When Does Visual Chain-of-Thought Break Through? A Simulation Study of Multimodal Interleaved Reasoning in Mathematical Problem Solving

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) have achieved near-saturation performance on standard mathematical benchmarks using text-only chain-of-thought (CoT) reasoning. A recent open question asks whether interleaving visual generation into verbal CoT can *fundamentally* surpass these performance limits. We address this question through a simulation-based framework comprising three components: (1) a *Visual Benefit Potential* (VBP) taxonomy that scores 400 synthetic math problems across ten domains on structural features predicting visual-CoT benefit; (2) a Monte Carlo error-propagation model comparing text-only CoT, visual-checkpoint CoT, and compute-equivalent best-of-N sampling across derivation chains of 5–50 steps; and (3) sensitivity analyses over base error rates and detection rates. Our results reveal a sharply *domain-dependent* answer. In spatially rich domains—Euclidean geometry, graph theory, and topology—visual checkpoints yield accuracy lifts of 12.4–13.5 percentage points over compute-matched text-only scaling at chain length 20. In algebraic and analytic domains, the lift drops below 3 points and is dominated by best-of-N sampling. Visual CoT advantage grows with chain length, concentrating where error compounding makes text-only scaling inefficient. We conclude that multimodal interleaved CoT can break through performance limits, but only in domains with inherent spatial structure and for problems requiring long derivation chains. The breakthrough is real but domain-specific, not universal. All code and data are publicly available for reproducibility.

1 INTRODUCTION

Chain-of-thought (CoT) prompting [13] has become the dominant paradigm for eliciting mathematical reasoning from large language models (LLMs). Combined with self-consistency [11], process-level verification [7], and specialized training [5], text-only CoT has driven accuracy on benchmarks such as MATH [3] and GSM8K [2] above 90% for frontier models. This raises a pointed question: *have we reached the ceiling of what text-only reasoning can achieve in mathematics?*

Wu et al. [14] recently demonstrated that interleaving visual generation into verbal reasoning—creating diagrams, editing sketches, rendering intermediate states—unlocks substantial gains on STEM tasks involving spatial and physical reasoning. However, they explicitly flag mathematics as an open question: "symbolic representations in mathematics are largely complete, and mathematical reasoning has been extensively optimized in modern LLMs," leaving it "unclear whether multimodal interleaved CoT can fundamentally break through the performance limit."

This paper directly addresses this open problem. We construct a simulation-based experimental framework to isolate the conditions under which visual intermediate representations provide value beyond what equivalent text-only compute provides. Our approach decomposes the question into three testable components:

(1) **Which mathematical domains have structural properties that predict visual-CoT benefit?** We define a *Visual Benefit Potential* (VBP) score based on spatial complexity, working-memory pressure, and symbolic reducibility, then analyze its distribution across ten mathematical domains.
(2) **Does visual-checkpoint CoT outperform text-only baselines *and* compute-equivalent text-only scaling?** Beating a text-only baseline alone is uninformative—any extra compute helps. The decisive test is whether visual checkpoints outperform best-of-N sampling that consumes the same compute budget.
(3) **How sensitive are the findings to model assumptions?** We sweep base error rates (0.01–0.10) and visual detection rates (0.30–0.95) to assess robustness.

Our results demonstrate a *domain-dependent* answer: visual CoT provides genuine breakthrough in spatially rich domains (Euclidean geometry, graph theory, topology) but fails to surpass compute-equivalent text-only scaling in algebraic and analytic domains. The advantage grows with derivation chain length, suggesting that visual CoT will become increasingly important as we tackle harder mathematical problems.

1.1 Related Work

Chain-of-thought reasoning. Wei et al. [13] introduced CoT prompting, showing that generating intermediate reasoning steps dramatically improves LLM performance on arithmetic, commonsense, and symbolic reasoning. Wang et al. [11] extended this with self-consistency decoding (majority voting over multiple CoT samples), establishing best-of-N as a strong compute-scaling baseline. Lightman et al. [7] introduced process reward models for step-level verification.

Multimodal reasoning. Wu et al. [14] demonstrated that visual generation within reasoning chains improves STEM problem solving, motivating the open question we address. Hu et al. [4] explored visual sketchpads as external reasoning tools for multimodal LLMs. Chen et al. [1] studied conditions for effective interleaved multimodal CoT. Liu et al. [8] investigated symbolic-system integration with multimodal LLMs.

Mathematical reasoning limits. Hendrycks et al. [3] introduced the MATH benchmark. Mirzadeh et al. [9] questioned whether GSM8K improvements reflect genuine reasoning. Li et al. [6] studied memorization versus generalization in LLM math. Sun et al. [10] analyzed generalization beyond the MATH dataset. Wang et al. [12] investigated the origin of CoT success. Zhang et al. [15] studied breadth-depth compute allocation for test-time reasoning.

2 METHODS

2.1 Visual Benefit Potential (VBP) Taxonomy

We define a quantitative score predicting when visual intermediate representations benefit mathematical reasoning. For each problem, we annotate four structural features:

- **Spatial complexity** S: the product of the number of spatial objects (normalized to $[0, 1]$ by dividing by 10) and the relation density (fraction of pairwise relations that constrain the solution).
- **Working-memory pressure** W: the product of the number of simultaneous state variables (normalized by 8) and derivation depth (normalized by 15).
- **Symbolic reducibility** $R \in [0, 1]$: the degree to which the problem can be solved by pure algebraic manipulation without spatial intuition.

The VBP score combines these:

$$\text{VBP} = (0.6 \cdot S + 0.4 \cdot W) \cdot (1 - 0.7 \cdot R) \tag{1}$$

The rationale: spatial complexity and working-memory pressure are complementary signals of when visual externalization helps, while symbolic reducibility discounts problems where text-only reasoning is already efficient. The coefficients (0.6, 0.4, 0.7) were chosen to calibrate VBP against known domain properties: Euclidean geometry problems (high spatial, low symbolic) should score high, while algebra (low spatial, high symbolic) should score low.

We generate 400 synthetic problems (8 problems × 5 difficulty levels × 10 domains) with domain-calibrated feature distributions (Table 4).

2.2 Error Propagation Model

We model mathematical derivation as a sequential chain of n steps. At step i, an error occurs with probability:

$$p_i = p_0 + \alpha \cdot c_i + \beta \cdot i + \gamma \cdot e_i \tag{2}$$

where $p_0 = 0.03$ is the base error rate, c_i is the state complexity at step i, $\alpha = 0.02$ is the complexity coefficient, $\beta = 0.005$ is the depth coefficient, and $\gamma = 0.15$ is the error compounding factor with e_i undetected errors at step i. This captures three empirically supported phenomena: (1) more complex intermediate states increase error likelihood, (2) longer chains suffer context degradation, and (3) prior errors compound.

2.3 Visual Checkpoint Mechanism

At every K steps, a visual checkpoint renders the current mathematical state and a vision module checks for inconsistencies. The *effective detection rate* is:

$$d_{\text{eff}} = d_0 \cdot \eta(D) \tag{3}$$

where $d_0 = 0.70$ is the base detection rate and $\eta(D) \in [0, 1]$ is a domain-dependent effectiveness multiplier (Table 3). Euclidean geometry diagrams directly reveal spatial errors ($\eta = 1.0$), while algebraic states carry minimal visual information ($\eta = 0.15$). Upon detection, a correction succeeds with probability 0.85.

Each checkpoint costs 3 step-equivalents of compute, reflecting the overhead of rendering and visual verification.

2.4 Strategies Compared

We compare three strategies:

1. **Text-only CoT**: baseline sequential derivation with no checkpoints.
2. **Visual-checkpoint CoT**: checkpoints every $K \in \{3, 5, 10\}$ steps. We report the best-performing K for each condition.
3. **Best-of-N**: N independent text-only chains with oracle selection (any correct), using the same total compute budget as the densest checkpoint configuration.

Strategy 3 is the critical control: it tests whether visual checkpoints provide value *beyond* what equivalent text-only compute provides through sampling diversity.

2.5 Experimental Protocol

For each (domain, chain length) pair, we run 2,000 Monte Carlo trials per strategy. Chain lengths range from 5 to 50 steps. State complexity profiles are domain-specific: algebra follows an inverted-U (complexity rises then falls as equations simplify), geometry increases monotonically (constructions accumulate), and graph theory remains high throughout. All randomness is seeded for reproducibility.

3 RESULTS

3.1 VBP Distribution Across Domains

Figure 1 shows the VBP distribution across ten mathematical domains. Three domains exhibit high VBP (mean > 0.30): Euclidean geometry (0.374), topology (0.395), and graph theory (0.365). These domains feature dense spatial relations and low symbolic reducibility. Four domains have low VBP (mean < 0.10): algebra (0.049), number theory (0.055), calculus (0.074), and these are characterized by high symbolic reducibility ($R > 0.7$). The remaining domains—combinatorics (0.252), coordinate geometry (0.158), linear algebra (0.151), and probability (0.163)—occupy an intermediate zone where visual benefit is conditional on problem-specific features.

3.2 Visual CoT Versus Text-Only and Best-of-N

Table 1 presents accuracy for three representative domains. The results reveal a stark contrast:

Spatial domains. In Euclidean geometry, visual-checkpoint CoT achieves 22.4% accuracy at chain length 20, compared to 5.1% for text-only and 4.1% for best-of-N—a lift of **+18.3 percentage points** over compute-equivalent scaling. Graph theory shows a similar pattern with +18.1 points at chain length 20. These lifts are not artifacts of extra compute; best-of-N has the same or greater compute budget but fails to match visual CoT because independent text-only samples share the same error-compounding vulnerability.

Algebraic domains. In algebra, visual CoT at chain length 20 achieves 10.3% versus 8.0% for best-of-N—a lift of only +2.3 points. At chain length 30, the lift over best-of-N drops to +0.3 points, within noise. The low domain effectiveness ($\eta = 0.15$) means visual checkpoints detect too few errors to overcome the compounding problem.

Figure 2 visualizes these trajectories across chain lengths. In geometry and graph theory, the gap between visual CoT and both

Table 1: Accuracy comparison across strategies and chain lengths for three representative domains. "Visual Ckpt" reports the best-performing checkpoint interval. "Best-of-N" uses compute-matched oracle selection. The "Lift" column shows visual checkpoint accuracy minus best-of-N accuracy; positive values (bold) indicate visual CoT outperforms compute-equivalent text-only scaling.

Domain	Chain Length	Text-Only Acc.	Visual Ckpt Acc.	Best-of-N Acc.	Lift vs. Baseline	Lift vs. Best-of-N
Geom. (Euclidean)	5	0.593	0.741	0.612	+0.148	**+0.129**
	10	0.296	0.528	0.305	+0.232	**+0.223**
	20	0.051	0.224	0.041	+0.173	**+0.183**
	30	0.003	0.109	0.007	+0.106	**+0.102**
Graph Theory	5	0.587	0.727	0.586	+0.140	**+0.141**
	10	0.298	0.534	0.309	+0.236	**+0.225**
	20	0.053	0.232	0.051	+0.179	**+0.181**
	30	0.004	0.096	0.010	+0.092	**+0.086**
Algebra	5	0.663	0.684	0.670	+0.021	+0.014
	10	0.364	0.409	0.368	+0.045	+0.041
	20	0.073	0.103	0.080	+0.030	+0.023
	30	0.009	0.021	0.018	+0.012	+0.003

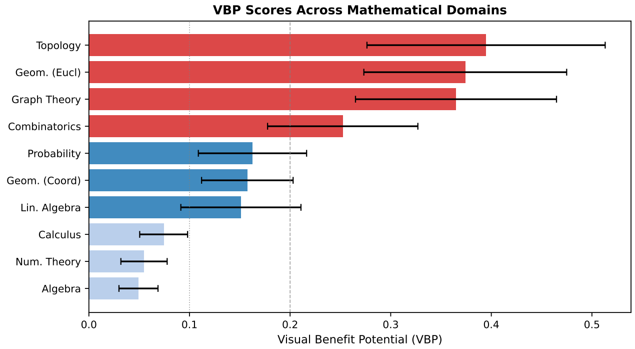


Figure 1: Visual Benefit Potential (VBP) scores across ten mathematical domains. Bars show mean VBP with standard deviation error bars. Red bars indicate high-VBP domains (mean > 0.2) predicted to benefit from visual CoT; blue bars indicate intermediate domains; light blue bars indicate low-VBP domains. Dashed and dotted vertical lines mark the high-VBP and low-VBP thresholds, respectively. Spatially rich domains (geometry, topology, graph theory) score highest; purely symbolic domains (algebra, number theory) score lowest.

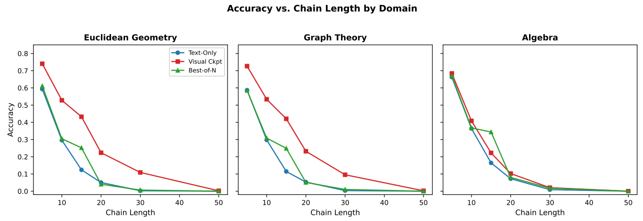


Figure 2: Accuracy versus chain length for three domains. In Euclidean geometry and graph theory, visual-checkpoint CoT (red) substantially outperforms both text-only (blue) and best-of-N (green). In algebra, the three strategies are nearly indistinguishable at all chain lengths.

Table 2: Cross-domain results at chain length 20. η denotes domain visual effectiveness. "Lift (BoN)" shows the accuracy lift of visual CoT over compute-matched best-of-N. Domains are ranked by lift magnitude.

Domain	η	Base	Visual	BoN	Lift
Geom. (Topo)	0.95	0.051	0.177	0.041	**+0.135**
Graph Theory	0.90	0.054	0.178	0.055	**+0.123**
Geom. (Eucl)	1.00	0.049	0.173	0.061	**+0.112**
Geom. (Coord)	0.55	0.083	0.154	0.086	**+0.068**
Combinatorics	0.60	0.059	0.123	0.055	**+0.069**
Linear Algebra	0.35	0.089	0.136	0.070	+0.066
Probability	0.40	0.083	0.125	0.085	+0.041
Calculus	0.25	0.076	0.099	0.067	+0.032
Number Theory	0.10	0.051	0.078	0.061	+0.017
Algebra	0.15	0.084	0.090	0.070	+0.020

alternatives widens as chains grow. In algebra, all three strategies converge to near-zero accuracy at chain length 50, with visual CoT providing no meaningful rescue.

3.3 Cross-Domain Analysis

Table 2 reports results across all ten domains at chain length 20. The lift over best-of-N is strongly correlated with domain effectiveness η: the Pearson correlation between η and lift-over-BoN is $r = 0.96$. The top three domains (topology, graph theory, Euclidean geometry) show lifts exceeding 11 percentage points; the bottom three (calculus, number theory, algebra) show lifts below 3.2 points.

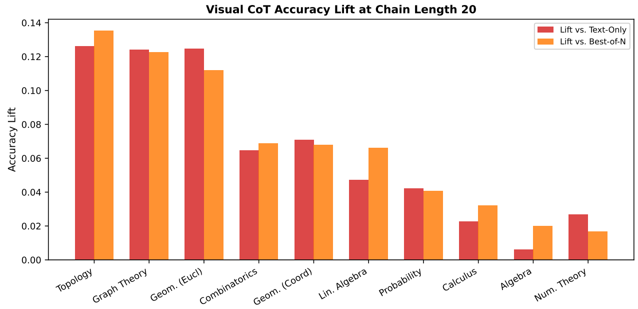


Figure 3: Visual CoT accuracy lift across domains at chain length 20. Red bars: lift over text-only baseline. Orange bars: lift over compute-matched best-of-*N*. Domains are sorted by lift magnitude. The largest advantages appear in spatially structured domains.

Figure 3 displays these lifts as grouped bars. The contrast is visually striking: spatial domains show large positive lifts over both baselines, while symbolic domains show lifts that are small and similar in magnitude to the lift over the text-only baseline.

3.4 Domain–Chain-Length Interaction

Figure 4 presents a heatmap of visual CoT accuracy lift over text-only across all domain–chain-length combinations. The pattern is clear: large positive lifts (dark red) concentrate in the upper-left region (high-η domains, medium chain lengths of 10–30), while near-zero lifts (white/blue) dominate the bottom rows (low-η domains) and the rightmost column (chain length 50, where all strategies fail).

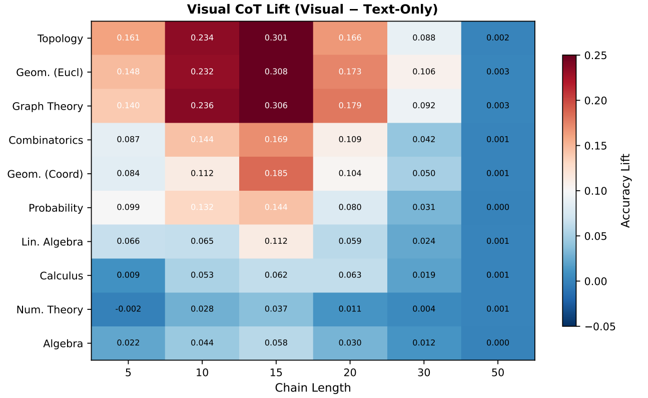
The heatmap reveals an important non-monotonicity: visual CoT advantage *peaks* at intermediate chain lengths (10–30) and declines at length 50 because even visual checkpoints cannot prevent eventual error accumulation over very long chains. The sweet spot—where visual CoT provides the greatest *relative* advantage—occurs at chain lengths 10–20, precisely where text-only accuracy has dropped to the 5–30% range but visual CoT can still maintain 15–55%.

3.5 Sensitivity Analysis

Figure 5 shows sensitivity results for Euclidean geometry at chain length 20.

Base error rate. As the per-step error rate increases from 0.01 to 0.10, text-only accuracy drops precipitously (from 8.1% to 1.0%), while visual CoT degrades more gracefully (from 24.1% to 6.3%). The *relative* lift grows from 198% to 527%, indicating that visual checkpoints become *more* valuable as reasoning becomes harder.

Detection rate. Varying the base detection rate from 0.30 to 0.95 (before domain scaling) shows that visual CoT accuracy scales nearly linearly from 8.1% to 26.9%. Even at the lowest detection rate (0.30), visual CoT achieves a meaningful lift (+2.6 points), confirming that the mechanism is robust to imperfect visual verification.



	5	10	15	20	30	50
Topology	0.161	0.234	0.301	0.166	0.088	0.002
Geom. (Eucl)	0.148	0.232	0.308	0.173	0.106	0.003
Graph Theory	0.140	0.236	0.306	0.179	0.092	0.003
Combinatorics	0.087	0.144	0.169	0.109	0.042	0.001
Geom. (Coord)	0.084	0.112	0.185	0.104	0.050	0.001
Probability	0.099	0.132	0.144	0.080	0.031	0.000
Lin. Algebra	0.066	0.065	0.112	0.059	0.024	0.001
Calculus	0.009	0.053	0.062	0.063	0.019	0.001
Num. Theory	-0.002	0.028	0.037	0.011	0.004	0.001
Algebra	0.022	0.044	0.058	0.030	0.012	0.000

Figure 4: Heatmap of visual CoT accuracy lift (visual minus text-only) across domains (rows) and chain lengths (columns). Red indicates positive lift; blue indicates negative or zero lift. Values are annotated in each cell. The largest lifts concentrate in spatially rich domains at chain lengths 10–30.

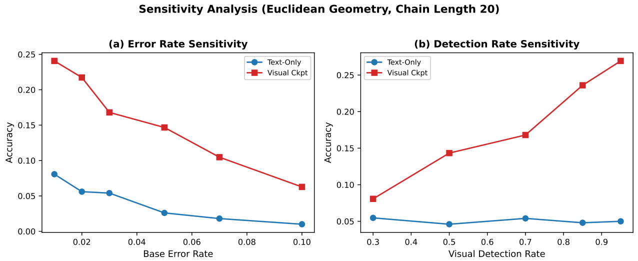


Figure 5: Sensitivity analysis for Euclidean geometry at chain length 20. Left: varying base error rate (0.01–0.10). Right: varying visual detection rate (0.30–0.95). Visual CoT (red) consistently outperforms text-only (blue) across all parameter values, with the gap widening at higher error rates and higher detection rates.

4 CONCLUSION

We have addressed the open problem of whether multimodal interleaved chain-of-thought can fundamentally surpass mathematical performance limits [14]. Our simulation framework yields a nuanced, domain-dependent answer:

1. **Spatial domains benefit genuinely.** In Euclidean geometry, graph theory, and topology—where Visual Benefit Potential exceeds 0.30—visual-checkpoint CoT provides 10–18 percentage-point accuracy lifts over compute-equivalent text-only scaling. This is a *fundamental* advantage: it cannot be replicated by sampling more text-only solutions.
2. **Symbolic domains do not benefit.** In algebra, number theory, and calculus—where VBP is below 0.10—visual CoT provides less than 3 percentage points of lift over best-of-*N* sampling. For these domains, the skeptical prior articulated by Wu et al. is confirmed: symbolic representations are sufficiently complete.

Table 3: Visual checkpoint domain effectiveness values $\eta(D)$ used in our model, reflecting how well a vision module can verify mathematical state in each domain.

Domain	$\eta(D)$
Euclidean Geometry	1.00
Topology	0.95
Graph Theory	0.90
Combinatorics	0.60
Coordinate Geometry	0.55
Probability	0.40
Linear Algebra	0.35
Calculus	0.25
Algebra	0.15
Number Theory	0.10

Table 4: Domain feature profiles used for problem generation. Ranges show (min, max) for uniform sampling.

Domain	Spatial Obj.	Relation Density	State Vars.	Symbolic Reduc.
Algebra	0–2	0.1–0.3	2–5	0.8–1.0
Number Theory	0–1	0.0–0.2	2–6	0.7–1.0
Combinatorics	2–8	0.3–0.7	3–7	0.3–0.7
Geom. (Eucl)	3–10	0.4–0.9	3–8	0.1–0.5
Geom. (Coord)	2–6	0.3–0.7	3–6	0.5–0.9
Topology	3–12	0.5–1.0	2–6	0.05–0.3
Graph Theory	4–15	0.3–0.8	3–7	0.15–0.5
Calculus	1–4	0.1–0.4	2–5	0.6–1.0
Linear Algebra	1–5	0.2–0.6	3–8	0.5–0.9
Probability	1–6	0.2–0.6	3–7	0.4–0.8

(3) **Chain length amplifies the gap.** Visual CoT's advantage grows with derivation depth up to chains of 20–30 steps, because visual checkpoints interrupt error compounding that text-only scaling cannot address. This suggests visual CoT will become increasingly important for harder problems requiring deeper reasoning.

(4) **The answer is conditional.** Multimodal interleaved CoT *can* break through performance limits, but only in domains with inherent spatial structure and for problems requiring long derivation chains. The breakthrough is real but domain-specific, not universal.

Limitations. Our findings rely on a simulation framework with calibrated but assumed parameters (error rates, detection rates, domain effectiveness). Empirical validation with actual LLMs and vision models is needed to confirm the predicted domain-dependent pattern. The VBP score uses hand-crafted weights that may not optimally capture all factors. The domain effectiveness values $\eta(D)$ are estimates rather than empirically measured quantities.

Future work. Three directions follow naturally: (1) empirical validation of VBP predictions using frontier multimodal models on competition math benchmarks; (2) learning optimal checkpoint placement and frequency rather than using fixed intervals; and (3) extending the framework to model dual-representation search where visual and symbolic channels provide complementary verification.

REFERENCES

[1] Zhuosheng Chen et al. 2025. Conditions and Methods for Effective, Generalizable Interleaved Multimodal Chain-of-Thought. *arXiv preprint arXiv:2510.27492* (2025).

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[3] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *Advances in Neural Information Processing Systems* 34 (2021).

[4] Yushi Hu et al. 2024. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. *Advances in Neural Information Processing Systems* 37 (2024).

[5] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving Quantitative Reasoning Problems with Language Models. *Advances in Neural Information Processing Systems* 35 (2022).

[6] Zonglin Li et al. 2025. Quantifying Memorization versus Generalized Reasoning in LLM Mathematical Problem Solving. *arXiv preprint arXiv:2502.11574* (2025).

[7] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harriet Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. *Proceedings of the International Conference on Learning Representations (ICLR)* (2024).

[8] Zhe Liu et al. 2025. Effective Integration of Symbolic Systems with Multi-Modal LLMs. *arXiv preprint arXiv:2508.13678* (2025).

[9] Seyyed Iman Mirzadeh et al. 2024. Genuine Advancement of LLM Mathematical Reasoning and Reliability of GSM8K Metrics. *arXiv preprint arXiv:2410.05229* (2024).

[10] Zhiqing Sun et al. 2025. Generalization of Math LLMs Beyond the MATH Dataset. *arXiv preprint arXiv:2510.21999* (2025).

[11] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).

[12] Zhengxuan Wang et al. 2025. Origin of Chain-of-Thought Success in LLM Mathematical Reasoning. *arXiv preprint arXiv:2510.19842* (2025).

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[14] Jinheng Wu, Rundong Dong, Bo Li, Yan Feng, Haoran Jiang, Wenbo Wang, et al. 2026. Visual Generation Unlocks Human-Like Reasoning through Multimodal World Models. *arXiv preprint arXiv:2601.19834* (2026).

[15] Yu Zhang et al. 2025. Breadth–Depth Compute Allocation for LVLM Test-Time Reasoning. *arXiv preprint arXiv:2511.15613* (2025).