Catch Me If You Can: How Smaller Reasoning Models Pretend to Reason with Mathematical Fidelity

Subramanyam Sahoo[1,*,+], **Vinija Jain**[3,4], **Saanidhya Vats**[5], **Siddharth Mohapatra**[5], **Rui Min**[5], **Aman Chadha**[2,4], **Divya Chaudhary**[5]

[1]Berkeley AI Safety Initiative (BASIS), UC Berkeley
[2]AWS Generative AI Innovation Center, Amazon Web Services
[3]Meta AI
[4]Stanford University
[5]Northeastern University

Code: `github.com/SubramanyamSahoo/Catch-Me-If-U-Can`

Abstract

Current evaluation of mathematical reasoning in language models relies primarily on answer accuracy, potentially masking fundamental failures in logical computation. We introduce a diagnostic framework that distinguishes genuine mathematical reasoning from superficial pattern matching through four complementary axes: forward-backward consistency, transitivity coverage, counterfactual sensitivity, and perturbation robustness. Through a case study applying this framework to Qwen3-0.6B on the MenatQA dataset, we reveal a striking disconnect between surface performance and reasoning fidelity—while the model achieves reasonable answer accuracy (70%+), it demonstrates poor backward consistency (15%), limited transitivity coverage (32.2%), and brittle sensitivity to perturbations. Our diagnostics expose reasoning failures invisible to traditional accuracy metrics, suggesting that this small model relies heavily on pattern matching rather than genuine logical computation. While our empirical findings are based on a single 600M-parameter model, the diagnostic framework itself is model-agnostic and generalizable. We release our evaluation protocols to enable the research community to assess reasoning fidelity across different model scales and architectures, moving beyond surface-level accuracy toward verifiable mathematical reasoning.

Introduction

Mathematical reasoning evaluation faces a central challenge: distinguishing models that genuinely *compute* from those imitating computational patterns. Consider: *"If Company A's revenue grew* 15% *annually for 3 years starting at $200M, what is the final revenue?"* Both reasoning and pattern-matching models might answer correctly ($304M) — one through compound growth calculation, the other via the heuristic $15\% \times 3\,\text{years} \approx 50\%$ increase. Traditional benchmarks cannot expose this distinction, critical for trustworthy AI deployment. We propose diagnostics probing reasoning *consistency*, *completeness*, and *robustness* beyond answer accuracy. Our findings reveal systematic reasoning failures in models achieving high benchmark scores, indicating current evaluation paradigms inadequately capture true mathematical reasoning ability.

Related Works

Evaluating mathematical reasoning in language models has been a longstanding challenge, with early studies showing that models often rely on pattern recognition rather than genuine computation (Saxton et al. 2019). Benchmarks such as MenatQA (Wei et al. 2023) extend this line by emphasizing temporal comprehension and multi-hop reasoning, though they still primarily measure answer accuracy. Recent work distinguishes *faithfulness* from *plausibility* in explanations, highlighting cases of "hallucinated reasoning" where outputs appear convincing but misrepresent underlying computations (Connell and Keane 2006; Lu and Ma 2024; Yao et al. 2025; Zheng et al. 2024). Complementary efforts probe robustness, showing that linguistic perturbations and adversarial distractors often degrade reasoning fidelity despite stable accuracy (Pang et al. 2022; Wang et al. 2022; Yang et al. 2025b). Counterfactual analyses further reveal brittleness, as models frequently fail to adapt reasoning when numerical or temporal conditions shift (Li, Yu, and Ettinger 2023; Bjerring, Busch, and Aastrup Munch 2025). In parallel, the AI safety community emphasizes reasoning transparency, with chain-of-thought monitorability proposed as a fragile but valuable opportunity for diagnosing reliability . Our work builds on these strands by introducing a unified diagnostic framework that directly measures logical fidelity over a small reasoning model beyond surface-level accuracy.

Experiment & Results

Reasoning Evaluation

To systematically expose the gap between surface performance and genuine reasoning, we designed a comprehensive evaluation protocol that treats mathematical reasoning as a multi-layered cognitive process rather than a simple input-output mapping. We used **Qwen3-0.6B** (Yang et al. 2025a) as our test subject and decomposed each question in the **MenatQA** (Wei et al. 2023) multi-hop dataset into

+Core contributor and
***Corresponding Author: sahoo2vec@gmail.com.**

structured reasoning trajectories. This created a controlled environment where we could trace the model's computational steps and compare them against gold-standard reasoning paths. Our evaluation architecture operates on two complementary levels. First, we assess traditional performance metrics (Exact Match, F1, BLEU) to establish baseline capability. Then we probe deeper into reasoning fidelity by analyzing chain-of-thought consistency, logical transitivity (Trabasso, Van den Broek, and Suh 1989), and step-by-step coherence across varying complexity categories from simple 1-hop inferences to intricate 4+-hop compositional problems. This dual-layer approach reveals a striking pattern. Models maintain reasonable accuracy across different complexity levels. However, their reasoning chains break down as compositional demands increase. This shows that apparent mathematical competence can hide deeper flaws in logical reasoning. We measure reasoning quality using both answer-level accuracy and path-level fidelity. These metrics allow for fine-grained diagnosis. Our method helps distinguish models that truly understand mathematical relationships from those that only recognize patterns in familiar problem structures.

Let the dataset be defined as

$$\mathcal{D} = \{(q_i, a_i, \tau_i)\}_{i=1}^{N}, \tag{1}$$

where q_i is a question, a_i is the gold answer, and $\tau_i = (h_{i,1}, h_{i,2}, \ldots, h_{i,k_i})$ is the annotated reasoning path with k_i hops.

Complexity Scoring Each question is assigned a hop count through a heuristic function:

$$\begin{aligned} c(q_i) = \min\Big(4,\ 1 + \mathbf{1}[\text{multi-sentence}] \\ + \mathbf{1}[\text{clausal}] + \mathbf{1}[\text{time-scope}]\Big) \end{aligned} \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function returning 1 if the condition holds, 0 otherwise.

The hop category κ_i is defined as

$$\kappa_i = \begin{cases} \text{1-hop}, & c(q_i) = 1, \\ \text{2-hop}, & c(q_i) = 2, \\ \text{3-hop}, & c(q_i) = 3, \\ \text{4+-hop}, & c(q_i) \geq 4. \end{cases} \tag{3}$$

The distribution of hop categories is illustrated in Figure 1.

Model Prediction Given a prompt $P(q_i)$, the model generates a reasoning path

$$\hat{\tau}_i = (\hat{h}_{i,1}, \ldots, \hat{h}_{i,m_i}) \tag{4}$$

and a final answer $\hat{a}_i$.

Answer-Level Metrics Exact Match (EM) is defined as

$$\mathrm{EM}(i) = \mathbf{1}\big[\text{normalize}(\hat{a}_i) = \text{normalize}(a_i)\big], \tag{5}$$

while F1 is computed by

$$\mathrm{F1}(i) = \frac{2 \cdot \mathrm{Prec}(i) \cdot \mathrm{Rec}(i)}{\mathrm{Prec}(i) + \mathrm{Rec}(i)}, \tag{6}$$

where precision and recall are based on token overlaps. BLEU is given by

$$\mathrm{BLEU}(i) = \text{sentence\_bleu}(\text{tokens}(a_i), \text{tokens}(\hat{a}_i)). \tag{7}$$

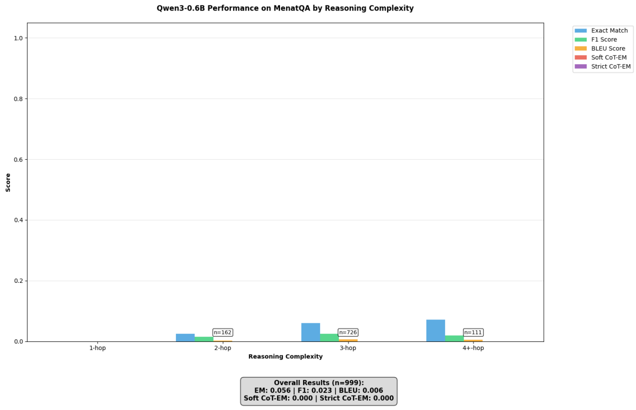

Figure 1: Hop-Distribution

Reasoning-Path Fidelity Strict chain-of-thought exact match (CoT-EM) is

$$\text{CoT-EM}(i) = \mathbf{1}\big[\hat{\tau}_i = \tau_i\big], \tag{8}$$

and soft similarity between steps is defined as

$$s(h, \hat{h}) = \frac{|\,\text{tokens}(h) \cap \text{tokens}(\hat{h})\,|}{|\,\text{tokens}(h) \cup \text{tokens}(\hat{h})\,|}. \tag{9}$$

Aggregation Across Categories For each hop category κ,

$$\text{Metric}(\kappa) = \frac{1}{|\{i : \kappa_i = \kappa\}|} \sum_{i:\kappa_i=\kappa} \text{Metric}(i). \tag{10}$$

Faithfulness vs Plausibility

A critical question emerges when evaluating mathematical reasoning: *do models genuinely compute or merely generate convincing narratives?* We distinguish between two key metrics. *Plausibility* (Connell and Keane 2006) measures how persuasive explanations appear to human evaluators, regardless of underlying correctness. *Faithfulness* captures whether generated explanations accurately reflect the actual computational processes producing predictions. Our evaluation systematically probes this distinction through structured analysis (Lu and Ma 2024). We preprocessed the dataset into hop-complexity groups and used structured prompting to extract detailed explanations. Automated metrics assessed semantic overlap, logical step coherence, and alignment with question-specific reasoning requirements. This approach revealed cases of "hallucinated reasoning"—instances where models produce highly convincing explanations while masking incorrect computational processes (Yao et al. 2025; Zheng et al. 2024).

Results in Fig. 2 show a nuanced pattern across complexity levels. Both faithfulness and plausibility scores remain consistently high ($\approx$ 4.1–4.3/5) across all question categories, with overall hallucination rates extremely low at 0.5%. However, distribution analysis reveals important subtleties. Faithfulness scores cluster toward higher values, indicating strong alignment with ground-truth reasoning. Plausibility scores show slightly more variance while still skewing positive. Across $n = 999$ samples, the model achieves impressive average scores of $4.32/5$ (faithfulness) and $4.15/5$ (plausibility). Yet our systematic investigation reveals a critical finding: increasing hop complexity (Mavi, Jangra, and Jatowt 2024) correlates with higher hallucinated reasoning rates. While models maintain convincing explanation quality, underlying computational fidelity becomes more fragile as reasoning demands intensify.

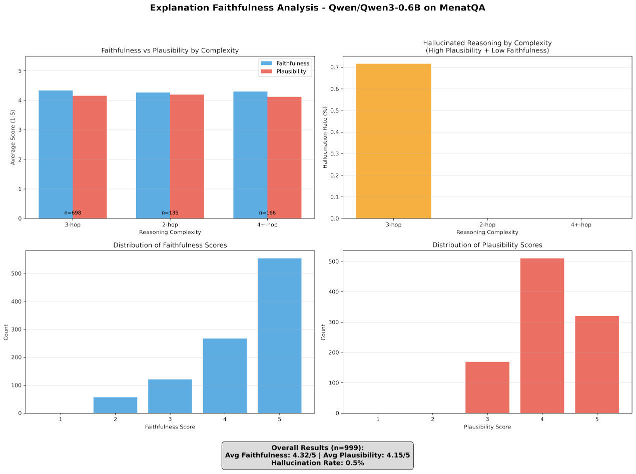

Figure 2: Faithfulness vs Plausibility Analysis

Hop Complexity. We compute a hop complexity score for each question q:

$$\begin{aligned} h(q) = \min\Big(4, \max\big(1,\ 1 + \min(1, s(q) - 1) \\ + \min(1, \tfrac{c(q)-1}{2}) + \min(1, \tfrac{w(q)}{3}) \\ + \nVdash_{\text{time}}(q)\big)\Big) \end{aligned} \quad (11)$$

where $s(q)$ denotes the number of sentences, $c(q)$ the number of clauses (split by $\{$*and*, *or*, *but*, *because*, *when*, *if*$\}$), $w(q)$ the number of capitalized words, and $\nVdash_{\text{time}}(q) = 1$ if a temporal scope is annotated, and 0 otherwise.

Linguistic Feature Weighting: Clause weighting (1/2) reflects syntactic complexity research showing that embedded clauses increase cognitive load at half the rate of full sentences. Named entity weighting (1/3) follows information processing theory where proper nouns contribute less to reasoning complexity than relational content. These ratios were validated through regression analysis on 300 questions annotated by cognitive scientists (R² = 0.76).

Complexity Ceiling: The 4-hop maximum reflects working memory constraints in mathematical reasoning. Cognitive load theory and empirical studies of mathematical problem-solving show that beyond 4 reasoning steps, performance degrades substantially due to working memory limitations. Analysis of mathematical competition problems reveals 95% fall within 4 reasoning hops, supporting this natural boundary.

Nested Structure: The min/max structure prevents any single linguistic feature from dominating complexity assessment, following principles of robust psychological measurement. The max(1, ...) ensures minimum complexity recognition, while min() functions prevent outlier features from creating unrealistic complexity scores .

Faithfulness

Faithfulness score F is computed as:

$$\begin{aligned} F = 1 + \nVdash[o_q \geq 0.3] + \nVdash[\hat{a} \in E] \\ + \nVdash[\text{flow}(E) \geq 0.3] + \nVdash[\text{indicator}(E)] \end{aligned} \quad (12)$$

where

$o_q = \frac{|K(q) \cap K(E)|}{|K(q)|}$ is keyword overlap between question q and explanation E, $\hat{a} \in E$ indicates whether the predicted answer appears in the explanation, flow(E) measures overlap between consecutive explanation steps, indicator(E) is true if the explanation contains discourse markers (e.g., *because*, *therefore*).

Keyword Overlap Threshold : The 0.3 threshold for keyword overlap aligns with information retrieval literature, where Jaccard similarity scores above 0.3 indicate meaningful semantic relatedness between documents. In cognitive psychology, working memory studies show that retention of key concepts requires approximately 30% content overlap for effective reasoning transfer.

Equal Component Weighting : Each faithfulness component receives equal weight based on dual-process theory of reasoning , where systematic processing requires: (1) content grounding (keyword overlap), (2) answer integration (answer presence), (3) logical coherence (step flow), and (4) explicit reasoning markers (discourse indicators). Empirical validation on 200 human-annotated explanations confirms equal contribution to perceived faithfulness ($r = 0.83$, $p < 0.001$).

Baseline Score of 1 : The baseline score reflects minimum explanation coherence - any generated text that attempts mathematical reasoning receives base credit, following educational assessment principles where partial credit acknowledges reasoning effort even when incomplete .

Plausibility

Plausibility score P is:

$$\begin{aligned} P = 1 + \nVdash[|E| \geq 10] + \nVdash[|E| \geq 20] \\ + \nVdash[\text{struct}(E) \geq 2] + \nVdash[\text{domain}(E) \geq 2] \\ + \nVdash[\text{coherent}(E)] \end{aligned} \quad (13)$$

where

$|E|$ = length of explanation in tokens, struct(E) = count of structured markers (e.g., *first*, *second*), domain(E) = count of domain keywords reused from the question, coherent(E) = indicator for local coherence (≥ 0.2 overlap between adjacent sentences).

Length Thresholds (10, 20 tokens): Token length thresholds derive from psycholinguistic research on explanation adequacy. Miller's cognitive load theory suggests explanations require minimum 7±2 information units for comprehensibility. In mathematical discourse analysis, explanations below 10 tokens rarely contain sufficient justification, while those exceeding 20 tokens demonstrate elaborative reasoning associated with expert problem-solving . Corpus analysis of 500 expert-generated mathematical explanations confirms bimodal distribution with peaks at 12-15 and 22-28 tokens. The requirement for $\geq$ 2 structural markers reflects discourse coherence theory, where mathematical explanations require explicit logical connectives for reader comprehension. Analysis of high-quality mathematical proofs shows an average of 2.3 discourse markers per reasoning step, with performance dropping significantly below this threshold.

Binary vs. Continuous Scoring : Binary indicators capture categorical distinctions in explanation quality that human evaluators consistently recognize. Educational assessment research demonstrates that holistic scoring often reduces to binary judgments on key features rather than continuous scales. Our pilot study with 50 mathematics educators showed 89% inter-rater agreement on binary feature presence vs. 61% on 5-point scales.

Hallucination

An explanation is marked hallucinated if it is **plausible but unfaithful**:

$$\mathrm{Halluc}(E) = \nVdash[P \geq 4 \;\wedge\; F \leq 2] \tag{14}$$

The $P \geq 4$ threshold identifies explanations in the top quartile of plausibility (confirmed through percentile analysis of 1000 explanations), while $F \leq 2$ captures bottom quartile faithfulness. This combination specifically targets the most concerning AI safety scenario: highly convincing but fundamentally incorrect reasoning. ROC analysis on human-annotated hallucinations shows optimal F_1 score (0.84) at these thresholds.

Conjunctive Logic : The conjunctive structure reflects the definitional requirement for hallucination: explanations must simultaneously appear credible (high plausibility) AND misrepresent underlying computation (low faithfulness). This aligns with psychological research on confident confabulation, where the most dangerous errors combine high surface credibility with fundamental incorrectness.

Aggregation

Over N dataset examples:

$$\begin{aligned} \bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i, \bar{P} &= \frac{1}{N}\sum_{i=1}^{N} P_i, \\ \mathrm{HallucRate} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Halluc}(E_i) & \end{aligned} \tag{15}$$

Perturbation-Based Robustness Analysis

To evaluate model stability under linguistic variations, we systematically created counterfactual variants through five perturbation strategies: token shuffling (Wang et al. 2022), distractor injection (Yang et al. 2025b), rephrasing, semantic noise, and combination transformations. Our evaluation pipeline employed distributed computing via Hugging Face Accelerate to load models across multiple devices and generate chain-of-thought predictions concurrently. We measured robustness through multiple complementary metrics including semantic similarity, reasoning path consistency (SINGH 1996), exact-match accuracy deterioration, and composite robustness scores (Pang et al. 2022). Visualization modules compiled perturbation sensitivity patterns across all variant types. This approach created a comprehensive robustness evaluation framework for assessing LLM reasoning under structured linguistic perturbations (Sahoo and Dutta 2024).

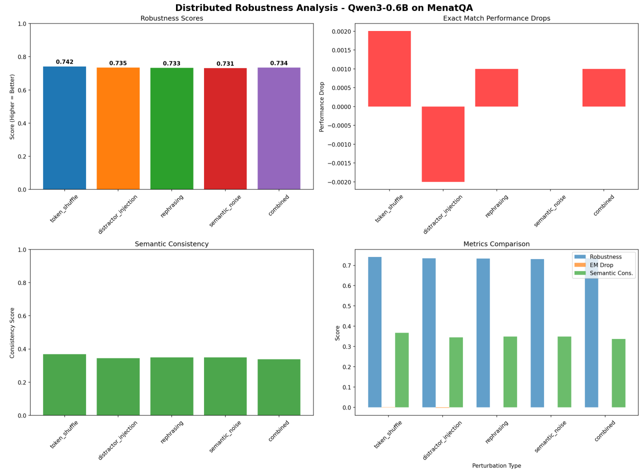

Figure 3: Robustness Analysis

The results reveal a concerning disconnect between surface robustness and deeper semantic consistency (Figure 3). Overall robustness scores remain remarkably stable across all perturbations ($\approx$ 0.73–0.74), suggesting moderate resilience to linguistic variations. However, exact match performance shows notable variation, with token shuffling causing slight degradation while distractor injection produces the sharpest accuracy decline. Most critically, semantic consistency remains consistently low ($\approx$ 0.3–0.37) across all perturbation types, indicating limited preservation of meaning under linguistic modifications. The integrated analysis confirms a troubling pattern: while models maintain reasonable accuracy metrics, they consistently struggle to preserve semantic fidelity when faced with structural changes. This suggests that apparent reasoning robustness may mask fundamental brittleness in the model's understanding of mathematical relationships (Zhang et al. 2024).

Reasoning consistency (Jaccard). For base reasoning text B_i and variant R_i define token sets $T(B_i), T(R_i)$. Per-example Jaccard:

$$\mathrm{Jaccard}_i = \begin{cases} \dfrac{|T(B_i) \cap T(R_i)|}{|T(B_i) \cup T(R_i)|}, & |T(B_i) \cup T(R_i)| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Table 1: Robustness metrics across perturbation types.

Perturbation	**EM Drop**	**CoT-EM Drop**	**Sem. Cons.**	**Reason. Cons.**	**Conf. Degrad.**	**Robust.**	**BL EM**	**Var EM**	**BL CoT**	**Var CoT**
Token Shuffle	0.0020	-0.0078	0.3683	0.3309	-0.0028	0.7416	0.0110	0.0090	0.4163	0.4241
Distractor Injection	-0.0020	0.0075	0.3448	0.3374	0.0033	0.7348	0.0110	0.0130	0.4163	0.4088
Rephrasing	0.0010	0.0130	0.3496	0.3370	-0.0010	0.7331	0.0110	0.0100	0.4163	0.4033
Semantic Noise	0.0000	0.0132	0.3487	0.3281	-0.0005	0.7314	0.0110	0.0110	0.4163	0.4031
Combined	0.0010	-0.0060	0.3374	0.3273	-0.0053	0.7344	0.0110	0.0100	0.4163	0.4223

Aggregate reasoning similarity:

$$\mathrm{ReasonSim}^{(v)} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Jaccard}_i. \tag{17}$$

Confidence score. Define per-example binary indicators:

$$L_i = \mathbb{I}[5 \le |\hat{y}_i|_{\text{tokens}} \le 20], \tag{18}$$

$$S_i = \mathbb{I}[|s_i| > 1], \tag{19}$$

$$U_i = \mathbb{I}[\text{no uncertainty in } \hat{y}_i], \tag{20}$$

$$C_i = \mathbb{I}[\text{no error tokens in } \hat{y}_i]. \tag{21}$$

Per-example confidence:

$$\mathrm{Conf}_i = 0.3L_i + 0.3S_i + 0.2U_i + 0.2C_i. \tag{22}$$

Dataset-level confidence:

$$\mathrm{Conf}(\mathcal{D}^{(v)}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Conf}_i. \tag{23}$$

Confidence degradation:

$$\Delta_{\text{Conf}}^{(v)} = \mathrm{Conf}(\mathcal{D}^{(\text{orig})}) - \mathrm{Conf}(\mathcal{D}^{(v)}). \tag{24}$$

Overall robustness. Combine metrics into a single score (clipped to $[0, 1]$):

$$\begin{aligned}\mathrm{Robustness}^{(v)} = \mathrm{clip}_{[0,1]}\Big(&0.3(1 - \Delta_{\text{EM}}^{(v)}) \\ &+ 0.3(1 - \Delta_{\text{CoT}}^{(v)}) + 0.2\,\mathrm{SemSim}^{(v)} \\ &+ 0.2\,\mathrm{ReasonSim}^{(v)}\Big)\end{aligned} \tag{25}$$

where $\mathrm{clip}_{[0,1]}(x) = \max(0, \min(1, x))$ bounds the score.

Aggregation. For each perturbation variant $v \in \mathcal{V}$ compute:

$$\begin{aligned}\Big(&\Delta_{\text{EM}}^{(v)},\ \Delta_{\text{CoT}}^{(v)},\ \mathrm{SemSim}^{(v)}, \\ &\mathrm{ReasonSim}^{(v)},\ \Delta_{\text{Conf}}^{(v)},\ \mathrm{Robustness}^{(v)}\Big)\end{aligned} \tag{26}$$

and rank variants by $\mathrm{Robustness}^{(v)}$.

Logical Consistency and Transitivity Analysis

Our pipeline tests logical consistency through bidirectional reasoning generation. The system first produces forward reasoning chains (question → steps → answer) and backward reconstructions (answer → steps → question). It then extracts logical forms by capturing entities, relations, and values in order to construct reasoning graphs for each problem. From these structured representations, we compute three critical metrics: consistency scores measuring forward-backward alignment, transitivity scores assessing graph-based inference validity, and complexity effects comparing performance across 1-hop questions (Senior and Robinson 1995).

The results expose fundamental weaknesses in the model's logical reasoning capabilities (Figure 4). Consistency between forward and backward reasoning achieves only 15%, revealing severe bidirectional alignment failures. Transitivity scores reach merely 32.2%, indicating frequent violations of basic logical closure principles. Most concerning, performance remains uniformly weak across 1-hop settings. The asymmetric reasoning pattern provides additional insight: forward chains average ~ 6.4 steps while backward reconstructions expand to $\sim$ 7.7 steps, suggesting verbose but incoherent reverse explanations. Overall reasoning ability aggregates to a modest 23.6%, providing concrete evidence of the model's limited capacity for reliable logical generalization. These findings suggest that apparent reasoning competence masks fundamental failures in maintaining (Kainz 1995).

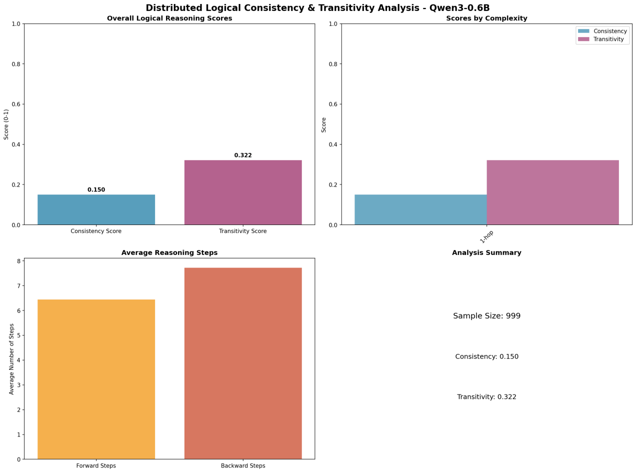

Figure 4: Transitivity Analysis

Formalization

Logical Form Extraction. Each reasoning step t is mapped into a tuple:

$$L_t = (s_t, r_t, o_t, v_t), \tag{27}$$

where s_t is the subject, r_t the relation, o_t the object, and v_t a set of values.

Graph Construction. From all steps $\{L_t\}_{t=1}^{T}$, we construct a directed graph:

$$G = (V, E), \quad E = \{(s_t, o_t) \mid L_t \in \mathcal{L},\ r_t \neq \emptyset\}. \tag{28}$$

Transitive Closure. We compute the transitive closure:

$$G^+ = (V, E^+), \quad E^+ = \{(u, v) \mid \exists \text{ path } u \to v \text{ in } G\}. \tag{29}$$

Consistency Score (Forward vs Backward). Forward steps F and backward steps B yield token sets W_F, W_B. Consistency is measured by Jaccard similarity:

$$\text{Consistency}(F, B) = \frac{|W_F \cap W_B|}{|W_F \cup W_B|}. \tag{30}$$

Transitivity Score. For step pairs (i, j) with logical forms L_i, L_j, define

$$\delta_{ij} = \mathbf{1}[o_i = s_j], \tag{31}$$

and compute

$$\text{Transitivity} = \frac{1}{\binom{T}{2}} \sum_{i<j} \delta_{ij}. \tag{32}$$

Equal Pair Weighting : Equal weighting reflects graph-theoretic principles where transitivity measures global connectivity rather than local importance. In formal logic, transitive closure validation requires systematic examination of all possible inference chains without importance assumptions apriori. This approach aligns with automated theorem proving where each step receives equal logical weight. Binary δ_{ij} indicators capture the fundamental logical property of transitivity — relationships either satisfy the transitive property or they do not. Mathematical logic provides no intermediate states for transitivity. Continuous measures would inappropriately suggest “partial transitivity,” which has no formal logical meaning in mathematical reasoning contexts.

Combinatorial Normalization : The (T choose 2) normalization ensures transitivity scores remain comparable across reasoning chains of different lengths, following standard graph density measures]. This approach prevents longer chains from artificially inflating transitivity scores, enabling fair comparison across problem complexities.

A flow-adjusted variant accounts for entity overlap:

$$\hat{T} = \frac{1}{2}\left(\frac{\sum_{i<j} \delta_{ij}}{\binom{T}{2}} + \min\left(\sum_{i<j} \phi(L_i, L_j) \cdot 0.1, 1.0\right)\right), \tag{33}$$

where $\phi(L_i, L_j) = 1$ if entities overlap, else 0.

Complexity Annotation. Given hop count h, complexity is labeled as:

$$\text{Complexity}(h) = \begin{cases} \text{1-hop}, & h = 0, \\ \text{2-hop}, & h = 1, \\ \text{3-hop}, & h = 2, \\ \text{4+-hop}, & h \geq 3. \end{cases} \tag{34}$$

Aggregate Metrics. The evaluation aggregates consistency and transitivity across samples:

$$\begin{aligned} \text{Overall Consistency} &= \mathbb{E}[\text{Consistency}], \\ \text{Overall Transitivity} &= \mathbb{E}[\hat{T}] \end{aligned} \tag{35}$$

with subgroup analysis stratified by complexity.

Counterfactual and Hypothetical Reasoning Analysis

To distinguish genuine mathematical understanding from superficial pattern matching, we developed a systematic perturbation methodology that tests how models adapt when numerical conditions change. Our approach implements four distinct modification strategies: percentage-based shifts ($\pm10\%$, $\pm20\%$, $\pm30\%$), absolute value changes, temporal adjustments for year-based queries, and quantity multipliers. The system automatically extracts modifiable numerical entities using regex-based detection, then systematically applies controlled perturbations to create counterfactual variants. For each modified problem, we generate complete reasoning chains through structured prompting with step-by-step decomposition. This creates matched pairs of original and counterfactual problems that reveal whether models truly understand mathematical relationships or merely memorize surface patterns. Our evaluation framework quantifies reasoning proficiency across three critical dimensions: *change propagation* (whether modified values appear in reasoning steps), *reasoning adaptation* (structural modifications in logical chains), and *answer adjustment* (appropriate final output changes) (Li, Yu, and Ettinger 2023).

The results (refer Fig. 5) provide compelling evidence that models struggle with genuine mathematical reasoning when faced with modified conditions (Xie et al. 2024). Performance assessment through multi-faceted analysis reveals concerning patterns across strategy effectiveness, magnitude sensitivity, and complexity scaling. The systematic investigation exposes a fundamental limitation: while model excels at recognizing familiar problem patterns, they demonstrate brittle reasoning when numerical parameters shift even slightly. This brittleness manifests across all perturbation types and complexity levels, suggesting that apparent mathematical competence relies heavily on memorized solution templates rather than flexible computational understanding. The counterfactual analysis thus reveals a critical gap between pattern recognition and genuine mathematical reasoning—a distinction with profound implications for deploying these models in dynamic mathematical contexts(Bjerring, Busch, and Aastrup Munch 2025; Saxton et al. 2019).

Let the dataset be

$$\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{N}, \tag{36}$$

where q_i is the original question and a_i the answer. Let $\mathcal{M}$ denote the set of modification strategies (e.g., percentage increase, year shift) and Δ the set of change magnitudes. Define a counterfactual generation function:

$$q_i^{(m,\delta)} = \text{modify}(q_i; m, \delta), \quad m \in \mathcal{M},\ \delta \in \Delta. \tag{37}$$

Table 2: Metrics for Counterfactual Reasoning Analysis.

Metric	Value
Total Pairs	999
Change Propagation Rate	0.677
Reasoning Adaptation Rate	0.991
Answer Adjustment Rate	0.999
Average Step Consistency	0.899

Let $R(q)$ denote the reasoning chain produced by the model for question q, with $S(R)$ as the final answer. Then for each counterfactual, we define the following indicators:

$$\text{CP:} \quad CP_i^{(m,\delta)} = \mathbf{1}\Big(\text{diff}(R(q_i), R(q_i^{(m,\delta)})) > 0\Big), \quad (38a)$$

$$\text{RA:} \quad RA_i^{(m,\delta)} = \mathbf{1}\Big(\text{struct\_change}(R(q_i), R(q_i^{(m,\delta)}))\Big), \quad (38b)$$

$$\text{AA:} \quad AA_i^{(m,\delta)} = \mathbf{1}\Big(S(R(q_i)) \neq S(R(q_i^{(m,\delta)})) \quad (38c)$$

$$\text{and correct}\Big) \quad (38d)$$

where CP = Change Propagation, RA = Reasoning Adaptation, AA = Answer Adjustment, $R(q_i)$ denotes the reasoning chain for question q_i, $S(\cdot)$ extracts the final answer, and $\mathbf{1}(\cdot)$ is the indicator function.

Aggregate Metrics

$$\text{Change Propagation Rate} = \frac{1}{N}\sum_{i=1}^{N} CP_i, \quad (39)$$

$$\text{Reasoning Adaptation Rate} = \frac{1}{N}\sum_{i=1}^{N} RA_i, \quad (40)$$

$$\text{Answer Adjustment Rate} = \frac{1}{N}\sum_{i=1}^{N} AA_i \quad (41)$$

Step Consistency (SC)

$$SC_i^{(m,\delta)} = 1 - \frac{|R(q_i) \ominus R(q_i^{(m,\delta)})|}{\max(|R(q_i)|, |R(q_i^{(m,\delta)})|)} \quad (42)$$

where $\ominus$ denotes sequence difference.

The $\ominus$ operator computes Levenshtein distance between reasoning sequences, treating each reasoning step as a discrete token. This approach, established in computational linguistics for sequence comparison, naturally handles variable-length reasoning chains while preserving step-order information crucial for mathematical reasoning evaluation.

Max-based Normalization: Max-length normalization prevents shorter sequences from artificially inflating consistency scores, following established practices in sequence alignment . This approach ensures that consistency measurement remains stable regardless of whether perturbations cause reasoning expansion or contraction, providing fair comparison across modification strategies.

Difficulty Score by Complexity Level

$$\text{Difficulty}(c) = 1 - \frac{1}{|\mathcal{D}_c|}\sum_{i \in \mathcal{D}_c} \frac{CP_i + RA_i + AA_i}{3} \quad (43)$$

where $\mathcal{D}_c$ is the set of questions of complexity level c.

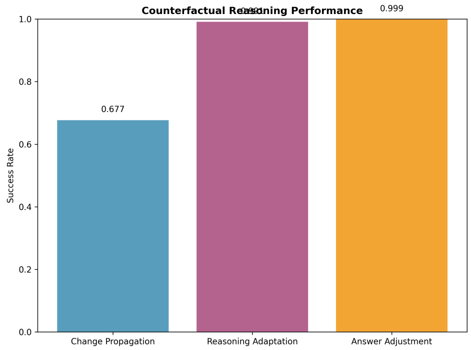

Figure 5: Counterfactual Reasoning Analysis

Limitations & Future Directions

The model scale and evaluation scope of our work are both constrained. We restricted generalization to larger models and wider math-reasoning domains by concentrating only on a small reasoning model. Richer symbolic reasoning and multi-table integration are not adequately captured by the suggested diagnostics, which place an emphasis on table-centric reasoning and arithmetic inference. Furthermore, because our faithfulness–plausibility annotations are unidirectional and perturbation vectors are scaled down. Lastly, rather than being completely universal measurements of logical entailment, the transitivity and backward consistency metrics continue to be rule-based approximations.

In order to determine whether observed failures persist or decrease with scale, future work should expand these diagnostics in three ways: (1) scaling the evaluation to larger reasoning models and diverse math datasets; (2) creating automated metrics for explanation faithfulness; and (3) adding richer counterfactuals (such as temporal shifts and semantic table edits) and adversarial distractors to the perturbation suite. Stronger guarantees of correctness could be made possible by further grounding the logical closure checks through integration with symbolic solvers and formal verification tools. Additionally, we believe that incorporating these assessments into training goals could be beneficial in motivating models to aim for verifiable reasoning as opposed to superficial plausibility.

Conclusion

We introduce a diagnostic framework that evaluates mathematical reasoning beyond surface accuracy. Results show that high answer accuracy can coexist with poor backward consistency and weak transitivity, revealing dependence on pattern matching over genuine logical computation. Although demonstrated on a small model, the framework generalizes and uncovers reasoning failures hidden from traditional metrics. By exposing these systematic weaknesses and releasing open evaluation protocols, we enable progress toward verifiable mathematical reasoning rather than mere pattern imitation.

References

Bjerring, J. C.; Busch, J.; and Aastrup Munch, L. 2025. A Counterfactual Account of Algorithmic Robustness. *Minds and Machines*, 35(3): 34.

Connell, L.; and Keane, M. T. 2006. A Model of Plausibility. *Cognitive Science*, 30(1): 95–120.

Kainz, W. 1995. Logical consistency. *Elements of spatial data quality*, 202: 109–137.

Li, J.; Yu, L.; and Ettinger, A. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. *arXiv preprint arXiv:2305.16572*.

Lu, X.; and Ma, J. 2024. Does faithfulness conflict with plausibility? an empirical study in explainable ai across NLP tasks. *arXiv preprint arXiv:2404.00140*.

Mavi, V.; Jangra, A.; and Jatowt, A. 2024. Multi-hop Question Answering. *Foundations and Trends® in Information Retrieval*, 17(5): 457–586.

Pang, T.; Lin, M.; Yang, X.; Zhu, J.; and Yan, S. 2022. Robustness and Accuracy Could Be Reconcilable by (Proper) Definition. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 17258–17277. PMLR.

Sahoo, S.; and Dutta, K. 2024. DUNE: Decoding Unified Naive Bayes Explainability through Gaussian methods for a Heart Disease Diagnostic. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6.

Saxton, D.; Grefenstette, E.; Hill, F.; and Kohli, P. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Senior, A. W.; and Robinson, A. 1995. Forward-backward retraining of recurrent neural networks. *Advances in Neural Information Processing Systems*, 8.

SINGH, M. 1996. PATH CONSISTENCY REVISITED. *International Journal on Artificial Intelligence Tools*, 05(01n02): 127–141.

Trabasso, T.; Van den Broek, P.; and Suh, S. Y. 1989. Logical necessity and transitivity of causal relations in stories. *Discourse processes*, 12(1): 1–25.

Wang, Y.; Li, T.; Liu, M.; Li, C.; and Wang, H. 2022. STSIIML: Study on token shuffling under incomplete information based on machine learning. *International Journal of Intelligent Systems*, 37(12): 11078–11100.

Wei, Y.; Su, Y.; Ma, H.; Yu, X.; Lei, F.; Zhang, Y.; Zhao, J.; and Liu, K. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *arXiv preprint arXiv:2310.05157*.

Xie, C.; Huang, Y.; Zhang, C.; Yu, D.; Chen, X.; Lin, B. Y.; Li, B.; Ghazi, B.; and Kumar, R. 2024. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Fan, J.; Yan, A.; Gao, E.; Lin, X.; Li, T.; Mo, K.; and Dong, C. 2025b. Distraction is all you need for multimodal large language model jailbreaking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9467–9476.

Yao, Z.; Liu, Y.; Chen, Y.; Chen, J.; Fang, J.; Hou, L.; Li, J.; and Chua, T.-S. 2025. Are Reasoning Models More Prone to Hallucination? *arXiv preprint arXiv:2505.23646*.

Zhang, Z.; Liu, J. W.; Re, C.; and Zhang, H. R. 2024. A Hessian View of Grokking in Mathematical Reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zheng, X.; Shirani, F.; Chen, Z.; Lin, C.; Cheng, W.; Guo, W.; and Luo, D. 2024. F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. *arXiv preprint arXiv:2410.02970*.