NeurIPS 2026 | 40th Conference on Neural Information Processing Systems

Farther the Shift, Sparser the Representation: Analyzing OOD Mechanisms in LLMs

Paper Code

Mingyu Jin[1], Yutong Yin[2], Jingcheng Niu[3,†], Qingcheng Zeng[2], Wujiang Xu[1], Xinyuan Song[7], Mengnan Du[4], Wei Cheng[5], Zhaoran Wang[2], Tianlong Chen[6], Dimitris N Metaxas[1]

[1] Rutgers University [2] Northwestern University [3] UKP Lab, TU Darmstadt [4] New Jersey Institute of Technology
[5] NEC Laboratories America [6] University of North Carolina at Chapel Hill [7] Emory University † Corresponding author

DIFFICULTY EXPERIMENTS / QWEN2.5-3B

Harder inputs concentrate activation energy

Across four tests of out-of-distribution (OOD) shift, harder inputs concentrate more of the final token's last-layer activation energy in the strongest 10% of dimensions.

Top-10% energy Share of squared activation in the strongest 10% of dimensions

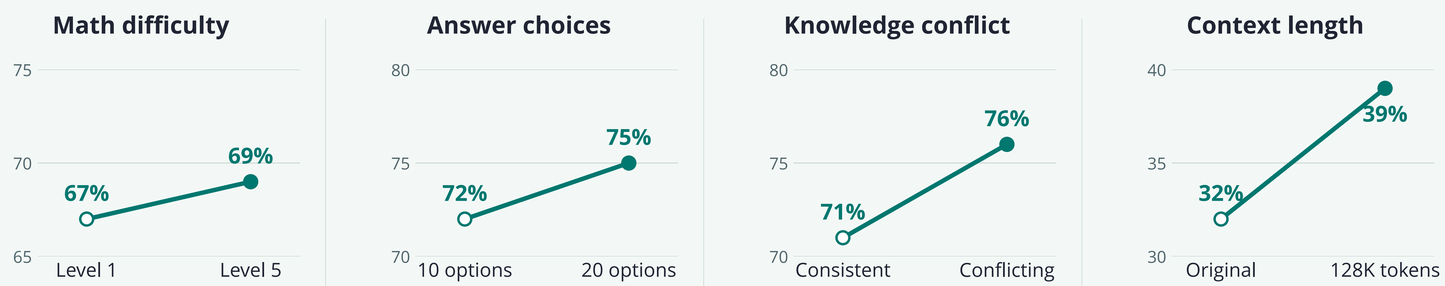


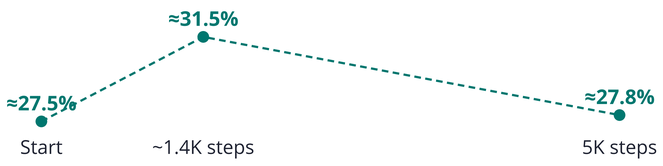
Figure 1 endpoint means, read from the plots and rounded to the nearest percentage point.
Each panel spans 10 percentage points. Individual problems vary within each condition.

TRAINING DYNAMICS / SYNTHETIC KNOWLEDGE GRAPH

Familiar inputs develop denser representations

Top-5% energy on fixed training examples

Trans-32-32 · Selected checkpoints from Figure 6



What changes with practice?

Early in training, energy concentrates in a few dimensions. As the model masters familiar examples, their energy spreads out again while accuracy stays near 75%.

FEW-SHOT REASONING / QWEN2.5-7B

Matching examples by difficulty

76.6%

SG-ICL accuracy on MATH-500

1.4 percentage points above Auto-CoT

Accuracy by prompting strategy

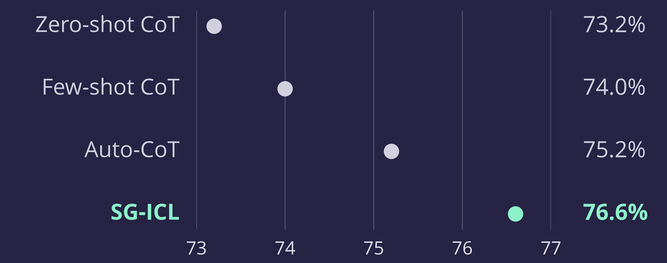


Figure 7 · CoT = chain-of-thought prompting

How SG-ICL selects demonstrations

Score the pool
Group examples by their top-k energy.

Retrieve by meaning
Find candidates semantically similar to the query.

Match the difficulty
Select candidates from the query's sparsity bin.