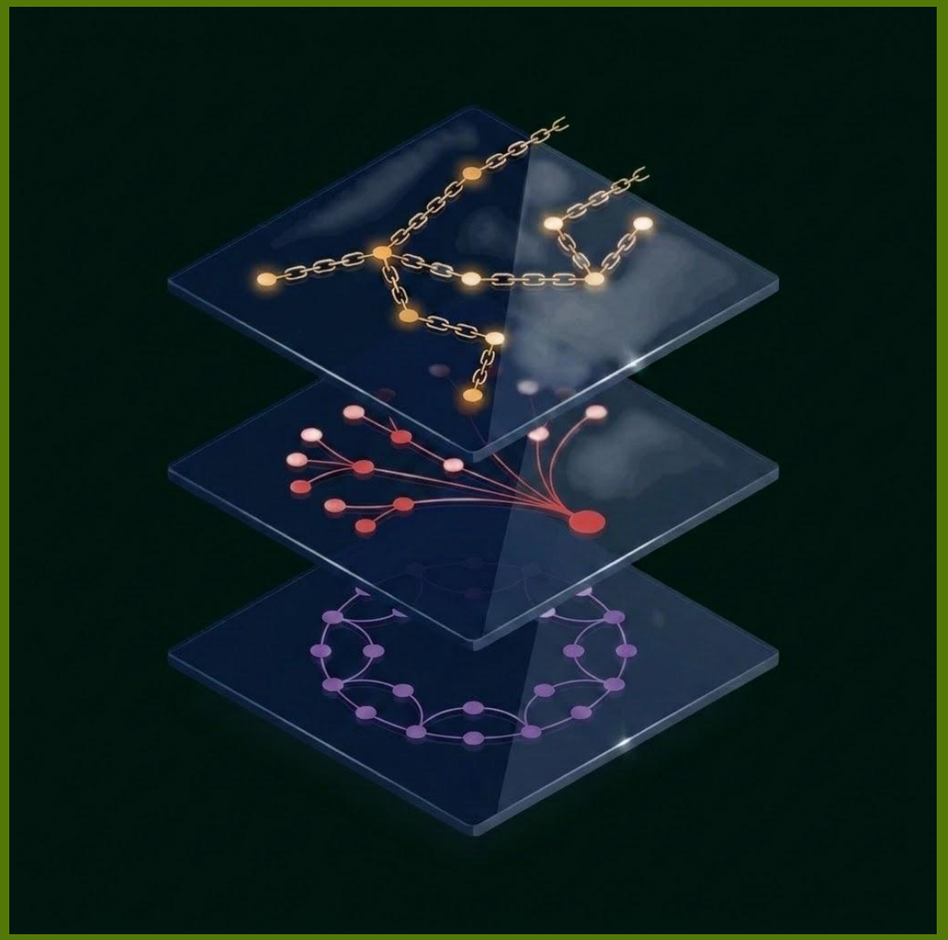
ICLR
International Conference On Learning Representations

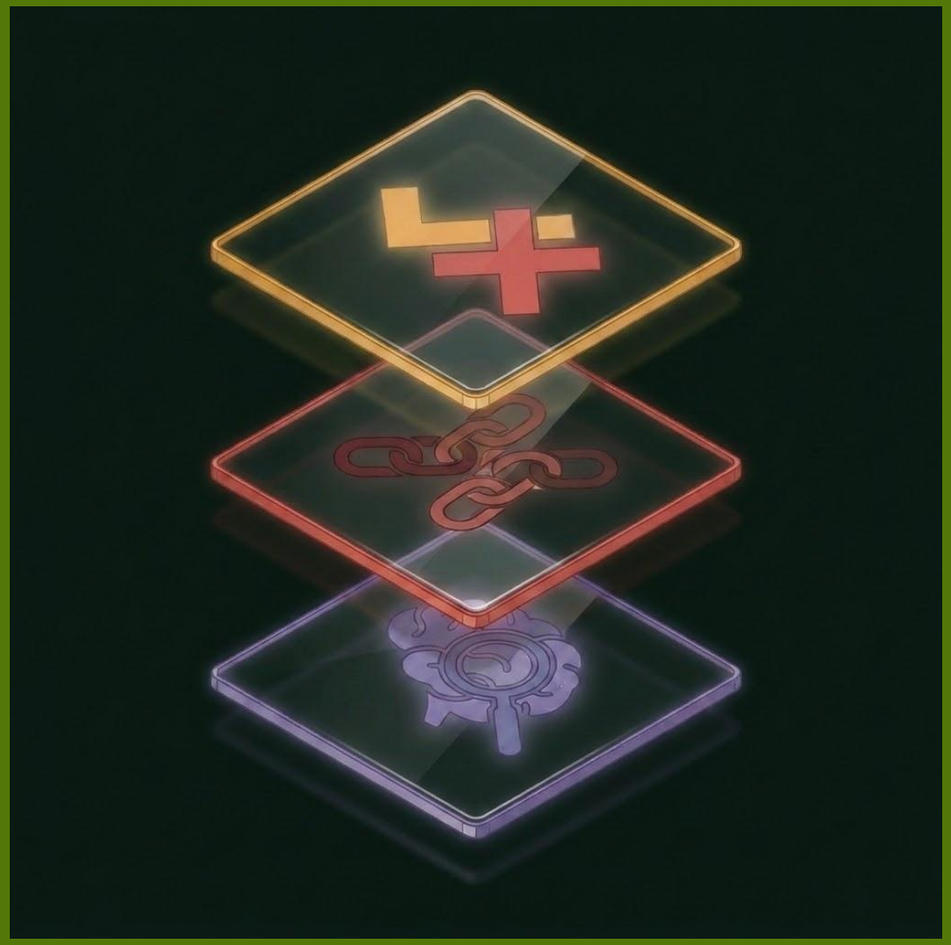
ICLR 2026 WORKSHOP · LOGICAL REASONING OF LLMS · RIO DE JANEIRO

Scaling Reasoning Depth Reveals Three Tiers of Failure in Multi-Model Mathematical Deduction



Harsh Rathva ·

Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India



THE PROBLEM

Why Accuracy Alone Is Misleading

Standard evaluation collapses qualitatively distinct failure mechanisms into a single "incorrect" label.

Truncation Failure

A model correctly decomposes a number theory problem and identifies the right strategy, but runs out of tokens before reaching the answer.

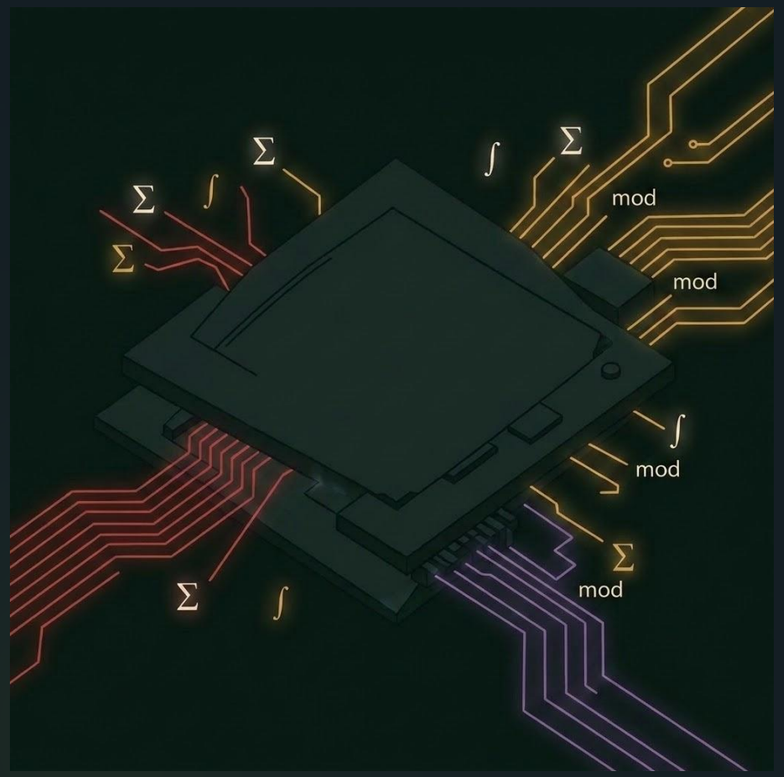
Meta-Cognitive Failure

A model proves its own answer wrong through self-verification, acknowledges the contradiction, and submits the wrong answer regardless.

Central thesis: Scaling reasoning depth shifts the dominant failure mode from superficial truncation artifacts to structural logical instability.

EXPERIMENTAL SETUP

Small-N, High-Signal Diagnostic Study



Benchmark

15 competition-level problems from AIME (2018–2022) and IMO Shortlist (2001–2002), spanning number theory, combinatorics, algebra, and geometry.

Models

- QwQ-32B (RL-optimized reasoning)
- DeepSeek-R1-Distill-32B (distilled from R1-671B)
- Phi-4-Reasoning-14B (high-density pretraining)

Protocol

- 4-bit quantization, NVIDIA H100 (80 GB), greedy decoding (T=0.0)
- Token budgets: 4,096 (primary) + 8,192 (ablation on P15)
- 45 independent reasoning traces + 90 cross-model verification attempts
- Diagnostic signal: finish reason (eos vs length), tokens consumed, extracted answer, verifier verdicts

CORE CONTRIBUTION

Three-Tier Failure Taxonomy

Three mechanistically distinct failure modes arranged along a depth-dependent progression.

Tier 1 — Capacity

8/11 errors (73%) · Finish reason: length

Valid deduction truncated by generation budget. Extracted answer is an intermediate computation, not a deliberate conclusion.

Anchor: P5 — two models truncated one step from the answer.

Confidence: 1.0. No failure signal.

Tier 2 — Correlated Deduction

2/11 errors (18%) · Finish reason: eos

Shared invalid logical reduction produces wrong answers regardless of available compute.

Anchor: P7 — all three models produced 196 instead of 756 via identical invalid reduction.

Tier 3 — Meta-Cognitive Override

1 controlled specimen · Finish reason: eos

Self-verification produces valid negation, then model overrides it during answer selection.

Anchor: P15 at 8,192 tokens — QwQ correctly refuted n=34, then submitted it. Ground truth: 797.

RESULTS

The 92% Finding — Finish Reason as a Zero-Cost Diagnostic

92%

EOS Accuracy

Of 12 naturally completing (eos) runs, 11 were correct — vs. 27% overall

0

Errors Fixed

Majority voting corrected zero errors and introduced one false correction

10

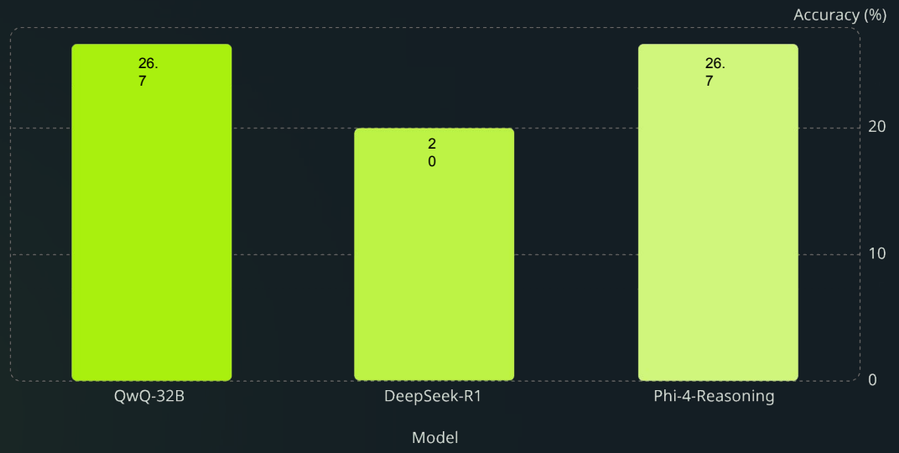
Errors Reinforced

Voting amplified shared errors rather than correcting them

1

Correct→Wrong

Voting converted one correct answer into an incorrect one



Individual Accuracy (4/15 problems)

QwQ-32B: 26.7% (4/15) · DeepSeek-R1-Distill-32B: 20.0% (3/15) · Phi-4-Reasoning-14B: 26.7% (4/15)

Perhaps the single most actionable finding for practitioners: finish reason alone separates Tier 1 from genuine reasoning failures at zero additional cost.

FAILURE CORRELATION

Why Majority Voting Cannot Help

Pairwise Jaccard Overlap

$J(A,B) = |E_A \cap E_B| / |E_A \cup E_B|$

QwQ–DeepSeek: J = **0.92** (11/12 shared failures)

QwQ–Phi-4: J = **0.83** (10/12)

DeepSeek–Phi-4: J = **0.77** (10/13)

With individual error rate p ≈ 0.75, even independent errors would yield ensemble error ≈ 0.84. High Jaccard similarity eliminates even the theoretical benefit of voting.

Tier 1 induces correlated truncation artifacts; Tier 2 induces shared deductive blind spots — both defeat voting.

VERIFICATION

Verification Amplifies Failure

63–72%

Compute Consumed

By cross-model verification — with zero correct-answer recoveries

0/90

Recoveries

Across all 90 verification attempts

Problem Mutation

DeepSeek fabricated an entirely different problem (P14) and verified against the fabrication.

Training-Data Injection

QwQ solved unrelated memorized problems (P15 at 8,192 tokens), suggesting answer 144 from a different problem.

Degenerate Generation

Phi-4 entered a meta-instruction loop consuming full budget without mathematical analysis.

False Negatives

P12: all three models derived 116 correctly, then all three verifiers marked it incorrect — 100% false-negative rate.

Echo-Chamber Verification

P7: QwQ verified DeepSeek's wrong answer (196) as correct at confidence 0.728, sharing the same invalid reduction.

TOKEN ABLATION

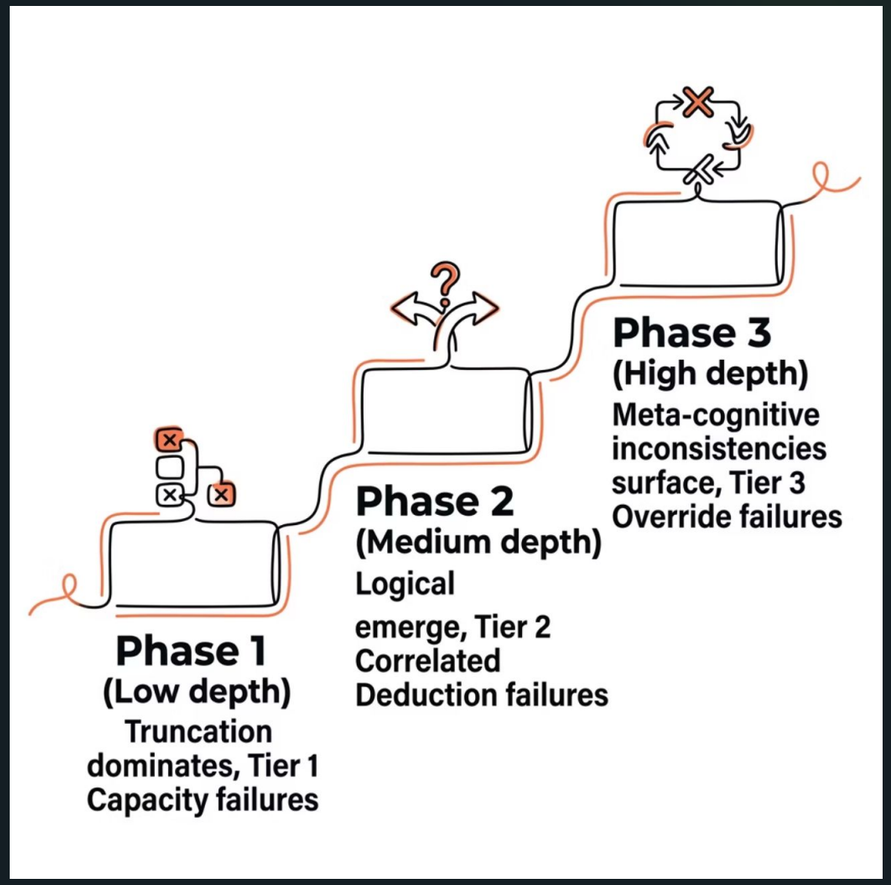
Depth Reveals Structure — Not Just Better Answers

P15 Ablation: QwQ-32B

Metric	4,096 tokens	8,192 tokens
Finish reason	length (truncated)	eos (completed)
Answer	16,777,192	34 (deliberate)
Self-verification	Impossible	Succeeded, then overridden
Tokens remaining	0	505
Failure tier	Tier 1 — Capacity	Tier 3 — Override

DeepSeek at 8,192 tokens correctly rejected its candidate after verification showed the condition failed. Meta-cognitive override is **model-specific**, not architecturally inevitable.

Conceptual Depth Model



Standard evaluations at fixed context length sample only one phase and may systematically mischaracterize model capability.

TAKEAWAYS

Failure-Tier-Aware Evaluation Is Required

1

Report Finish Reason

eos vs length separates Tier 1 from genuine failures at zero cost. Models scoring 27% may possess substantially greater competence.

2

Voting Requires Independence

Pairwise Jaccard ($J > 0.77$) diagnoses violated independence. Each tier induces correlation defeating voting.

3

Verification Is Not Auditing

Shared blind spots cause verification to confirm shared errors — 63–72% of compute, zero recoveries.

4

Meta-Cognitive Override Exists

Models produce valid self-refutation then override it, producing wrong answers indistinguishable at the output layer.

Confidence signal warning: Tier 1: P5 reported confidence 1.0 on extraction artifacts · Tier 2: P7 reported verification confidence 0.728 on a shared error · Tier 3: QwQ submitted a wrong answer indistinguishable from a correct one at the output layer

Future directions: Larger benchmark with post-cutoff problems · fp16/bf16 replication · Blind verification · Mechanistic interpretability · Tool-augmented experiments

Thank you. Contact: u24ai036@aid.svnit.ac.in