

an application of LLMs to specific tasks but rather builds upon the insights that LLMs’ reasoning ability can be derived directly from the experience gained by analyzing incorrect reasoning chains, without relying on human priors.

To emphasize our insights, we share three key observations derived from applying gpt-3.5-turbo with a temperature of 0.7 and a top-p value of 0.7 on the Game of 24 dataset below.

The prompt with experience encourages LLMs to explore more logic in the responses. As shown in Table 5 when no experience is included in the prompt, the model generates the same reasoning step five times. This observation shows one of the common problems of LLMs, which is the lack of self-motivation to explore different reasoning logics. Thus, despite their strong potential for reasoning, LLMs may become trapped in a cycle of beginning with the simplest reasoning step, which may never culminate in finding the final solution. The ‘Obtained reasoning chain’ part of Table 5 presents the wrong reasoning chain. We believe that as Tree of Thoughts Yao et al. (2024) generates multiple responses as thoughts to build the nodes of the Tree, such a duplicated reasoning step may lead to the failure of this algorithm in some cases. However, in the second iteration of BoT, the introduction of experience into the prompt leads to the generation of distinct initial reasoning steps, as illustrated in Table 6. The final reasoning chain can ultimately arrive at the correct solution by commencing with a wider range of potential logic. Weng et al. (2023)

LLMs avoid making similar mistakes emphasized by the experience of the prompt. Without including experience, which contains the error analysis, in the prompt, LLMs will make many mistakes, such as deviations from the task’s rules and regressing to the initial reasoning step in the final process, as shown by Table 5’s ‘Obtained reasoning chain’ part. After analyzing this reasoning chain and incorporating the feedback as the experience into the prompt for the second iteration of BoT, it becomes evident from Table 6 that LLMs will fully learn from the experience before engaging in reasoning. First, none of the responses replicate the same erroneous reasoning step, as illustrated by the diverse initial reasoning steps in the ‘Five responses from the gpt-3.5-turbo’. Second, LLMs successfully circumvent all previously identified mistakes by rigorously adhering to task rules, eliminating incorrect reasoning, and executing logical reasoning steps. Third, it eventually leads to the correct solution for the ‘1 1 4 6’ Game of 24 task. Other work, such as Weng et al. (2023); Madaan et al. (2023); Zheng et al. (2023), also highlighted the importance of enhancing the prompt with the feedback, which is self-evaluation of previous answers. Nevertheless, BoT is the pioneering work founded on the insight that embracing error analysis for learning empowers LLMs to attain formidable reasoning capabilities.

Without human annotations, LLMs automatically organize an effective reasoning chain toward solving complex problems based on experience containing error analysis. BoT is the pioneering initiative that proposes an automated prompting framework, leveraging the insight that LLMs can acquire effective reasoning skills for problem-solving solely through error analysis and guidance, all without the need for human annotations. As shown in Table 5 which shows the first iteration of BoT, the initial prompt only contains the basic task guidance and the question without any in-context learning examples like what in CoT. Even though the reasoning chain obtained by LLMs with such a prompt contains numerous errors and invalid reasoning steps, its error analysis and advice can be included as the experience in the input prompt to benefit the reasoning for the second iteration, as shown in Table 6. It can be observed that with no prior human guidance on correct reasoning procedures, LLMs can acquire knowledge from experience that includes error analysis and guidance from previously generated reasoning chains, leading to a progressive improvement in reasoning for problem-solving.

Therefore, our BoT provides long-term guidance for research as it demonstrates the importance of recalling error analysis and advice when enabling LLMs to generate effective reasoning processes for complex tasks. With these insights, the research of prompt engineering on inducing the reasoning ability of LLMs can focus on how to generate experience instead of introducing more human priors.

C THOUGHT STRUCTURES GENERATION

BoT is an automated prompting framework that iteratively accumulates the experience derived from the analysis of reasoning chains. Consequently, BoT is generalized to various thought generation methods and LLMs capable of generating and evaluating reasoning steps. And the performance of