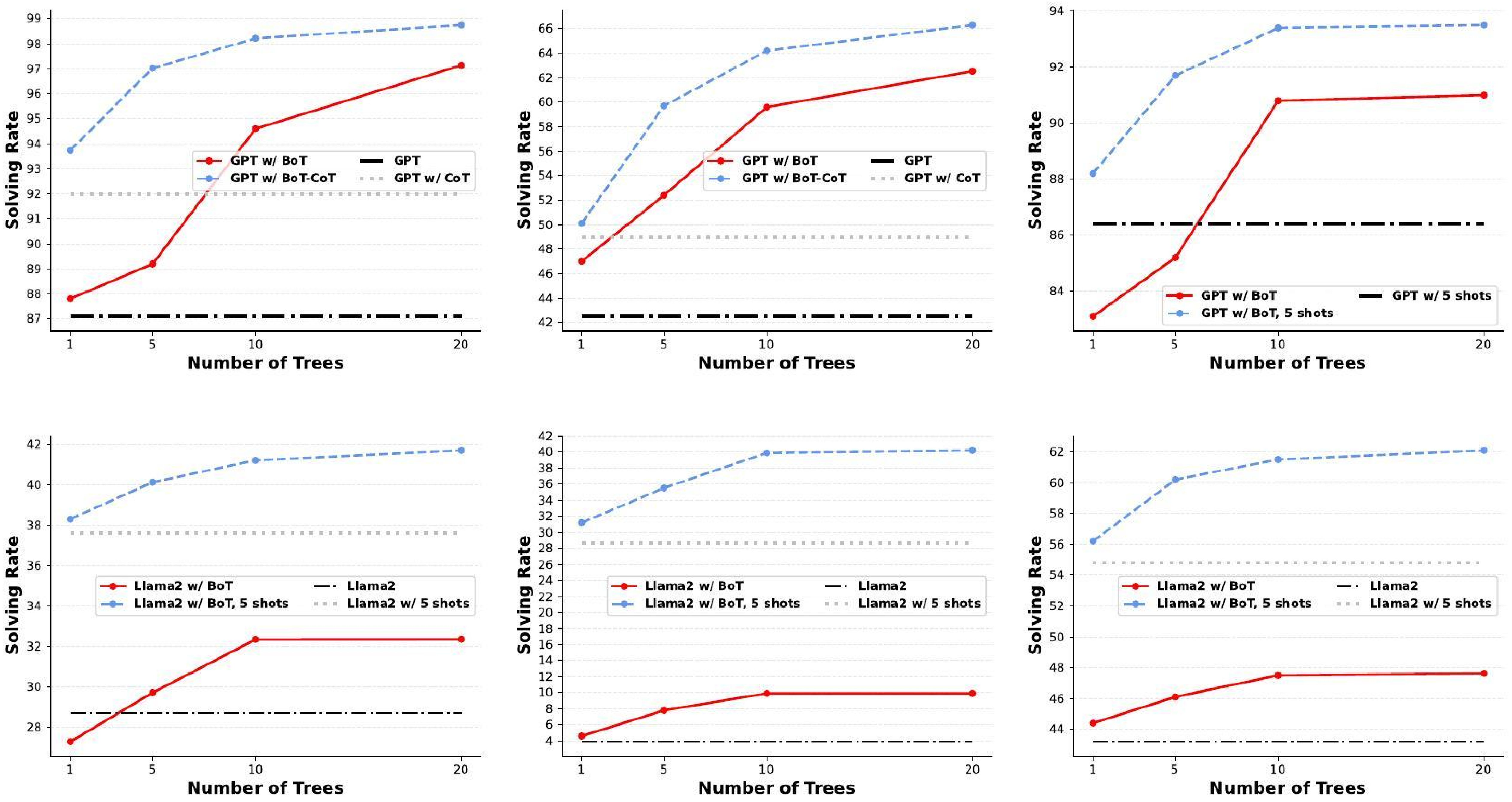
Table 1: Utilizing BoT with GPT-4, even without human annotations, yields a notable performance enhancement. Once the simple initial prompt of BoT contains CoT examples, the corresponding approach BoT+CoT exhibits even higher solving rates. Our framework is also evaluated against leading methods such as Model Selection Zhao et al. (2023), PHP Zheng et al. (2023), and CSV Zhou et al. (2023a), each achieving state-of-the-art (SOTA) performance on the SVAMP, AQuA, and GSM8K & MATH datasets, respectively.

Methods	No need Human Annotation	Datasets				Average
		SVAMP	GSM8K	AQuA	MATH	
SOTA	✗	93.7	97	79.9	84.3	88.7
Standard	✓	68.7	87.1	40.6	42.5	59.7
CoT	✗	77.6	92	74.0	48.93	73.1
Zero-shot CoT	✓	74.3	89.6	73.2	47.7	71.2
Complex-CoT	✗	90.5	94.9	77.5	50.4	78.3
PHP Complex-CoT	✗	91.9	95.5	79.9	53.9	80.3
BoT	✓	92.7 (↓ 1)	**97.1** (↑ 0.1)	**81.4** (↑ 2.5)	62.5 (↓ 21.8)	83.7 (↓ 7.6)
BoT + CoT	✗	**94.9** (↑ 1.2)	**98.7** (↑ 1.7)	**84.9** (↑ 5)	66.3 (↓ 18)	86.2 (↓ 2.5)



(a) GSM8K. **(b)** MATH. **(c)** MMLU.

Figure 3: Evaluating solve rates by applying BoT and BoT+CoT in GPT-4 OpenAI (2023) and Llama2 Touvron et al. (2023).

1.3% on average in GSM8K and AQuA datasets. We argue that the CoT examples can be regarded as the success cases in the *experience*, directly guiding the subsequent thought structures generation of BoT. Thus, cooperating with the iteration refinement, BoT+CoT reaches a new SOTA. It also deserves to show that because BoT can gradually collect analysis of various reasoning chains (bad or good) as *experience*, it is consistently close to the BoT+CoT. However, BoT and BoT+CoT, especially BoT, are at least 18% lower than SOTA in MATH. This observation means weak LLMs may not perform well with BoT due to their lower ability to analyze reasoning chains for an effective *experience*, as supported by Fig. 3

Fig. 3 presents that with BoT, GPT-4 and Llama2 are respectively improved by 11.6% and 4.4% on average in three datasets. The two numbers show a clear trend that when the LLM is weaker, BoT's performance drops significantly. With powerful GPT-4, as presented in Fig. 3, BoT and BoT-CoT behave similarly to those shown in Table. 1. Additionally, their performance escalates along a similar trend as the number of trees varies from 1 to 20. As Llama2 is weaker, BoT is unable to benefit from its analysis to perform the *experience*-driven iteration process, which is particularly shown by Fig. 3 (a). When provided with valid success cases, i.e., 5-shots, BoT, through progressive refinement, can still help Llama2 to solve more problems than the baseline even though the improvement is limited.