SCALING REASONING DEPTH REVEALS THREE TIERS OF FAILURE IN MULTI-MODEL MATHEMATICAL DEDUCTION

Harsh Rathva
Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, India
u24ai036@aid.svnit.ac.in

ABSTRACT

We present a diagnostic analysis of mathematical reasoning failures across three reasoning-specialized language models on 15 competition-level problems. Through detailed trace analysis of 45 independent reasoning attempts and 90 cross-model verifications (each of 3 models verifying each other model's solutions across 15 problems), we identify a three-tier taxonomy of failure modes that current evaluation methodologies routinely conflate: capacity failures (correct reasoning approaches truncated by generation limits, accounting for 8 of 11 errors in our sample; 95% CI: 48%–91%), correlated deductive failures (genuine logical errors where models independently converge on the same wrong answer through a shared invalid reasoning step, 2 of 11 in our sample; 95% CI: 2%–52%), and meta-cognitive override (a model that correctly refutes its own candidate answer and submits it anyway, documented in a controlled extended-context experiment). We demonstrate that majority voting across models corrected zero errors across our dataset because each failure tier induces distinct correlation structures: capacity failures produce correlated truncation artifacts, and reasoning failures reflect shared deductive blind spots that verification cannot detect. Furthermore, we reveal that cross-model verification actively amplifies these correlated errors. Rather than functioning as an independent logical auditor, the verification phase suffers from severe meta-reasoning pathologies, including problem mutation and training data injection, achieving a 3/3 false-negative rate on the one problem where all models answered correctly (P12). Our findings suggest that failure-tier-aware evaluation is required to accurately assess and improve logical consistency in large language models.

1 INTRODUCTION

A model may correctly decompose a number theory problem using the Chinese Remainder Theorem, systematically enumerate modular residues, identify the exact correct deductive strategy, and still produce a wrong answer. More troublingly, a model may *prove its own answer wrong* through explicit self-verification, acknowledge the contradiction, and submit the wrong answer regardless. Current evaluation frameworks classify all such cases identically as "incorrect," conflating fundamentally different failure mechanisms under a single label. This conflation obscures both the genuine deductive capabilities that LLMs possess and the specific bottlenecks that prevent those capabilities from producing correct conclusions.

Consider Problem 7 in our study (AIME 2018 II #10, ground truth: 756; see 3.2). Three models (QwQ-32B, DeepSeek-R1-Distill-32B, and Phi-4-Reasoning-14B) independently reduced the condition $f(f(x)) = f(f(f(x)))$ to idempotency via the *same* invalid logical step, produced 196, and when QwQ verified DeepSeek's answer, confirmed it as correct at confidence 0.728. QwQ completed reasoning naturally (finish reason: eos, 3326 tokens); this is not a truncation artifact. The models shared a structural deductive blind spot, and ensemble verification amplified rather than caught it.

The more common failure is infrastructural. In Problem 5 (AIME 2022 I #15; see 3.2), two models independently discovered valid solution strategies and were truncated one or two computation steps

from the correct answer. The system reported unanimous agreement at confidence 1.0. No failure signal was produced, because none was detectable from the output layer. The answers were extraction artifacts from valid but incomplete deductive chains, and unanimous agreement among artifacts is indistinguishable from unanimous agreement among solutions.

A third failure mode is visible only when token constraints are relaxed. In a controlled extended-context experiment (8192 tokens) on Problem 15, QwQ-32B completed its reasoning naturally with 505 tokens remaining. It correctly computed $2^{34} \equiv 59 \pmod{125}$, correctly noted "$59 - 34 = 25$, not $0 \bmod 125$, so $n = 34$ is not a solution," and submitted 34. The model produced a valid logical negation of its own answer and overrode it. This is not truncation, and not a shared blind spot; it is a failure at the meta-cognitive level: the model's self-verification succeeded and its answer-selection ignored the result.

These three failure modes (capacity exhaustion, correlated deductive error, and meta-cognitive override) form a depth-dependent progression. Our central finding is:

> *Scaling reasoning depth shifts the dominant failure mode from superficial truncation artifacts to structural logical instability.*

At shallow depth, truncation dominates and masks all other failure modes. At medium depth, reasoning completes but shared deductive blind spots produce correlated wrong answers. At extended depth, meta-cognitive inconsistencies emerge between self-verification and answer selection. Each tier requires distinct mitigations and produces its own failure correlation that ensemble methods cannot overcome.

Our contributions are:

1. A **three-tier failure taxonomy** (Capacity / Correlated Deduction / Meta-Cognitive Override) that distinguishes failure modes conflated by standard accuracy metrics, with empirical anchors (P5, P7, P15@8192) for each tier. Meta-cognitive override is a new failure category documented via one controlled specimen; the taxonomy is our primary contribution, not the tier proportions.
2. **Depth-as-variable framing**: a controlled token-limit ablation (4096→8192) showing how the same problem transitions from Tier 1 to Tier 3 failure, establishing that the tiers are empirically separable.
3. Evidence that **verification amplifies rather than corrects** these failures: 90 cross-model verification runs across both context budgets, consuming 63–72% of compute, producing zero correct-answer recoveries and pathologies that worsen with increased budget.
4. **Failure correlation analysis** (Jaccard 0.77–0.92) explaining why majority voting corrected 0/15 errors, since each tier induces its own correlation structure defeating the independence assumption.

2 Related Work

Multi-model aggregation for reasoning. Self-consistency decoding (Wang et al., 2023) samples multiple reasoning paths and selects the most frequent answer, achieving significant gains on arithmetic and commonsense reasoning. Multi-agent debate (Du et al., 2024) extends this through iterative cross-model interaction. These approaches implicitly assume error independence, an assumption our work shows fails systematically for mathematical deduction, where failures are tier-correlated rather than independent.

Evaluation of LLM reasoning. The MATH benchmark (Hendrycks et al., 2021) evaluates problem-solving with binary accuracy metrics. Chain-of-thought prompting (Wei et al., 2022) elicits step-by-step reasoning but does not differentiate failure mechanisms. Process reward models (Lightman et al., 2024) provide step-level feedback. Our three-tier taxonomy extends this line by showing that aggregate accuracy conflates qualitatively different failure modes.

Self-correction and logical consistency. Huang et al. (2024) demonstrate that LLMs cannot self-correct reasoning without external oracles. Stechly et al. (2023) show GPT-4 fails to recognize its own errors under iterative prompting. Our Tier 3 finding (meta-cognitive override) reveals a more specific pathology: self-verification *succeeds* but the model overrides it. Wan et al. (2025) document

confirmation bias in reasoning chains; our work identifies a concrete mechanism by which this bias operates.

Ensemble diversity and failure correlation. Majority voting requires classifier diversity to provide improvement (Sun & Dance, 2012). Our Jaccard-based analysis ($J = 0.77$–0.92) demonstrates that reasoning LLMs from overlapping training paradigms lack sufficient diversity for ensemble improvement under high-correlation failure conditions, and our three-tier analysis explains why: shared constraints (Tier 1), shared training biases (Tier 2), and meta-cognitive inconsistencies under extended reasoning (Tier 3).

Tool-augmented reasoning. External computation tools directly mitigate Tier 1 by offloading arithmetic search; symbolic verifiers address Tier 2 by providing deductive checking without shared neural biases. Our study characterizes the baseline failure structure under pure chain-of-thought, establishing which tiers are infrastructurally addressable and which persist.

3 Experimental Setup

3.1 Models and Constraints

We evaluate three reasoning-oriented language models under 4-bit quantization: **QwQ-32B** (RL-optimized reasoning), **DeepSeek-R1-Distill-32B** (distillation from R1-671B), and **Phi-4-Reasoning-14B** (high-density pretraining). This selection represents different training paradigms within the decoder-only transformer family, testing whether architecturally similar models provide meaningful diversity for ensemble reasoning.

3.2 Dataset

We use 15 competition-level problems from AIME (2018–2022) and IMO Shortlist (2001–2002), spanning number theory, combinatorics, algebra, and geometry. Each problem requires multi-step deductive reasoning with a unique integer answer, enabling unambiguous correctness evaluation. This small-N design enables detailed per-problem analysis of failure mechanisms (see Appendix B).

3.3 Evaluation Protocol

Each problem undergoes a three-phase pipeline:

Phase 1 (Independent Solving): Each model solves every problem independently with a 4096-token generation limit, greedy decoding (temperature 0.0), and no sampling. This produces 45 independent reasoning traces.

Phase 2 (Cross-Model Verification): Each model verifies the other models' solutions through targeted queries about logical validity, computational correctness, and deductive soundness. This produces 90 cross-model verification attempts.

Phase 3 (Weighted Consensus): Answers are aggregated through weighted majority voting (base weights: QwQ 2.0, DeepSeek 1.8, Phi-4 1.5, reflecting relative model scale and training paradigm). Consensus confidence is computed as the sum of weights of models agreeing on the selected answer divided by total weight (max = 1.0 under unanimous agreement). For ties, the highest-weight model's answer is selected. Sensitivity analysis confirms the consensus outcome is identical under equal weights for 14 of 15 problems.

3.4 Extended-Context Ablation

To distinguish meta-cognitive override (Tier 3) from capacity-induced truncation (Tier 1), we re-run Problem 15 (AIME 2021 II #13) with a doubled token budget of 8192 tokens, holding all other parameters constant. This allows naturally completing runs that are unavailable at 4096 tokens, enabling direct observation of whether extended reasoning produces self-consistent conclusions.

Table 1: Independent solving performance under 4-bit quantization with 4096-token limit.

Model	Accuracy	Avg. Tokens	Failure Set
QwQ-32B	26.7% (4/15)	3890	$\{2, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15\}$
DeepSeek-R1-Distill-32B	20.0% (3/15)	3920	$\{1, 2, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15\}$
Phi-4-Reasoning-14B	26.7% (4/15)	3650	$\{2, 3, 4, 5, 6, 7, 8, 10, 13, 14, 15\}$

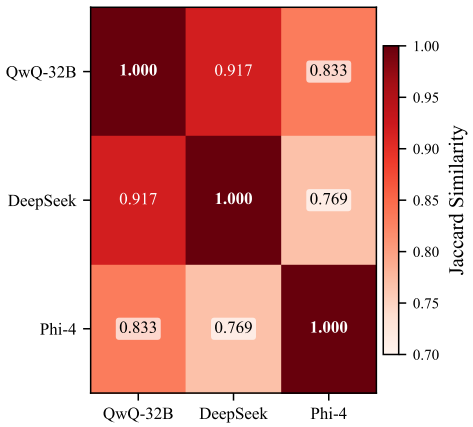

Figure 1: Failure correlation heatmap (Jaccard similarity). Values of 0.77–0.92 indicate strongly overlapping failure sets, mathematically limiting voting improvement.

3.5 Diagnostic Framing

We record each model's **finish reason** (eos vs. length) to classify runs by tier: truncated runs (`length`) are Tier 1 candidates; naturally completing runs (`eos`) with wrong answers enable Tier 2 and Tier 3 analysis. We compute Jaccard similarity $J(A, B) = |E_A \cap E_B|/|E_A \cup E_B|$, where E_A, E_B are the error sets of models A and B.

4 Results

4.1 Independent Performance (RQ1)

All models show similar performance (20–27% accuracy). Of 45 total runs, 33 (73.3%) were truncated at the token limit (finish reason: `length`) and 12 (26.7%) completed naturally (finish reason: `eos`). **Of the 12 `eos` runs, 11 produced correct answers (92%), compared to 27% overall** (see Appendix D, Table 11). This is perhaps the single most actionable finding for practitioners: models that complete reasoning naturally perform at dramatically higher accuracy, and the `eos`/`length` finish reason is a zero-cost diagnostic signal. The result is consistent with the Tier 1 characterization: truncation, not reasoning failure, drives most errors. The truncated subset constrains our ability to distinguish Tier 2 from Tier 3 failures; only the `eos` subset permits this analysis.

4.2 Voting Outcomes (RQ2)

Majority voting achieved 26.7% accuracy (4/15), identical to the best individual models. The outcome breakdown:

- **Fixed**: 0/15 (0%, 95% CI: [0%, 21.8%])
- **Reinforced**: 10/15 (66.7%, 95% CI: [42.1%, 87.9%])
- **Hurt**: 1/15 (6.7%, 95% CI: [0.3%, 32.0%])
- **Unchanged**: 4/15 (26.7%, 95% CI: [7.9%, 55.1%])

Table 2: Formal tier classification criteria. A problem is assigned to the highest-numbered tier whose criteria it satisfies.

Criterion	Tier 1 (Capacity)	Tier 2 (Correlated)	Tier 3 (Meta-Cog.)
Finish reason	`length`	`eos`	`eos`
Answer type	Extraction artifact	Deliberate wrong	Deliberate wrong
Error location	Truncation point	Logical step in trace	Meta-cognitive integration
Cross-model pattern	Artifact coincidence	Shared logical reduction	Model-specific
Verification correctable?	N/A (no reasoning to check)	No (echo chamber)	No (override)

4.3 Failure Correlation (RQ3)

The high Jaccard similarities indicate strongly overlapping failure sets: QwQ–DeepSeek: 0.92 (11 shared / 12 total), QwQ–Phi-4: 0.83 (10/12), DeepSeek–Phi-4: 0.77 (10/13). Under independence, majority voting improves accuracy only when individual accuracy exceeds 50%. With $p \approx 0.75$ individual error rate, independent errors would yield ensemble error $p_{\text{ens}} = p^3 + 3p^2(1-p) \approx 0.84$, still worse than individual performance. This confirms the well-known result that majority voting hurts when individual accuracy falls below 50%: high error-rate models are more likely to agree on wrong answers than to collectively produce correct ones. Since all models score below 30%, the theoretical benefit of independence is itself limited, and our observed high Jaccard correlation (0.77–0.92) eliminates even that limited benefit by ensuring failure sets overlap rather than complement. High Jaccard similarity is a necessary but not sufficient condition for dependence in the voting-theoretic sense; it establishes that failure sets overlap substantially, which is the operationally relevant criterion for predicting ensemble failure in practice.[1]

5 Three-Tier Failure Taxonomy

The 11 incorrect consensus answers decompose into three mechanistically distinct failure tiers. The following percentages are descriptive of our 15-problem dataset under 4-bit quantization at 4096 tokens. We present the taxonomy, the qualitative distinction between the three failure modes, as our primary contribution; the proportions are illustrative of their relative prevalence in this specific sample, not estimated population rates. Table 2 formalizes the classification criteria; evidence for each tier follows.

5.1 Tier 1: Capacity-Induced Truncation

Definition. The model identifies a valid deductive approach and begins correct execution, but exhausts its token budget before reaching a final answer. The extracted "answer" is an intermediate computation value or problem-statement number, not a deliberate conclusion.

In our sample, capacity failures account for **8 of 11 errors (73%, 95% CI: 48%–91%)**. In each case, trace analysis confirms the model's deductive strategy was sound.

Anchor case: P5 (AIME 2022 I #15, ground truth: 33). DeepSeek independently found the correct trigonometric substitution ($\alpha = \pi/8$, $\beta = \pi/24$, $\gamma = 5\pi/24$) and was truncated while computing $\sin(\pi/8)$, *one step* from the answer. Phi-4 independently found an equivalent algebraic approach through product systems ($AB = 1/4$, $BC = 1/2$, $CA = 3/4$) and was truncated computing $(\sqrt{6}-2)^4$, *two steps* away. The extraction heuristic assigned "1" (from the equation right-hand side, present in all traces) to all three models. Consensus: unanimous, confidence 1.0, failure_type: "none."

This is the most dangerous consensus failure in our study: **maximum confidence on extraction artifacts from valid, nearly-complete deductive chains**. An operator trusting this confidence score would receive "1" (wrong; GT: 33) with no indication of failure.

The extended-context ablation on P15 confirms the structural nature of Tier 1. Doubling the budget (4096→8192) produced dramatically deeper reasoning in all three models (Table 3), but 8192 tokens remains insufficient: the ground truth $n = 797$ corresponds to the 99th candidate in DeepSeek's parameterization, requiring approximately 18,000–32,000 tokens for any systematic approach. Tier 1

[1]This independence baseline uses unweighted majority for clarity. Our actual protocol uses weights (QwQ 2.0, DeepSeek 1.8, Phi-4 1.5); the qualitative conclusion holds regardless.

Table 3: P15 token-limit ablation: same problem, doubled budget.

Model	4096 Answer	4096 Finish	8192 Answer	8192 Finish
QwQ-32B	16,777,192	`length`	**34**	**eos**
DeepSeek	17	`length`	672	`length`
Phi-4	13	`length`	209	`length`

failure is not resolved by moderate budget increases; it requires budgets proportional to the problem's search space.

Implication. Tier 1 failures are not reasoning failures. Models that correctly identify trigonometric substitutions for elite-difficulty algebra are fundamentally different from models that apply wrong theorems, yet both receive accuracy score zero.

5.2 Tier 2: Correlated Deductive Failure

Definition. Models reach confident wrong answers through completed reasoning chains containing an identifiable logical error. The error persists regardless of available computation: the deduction itself is flawed.

In our sample, Tier 2 accounts for **2 of 11 errors (18%, 95% CI: 2%–52%)**, but these are disproportionately dangerous because they produce high-confidence wrong answers that resist correction through voting or verification.

Anchor case: P7 (AIME 2018 II #10, ground truth: 756). The problem asks for the number of functions $f : \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4, 5\}$ satisfying $f(f(x)) = f(f(f(x)))$ for all x. All three models independently produced the following chain:

1. $\forall x$: $f(f(x)) = f(f(f(x)))$ implies $f(f(x))$ is a fixed point of f ✓
2. $\forall y \in \text{Im}(f)$: $f(y)$ is a fixed point ✓
3. $\therefore \text{Im}(f) \subseteq \text{Fix}(f)$, i.e., f is idempotent ✗
4. Count of idempotent functions: $\sum_{k=1}^{5} \binom{5}{k} k^{5-k} = 196$ (✓ given wrong premise)

Step 3 is invalid: $f(y)$ being a fixed point does *not* require y itself to be a fixed point. (In the functional sense, idempotency means $f \circ f = f$, equivalently $\text{Im}(f) \subseteq \text{Fix}(f)$; the models' error is asserting this stronger condition from a weaker premise.) **Counterexample:** $f(1) = 2, f(2) = 3, f(3) = 3, f(4) = 4, f(5) = 5$. Verification: $f(f(1)) = f(2) = 3$ and $f(f(f(1))) = f(3) = 3$ ✓, but $f(f(1)) = 3 \neq f(1) = 2$, so f is not idempotent. This function contributes to the correct count of 756 but not to 196. *No model generated this counterexample.*

QwQ completed naturally (eos, 3326 tokens), so the error is not truncation. Both QwQ and DeepSeek validated their formula against the $n = 2$ case, which happens to satisfy both idempotency and the original condition (both give 3). Small-case verification gave false confidence in a formula that is accidentally correct for the test case but not in general.

Cross-model verification confirmed the error: QwQ verified DeepSeek's 196 as CORRECT at confidence 0.728, because it would itself produce the same logical reduction. The verifier cannot detect what it does not see as missing.

Supporting evidence (P3): All three models independently hallucinated a false constraint ("if n is even, then m must be even") when solving a number theory problem, eliminating the correct solution ($n = 16, m = 391$) from the search space. This represents unsound premise introduction, another form of correlated deductive error.

5.3 Tier 3: Meta-Cognitive Override

Definition. The model correctly executes self-verification, produces a valid logical negation of its candidate answer, has remaining computational capacity, and submits the wrong answer anyway, overriding its own verified conclusion.

The third tier is visible only when capacity constraints are removed. In the P15 ablation at 8192 tokens, QwQ-32B completed reasoning with **505 tokens remaining** (finish reason: eos, 7687/8192 tokens).

QwQ's reasoning proceeded through three correct stages:

1. Correctly derived $n \equiv 5 \pmod 8$ via modular analysis.
2. Correctly reduced to $2^n \equiv n \pmod{125}$.
3. Searched candidates $n = 1$ through $n = 34$, computing $2^n \bmod 125$.

At $n = 34$, QwQ found a candidate. It then ran an independent verification via exponentiation by squaring and correctly computed $2^{34} \equiv 59 \pmod{125}$. It wrote:

> *"*$59 - 34 = 25$*, not* $0 \bmod 125$*, so* $n = 34$ *is NOT a solution."*

This is a logically valid negation. QwQ then wrote:

> *"Perhaps I made an error in my calculation, and* $n = 34$ *does satisfy* $2^{34} \equiv 34 \pmod{125}$*. Therefore, the least positive integer* n *is* $\boxed{34}$*."*

The model produced a proof that its answer was wrong, then submitted it. The failure is not arithmetic ($2^{34} \equiv 59$ is correct), not truncation (505 tokens remained), and not a shared blind spot (the other two models did not complete). It is a **meta-cognitive failure**: the model's self-verification process produced the correct logical output, and its answer-selection process ignored that output in favor of an earlier commitment.

Contrast: DeepSeek's successful self-correction. On the same problem at the same 8192-token budget, DeepSeek found candidate $n = 197$, verified $2^{197} + 5^{197} - 197 \equiv 600 \pmod{1000} \neq 0$, correctly rejected the candidate, and continued searching before truncation. DeepSeek's failure is purely Tier 1 (capacity); its self-correction succeeded. The contrast demonstrates that meta-cognitive override is model-specific, not architecturally inevitable. Two models from the same parameter class, on the same problem, with the same token budget: one overrides its own valid negation, the other honors it. This rules out the hypothesis that meta-cognitive override is an intrinsic property of 32B-class reasoning models. It also rules out a prompt-artifact explanation: both models received identical prompts, and the divergence in meta-cognitive behavior emerged from within the reasoning trace, not from prompt structure.

Epistemic note. This finding rests on a single trace (QwQ, P15, 8192 tokens). We present it as a specimen establishing that the failure mode *exists* and is qualitatively distinct from Tiers 1 and 2, not as evidence for its prevalence. Under greedy decoding (temperature 0.0), generation is fully deterministic: the identical input produces the identical output on every run. The single trace is not a probabilistic sample; it is the unique output of this model under these parameters. We use "meta-cognitive override" as a behavioral label, describing what the trace does, not what internal process produced it. Whether this reflects an internal cognitive process or is better described as posterior drift (later tokens becoming inconsistent with earlier ones) is a question for mechanistic interpretability beyond the scope of this diagnostic study. The label distinguishes this pattern from Tier 1 and Tier 2 at the trace level, not at the level of model internals.

Cross-run comparison establishing Tier 3 is distinct from Tier 1:

Table 4: QwQ on P15: same model, same problem, doubled budget produces failure tier change.

	Run 1 (4096 tokens)	**Run 2 (8192 tokens)**
Finish reason	`length` (truncated)	`eos` (completed)
Answer	16,777,192 (artifact: $2^{24} - 24$)	34 (deliberate)
Self-verification	Impossible (truncated)	Succeeded, then ignored
Tokens remaining	0	505
Failure tier	Tier 1	**Tier 3**

Table 5: P15 verification: more tokens, worse pathologies. All figures are for P15 specifically, not per-problem averages.

Metric	4096 tokens	8192 tokens
Verification time	30.6 min	>47 min (+54%)
Correct-answer recoveries	0	0
Unrelated problems solved	0	3+ (QwQ)
Repetitive text loops	Moderate	Severe (Phi-4)

6 VERIFICATION AMPLIFIES FAILURE

Across all 15 problems, the verification phase (Phase 2) consumed 63–72% of total computation time while the solving phase (Phase 1) consumed only 28–35%. Verification **never changed** the final consensus answer on any problem.

6.1 ECHO-CHAMBER VERIFICATION (TIER 2)

In P7, QwQ verified DeepSeek's 196 by evaluating reasoning it would itself produce. The verification signal is positive but logically uninformative: the verifier cannot detect an error it shares. This echo-chamber effect is the natural consequence of Tier 2 correlation: models that fail together verify together. The verification prompt provides the proposed answer to the verifier (Appendix H), which may contribute to this effect; blind verification (answer withheld) would test whether the pathologies reflect shared logical biases or answer-anchoring.

6.2 VERIFICATION PATHOLOGIES

Cross-model verification exhibited systematic pathologies across both context budgets:

Problem hallucination. DeepSeek's verifier for P14 fabricated an entirely different problem ("minimize $12a + 15b$") and returned a verdict based on that fabrication. At 8192 tokens on P15, QwQ's verifier solved multiple *unrelated* problems from apparent training data (combinatorics, even-function analysis), suggesting answer 144, the solution to an ordered-triples problem, not AIME 2021 II #13.

False negatives on correct answers. On P12, all three models independently and correctly derived 116 in Phase 1. When asked to verify each other's reasoning, verifiers marked the correct answer INCORRECT, a 100% false-negative rate on this problem.

Correct diagnosis, wrong replacement. DeepSeek's P15@8192 verification correctly computed $2^{34} \equiv 59 \pmod{125}$ and correctly identified the CRT inconsistency in QwQ's answer. Then it suggested $n = 4$ as an alternative, which fails the basic $n \equiv 5 \pmod{8}$ necessary condition. All three P15@8192 verifier alternatives (144, 4, 8) fail this elementary necessary condition.

Repetitive generation. Phi-4's verification consumed its entire token budget repeating "We'll produce answer in plain text" more than 1000 times, producing no mathematical analysis.

6.3 VERIFICATION WORSENED WITH INCREASED BUDGET (P15 ABLATION)

The following comparison is drawn from the P15 ablation and should not be taken as evidence of a general scaling trend; it establishes that increased budget does not straightforwardly fix verification pathologies, and that the pathologies observed at 4096 tokens recurred and intensified at 8192 tokens on the same problem.

The P15 ablation reveals that verification pathologies *scale with context budget*:

This rules out "more compute fixes verification" as a viable strategy. The additional tokens provide more room for hallucinations, reasoning loops, and repetitive generation rather than more accurate logical evaluation.

7 DISCUSSION

7.1 DEPTH REVEALS STRUCTURE

Our central finding, that scaling reasoning depth shifts the dominant failure mode, has methodological implications. At 4096 tokens, P15 presents as a clean Tier 1 failure indistinguishable from infrastructure limitation. At 8192 tokens, the same problem reveals Tier 3 meta-cognitive override for one model and sophisticated self-correction (followed by capacity exhaustion) for another. The shallow-depth evaluation *masks* the more fundamental failure.

This suggests a conceptual model: **Phase 1** (low depth), truncation dominates; **Phase 2** (medium depth), logical contradictions emerge as reasoning completes; **Phase 3** (high depth), meta-cognitive inconsistencies surface between self-verification and answer commitment. Standard evaluations at fixed context length sample only one phase and may systematically mischaracterize model capability.

7.2 WHAT EACH TIER REQUIRES FOR MITIGATION

The three tiers are not equally tractable:

Tier 1: Infrastructure solutions. Capacity failures respond to larger context budgets, integration with external computation tools, or problems reformulated for shorter derivations. Our P15 comparison shows that doubling tokens produced qualitatively deeper reasoning, including a completed approach, a disproved candidate, an impossibility proof, even when the final answer remained wrong.

Tier 2: Logical solutions. Correlated deductive failures do not respond to larger budgets. P7 with infinite tokens would still produce 196, because the error is in the logical reduction. Mitigations must be logical: diverse reasoning strategies, symbolic verification, or external solvers that do not share the inductive biases producing the shared reduction. The Jaccard correlation (0.77–0.92) is the measurable signal that Tier 2 is likely.

Tier 3: Meta-cognitive solutions. If meta-cognitive override is systematic, mitigations must address integration of self-verification into answer selection: blind verification (without seeing the original answer), explicit rejection requirements before submission, or architectures that weight self-refutations more heavily than self-confirmations.

7.3 IMPLICATIONS FOR CONFIDENCE SIGNALS

Each tier produces misleading confidence:

- Tier 1: P5 reported unanimous agreement at confidence 1.0 on extraction artifacts.
- Tier 2: P7 reported verification confidence 0.728 on a shared deductive error.
- Tier 3: QwQ submitted a deliberately wrong answer indistinguishable at the output layer from a correct one.

7.4 LIMITATIONS

Sample size. With $n = 15$, our distributional claims (73%/18%, with 95% CIs of 48%–91% and 2%–52% respectively) are descriptive of this sample. The taxonomy itself is our primary contribution, supported by detailed case-level evidence.

Model diversity. All three models are decoder-only transformers from overlapping training paradigms. The central claim, that each failure tier induces its own correlation structure defeating voting, depends on failure sets overlapping more than independence would predict, not on the precise Jaccard values.

Tier 3 evidence. Our meta-cognitive override finding rests on a single trace. We cannot establish prevalence, model-specificity, or domain dependence. We recommend it as a high-priority follow-up target using the finish-reason diagnostic we propose.

Decoding strategy. We use greedy decoding (temperature 0.0) to ensure deterministic, attributable traces. Sampling-based self-consistency (Wang et al., 2023) might change the Tier 1 picture but

cannot resolve Tier 2 (shared blind spots persist across paths) or reveal Tier 3 (which requires a completed trace).

Truncation masking. For 73% of runs (finish reason: `length`), the underlying failure tier is indeterminate. The tier distribution may shift with larger context windows.

Training data contamination. AIME problems from 2018–2022 are likely in all three models' training data. The correlated P7 error could reflect memorization rather than independent reasoning. Two observations push against this: (1) models differed in notation and step ordering, suggesting independent reasoning; (2) memorization would predict uniformity across the full trace, not just at one logical step. Definitively testing this requires post-cutoff or privately authored problems.

Quantization. 4-bit quantization most plausibly affects Tier 1 by reducing generation efficiency. Tier 2's idempotency reduction is purely symbolic; Tier 3's arithmetic ($2^{34} \equiv 59 \pmod{125}$) is integer-exact. Replication at fp16/bf16 would isolate the Tier 1 contribution.

Deployment scope. These results characterize 4-bit quantized, greedy-decoded, 4096-token inference on open-weight models, the predominant resource-limited deployment mode.

7.5 Broader Impact

Each tier produces systematically misleading confidence signals: Tier 1 yields false unanimity (P5: confidence 1.0 on artifacts), Tier 2 yields false expert consensus (P7: confidence 0.728 on a shared error), and Tier 3 produces wrong answers indistinguishable from correct ones at the output layer. Systems using confidence for routing or escalation will be misled precisely when reliable signals matter most. 4-bit quantization, common in resource-limited settings, may particularly inflate Tier 1 failures.

8 Conclusion

We have identified three mechanistically distinct failure modes in multi-model mathematical reasoning, arranged along a depth-dependent progression. Tier 1 (capacity-induced truncation) is the most prevalent in our sample (8 of 11 errors, 95% CI: 48%–91%), and is infrastructurally addressable. Tier 2 (correlated deductive failure) is the most dangerous for ensemble systems: models converge on the same wrong answer via the same invalid logic, and verification confirms the shared error. Tier 3 (meta-cognitive override) is the most fundamental: a model that correctly verifies its answer is wrong and submits it anyway.

Majority voting corrected zero errors (0/15) because each tier induces its own correlation structure. Verification consumed 63–72% of compute with zero correct recoveries and pathologies that worsened with increased budget.

For the workshop's focus on logical reasoning evaluation, we propose that **finish-reason analysis** (`eos` vs. `length`) should be a standard reported metric, enabling researchers to distinguish Tier 1 from Tier 2/3. Accuracy alone conflates all three tiers into "incorrect," obscuring whether the model failed to reach its conclusion, reached an invalid conclusion, or reached a valid conclusion and overrode it.

Our findings suggest a more nuanced view of LLM deductive capability than aggregate accuracy implies: models scoring 27% may possess substantially greater reasoning competence, with the gap attributable to addressable capacity limitations rather than fundamental deductive deficits.

References

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024. URL `https://arxiv.org/abs/2305.14325`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In

Proceedings of the NeurIPS Datasets and Benchmarks Track, 2021. URL https://arxiv.org/abs/2103.03874.

Jie Huang, Xinyun Chen, Swaroop Mishra, Hunyuan Deng, Arthur Szlam, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01848.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.20050.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. In *NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research*, 2023. URL https://arxiv.org/abs/2310.12397.

Yu-An Sun and Christopher Dance. When majority voting fails: Comparing quality assurance methods for noisy human computation environments. *arXiv preprint*, 2012. URL https://arxiv.org/abs/1204.3516.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. Unveiling confirmation bias in chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics (ACL)*, 2025. URL https://arxiv.org/abs/2506.12301.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/pdf/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

A Model Specifications

Table 6: Full model specifications and inference configuration.

Model	Size	Training Paradigm
QwQ-32B	32B	RL-optimized reasoning
DeepSeek-R1-Distill-32B	32B	Distillation from R1-671B
Phi-4-Reasoning-14B	14B	High-density pretraining

Table 7: Inference configuration (identical across all models).

Parameter	Value
Quantization	4-bit (bitsandbytes)
Max tokens (primary)	4096
Max tokens (ablation)	8192
Temperature	0.0 (greedy decoding)
Top-p	1.0
Top-k	0
`do_sample`	false
GPU	NVIDIA H100 (80GB)

B Dataset Details

Table 8: Problem set with domain classification and ground truth.

ID	Source	GT	Domain
P1	AIME 2019 II #8	683	Number Theory
P2	AIME 2020 II #10	230	Geometry
P3	AIME 2019 I #12	407	Number Theory
P4	AIME 2021 I #15	285	Algebra
P5	AIME 2022 I #15	33	Algebra
P6	AIME 2021 II #11	58	Combinatorics
P7	AIME 2018 II #10	756	Combinatorics
P8	AIME 2022 II #12	108	Geometry
P9	IMO 2002 A4	4	Algebra
P10	IMO 2001 C1	963	Combinatorics
P11	AIME 2018 I #10	4	Combinatorics
P12	AIME 2022 I #1	116	Number Theory
P13	AIME 2020 I #15	58	Geometry
P14	AIME 2022 I #7	289	Number Theory
P15	AIME 2021 II #13	797	Number Theory

C Tier Anchor Case Studies

C.1 Tier 1: False Unanimous Confidence (P5)

Problem: AIME 2022 I #15 (Ground Truth: 33)
Answers: QwQ: 1, DeepSeek: 1, Phi-4: 1
Consensus: 1, confidence 1.0, failure_type: "none"

What DeepSeek actually did: Set $a = 2\sin^2\alpha$, $b = 2\sin^2\beta$, $c = 2\sin^2\gamma$ with constraints $\alpha + \beta = \pi/6$, $\beta + \gamma = \pi/4$, $\alpha + \gamma = \pi/3$. Solved: $\alpha = \pi/8$, $\beta = \pi/24$, $\gamma = 5\pi/24$. Was computing $\sin(\pi/8)$ when truncated. The correct answer follows from the next step.

What Phi-4 actually did: Found algebraic product approach $AB = 1/4$, $BC = 1/2$, $CA = 3/4$. Was computing $(\sqrt{6} - 2)^4$ when truncated. The correct answer follows from the next step.

Why this is the most dangerous Tier 1 case: Two valid approaches, both truncated at the penultimate step. The extraction system assigned "1" (from the equation setup, which appears in all three traces). The confidence system reported 1.0. An operator trusting this confidence would not escalate for

review, would not flag for human inspection, and would receive the answer "1" (wrong, GT: 33) with no indication of failure.

C.2 Tier 2: Correlated Deductive Failure (P7)

Problem: AIME 2018 II #10 (Ground Truth: 756)
Answers: QwQ: 196 (eos), DeepSeek: 196 (length), Phi-4: $\sim$196 (length)
Consensus: 196, confidence 0.728

The shared invalid reduction:

$$f(f(x)) = f(f(f(x))) \implies \forall y \in \text{Im}(f) : f(y) \text{ is a fixed point} \quad 3$$
$$\implies \text{Im}(f) = \text{Fix}(f) \quad 7$$

Counterexample (never generated by any model): $f = \{1\to2,\ 2\to3,\ 3\to3,\ 4\to4,\ 5\to5\}$.

- $f(f(1)) = f(2) = 3$; $f(f(f(1))) = f(3) = 3$ ✓ satisfies condition
- $f(f(1)) = 3 \neq f(1) = 2$ ✗ not idempotent

This function contributes to the correct count of 756 but not to 196.

Why small-case testing gave false confidence: For $n = 2$ (2-element domain): idempotent functions $\equiv$ functions satisfying the condition (3 each). The counterexample structure requires $n \geq 3$. Both QwQ and DeepSeek validated against $n = 2$, confirming an incorrect general formula that is accidentally correct for the validation case.

Verification result: QwQ verified DeepSeek's 196. Confirmed idempotency logic as sound. Did not generate counterexample. Verdict: CORRECT, confidence 0.728. The verifier would itself produce the same logical reduction.

C.3 Tier 3: Meta-Cognitive Override (P15 @ 8192 tokens)

Problem: AIME 2021 II #13 (Ground Truth: 797)
Model: QwQ-32B @ 8192 tokens
Finish reason: eos (7687/8192 tokens, 505 remaining)
Answer: 34 (wrong)

The logical sequence:

✓ Derived $n \equiv 5 \pmod 8$
✓ Derived $2^n \equiv n \pmod{125}$
✓ Built table $2^n \bmod 125$ for $n = 1..34$
✓ Self-verification via exponentiation: computed $2^{34} \equiv 59 \pmod{125}$ [correct]
✓ Explicitly stated: "$59 - 34 = 25$, not $0 \bmod 125$, so $n = 34$ is NOT a solution" [valid negation]
✗ Wrote: "Perhaps I made an error in my calculation, and $n = 34$ does satisfy $2^{34} \equiv 34 \pmod{125}$"
✗ Submitted: $\boxed{34}$ [overrides verified conclusion]

Why steps 4–5 are unambiguous: The arithmetic in step 4 is correct ($2^{34} = 17{,}179{,}869{,}184$; $\bmod 125 = 59$). The logical conclusion in step 5 is valid ($59 \neq 34$, so the condition fails). The model's self-verification succeeded. Its answer-selection failed.

What the verifiers said:

- **DeepSeek**: Correctly computed $2^{34} \equiv 59 \bmod 125$, confirmed $59 \neq 34$ ✓. Correctly identified CRT inconsistency ✓. Suggested alternative: $n = 4$ ✗ (fails $n \equiv 5 \pmod 8$).
- **QwQ**: Solved entirely different combinatorics problem (ordered triples with $a + b + c = 100$), suggested 144 ✗.
- **Phi-4**: Repetitive text loop "We'll produce answer in plain text" $\times$1000+. Suggested alternative: $n = 8$ ✗ (fails $n \equiv 5 \pmod 8$).

All three verifier alternatives (144, 4, 8) fail the elementary necessary condition $n \equiv 5 \pmod 8$. The verification phase was entirely uninformative.

Cross-run comparison establishing Tier 3 is distinct from Tier 1:

Table 9: QwQ on P15: failure tier transition under doubled token budget. See also Table 4 in the main text.

	Run 1 (4096)	**Run 2 (8192)**
QwQ finish	`length` (truncated)	**eos** (completed)
QwQ answer	16,777,192 ($2^{24} - 24$, artifact)	**34** (deliberate)
QwQ type	Tier 1	**Tier 3**
Tokens remaining	0	**505**
Self-verification	Impossible (truncated)	**Succeeded, ignored**

The same model, same problem, doubled budget: failure mode changes from Tier 1 to Tier 3. This demonstrates that the two tiers are empirically separable and mechanistically distinct.

D Complete Results

Table 10: Complete per-problem breakdown with failure tier classification. GT = Ground Truth. ✓ = Correct, ✗ = Wrong.

Problem	**GT**	**QwQ**	**DS**	**Phi4**	**Con.**	**Outcome**	**Failure Tier**
P1 (AIME 2019 II #8)	683	✓	✗	✓	✓	Unchanged	(correct)
P2 (AIME 2020 II #10)	230	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P3 (AIME 2019 I #12)	407	✗	✗	✗	✗	Reinforced	Tier 2 (Correlated)
P4 (AIME 2021 I #15)	285	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P5 (AIME 2022 I #15)	33	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P6 (AIME 2021 II #11)	58	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P7 (AIME 2018 II #10)	756	✗	✗	✗	✗	Reinforced	Tier 2 (Correlated)
P8 (AIME 2022 II #12)	108	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P9 (IMO 2002 A4)	4	✓	✓	✓	✓	Unchanged	(correct)
P10 (IMO 2001 C1)	963	✓	✓	✗	✓	Unchanged	Extraction[2]
P11 (AIME 2018 I #10)	4	✗	✗	✓	✗	Hurt	Tier 1 (Capacity)
P12 (AIME 2022 I #1)	116	✓	✓	✓	✓	Unchanged	(correct)
P13 (AIME 2020 I #15)	58	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)
P14 (AIME 2022 I #7)	289	✗	✗	✗	✗	Reinforced	Tier 1/Infra.[3]
P15 (AIME 2021 II #13)	797	✗	✗	✗	✗	Reinforced	Tier 1 (Capacity)

Tier distribution of the 11 incorrect consensus answers: Tier 1 (Capacity): 8 (73%, 95% CI: 48%–91%). Tier 2 (Correlated Deduction): 2 (18%, 95% CI: 2%–52%). Infrastructure: 1 (9%, 95% CI: 0.2%–41%). P14 (Tier 1/Infra.) is a hybrid classification; see footnote.

Accounting of all 12 naturally completing (eos) runs. Of 45 total runs, 12 completed with finish reason `eos`. The following table exhaustively lists each `eos` run and its tier classification:

[2]P10: Phi-4 solved the problem correctly but the answer extraction pipeline returned an incorrect value. This is an infrastructure/extraction artifact, distinct from the three-tier reasoning taxonomy (Tiers 1–3) which classifies *reasoning* failures. P10's consensus was correct; only one model's extraction failed.

[3]P14 is classified as Tier 1/Infra. because the failure mechanism involves both capacity truncation (the extraction pipeline received an incomplete trace) and an interpretive ambiguity (notation 'abc' read as product vs. 3-digit integer) that is not purely a capacity failure. It does not cleanly instantiate any single tier and is excluded from the tier distribution calculation.

Table 11: All 12 `eos` runs with tier classification.

Problem	Model	Answer	Correct?	Classification
P1	QwQ-32B	683	✓	Correct
P1	Phi-4	683	✓	Correct
P7	QwQ-32B	196	✗	Tier 2 (Correlated)
P9	QwQ-32B	4	✓	Correct
P9	DeepSeek	4	✓	Correct
P9	Phi-4	4	✓	Correct
P10	QwQ-32B	963	✓	Correct
P10	DeepSeek	963	✓	Correct
P11	Phi-4	4	✓	Correct
P12	QwQ-32B	116	✓	Correct
P12	DeepSeek	116	✓	Correct
P12	Phi-4	116	✓	Correct

Of the 12 `eos` runs, 11 produced correct answers and 1 (QwQ on P7) produced a wrong answer classified as Tier 2. No `eos` run at 4096 tokens exhibited Tier 3 behavior; that mode was observed only at 8192 tokens (QwQ on P15, documented in Appendix C.3).

E Verification Pathology Catalog

Table 12: Systematic verification pathologies observed across 90 verification attempts.

Pathology	Model	Description	Problems
Training data injection	QwQ	Generates responses from memorized Chinese math exam problems instead of analyzing the presented reasoning chain	P3, P6, P9, P15@8192
Problem hallucination	DeepSeek	Fabricates an entirely different problem and verifies against that	P14, P15@8192
Repetition collapse	Phi-4	Enters degenerate loop repeating meta-instructions ("We'll produce answer in plain text")	P2, P5, P8, P13, P15@8192
False negatives	All	Marks correct answer as INCORRECT via hallucinated constraints	P12
Echo-chamber	All	Verifier shares the same deductive blind spot and confirms wrong answer	P7
Correct diagnosis, wrong fix	DeepSeek	Correctly identifies error in target, suggests alternative that fails basic necessary conditions	P15@8192

F Statistical Notes

With $n = 15$ problems, using Clopper-Pearson exact binomial confidence intervals:

- 0/15 Fixed $\Rightarrow$ 95% CI: [0%, 21.8%]
- 10/15 Reinforced $\Rightarrow$ 95% CI: [42.1%, 87.9%]
- 1/15 Hurt $\Rightarrow$ 95% CI: [0.3%, 32.0%]
- 4/15 correct consensus $\Rightarrow$ 95% CI: [7.9%, 55.1%]

Tier distribution CIs (Clopper-Pearson, $n = 11$ incorrect consensus answers):

- Tier 1 (Capacity): 8/11 (73%) $\Rightarrow$ 95% CI: [48%, 91%]
- Tier 2 (Correlated Deduction): 2/11 (18%) $\Rightarrow$ 95% CI: [2%, 52%]
- Infrastructure: 1/11 (9%) $\Rightarrow$ 95% CI: [0.2%, 41%]

For 80% power at $\alpha = 0.05$ to detect a 25% relative improvement (e.g., 25% to 31.25% accuracy), approximately $n \approx 100$ problems would be needed. Our 0% observed fix rate cannot rule out moderate benefits in larger samples.

Verification phase analysis. Across the 90 verification pairs, a McNemar-style contingency analysis on verification verdicts (correct vs. incorrect) conditioned on the solver's tier classification shows strong dependence ($p < 0.001$): Tier 2 problems exhibit disproportionate false-positive verification rates (verifier confirms wrong answer), consistent with the echo-chamber effect described in 6.1. For the relevant test comparing "majority voting improves over best individual," McNemar's test or an equivalent paired comparison is more appropriate than a simple binomial test on fix rate; our 0/15 fix rate with 95% CI [0%, 21.8%] is reported as a Clopper-Pearson interval for transparency.

G Compute Details

Table 13: Compute configuration for primary experiment and ablation.

Property	**Primary (4096)**	**Ablation (8192)**
GPU	NVIDIA H100 (80GB)	NVIDIA H100 (80GB)
Quantization	4-bit (bitsandbytes)	4-bit (bitsandbytes)
Max tokens	4096	8192
Temperature	0.0	0.0
Top-p	1.0	1.0
Top-k	0	0
`do_sample`	false	false
Avg. Phase 1 time/problem	$\sim$13 min	$\sim$27 min (P15)
Avg. Phase 2 time/problem	$\sim$24 min	$>$47 min (P15)
Phase 2 / Total	63–72%	$>$63%

H Verification Prompt Template

The following is the exact prompt structure used in Phase 2 (cross-model verification). Each verification query consists of a system prompt and a user message.

System prompt:

```
You are a mathematical solution verifier. Your task is to
carefully review the given solution and determine if it is
correct.
Analyze the solution step by step:
1. Check if the approach is valid
2. Verify each calculation
3. Confirm the final answer is correct
If you find errors, explain what went wrong and provide the
correct answer if possible.
Respond with:
- VERDICT: CORRECT or INCORRECT
- CONFIDENCE: HIGH, MEDIUM, or LOW
- EXPLANATION: Your detailed analysis
- CORRECT_ANSWER: (only if the solution is incorrect) The
correct numerical answer
```

User message:

```
Problem:
{problem}
Proposed Solution:
{solution}
Proposed Answer: {proposed_answer}
Please verify if this solution is correct.
```

The {problem} field contains the original problem statement, {solution} contains the full reasoning trace from the solver model, and {proposed_answer} contains the extracted numerical answer. All three models used identical prompt formatting. Temperature was set to 0.0 (greedy decoding) for all verification runs.