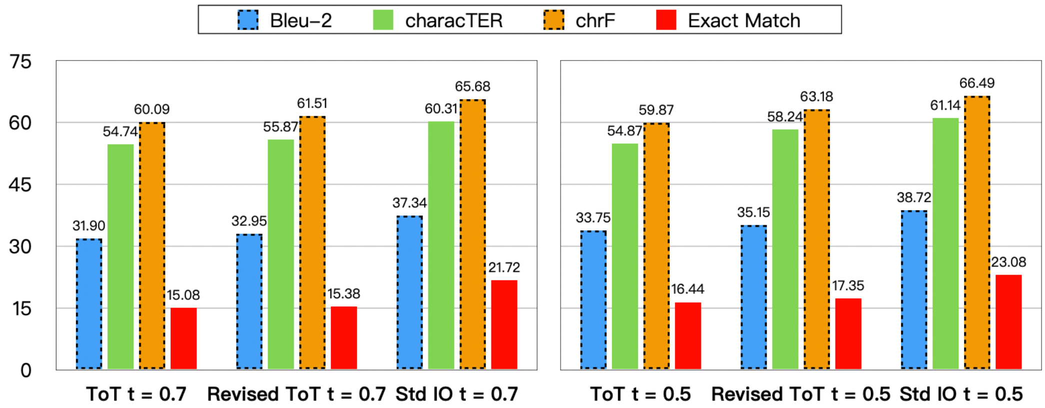

Figure 7: The combined average results for solving English into Foreign language and Foreign language into English puzzles. Revised ToT is Tree-of-Thought without step limit. The baseline Standard Input-Output Prompting (Std IO) with two different temperature settings (t = 0.5 and 0.7) appears to outperform Tree-of-Thought Prompting (ToT) with both t = 0.5 and t = 0.7 on the Competition Data of the Puzzling Machine. While unlimited thought-generating steps do seem to improve the results of the two ToT approaches, they still remain below the baseline.

(Wang et al., 2016), but also exhibit superior performance improvement in terms of Exact Match, where EM is calculated as 1 if the prediction and reference sentences match and 0 otherwise (Şahin et al., 2020b). While the Tree-of-Thought method with a lower temperature (temperature $=$ 0.5) demonstrates better results compared to the higher temperature setting (temperature $=$ 0.7), using unlimited steps produces more precise answers than using limited maximum steps. Nevertheless, even with the best version of the Tree-of-Thought method we conducted (unlimited steps, temperature $=$ 0.5), the performance still falls short of the baseline. This phenomenon can also be observed within detailed result, such as translating English into Foreign and vice versa. (See Figure 8 and 9)

5 Discussion

We have conducted an investigation into the Tree-of-Thought methodology for addressing linguistic challenges utilizing GPT-3.5 turbo. Our analysis of the outcomes reveals that this approach does not outperform the conventional Standard Input-Output Prompting method. To dive deeper into this, we have examined different factors that could lead to this result.

5.1 Prompt

Before embarking on the final six rounds of experiments, we conducted preliminary testing on GPT-3.5 turbo using Tree-of-Thought with various candidate thought-proposing prompts. During these testing rounds, we observed instances

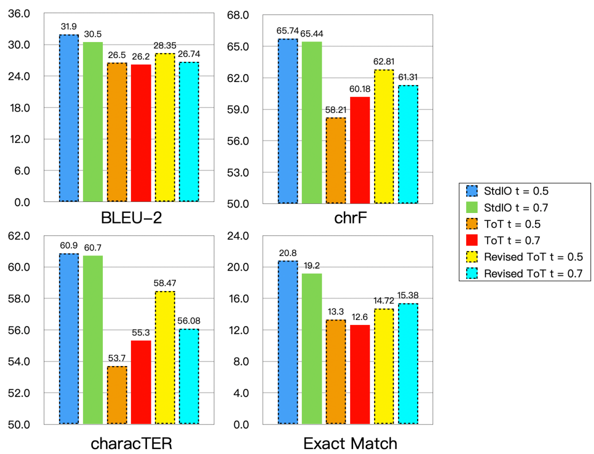

Figure 8: The scores for solving English into foreign language puzzles. It shares a similar trend to the average results. The results from the two Standard Input Output methods still surpass all variations of the Tree-of-Thought approach with different parameters.

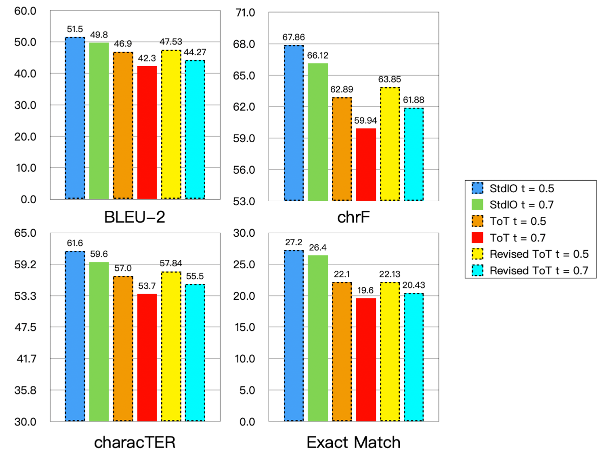

Figure 9: The scores for solving foreign language into English puzzles. The two Standard Input Output methods are still in the lead. It is worth noting that both the Standard Input Output and Tree-of-Thought methods translate foreign language into English more accurately than they do English into the foreign language.

where GPT occasionally exhibited confusion between translation and rephrasing. It turns out we accidentally queried GPT-3.5 turbo with 'Translate the following source language sentences into target language(English).' instead of 'Please solve the following translation puzzles.' we used in our later prompts. Thus make one-third of total 428 instances of where we queried translations from the target language (English) into the source language, rephrasing rather than translating. This underscores the significance of precise and concise prompts, particularly in tasks that involve iterative prompting of GPT. Even though the prompts we used elicit candidates with the correct format, it is possible that the prompts used might not have been precise enough to elicit reasonable candidates from the model.

5.2 Evaluation

Another factor to consider is the evaluation method employed. We used the Standard Input Output Prompting method with few-shot exemplars, as described in Figure 6. However, this prompting method might be overly simplistic, potentially missing the ability to recognize contradictions introduced by new candidates within the answers to sub-problems. Consequently, enhancing the sensitivity of the evaluation becomes a plausible solution to improve the Tree-of-Thought's effectiveness in solving linguistic problems. Approaches like Chain-of-Thought proposed by (Wei et al., 2022) and Multiagent Debate suggested by Du et al. (2023) offer promising avenues to enhance GPT's reasoning capabilities and could lay the foundation for accurate and sensitive evaluation.

5.3 Large Language Models

We also cannot overlook the possibility that GPT-3.5 turbo might not possess the required robustness to discern obscure patterns behind linguistic puzzles, especially when compared to tasks with explicit rules to follow, such as the Game of 24 and Mini crosswords examined by (Yao et al., 2023) in their study. This comparison is further accentuated when we juxtapose GPT-3.5 turbo with newer models like GPT-4, utilized in experiments conducted by (Yao et al., 2023).

5.4 Structure of Tree-of-Thought Prompting

One speculation is that the human thinking process that Tree-of-Thought attempts to emulate might not be well-suited for solving linguistic puzzles. When dealing with linguistic puzzles that involve hidden and intricate patterns, the approach to solving them might not be as straightforward as tackling them one by one through trial and error. It is possible that a deeper analysis of the Known Set or Train Set to uncover hidden patterns and rules is crucial and should be given priority. There might be an alternative prompting method that could be more effective in addressing linguistic problems.

6 Conclusion

This paper has elucidated the novel application of the Tree-of-Thought Prompting method aimed at deciphering linguistic challenges and augmenting the reasoning prowess of language models. Beyond just theoretical implications, the practical manifestations of this method are manifold. It not only elevates the accuracy of language models but also optimizes their response time, making them more adept at real-time tasks. Furthermore, its versatility allows for potential applications across a gamut of domains, ranging from mathematical computations to discerning common sense and even to understanding symbolic representations.

However, as with any pioneering technique, the journey of experimentation is often punctuated by revelations. Our hands-on experience with the GPT-3.5 model has shed light on a few inherent challenges associated with the Tree-of-Thought Prompting approach. Utilizing the methodology outlined in the Tree of Thoughts approach proposed by Yao et al. (2023) is highly likely to present challenges when attempting to tackle the issues raised by the Rosetta Stone proficiently. Unless the evaluation method is redefined, or until we can assist the model in discerning the latent intricacies underlying the language, it remains plausible that the linguistic challenge transcends a purely linear paradigm.

References

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8.

Ivan Derzhanski and Thomas Payne. 2010. The linguistics olympiads: Academic competitions in linguistics for secondary school students. *Linguistics at school: language awareness in primary and secondary education*, pages 213–26.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv preprint arXiv:2202.01344*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020a. Puzzling machines: A challenge on learning from small data. *arXiv preprint arXiv:2004.13161*.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020b. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.