

are limited compared to the more substantial enhancements observed in simpler problem categories. After directly adding 5 correct CoT examples into the prompt, GPT-4 BoT + CoT demonstrates a significant performance boost, surpassing GPT-4 BoT by 7.7% and 11.5% in Precalculus and Intermediate Algebra domains, respectively. This basic conclusion from these observations is that to guarantee the top performance of BoT in complex mathematical problems, relying on trial-and-error analysis to learn how to reason is not sufficient; instead, the correct answers should also be provided in the prompt for LLMs.

While GPT3.5 with BoT may initially fall behind GPT-4 CoT, leveraging GPT-4 as the evaluator and analyzer to generate experience allows GPT-3.5 BoT (GPT-4) to outperform GPT-4 Complex CoT. With the GPT3.5, which has less capacity than GPT4, as the LLM, the solving rate obtained by BoT is at least 7.7% (on Algebra) lower than GPT4 ComplexCoT. It is evident that when less powerful LLMs produce lower-quality trial-and-error analyses, the BoT is unable to outperform GPT4 ComplexCoT. Thus, after using the GPT4 in the experience generation part while GPT3.5 is only used to generate reasoning steps, GPT3.5 BoT (GPT4) shows a significant improvement in all categories, leading to a solving rate of 55.8%, which outperforms GPT4 ComplexCoT by 5.5% and is even 1.9% higher than the current state-of-the-art GPT4 PHP+ComplexCoT. These observations further demonstrate that the accumulation of experience over iterations in the prompt constitutes the primary factor contributing to the success of the BoT framework.

G REASONING RESULTS OF “GAME OF 24”

First, in Table 5 - Table 9 we present the detailed prompts that BoT used during the reasoning process, thus providing a comprehensive understanding of what BoT does within each iteration. Then, starting from Table 10 we show some exact examples containing the whole reasoning process of BoT. Following the basic settings shown in the experiment section, these experiments are obtained using BoT with the GPT-3.5-turbo model.

Table 5: Reasoning steps generated by gpt-3.5-turbo when no experience is included in the input prompt. We first let the model generate one step of reasoning five times to check the diversity and then present the final reasoning chain after finishing the first iteration of BoT.

Prompt of Game of 24 without experience
<p>In the game of 24, you are given four numbers, and the goal is to use basic arithmetic operations (+, -, *, /) to combine these numbers and obtain a result of 24. You can only use each number once, and parentheses can be used to change the order of operations.</p> <p>Analysis format of each step: Step , Current set: , Selected two numbers: , Operation: , Computed new number: , Remaining numbers: , New set: .</p> <p>The given four numbers are: 1 1 4 6.</p> <p>Let's think step by step.</p> <p>Recall historical reasoning experience (Ignore when experience is empty):</p> <p>#####</p> <p>Pay attention to analysis and conclusion in the experience to avoid making similar mistakes by following the advice.</p> <p>Below is a list of ordered reasoning steps, accompanied by their evaluated scores (A higher score means the reasoning step is more likely to complete the task.):</p> <p>_____</p> <p>Based on listed reasoning steps only within the above "_____" (i.e., Not the ones in the experience block), please make one step of reasoning to generate only one subsequential possible reasoning step.</p>
Five responses from the gpt-3.5-turbo
<p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.,</p> <p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.,</p> <p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.,</p> <p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.,</p> <p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.,</p>
Obtained reasoning chain after aggregation
<p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: +, Computed new number: 2, Remaining numbers: 4 6, New set: 2 4 6.. Evaluate Score: 0.7</p> <p>Step 2, Current set: 1 1 4 6, Selected two numbers: 1 4, Operation: *, Computed new number: 4, Remaining numbers: 1 6, New set: 4 1 6.. Evaluate Score: 0.7</p> <p>Step 1, Current set: 1 1 4 6, Selected two numbers: 1 1, Operation: *, Computed new number: 1, Remaining numbers: 4 6, New set: 1 4 6. Evaluate Score: 0.7</p>

H REASONING RESULTS OF “GSM8K”

BoT uses similar basic prompts and the specific format as shown in Table 5 - Table 9. Only the task prompt will be changed, as shown in Table 15. Then, starting from Table 16 we show some exact examples containing the whole reasoning process of BoT. Following the basic settings shown in the experiment section, these experiments are obtained using BoT with the GPT-3.5-turbo model.