Parallel Continuous Chain-of-Thought with Jacobi Iteration

Haoyi Wu, Zhihao Teng, Kewei Tu*
School of Information Science and Technology, ShanghaiTech University
{wuhy1, tengzhh2022, tukw}@shanghaitech.edu.cn

Code Paper

EMNLP 2025 Suzhou, China 中国苏州

ORAL session @ A109, Wed, November 5, 16:30

What is **PCCoT**

TL;DR We propose Parallel Continuous Chain-of-Thought (PCCoT), which

- performs Jacobi iteration on the latent thought tokens;
- improves both training and inference efficiency of continuous CoT.

Relation

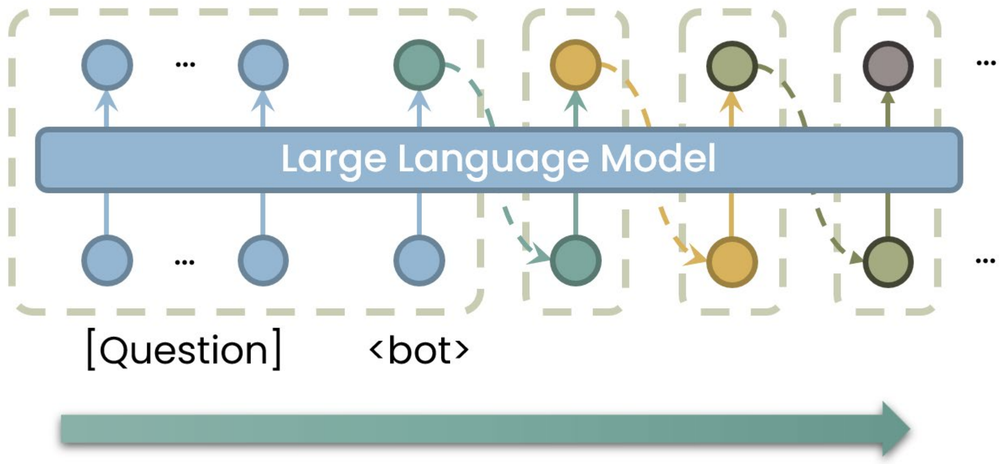
to existing approaches

There are two hyperparamers:

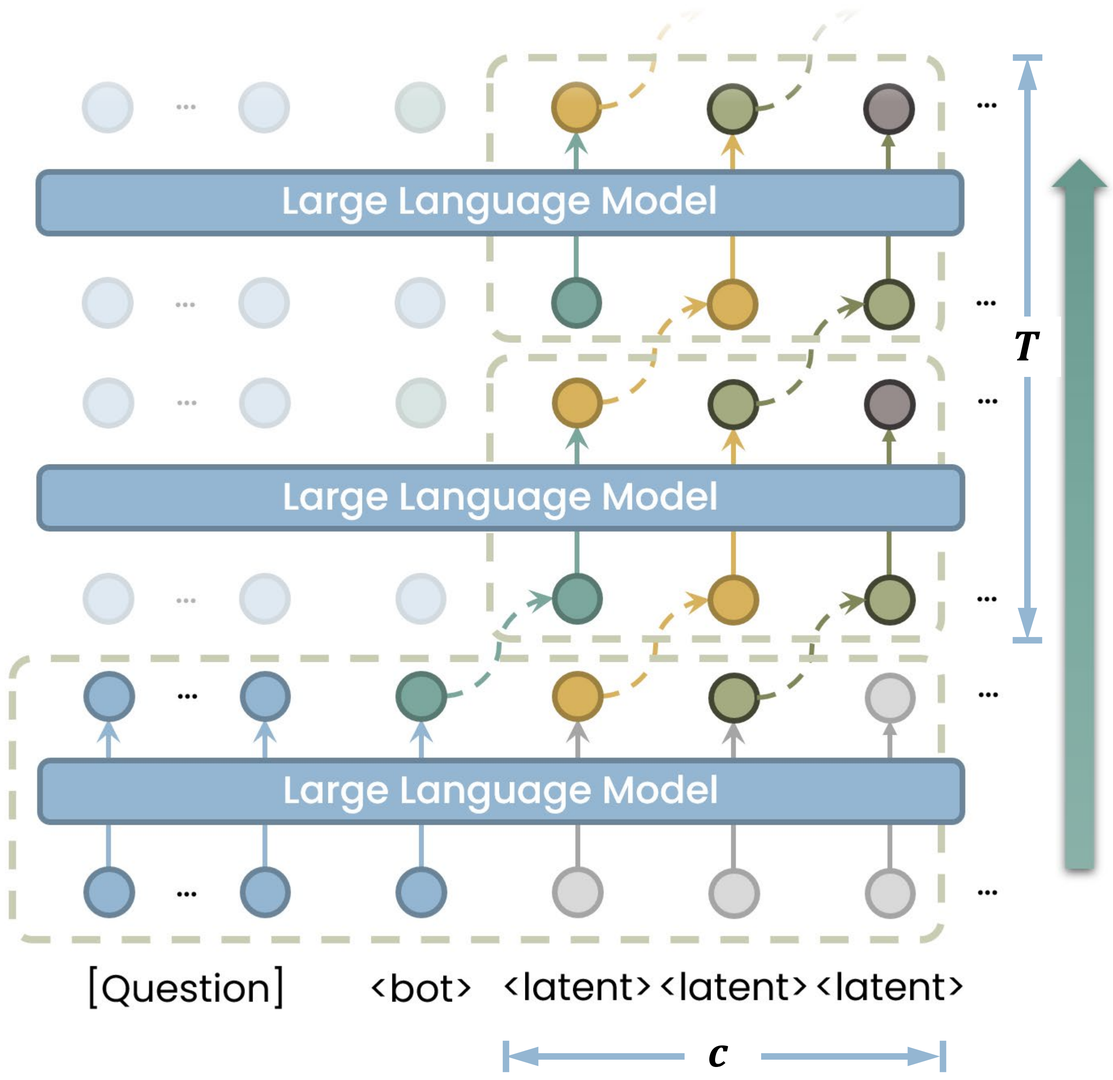
- c: the number of latent thought tokens;
- T: the number of extra iterations.

with different settings of c and T, PCCoT can be reduced to these existing approaches:

Approach	c & T
Implicit CoT (iCoT) (Deng et al., 2024)	$c = 0$
Pause Tokens (Goyal et al., 2024)	$c > 0, T = 0$
Continuous CoT (Hao et al., 2024)	$T \geq c > 0$



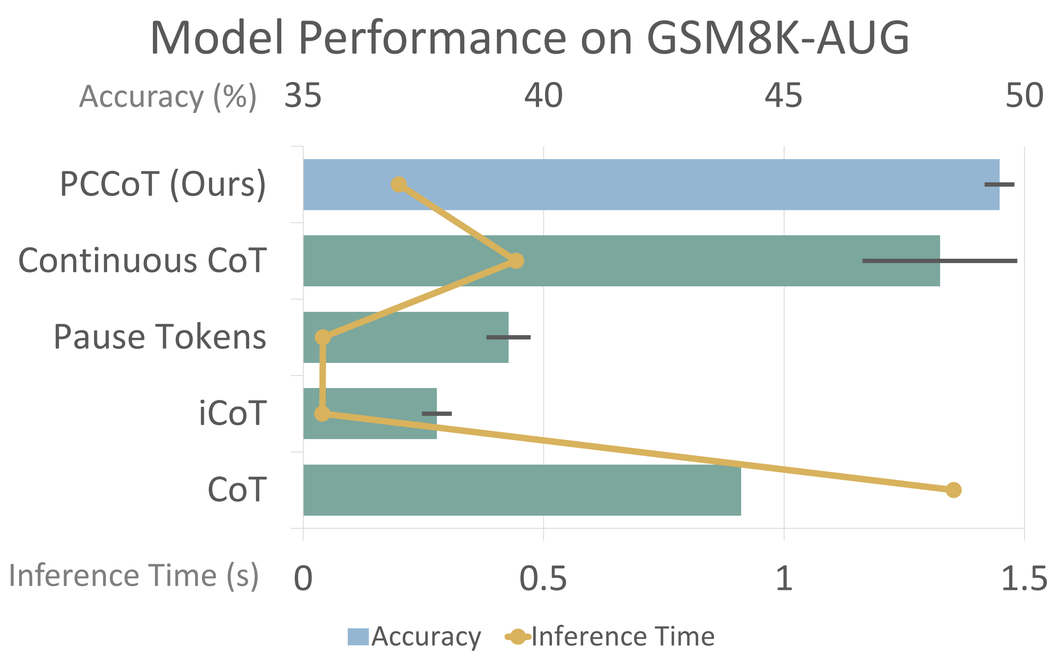
a. Continuous CoT



b. Parallel Continuous CoT (PCCoT)

Performance

PCCoT achieves better performance while saving nearly 50% of the training and inference time compared to continuous CoT. PCCoT also shows better stability and robustness.



* Experiment on GPT-2 Small, finetuned with LoRA on GSM8K-AUG. The results are averaged over 3 random runs with standard deviations as error bars. The inference time is measured with a batch size of 100. The training runs on 2 H800 GPUs and the inference runs on 1 A6000 GPU. Model training follows that of CODI (Shen et al. 2025).

Analysis

We further plot test set accuracy of PCCoT with different numbers of T and c. Interestingly, we find that increasing the number of iterations does not necessarily improve the performance. With a large value of T, training becomes unstable, which leads to a large standard deviation.