LLMs. The experiments conducted on the MATH dataset employed prominent large language models (LLMs), namely, GPT-3.5-Turbo, hereafter abbreviated as GPT3.5, and GPT-4, denoted as GPT4 for brevity. We directly utilized the release APIs of OPENAI.

Competitors.

- GPT4 ComplexCoT. This is the GPT4 model employing greedy decoding (i.e. temperature = 0) with the ComplexCoT Fu et al. (2022) prompting method. The reasoning examples utilized in the prompt for reasoning are derived from the corresponding Complex CoT publication Fu et al. (2022). As greedy decoding is used, we do not follow the self-consistency method Wang et al. (2022) to sample reasoning paths.
- GPT3.5. With the standard prompt, the GPT3.5 model is used to generate the answer.
- GPT3.5 ComplexCoT. Similar to the GPT4 ComplexCoT but change the model to GPT3.5.
- GPT4 PHP+ComplexCoT. This is the GPT4 model employing greedy decoding (i.e. temperature = 0) with the PHP Zheng et al. (2023)+Complex CoT Fu et al. (2022). Specifically, in the PHP Zheng et al. (2023) framework, the Complex CoT prompt is used to generate initial base answers, from which the PHP+Complex CoT can then develop the subsequent answer generation prompts. Thus, at the beginning of the interaction, by passing a concatenation of the base prompt of Complex CoT and the current question to the LLM, the base answer can be generated. Then, relying on the Complex CoT prompts revised into the PHP version with additional hint sentences, the progressive-hint prompting framework is performed on this base answer to update the hint over interactions to generate the right answer. We refer to this as the PHP+Complex CoT corresponding to the Progressive-Hint Prompting Complex CoT (PHP-Complex CoT) in the original work Zheng et al. (2023). The number of shots from Complex CoT is 8.
- GPT4 BoT wo/ experience. The GPT4 model is used to perform reasoning with the proposed BoT framework without the experience accumulation. The basic settings of BoT follow those presented in the main paper. Therefore, after one iteration, the aggregated chain will be used as the solution.
- GPT4 BoT. The GPT4 is used to perform reasoning with the full version of BoT as shown in the main paper.
- GPT4 BoT + CoT. Apart from the BoT framework, 5 reasoning examples from the CoT Wei et al. (2022) publication are included in the prompt. Therefore, in each iteration, the prompt contains not only experience but also additional 5 CoT reasoning examples.
- GPT3.5 BoT. Similar to the GPT4 BoT but change the model to GPT3.5.
- GPT3.5 BoT (GPT4). In this experiment, we utilize the GPT3.5 to perform reasoning, thus generating thought chains in the Thought Structure Generation. However, when performing the thought evaluation and the experience generation in the aggregated Thought Chain Analysis, the GPT4 model is used to get the evaluation and the analysis feedback.

We obtain the following additional observations from the results in Figure 5

The top performance of BoT on challenging problems derives from the accumulation of *experience*. BoT-related methods, such as GPT4 BoT and GPT4 BoT + CoT, consistently achieve the highest problem-solving rate on different sub-categories of MATH. Specifically, GPT4 BoT outperforms the current best GPT4 PHP+ComplexCoT by 8.6%, while GPT4 BOT + CoT is even 12.4% higher. In all seven categories, GPT4 BoT is at least 0.8% higher than GPT4 PHP+ComplexCoT, and the corresponding number on the Algebra problems is even 12.5%. Similar for GPT3.5 BoT and GPT3.5 BoT + CoT. However, when no experience is accumulated in the BoT framework, the performance drops significantly on all mathematical problems, as shown by the GPT4 BoT wo/ experience.

In addition to experience with error analysis, including correct examples, such as simple CoT instances, is essential for improving the problem-solving efficiency of the BoT in challenging mathematical problems.. GPT4 BoT outperforms the GPT4 PHP+ComplexCoT by a large margin on the first five sub-categories of MATH problems. Nevertheless, in the domains of Precalculus and Intermediate Algebra, which demand more intricate reasoning and complex logical steps for solutions, BoT exhibits only a marginal improvement of 0.8% and 2.4%, respectively. These gains