SoftCoT: Soft Chain-of-Thought for Efficient Reasoning with LLMs

Yige Xu, Xu Guo, Zhiwei Zeng, Chunyan Miao

Nanyang Technological University, Singapore

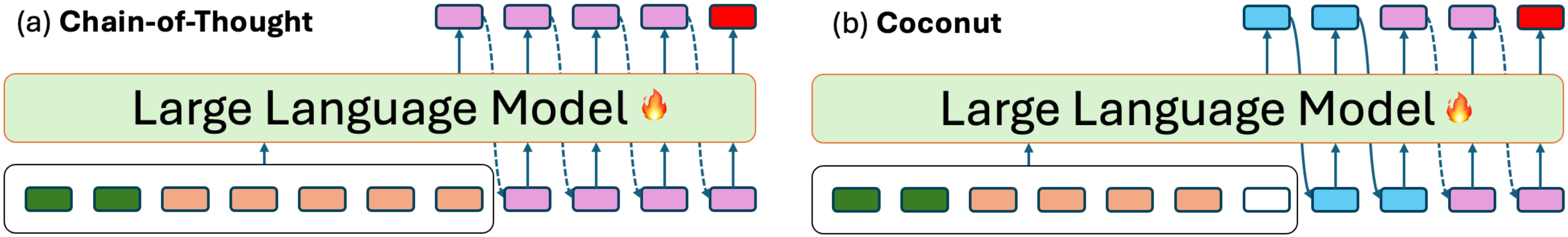
Research Background

- Chain-of-Thought reasoning has become one of the basic ability of LLMs.
- Three primary concerns:
  - Consistency and Stability: CoT can vary significantly with minor changes in prompts. [1,2]
  - Robustness: CoT's effectiveness depends on the quality of intermediate thoughts. [3]
  - Efficiency: CoT often requires substantial computational resources. [4]

SoftCoT

Continuous Space Reasoning

- Generate soft thought tokens according to the hidden of last-token last-layer
- Facilitates the reasoning chain generation
- Optimal latent-space exploration
  – Coconut [3], CCoT [5]



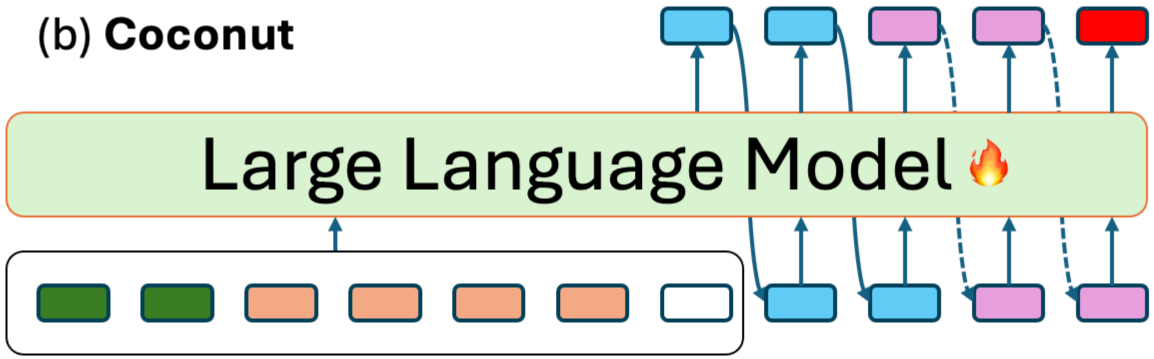
Motivation

Current latent-space reasoning approaches consider latent-space reasoning as a new task and fine-tunes the whole LLM [3,5], which results in …

- Catastrophic forgetting problem on SOTA LLMs
- Auto-regressively generate the soft thought tokens

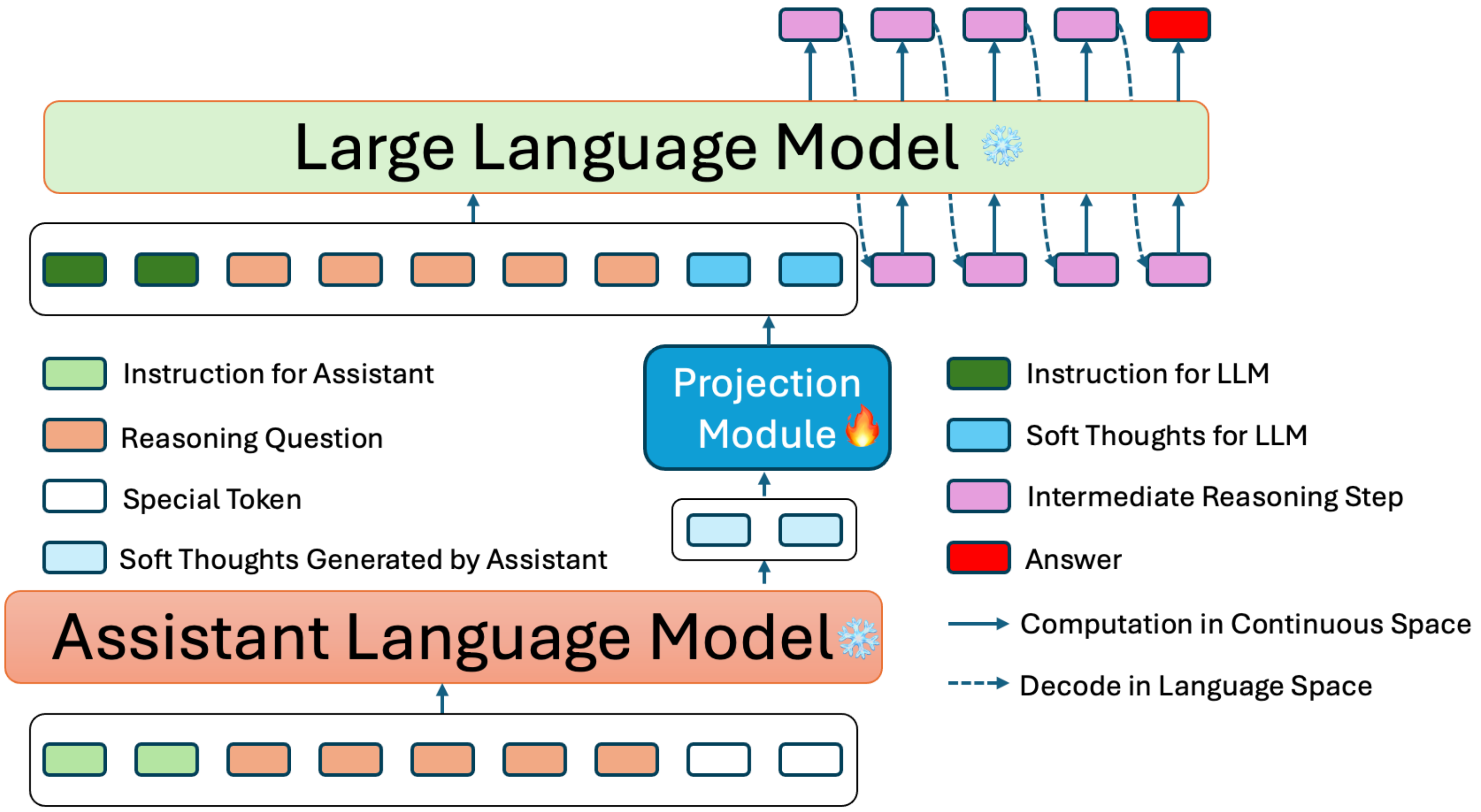
Can we ***freeze the LLM*** for mitigating the catastrophic forgetting problem?

Challenge: the fixed LLM struggle to generate learnable soft thought tokens.



How to generate the learnable soft thought tokens?

SoftCoT: Overall Architecture



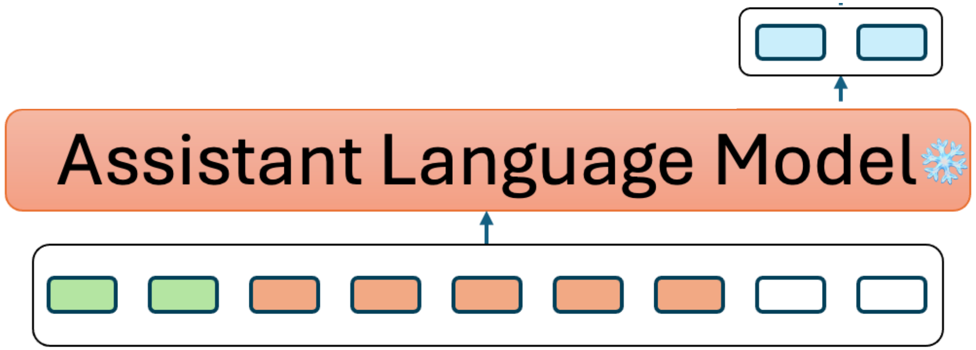
SoftCoT: Soft Thought Tokens Generataion

- Soft Thought Tokens Generation
  - Use auxiliary assistant model to produce the soft thoughts

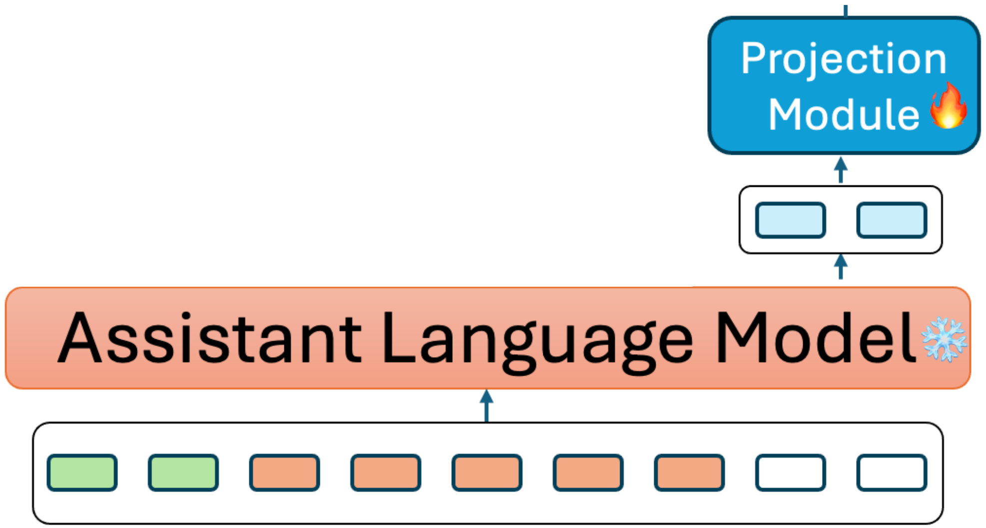
$$\mathbf{x}_{\text{assist}} = \text{concat}\left[\mathcal{I}_{\text{assist}}, \mathcal{Q}, \texttt{[UNK]}_{1:N}\right]$$

$$\mathbf{h}^{\text{assist}} = \text{Assistant}(\mathbf{x}_{\text{assist}}),$$

$$\mathbf{t}_{\text{assist}} = \mathbf{h}^{\text{assist}}_{|\mathcal{I}|+|\mathcal{Q}|+1:|\mathcal{I}|+|\mathcal{Q}|+N}.$$



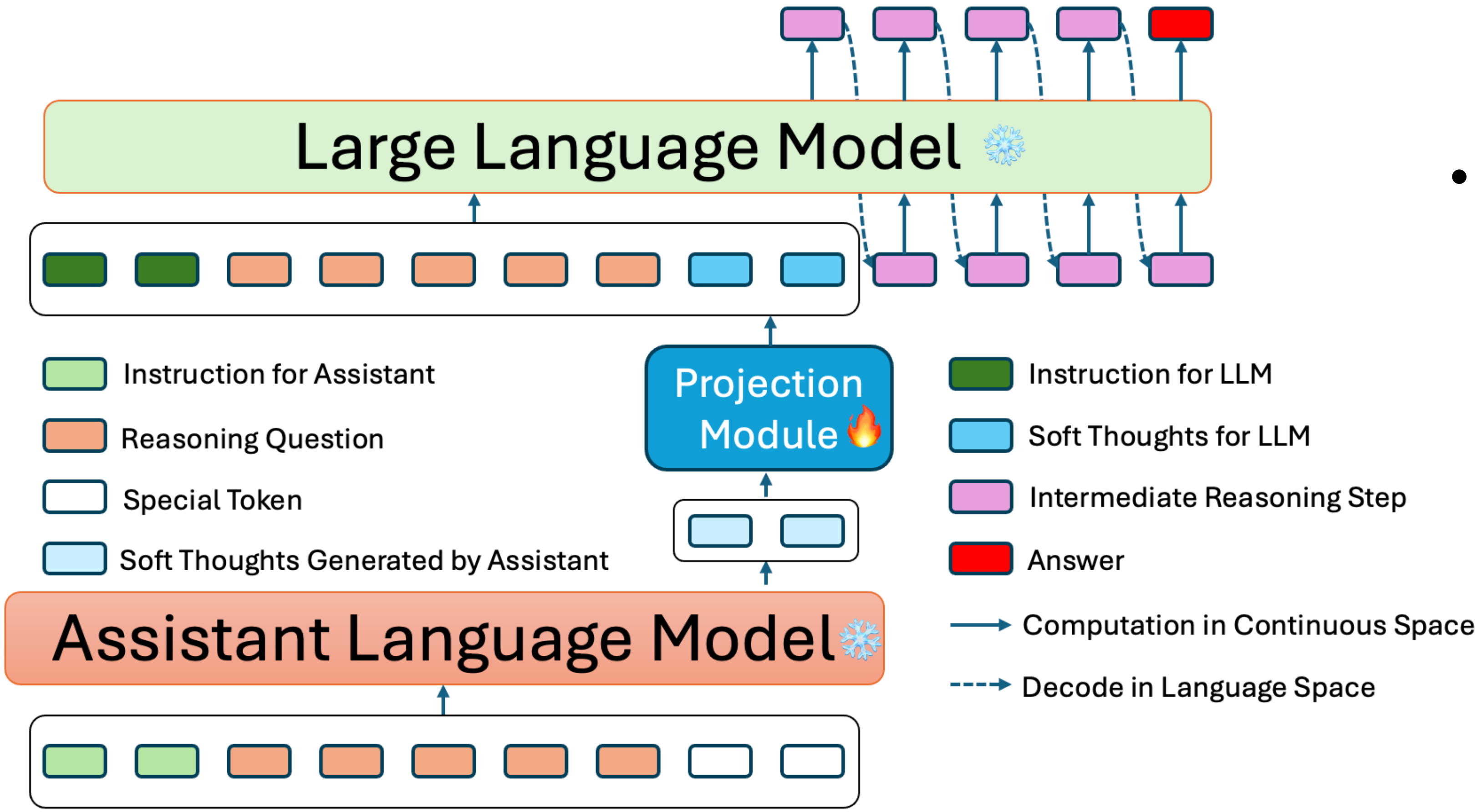
SoftCoT: Soft Thought Tokens Projection



- Soft Thought Tokens Projection
  - Maps the assistant-generated soft thoughts from the assistant model's embedding space to the LLM's embedding space.
  - Only the parameters in the projection module are trainable.

$$\mathcal{T}_{\text{soft}} = \text{Linear}_\theta(\mathbf{t}_{\text{assist}}),$$

SoftCoT: LLM Reasoning



- LLM Reasoning with SoftCoT
  - Apply the soft thoughts to aid LLMs in CoT reasonings.

$$\mathbf{x}_{\text{LLM}} = \text{concat}\left[\mathcal{I}_{\text{LLM}}, \mathcal{Q}, \mathcal{T}_{\text{soft}}\right],$$

$$\bar{\mathcal{R}} = \text{LLM}(\mathbf{x}_{\text{LLM}}),$$
$$\bar{\mathcal{A}} = \text{LLM}(\mathbf{x}_{\text{LLM}}, \bar{\mathcal{R}}),$$
$$\hat{\mathcal{A}} = \mathcal{E}(\bar{\mathcal{A}}),$$

Comparison with baselines

Model	GSM8K	ASDiv-Aug	AQuA	StrategyQA	DU	Avg.
	Mathematical			Commonsense	Symbolic	
GPT-2						
Coconut (Hao et al., 2024)	$34.10^{*}_{\pm 1.50}$	$38.92^{\dagger}_{\pm 0.00}$	$22.83^{\dagger}_{\pm 0.00}$	-	-	-
LLaMA-3.1-8B-Instruct						
Zero-Shot CoT	$79.61_{\pm 0.81}$	$86.78_{\pm 0.63}$	$54.65_{\pm 2.43}$	$65.63_{\pm 3.31}$	$54.40_{\pm 2.40}$	68.21
Zero-Shot CoT-Unk	$79.95_{\pm 0.59}$	$86.90_{\pm 0.41}$	$55.28_{\pm 1.88}$	$66.16_{\pm 2.70}$	$54.16_{\pm 1.46}$	68.49
Zero-Shot Assist-CoT	$80.76_{\pm 1.53}$	$86.96_{\pm 0.46}$	$55.83_{\pm 2.98}$	$66.55_{\pm 3.99}$	$58.24_{\pm 3.56}$	69.67
LoRA Fine-Tuning	$75.66_{\pm 0.00}$	$86.67_{\pm 0.00}$	$52.36_{\pm 0.00}$	-	-	-
Coconut (Hao et al., 2024)†	$76.12_{\pm 0.00}$	$86.80_{\pm 0.00}$	$53.15_{\pm 0.00}$	-	-	-
SoftCoT (Ours)	$\mathbf{81.03}_{\pm 0.42}$	$\mathbf{87.19}_{\pm 0.40}$	$\mathbf{56.30}_{\pm 1.67}$	$\mathbf{69.04}_{\pm 1.23}$	$\mathbf{59.04}_{\pm 1.93}$	**70.52**

- Supervised LoRA Fine-Tuning performs worse than zero-shot CoT, which make Coconut not applicable to SOTA LLMs
- Assistant model is effective to facilitate CoT reasoning
- SoftCoT consistently benefits from the supervised training

Generalization to Other LLM Backbones

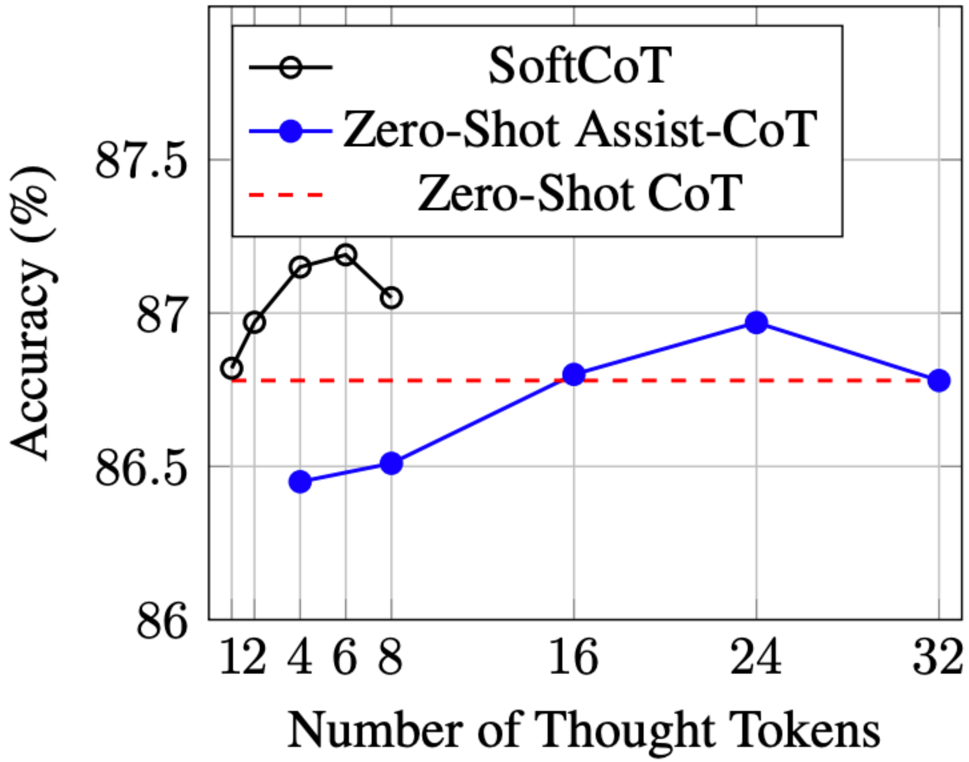
Model	GSM8K	ASDiv-Aug	AQuA	StrategyQA	DU	Avg.
	Mathematical			Commonsense	Symbolic	
Zero-Shot CoT	$83.70_{\pm 0.78}$	$87.19_{\pm 0.28}$	$64.53_{\pm 3.27}$	$49.65_{\pm 3.18}$	$66.40_{\pm 2.26}$	70.29
Zero-Shot CoT-Unk	$84.12_{\pm 0.71}$	$86.94_{\pm 0.89}$	$64.72_{\pm 2.06}$	$50.74_{\pm 1.90}$	$66.48_{\pm 1.43}$	70.60
Zero-Shot Assist-CoT	$84.85_{\pm 0.35}$	$88.63_{\pm 1.05}$	$64.96_{\pm 2.83}$	$52.71_{\pm 2.65}$	$67.04_{\pm 2.84}$	71.64
LoRA Fine-Tuning	$81.80_{\pm 0.00}$	$86.80_{\pm 0.00}$	$62.60_{\pm 0.00}$	-	-	-
Coconut (Hao et al., 2024)	$82.49_{\pm 0.00}$	$86.90_{\pm 0.00}$	$63.39_{\pm 0.00}$	-	-	-
SoftCoT (Ours)	$\mathbf{85.81}_{\pm 1.82}$	$\mathbf{88.90}_{\pm 1.01}$	$\mathbf{72.44}_{\pm 2.19}$	$\mathbf{60.61}_{\pm 1.55}$	$\mathbf{67.52}_{\pm 2.92}$	**75.06**

Results on Qwen2.5-7B-Instruct

- SoftCoT is effective across different LLM architectures

Model Analysis – Number of Thought Tokens

- Soft thoughts reduce the required number of thought tokens



Model Analysis – Size of Assistant Model

Method	0.5B	1.5B	7B
Zero-Shot CoT	83.70	83.70	83.70
Zero-Shot Assist-CoT	84.78	84.85	84.90
SoftCoT	**85.76**	**85.81**	**85.84**

Table 5: Performance on GSM8K with different sizes of assistant model on Qwen2.5 series.

- The scale of the assistant model has limited impact on the accuracy of the final answer

Model Analysis – Self-Consistency

Model	GSM8K		ASDiv-Aug		AQuA		StrategyQA		DU	
	$N=1$	$N=10$	$N=1$	$N=10$	$N=1$	$N=10$	$N=1$	$N=10$	$N=1$	$N=10$
Zero-Shot CoT	$79.61_{\pm 0.81}$	$90.36_{\pm 0.40}$	$86.78_{\pm 0.63}$	$89.23_{\pm 0.17}$	$54.65_{\pm 2.43}$	$63.23_{\pm 0.86}$	$65.63_{\pm 3.31}$	$70.13_{\pm 0.47}$	$54.40_{\pm 2.40}$	$65.76_{\pm 1.54}$
Zero-Shot Assist-CoT	$80.76_{\pm 1.53}$	$90.43_{\pm 0.69}$	$86.96_{\pm 0.46}$	$89.48_{\pm 0.36}$	$55.83_{\pm 2.98}$	$63.62_{\pm 0.99}$	$66.55_{\pm 3.99}$	$70.48_{\pm 0.68}$	$58.24_{\pm 3.56}$	$65.84_{\pm 1.93}$
SoftCoT (Ours)	$\mathbf{81.03}_{\pm \mathbf{0.42}}$	$\mathbf{90.63}_{\pm \mathbf{0.39}}$	$\mathbf{87.19}_{\pm \mathbf{0.40}}$	$\mathbf{89.75}_{\pm \mathbf{0.29}}$	$\mathbf{56.30}_{\pm \mathbf{1.67}}$	$\mathbf{65.51}_{\pm \mathbf{0.72}}$	$\mathbf{69.04}_{\pm \mathbf{1.23}}$	$\mathbf{71.14}_{\pm \mathbf{0.10}}$	$\mathbf{59.04}_{\pm \mathbf{1.93}}$	$\mathbf{67.36}_{\pm \mathbf{1.12}}$

Table 4: Self Consistency for SoftCoT on LLaMA-3.1-8B-Instruct. "N" indicates the number of reasoning chains.

- SoftCoT introduces an independent improvement mechanism, which can be effectively combined with self-consistency for enhanced reasoning performance

Takeaway messages

- We address the need for efficient CoT reasoning on continuous space within SOTA LLMs
  – Freezing the backbone LLM to mitigates the catastrophic forgetting problem.
  – Creating a learnable projection module to map the assistant-generated soft thoughts from the assistant model’s embedding space to the LLM’s embedding space.
- SoftCoT has demonstrated that
  – it enables reasoning on continuous space and has a better downstream performance than baselines.
  – it can be scaled to multiple LLM architectures
  – it can be scaled to existing test-time scaling methods such as self-consistency.

References

- [1] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, Yoav Goldberg. ***Measuring and Improving Consistency in Pretrained Language Models***. TACL 2021.
- [2] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, Denny Zhou. ***Self-Consistency Improves Chain of Thought Reasoning in Language Models***. ICLR 2023.
- [3] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, Yuandong Tian. ***Training Large Language Models to Reason in a Continuous Latent Space***. *arXiv preprint: 2412.06769*.
- [4] Zhenglin Wang, Jialong Wu, Yilong Lai, Congzhi Zhang, Deyu Zhou. ***SEED: Accelerating Reasoning Tree Construction via Scheduled Speculative Decoding***. COLING 2025.
- [5] Jeffrey Cheng, Benjamin Van Durme. ***Compressed Chain of Thought: Efficient Reasoning Through Dense Representations***. *arXiv preprint: 2412.13171*.