4.2 GAME OF 24

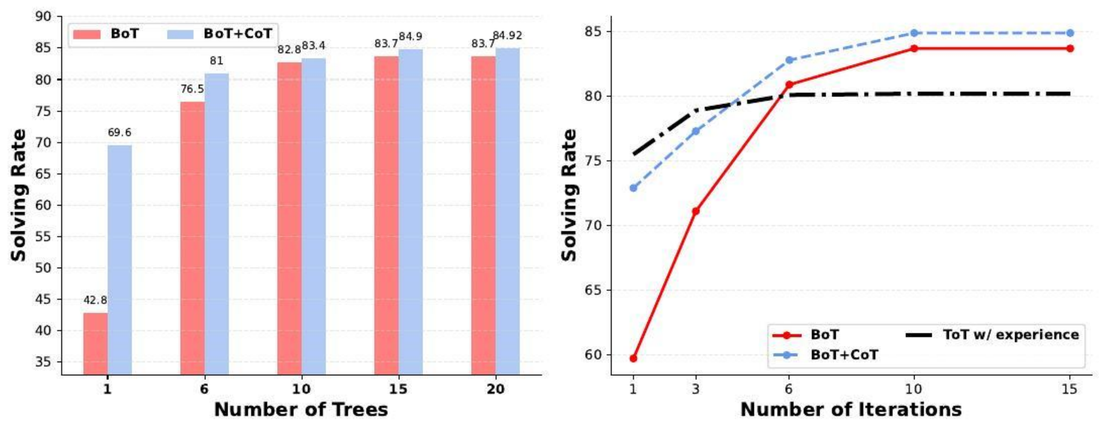

Figure 4: Comparison of three approaches across varying numbers of trees and iterations.

Method	Solving rate
Standard	7.3
Standard (best of 100)	33
CoT prompt	4
CoT prompt (best of 100)	49
CoT-SC (k=100)	9
ToT	74
BoT	83.7
BoT+CoT	84.9

Table 2: Results on `Game of 24` where the settings of different approaches follow those in ToT Yao et al. (2024).

Table 3: Showing aggregated thought chains and obtained *experiences* in iterations 1, 5, and 8. The given four numbers are: 2, 7, 8, 9.

t-th iteration	Two numbers	Arithmetic operation	New number set	Experience	Judgement
$\mathbf{F}^1$	2, 8	multiplication	16, 7, 9	The new set does not bring us closer to the target of 24. Try other numbers and operations.	Possible but more subsequent steps are required
	9, 7	addition	7, 16, 16	This step does not follow the rules of combining the remaining numbers and the obtained new number into a new set. Adjust the new set.	
	16, 7	multiplication	16, 112	Too many numbers in the new set. More steps are required to reach the target of 24.	
$\mathbf{F}^5$	9, 7	addition	16, 2, 8	The "Evaluation Score: 0.5" is low. Increase the score.	Possible but should revise some steps
	16, 8	addition	2, 24	It is not possible to further manipulate the numbers to reach 24. Choose different numbers.	
	2, 24	subtraction	22	The new set is not correct. Can choose other two numbers.	
$\mathbf{F}^8$	9, 7	addition	16, 2, 8	-	Possible
	16, 2	multiplication	32, 8	-	
	32, 8	subtraction	24	-	

Due to the hardness of the `Game of 24` problem, GPT-4 and Llama2 both perform badly on this task, even incorporating the CoT, and CoT-SC approaches. The Llama2 model even fails to follow the correct rules of addressing the problem, making the solve rate even lower. Especially when applying BoT, which relies on the *experience*, to Llama2, all results are lower than 5% without significant improvement. Thus, we only report the performance of BoT with GPT-4. To maintain a fair comparison, we follow the settings proposed by ToT Yao et al. (2024).

As shown in Table 2, BoT without human annotations is 9.7% higher than ToT, which relies on one example showing all possible next steps. Besides, BoT+CoT, which contains 5 CoT shots in the initial prompt, is 1.2% higher than BoT. Such a close performance between BoT and BoT+CoT is attributed to the boosting mechanism, which progressively revises weak thoughts, as discussed in subsection 4.1. Adopting an *experience*-driven iterative process, BoT exhibits enhanced performance as the number of trees M and the number of iterations T increment. Also shown by Fig. 4, compared to BoT+CoT, BoT relies more on M and T as it requires to collect *experience* from a better thought chain or longer iterations. Another observation is that when enabling ToT to operate iteratively with the prompt enriched by *experience*, the problem-solving rate escalates from 72.5% in the initial iteration to 80.2% by the 10-th iteration. This demonstrates that *experience* – the analysis of previous reasoning chains can be used by LLMs to significantly improve the solve rate. However, the score obtained by ToT is still 3.5% lower than BoT. This is attributed to the fact that the aggregation stage