Logical Phase Transitions: Understanding Collapse in LLM Logical Reasoning

ACL 2026

Xinglang Zhang[1]*, Yunyao Zhang[1]*, Zeliang Chen[1], Junqing Yu[1], Wei Yang[1], Zikai Song[1]†

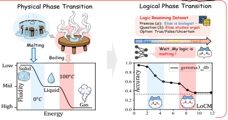
[1] Huazhong University of Science and Technology, Wuhan, China

1 Motivation

Smooth Degradation or Sudden Collapse?

In physical systems, state changes are often not smooth: as energy increases, matter can suddenly shift from **solid** → **liquid** → **gas**. We ask whether LLM logical reasoning shows a similar pattern.

LLM reasoning accuracy remains stable at low LoCM, drops sharply in **transition intervals**, then re-stabilizes at a lower level.



2 LoCM & LPTs

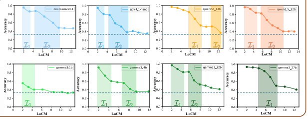
Measuring Complexity and Revealing Collapse

Logical Complexity Metric

$$LoCM(\phi) = f\left(\sum_{o\in O} \omega(o)\ freq(o,\phi) + \gamma h\right)$$

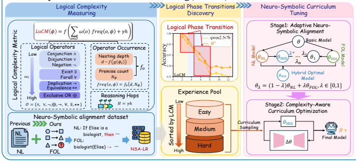
LoCM provides a ruler for **measuring** the structural complexity of logical reasoning.

LoCM *reveals* **LPT**



3 Method

Neuro-Symbolic Curriculum Tuning(NSCT)



Stage 1: Neuro-Symbolic Alignment

Align natural-language semantics with formal logical structure.

Stage 2: Complexity-Aware Curriculum Optimization

Expose the model to higher complexity at the right time.

NSCT trains with the phase boundary to mitigate reasoning collapse.

4 Results

Experimental Findings

Table 2: Performance comparison relative to Original. Green = improvement; Red = degradation; **Bold** = best.

Method	Dataset	Original	ProntoQA-tuned	ProofWriter-tuned	FOLIO-tuned	ProverQA-tuned	θ^* (Ours)
Naive	ProntoQA	55.20	55.40 (+0.20)	55.40 (+0.20)	56.20 (+1.00)	54.80 (-0.40)	**56.80** (+1.60)
	ProofWriter	44.16	44.00 (-0.16)	44.33 (+0.17)	43.80 (-0.36)	44.30 (+0.14)	**44.66** (+0.50)
	FOLIO	60.78	59.80 (-0.98)	59.80 (-0.98)	61.27 (+0.49)	59.80 (-0.98)	**62.25** (+1.47)
	ProverQA	54.13	54.07 (-0.06)	53.20 (-0.93)	53.80 (-0.33)	54.73 (+0.60)	**55.47** (+1.34)
	NSA-LR	49.55	49.09 (-0.46)	49.54 (-0.01)	49.09 (-0.46)	50.45 (+0.90)	**50.91** (+1.36)
	Naive Ave	52.76	52.47 (-0.29)	52.45 (-0.31)	52.83 (+0.07)	52.82 (+0.06)	**54.02** (+1.26)
CoT	ProntoQA	67.60	69.40 (+1.80)	70.60 (+3.00)	71.40 (+3.80)	71.20 (+3.60)	**72.00** (+4.40)
	ProofWriter	55.16	54.33 (-0.83)	54.00 (-1.16)	56.33 (+1.17)	53.50 (-1.66)	**60.71** (+5.55)
	FOLIO	66.17	65.68 (-0.49)	63.23 (-2.94)	63.72 (-2.45)	**66.67** (+0.50)	65.20 (-0.97)
	ProverQA	60.70	55.70 (-5.00)	60.80 (+0.10)	60.70 (-0.00)	61.50 (+0.80)	**64.20** (+3.50)
	NSA-LR	57.70	58.20 (+0.50)	54.70 (-3.00)	58.63 (+0.93)	59.62 (+1.92)	**65.00** (+7.30)
	CoT Ave	61.47	60.66 (-0.81)	60.67 (-0.80)	62.16 (+0.69)	62.50 (+1.03)	**65.42** (+3.95)

Table 3: Comparison with recent training-free reasoning methods on NSA-LR.

Method	Low	Medium	High	Overall
Naïve	60.6	49.8	38.6	49.6
+ NSCT	60.9 (+0.3)	50.2 (+0.4)	41.6 (+3.0)	50.9 (+1.3)
CoT	75.5	58.4	39.4	57.7
+ NSCT	84.0 (+8.5)	64.2 (+5.8)	46.8 (+7.4)	65.0 (+7.3)
ToT	82.0	64.8	42.4	63.2
+ NSCT	88.6 (+6.6)	68.3 (+3.5)	48.5 (+6.1)	68.4 (+5.2)
DetermLR	84.0	63.9	43.8	63.9
+ NSCT	87.5 (+3.5)	67.9 (+4.0)	49.2 (+5.4)	68.2 (+4.3)
SymbCoT	83.8	63.7	46.9	64.8
+ NSCT	89.2 (+5.4)	68.5 (+4.8)	50.5 (+3.6)	69.5 (+4.7)

Finding 1: LPTs universally occur across models. ✅

Finding 2: NSCT improves model's logical reasoning across benchmarks ✅

Finding 3: NSCT generalizes to inference-time strategies ✅

★ Xinglang is seeking for 27 Spring Research Intern opportunities and 2028 Fall Phd positions.

Scan here to contact me !