WORLD LAUREATES FORUM
世界顶尖科学家论坛
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
2026
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

CoT Vectors: Transferring and Probing the Reasoning Mechanisms of LLMs

Li Li[1], Ziyi Wang[1], Yongliang Wu[1], Jianfei Cai[2], Xu Yang[1]*

[1] Southeast University, China [2] Monash University, Australia * Corresponding author

01 ABSTRACT

Chain-of-thought (CoT) prompting improves multi-step reasoning, but longer prompts increase inference cost and CoT fine-tuning updates many weights. We encode task-level reasoning as a compact vector added to one transformer layer. Direct activation-gap extraction reveals strong layer dependence. A learned vector, trained while the language model remains frozen, raises average accuracy by 4.0 points on Qwen2.5-Math-7B and 1.6 points on LLaMA-3.1-8B-Instruct across four benchmarks.

Research question

Can a reusable latent shift transfer multi-step reasoning without updating the model?

02 METHOD

Extracted vector: average answer-token activation gaps between (Q, CoT, A) and (Q, A) support examples.
Learnable vector: a frozen teacher sees (Q, CoT, A); a student sees (Q, A) plus the vector. Optimize answer-state alignment and cross-entropy (λ = 0.5).
Inference: add the vector at one layer during each generation step.

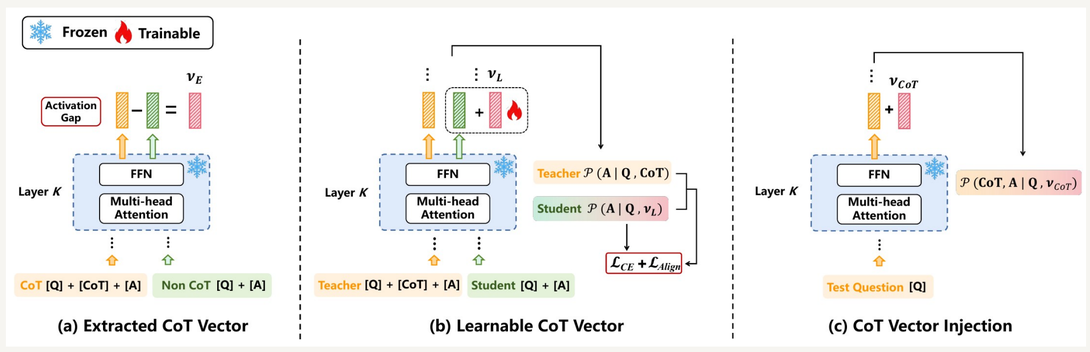


Figure 1. Original method schematic: activation-gap extraction, teacher–student learning, and single-layer injection.

Evaluation protocol

Qwen2.5-Math-7B and LLaMA-3.1-8B-Instruct; GSM8K, MATH-Easy, MATH-Hard, and MMLU-Pro. Each test uses 1,000 sampled examples. Support sets include up to 3,000 examples per math task and 70 MMLU-Pro examples. Compare zero-shot CoT, extracted vectors, learned vectors, and LoRA.

03 RESULTS

Best-layer CoT Vector results (accuracy, %). Learnable vectors use one hidden-size vector and keep all base-model parameters frozen.

Table 1. Accuracy (%); M-E/H = MATH Easy/Hard

Model	Method	# Par.	GSM8K	M-E	M-H	MMLU-P	Avg.
Qwen 7B	Baseline	—	74.6	69.9	47.9	33.2	56.4
	Extracted	—	78.2	72.0	49.7	35.3	58.8
	Learnable	**3.6K**	**83.5**	**71.9**	**50.9**	**35.1**	**60.4**
	LoRA	10.0M	79.0	70.4	48.2	33.8	57.9
LLaMA 8B	Baseline	—	77.4	62.0	34.6	44.6	54.6
	Extracted	—	78.6	63.2	35.7	45.5	55.8
	Learnable	**4.2K**	**78.2**	**63.8**	**36.4**	**46.2**	**56.2**
	LoRA	13.6M	78.6	63.5	36.3	45.5	56.0

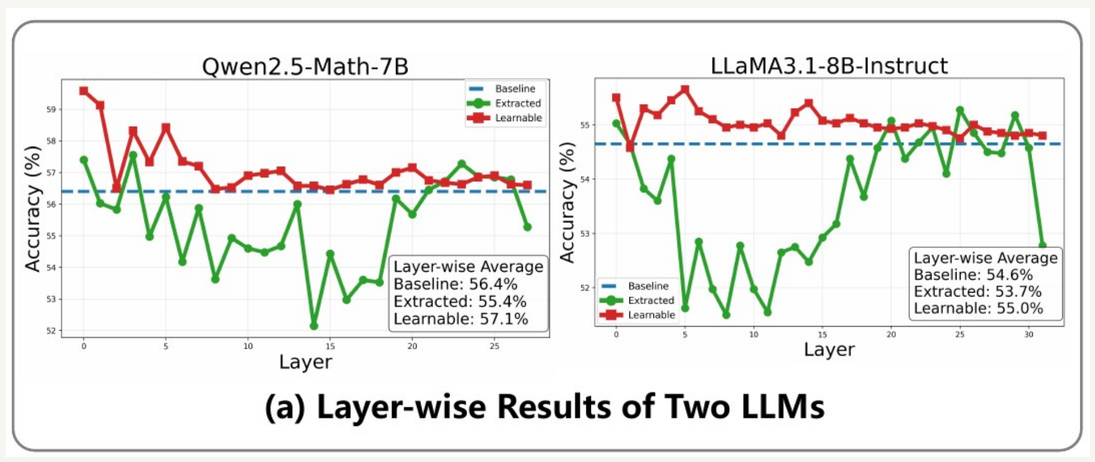


Figure 2. Extracted vectors vary sharply by layer; learned vectors remain comparatively stable. Curves average four benchmarks.

Key findings

Qwen: 60.4% mean accuracy with 3.6K trainable parameters, versus 56.4% for zero-shot CoT and 57.9% for 10.0M-parameter LoRA. LLaMA: 56.2% with 4.2K parameters, versus 54.6% baseline and 56.0% LoRA. The largest reported Qwen GSM8K gain is +8.9 points (83.5% vs. 74.6%).

04 DISCUSSION

The extracted vector works best in shallow and deep layers but can harm accuracy in middle layers. PCA and t-SNE analyses associate this instability with more dispersed, sample-specific mid-layer representations. Learned vectors are steadier, yet strong late-layer steering may overfit. Reported best-layer scores require layer selection.

05 CONCLUSIONS

A single learned hidden-state vector can distill CoT supervision into portable reasoning guidance with a frozen LLM. The performance and layer analyses support a perception–reasoning–expression account of the model's internal organization. Adaptive or finer-grained vectors may better capture diversity within a task.

CONTACT

Li Li · lilyli@seu.edu.cn

REFERENCE

Li et al., CoT Vectors, arXiv:2510.00579 (2025) · arxiv.org/abs/2510.00579