Confidence and Correctness Are Indistinguishable to a Prompted Model on Multimodal Geometry

Anonymous Authors[1]

Abstract

A self-improving model is usually steered by one of two signals, its own confidence or whether its answers are actually correct. Under reinforcement learning this choice matters, since optimizing toward confidence inflates overconfidence and can cost accuracy while ground-truth rewards do not. We ask whether the same gap appears in a lightweight, no-training, in-context loop. Two vision-language models, Claude Haiku and Pixtral-12B, each run the loop twice on diagram-based plane-geometry problems from MathVerse, one arm carrying forward the examples the model was most confident in, the other the examples that were actually correct. Across four rounds the selection signal does not change mathematical reasoning. Accuracy and calibration are statistically indistinguishable between the two arms, and so are confidence on wrong answers and abstention. The contrast with reinforcement learning is the point. A confidence-versus-correctness distinction that DCPO found consequential under RLHF does not transfer to prompt design, where it changes nothing we measured. What varies instead is the model. At nearly identical accuracy, confidence on wrong answers is 0.70 for Claude and 0.94 for Pixtral, and verified-correct feedback never reduces it. Overconfidence on errors is a property of the model, not of the selection signal or the task.

1. Introduction

Self-improvement loops let a model get better at a task from its own outputs. It attempts problems, keeps some by a reward signal, and conditions on the kept examples going forward. A recurring concern is calibration. As a model self-improves it tends to grow more confident faster than it grows more correct, so it becomes overconfident even when wrong (Huang et al., 2025). The reward signal is implicated. Reward models in RLHF give higher scores to confident-sounding answers regardless of correctness (Leng et al., 2024), and under reinforcement learning with verifiable rewards the reasoning objective and the confidence objective pull against each other, so separating them is needed to keep both accuracy and calibration (Ma et al., 2026).

That prior work establishes the effect in a training setting, where weights are updated. We ask a narrower, measurement question. Does the choice of self-improvement signal matter when there is no training at all? If a model carries forward the examples it was most *confident* in versus the examples that were actually *correct*, does its downstream accuracy, calibration, or willingness to abstain change, in a purely in-context loop?

This matters because in-context self-improvement, accumulating few-shot exemplars with no fine-tuning, is cheap, common, and increasingly how agents learn on the fly. If the confidence-versus-correctness gap reported under reinforcement learning also appears here, it is a reason to be careful about confidence-filtered prompting. If it does not, the effect is specific to the training setting and does not carry over to in-context prompting.

We run the comparison on two vision-language models over diagram-based plane-geometry problems from MathVerse, four rounds each, two arms per model. The headline is negative. The two signals are indistinguishable on accuracy, calibration, confidence-on-wrong, and abstention, in every model. The variation that does exist is across models, not across arms. Models with the same accuracy differ enormously in how confident they are when wrong, and verified-correct feedback does not pull that confidence down. We report the negative result as a negative result, with its sample-size and design caveats stated plainly, and read it as a reliability caution for self-evolving math agents that rely on verifiable feedback.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Related Work

Calibration and its measurement. We quantify calibration with Expected Calibration Error (Guo et al., 2017), which partitions the predictions into M equal-width confidence bins $B_1, \ldots, B_M$ and averages the gap between confidence and accuracy within each bin,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ are the empirical accuracy and mean predicted confidence in bin m, n is the number of predictions, and we use $M{=}10$. The role of calibration in self-improvement is established by Huang et al. (2025), who find that across basic prompting, chain-of-thought, and tuning-based self-improvement, models grow more accurate and simultaneously more overconfident, with ECE rising over iterations, while chain-of-thought helps accuracy but not overconfidence. Their fix, calibrating after a run of self-improvement iterations, beats calibrating before or between steps. We respond to the diagnostic gap they leave. They show self-improvement breeds overconfidence, and we test whether the overconfidence-driven *selection* of examples is what does it, by comparing a most-confident arm against a verified-correct arm.

Separating the confidence and correctness signals. Two results motivate treating confidence and correctness as separable signals. Leng et al. (2024) locate the mechanism in RLHF, where the reward model assigns higher rewards to more confident-sounding answers independent of correctness, so the policy learns to sound confident. That is precisely why one would want to reward correctness without rewarding confidence. DCPO (Ma et al., 2026) acts on that separation under reinforcement learning with verifiable rewards. It finds the reasoning-objective and confidence-objective gradients have a negative Fisher-metric inner product, so a single shared objective cannot calibrate confidence without hurting reasoning, and it routes separate rewards to reasoning and confidence tokens with a masked-gradient strategy. Our two arms apply the same confidence-versus-correctness separation, but over which examples a prompt carries rather than which gradients update the weights. As a measurement study we also report a tail these methods do not, namely confidence on incorrect answers and the loss of abstention.

Self-training and multimodal math. Bootstrapping a model from its own outputs is established. STaR (Zelikman et al., 2022) fine-tunes on self-generated rationales that reach correct answers, and self-consistency (Wang et al., 2023) aggregates sampled chains of thought. Recent work questions how far self-training goes without external signal (Shafayat et al., 2025). MathVerse (Zhang et al., 2024b) is a multimodal math benchmark whose version splits let us select items where the diagram is load-bearing, and work on calibrating vision-language models documents that they are miscalibrated to begin with (Xiao et al., 2026). We use the multimodal setting because geometry answers are cleanly verifiable, which the oracle arm needs. Recent multimodal-geometry work targets the visual-grounding bottleneck directly, including MAVIS (Zhang et al., 2025a), SVE-Math (Zhang et al., 2025b), and GeoDANO (Cho et al., 2025), relevant here because our carried exemplars contain no diagram.

Calibration under prompting and in-context learning. Closest to our setting is calibration when a model adapts in-context rather than through weight updates. Zhang et al. (2024a) report a calibration–accuracy trade-off in the low-shot regime ($k{<}4$) that eases with more examples ($k{=}8$) and larger models, which bears directly on our three-shot setting. Li et al. (2023) find in-context learners miscalibrated in low-resource setups and propose self-ensembling. Chhikara (2025) document an overconfidence gap and distractor effects on in-context calibration. Memory-based in-context self-evolution such as ReasoningBank (Ouyang et al., 2025) improves agents by accumulating prompt-level experience rather than updating weights, the regime our loop operates in. These motivate both our shot-count caveat and our treatment of verbalized confidence as a known-imperfect signal.

3. Method

Benchmark. We use MathVerse (Zhang et al., 2024b), restricted to its vision-dependent versions (Vision Dominant, Vision Intensive) so a text-only shortcut is unlikely to suffice, and to multiple-choice items so correctness is an exact letter match. After filtering, the pool is entirely plane geometry sourced from GeoQA, with diagram-plus-question problems of moderate, school-to-competition-entry difficulty, not olympiad level. The subfields present are Angle, Length, Area, and Applied. We draw a fixed evaluation set of $N{=}40$ problems and a disjoint practice pool. Both arms of a model see the identical evaluation set, so the only thing that varies is the reward.

Models. We test two vision-language models, Claude Haiku (`claude-haiku-4-5`) and Pixtral-12B (`pixtral-12b-2409`). Each problem is presented as the diagram image plus the question text. The model reasons briefly, then emits a final answer (a letter, or an explicit abstention) and an integer confidence from 0 to 100 read as its probability of being correct.

Experimental details. Evaluation runs at temperature 0, and the practice phase that harvests exemplars runs at tem-

perature 0.7. Each round draws 30 practice problems disjoint from the 40-item evaluation set and from earlier practice slices. From the solved practice attempts, the confidence arm keeps those with self-reported confidence at least 70 and the correctness arm keeps those matching the gold letter. The prompt for the next round is seeded with the three most recently kept exemplars, presented as text-only question–solution pairs (no diagram is carried, to keep context light). When fewer than three have been kept, fewer are used. Diagram images are passed at native resolution, identical across arms and models. The 40-item evaluation set is fixed by a single random seed, so it is the same set, in the same order, for both arms, both models, and all four rounds. Each model's two arms are therefore evaluated on exactly the same problems, which is what licenses the paired item-level tests below. There are no weight updates and no decoding changes between arms, and the arms differ only in the keep rule.

The two arms. Each model runs the same loop twice, differing only in which solved practice examples are kept and carried forward. The *confidence* arm keeps the examples the model was most confident in (confidence at least 70). The *correctness* arm keeps the examples whose answer matches the gold label. We run four evaluation rounds. Round 0 is a baseline with no carried examples. Before each later round the prompt is seeded with up to three of the most recently kept examples as few-shot demonstrations. There are no weight updates, and the only state carried between rounds is the few-shot exemplar set.

Metrics. True correctness is always scored by the gold label, in both arms, so the confidence arm's selection signal never contaminates the reported accuracy. Per round we report accuracy with Wilson 95% confidence intervals, Expected Calibration Error over ten equal-width confidence bins, mean confidence, confidence on wrong answers (mean confidence over gold-incorrect answers), and abstention rate.

4. Results

The selection signal changes neither reasoning nor calibration. Selecting self-improvement examples by confidence versus by correctness produced no measurable difference, in either model, on any metric (Table 1). The model's mathematical reasoning is unchanged. Mean accuracy across rounds was 0.717 (confidence) versus 0.711 (correctness) for Claude and 0.703 versus 0.693 for Pixtral, and a paired McNemar test on per-item correctness, pairing each problem against itself across the two arms on the shared evaluation set, finds no difference (p=0.69 for Claude, p=0.30 for Pixtral), with the paired 95% bootstrap interval on the accuracy difference spanning zero and bounded to a few points ($[-0.08, +0.04]$ for Claude, $[-0.13, +0.03]$ for Pixtral). Its calibration is equally unchanged. The Expected Calibration Error difference between arms has a bootstrap 95% interval of $[-0.04, +0.07]$ for Claude and $[-0.05, +0.09]$ for Pixtral, both spanning zero, and Brier score and negative log-likelihood differ by at most 0.02 and 0.16 (Table 1). Confidence on wrong answers and abstention are likewise indistinguishable between arms (Figure 1). Accuracy also did not rise across rounds within either arm, so we read the arm comparison as a contrast between the two selection signals within this in-context regime rather than a claim about in-context learning in general.

This sharpens the prior claim rather than contradicting it. DCPO (Ma et al., 2026) reports a real accuracy gap between confidence-optimized and correctness-optimized policies, but under reinforcement learning, where weights are updated. Applying the same contrast in a no-training in-context loop, the gap disappears. A distinction that was consequential when it shaped weights does not carry over to prompt design, where the choice of which examples to keep makes no measurable difference.

The real variation is across models. At nearly the same accuracy (0.67 to 0.69), the two models differ sharply in confidence (Figure 2). Confidence on wrong answers is 0.70 for Claude and 0.94 for Pixtral, and mean confidence follows suit (0.78 versus 0.94). Pixtral is almost as confident on the answers it gets wrong as on the ones it gets right. Overconfidence here is a model property, not a property of the task or the reward, and correctness-filtered feedback does not reduce it.

Performance varies by kind of problem, not by arm. Breaking results down by geometric subfield (Figure 3, Table 2), the models are consistently worst at Angle problems (Claude 0.62, Pixtral 0.59 to 0.62) and better at Length (Claude 0.84, Pixtral 0.71 to 0.86). Whether a model was fed its most-confident examples or its verified-correct examples did not change how good it was at a given subfield. The danger compounds. For Claude, Angle is also where confidence on wrong answers is highest (0.75, versus 0.54 on Length), so the model is most confidently wrong exactly where it is weakest. We restrict subfield claims to Angle (n=92 pooled) and Length (n=44). Area and Applied have only three problems each and are reported but not relied on.

Abstention. Claude essentially never abstains in either arm. Pixtral abstains 11% of the time under the confidence arm and 17% under the correctness arm, with the correctness-arm rate drifting up to 27.5% by round 3. The difference is suggestive but not significant at this sample size (two-proportion test, $p \approx 0.15$), and we do not claim it.

Table 1. Mean over four rounds, by model and arm. Differences between the confidence and correctness arms are within noise on every metric. Paired McNemar on accuracy gives p=0.69 (Claude) and p=0.30 (Pixtral), and the bootstrap 95% CI on the ECE difference spans zero for both. CW = confidence on wrong answers, Abst = abstention rate, NLL = negative log-likelihood.

Model	Arm	Acc	ECE	Brier	NLL	CW	Abst
Claude	confidence	.717	.128	.192	.586	.720	.006
Claude	correctness	.711	.131	.210	.652	.747	.025
Pixtral	confidence	.703	.268	.262	2.491	.915	.113
Pixtral	correctness	.693	.286	.281	2.332	.945	.169

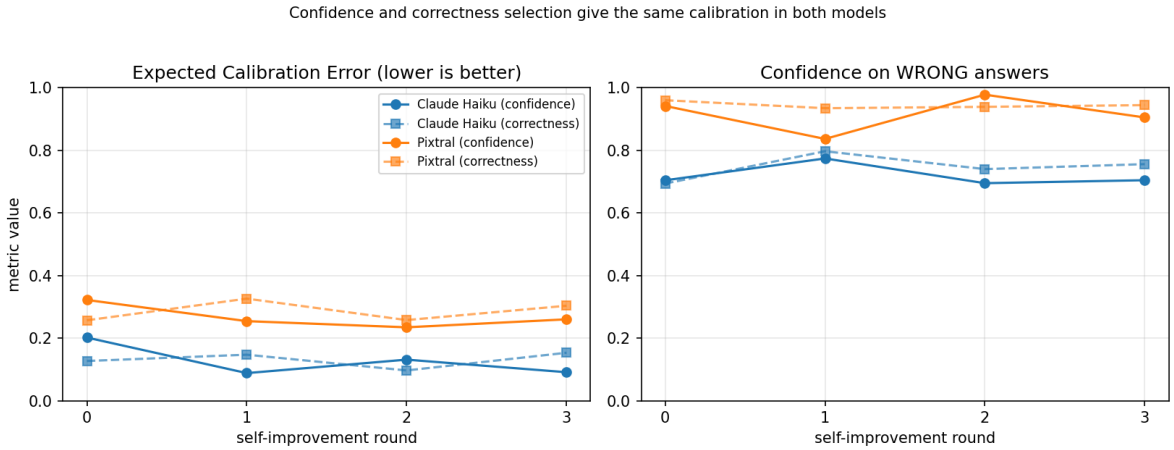


Figure 1. Calibration error (left) and confidence on wrong answers (right) across self-improvement rounds. Within each model the confidence-filtered (solid) and correctness-filtered (dashed) arms overlap, so the selection signal makes no difference. Pixtral (orange) sits above Claude (blue) on both, and confidence-on-wrong stays high and flat in every condition.

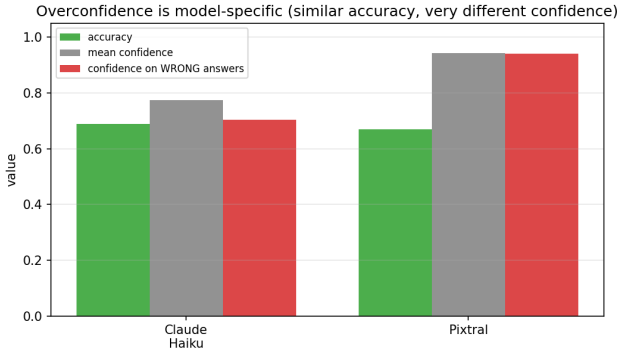


Figure 2. At similar accuracy, confidence diverges sharply by model. Confidence on wrong answers is 0.70 for Claude and 0.94 for Pixtral. Baseline round, confidence arm.

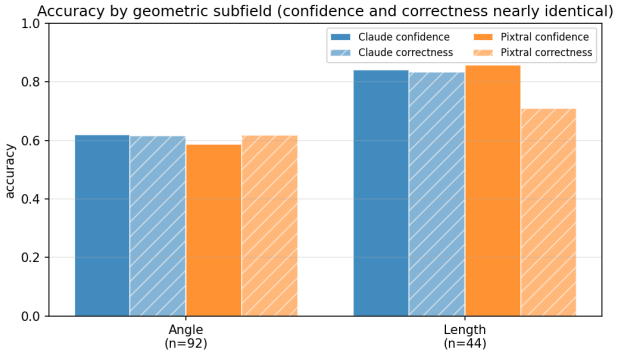


Figure 3. Accuracy by geometric subfield. Angle is hardest for both models. The confidence (solid) and correctness (hatched) arms are nearly identical within each model.

To answer the question we set out with, providing a model its most-confident answers versus its verified-correct answers does not change its multimodal mathematical reasoning, at least for two models, few-shot prompting with about forty samples, and three carried examples per round.

5. Discussion

In a purely prompting setting, selecting the carried examples by the model's confidence versus by their verified correctness had no measurable effect on how either model interpreted or reasoned about its multimodal mathematical input, neither on accuracy, nor on calibration, nor on the per-subfield pattern of which problems it solved. This is the result we emphasize, and it is significant because of the

Table 2. Accuracy and confidence on wrong answers (CW) by geometric subfield (confidence arm, where the correctness arm is within noise on every cell). Angle is the hardest subfield for both models and, for Claude, also where confidence-on-wrong is highest. †Area and Applied have only 3 problems each (×4 rounds), reported but not relied on. "–" marks no wrong answers to average.

		Claude		Pixtral	
Subfield	n	Acc	CW	Acc	CW
Angle	92	.62	.75	.59	.91
Length	44	.84	.54	.86	.97
Area†	12	.92	.95	1.00	–
Applied†	12	.82	.75	.80	.90

contrast with DCPO (Ma et al., 2026). Under reinforcement learning the confidence-versus-correctness distinction was consequential. Rewarding confidence pulled against reasoning, so the two signals had to be separated to preserve accuracy and calibration together. Transplanting the same distinction into prompt design, which examples a model carries forward, produced none of that effect here. A separation that mattered when it shaped weights was inert when it only shaped the few-shot context. Concretely, neither the model's reasoning (accuracy, McNemar $p > 0.3$, with the paired 95% interval on the accuracy difference spanning zero) nor its calibration (an ECE difference whose 95% interval spans zero) moves with the selection signal. By every measure we took, the two arms are the same model.

What the experiment shows cleanly is about calibration, not self-improvement. Overconfidence on wrong answers is large, persists across rounds, and is not reduced by feeding the model verified-correct examples. Its magnitude is a model property. Claude is moderately overconfident and Pixtral substantially more so, confident on almost every error. For self-evolving math agents that rely on verifiable feedback to stay reliable, this is a caution. Verifiability of the reward is not the same as calibration of the agent. An agent can be handed perfect feedback and still report high confidence on exactly the problems it gets wrong.

We are careful about what this does not show. We did not reduce overconfidence and observe an accuracy effect. Overconfidence stayed high in both arms, so we cannot speak to whether reducing it would help. We did not run weight updates, so we do not contradict DCPO on its own terms. And not observing a difference between arms is not the same as there being none. The statement is bounded by our sample size and our lightweight loop.

6. Limitations

The design is deliberately light, and several choices bound the claims.

Few shots, text-only exemplars. Only three examples were carried to the next round, not the full set of kept examples, which limits how much of the previous round's learning the loop can use. Prior work on in-context calibration finds a calibration–accuracy trade-off precisely in this low-shot regime ($k<4$) that eases by $k=8$ (Zhang et al., 2024a), so a shot-count ablation is the most important next experiment. The carried exemplars are also text-only, with the diagram dropped to keep context light, so on a task where our own subfield breakdown points to perception as the bottleneck (Angle problems), the loop cannot transmit the signal that is actually failing, and the null is best read as scoped to a perception-light loop rather than to multimodal in-context learning in general.

Sample size. The evaluation set is $N=40$ per round, single seed, so round-to-round movement is within noise and small subfields (Area, Applied, three problems each) cannot support claims.

Prompting, not training. The loop is prompting only, with no weight updates. This is intentional, since the point is to contrast in-context self-improvement against the reinforcement-learning setting of prior work, but it means the null cannot be read as a statement about training.

Multiple-choice grading. Grading is on multiple-choice items, so reported accuracy includes lucky guesses, which can inflate it and also contaminate confidence-on-wrong, since a reasoning failure that happens to guess the right letter is scored correct and never enters that average. A future version should use the benchmark's non-multiple-choice items.

Self-reported confidence. The confidence score is the model's own self-reported number, known to be unreliable for in-context learners (Li et al., 2023). Token-level option probabilities would be a sturdier basis, and a future version could also elicit and combine several component judgments.

Abstention is not incentivized. Abstention is offered but never rewarded, so the model has no incentive to abstain, which may explain why abstention is rare. Incentivizing it in the prompt design is a natural next step.

Benchmark scope. The filtered benchmark is entirely GeoQA-sourced plane geometry, not solid geometry, functions, or olympiad-level material.

7. Conclusion

Steering an in-context self-improvement loop by a model's confidence versus by verified correctness made no measurable difference to accuracy, calibration, confidence-on-

wrong, or abstention, across two vision-language models on diagram-based plane geometry. The confidence-versus-correctness gap reported under reinforcement learning does not reproduce when the same distinction is applied to prompt design rather than to weight updates.

The variation that does exist is across models. At matched accuracy, confidence on wrong answers is 0.70 for Claude and 0.94 for Pixtral, and verified-correct feedback does not bring it down. For self-evolving math agents that depend on verifiable feedback, the takeaway is that a verifiable reward does not buy a calibrated agent. Reducing that overconfidence, and testing whether reducing it improves accuracy under a real training loop, is the experiment this measurement motivates.

Impact Statement

We measure how a self-improving vision-language model reasons about mathematical diagrams. The central finding is null. Steering the in-context loop by the model's confidence versus by verified correctness does not change mathematical-reasoning accuracy or calibration (ECE) to a statistically significant degree in either model tested, and breaking results down by problem type shows performance tracking the problem (angle problems are hardest), not the selection mode. The reliability message is that a verifiable reward does not by itself yield a calibrated agent. The models report high confidence precisely on the answers they get wrong, so self-evolving math agents that trust verifiable feedback should pair it with explicit calibration and abstention. The study uses only a public geometry benchmark with no human subjects, and we see no other societal consequences that must be highlighted here.

References

Chhikara, P. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*, 2025.

Cho, S., Qin, Z., Liu, Y., Choi, Y., Lee, S., and Kim, D. GeoDANO: Geometric VLM with domain agnostic vision encoder. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Huang, L., Li, D., Liu, H., and Cheng, L. Beyond accuracy: The role of calibration in self-improving large language models. *arXiv preprint arXiv:2504.02902*, 2025.

Leng, J., Huang, C., Zhu, B., and Huang, J. Taming overconfidence in LLMs: Reward calibration in RLHF. *arXiv preprint arXiv:2410.09724*, 2024.

Li, C., Zhou, H., Glavaš, G., Korhonen, A., and Vulić, I. Large language models are miscalibrated in-context learners. *arXiv preprint arXiv:2312.13772*, 2023.

Ma, Z., Wen, X., Cao, B., Lu, Y., Lin, H., Yang, J., He, M., Han, X., and Sun, L. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards. *arXiv preprint arXiv:2603.09117*, 2026.

Ouyang, S., Yan, J., Hsu, I.-H., Chen, Y., Jiang, K., Wang, Z., Han, R., Le, L. T., Daruki, S., Tang, X., Tirumalashetty, V., Lee, G., Rofouei, M., Lin, H., Han, J., Lee, C.-Y., and Pfister, T. ReasoningBank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Shafayat, S. et al. Can large reasoning models self-train? *arXiv preprint arXiv:2505.21444*, 2025.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Xiao, W., Xu, X., and Gan, L. VL-Calibration: Decoupled confidence calibration for large vision-language models reasoning. *arXiv preprint arXiv:2604.09529*, 2026.

Zelikman, E., Wu, Y., Mu, J., and Goodman, N. D. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhang, H., Zhang, Y.-F., Yu, Y., Madeka, D., Foster, D., Xing, E., Lakkaraju, H., and Kakade, S. A study on the calibration of in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024a.

Zhang, R., Wei, X., Jiang, D., Zhang, Y., Guo, Z., Tong, C., Liu, J., Zhou, A., Wei, B., Zhang, S., Gao, P., and Li, H. MAVIS: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations (ICLR)*, 2025a.

Zhang, R. et al. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024b.

Zhang, S., Chen, A., Sun, Y., Gu, J., Zheng, Y.-Y., Koniusz, P., Zou, K., van den Hengel, A., and Xue, Y. Open eyes, then reason: Fine-grained visual mathematical understanding in MLLMs. *arXiv preprint arXiv:2501.06430*, 2025b.