MATHEMATICAL TRANSFER IN LLMS FOLLOWS REASONING APPROACH MORE THAN TOPIC

Anonymous authors
Paper under double-blind review

ABSTRACT

When selecting mathematical training data for LLMs, a natural organizing principle is topic: probability examples for probability targets. An alternative is reasoning approach: worked solutions that share a solution method with the target, even when the mathematical domain differs. We ask which relation produces greater transfer after fine-tuning. We evaluate two counterbalanced 2×2 designs: probability and combinatorics crossed with invariant reasoning and double counting (2,000 problems), and number theory and geometry crossed with complement and pigeonhole reasoning (800 problems). In each design, every cell serves as the held-out target in turn: same-approach (SA) sources share the target's method but change the topic, while same-topic (ST) sources share the topic but change the method. Every source appears once in each role, so additive source-quality effects cancel from the equally weighted aggregate contrast. Across five base models and three training seeds per design, SA outperforms ST in all 40 seed-pooled model–target comparisons. Model-level advantages range from 8.2 to 16.2 percentage points in the primary design (mean: 10.8) and from 12.0 to 16.0 in the second design (mean: 14.3); all ten model-level 95% confidence intervals exclude zero. In both designs, ST sources are *more* similar to targets under embedding and lexical measures, so the SA advantage runs opposite to the measured ordering of statement-level resemblance. These findings identify reasoning approach as a more effective matching criterion than topic for mathematical transfer across the evaluated topic–approach combinations.

1 INTRODUCTION

What should mathematical training data share with the problems a language model will be asked to solve? The most common organizing principle is *topic*: to improve performance on probability, train on probability; to improve on combinatorics, train on combinatorics. But mathematical problems can also be organized by *reasoning approach*, where a worked solution may exploit a conserved quantity (invariant reasoning) or establish two counts of the same collection (double counting), regardless of whether the objects are probabilistic or combinatorial. These two axes can disagree: a source dataset can match the target's topic while using a different solution method, or match the method while changing the topic. Which relationship produces greater transfer after fine-tuning?

Prior work gives reason to expect that approach matters. In human problem-solving, experts categorize problems by deep structural principles rather than surface features, and structural correspondences are central to theories of analogical transfer (Chi et al., 1981; Gick & Holyoak, 1983; Gentner, 1983). In LLMs, skill-matched exemplars improve mathematical prompting over topic-matched alternatives (Didolkar et al., 2024), and skill-aware data selection improves training efficiency (Chen et al., 2023; He et al., 2026; Zhang et al., 2026). These findings motivate our hypothesis, but leave a specific question open: after *parameter updates*, does supervision from a different topic that shares the target's approach outperform the same-topic supervision that uses a different approach?

Answering this question requires a design that separates the *relationship* between source and target from the identity of the source data. Comparing a single "approach-matched dataset" against a single "topic-matched dataset" would confound the relationship with the particular examples chosen. We therefore construct a 2×2 factorial crossing two topics (probability, combinatorics) with two approaches (invariant, double counting) and rotate the held-out target through all four cells (Figure 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

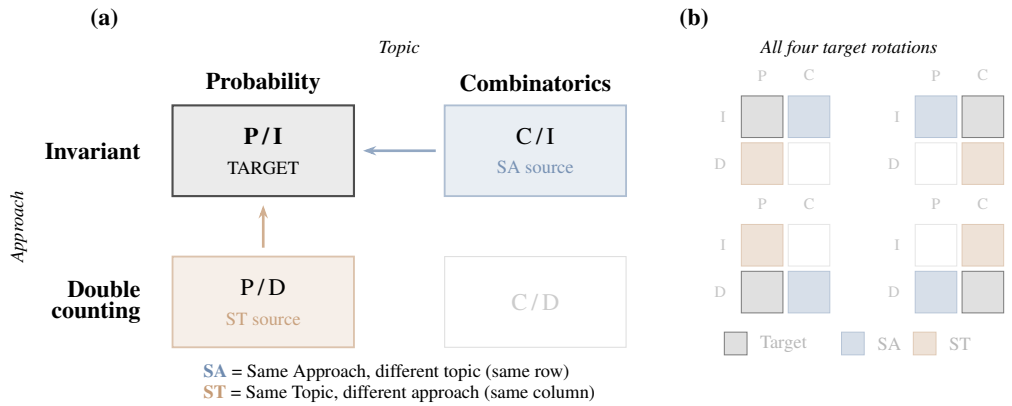


Figure 1: **The** 2×2 **rotational design (a)** Each cell pairs a topic (column) with a solution approach (row). When one cell is held out as the target (here: probability/invariant), its SA source occupies the same row—shared approach, different topic—while its ST source occupies the same column—shared topic, different approach. **(b)** The target rotates through all four cells.

For each target, its same-approach (SA) source shares the approach but changes the topic, while its same-topic (ST) source shares the topic but changes the approach. Because every source cell appears exactly once as SA and once as ST, additive source-quality differences cancel from the aggregate contrast.

Across five base models, three training seeds, and all four target rotations, SA outperforms ST in every one of the 20 model–target comparisons, with model-level advantages ranging from 8.2 to 16.2 percentage points (mean: 10.8). In 19 of 20 comparisons, SA also improves over the unadapted base model, confirming that the advantage reflects genuine positive transfer, not merely less harmful adaptation. A statement-similarity analysis strengthens the finding: same-topic problems are *more* similar to targets under both embedding and lexical measures, yet they produce weaker transfer. The SA advantage, therefore, runs *opposite* to the measured surface resemblance.

The ordering also replicates in an 800-problem design crossing number theory and geometry with complement and pigeonhole reasoning. Across the same five models and three seeds, all 20 seed-pooled model–target contrasts again favor SA, extending the finding beyond the topic and approach choices of the primary design.

Our contribution is a controlled, counterbalanced comparison of topic-matched and approach-matched mathematical supervision, replicated across five models, three seeds, and four target rotations. The results identify the reasoning approach as a more effective axis than topic for selecting transfer-oriented fine-tuning data in the evaluated mathematical setting.

2 Related Work

A foundational finding in cognitive science is that experts and novices organize problems differently: experts classify physics problems by underlying principles (e.g., conservation of energy), while novices group them by surface features (e.g., inclined planes) (Chi et al., 1981). This distinction extends to transfer. Gick & Holyoak (1983) show that analogical transfer depends on abstracting a shared schema from surface-dissimilar examples, and Gentner (1983) formalizes this as structure mapping—the claim that transfer succeeds when relational structure, not the surface objects, is preserved between the source and the target. These results motivate the hypothesis that the shared reasoning approach may support mathematical transfer in LLMs more effectively than shared topic. The cognitive evidence motivates the comparison, but does not predetermine its outcome, since human categorization and learning from gradient updates on worked solutions are different processes.

Within the LLM literature, the closest existing comparison is Didolkar et al. (2024), who derive mathematical skill labels and find that skill-matched exemplars improve in-context prompting over topic-matched alternatives. Our study differs in two respects: the intervention is fine-tuning (parameter

updates) rather than in-context demonstration, and the comparison uses a counterbalanced rotation so that every source cell serves in both roles. Guo et al. (2024) further examines mathematical skill transfer under both in-context learning and instruction tuning. Importantly, while their in-context results show skill-sensitive effects, some instruction-tuning conditions produce similar changes across skills, suggesting that transfer patterns observed in prompting need not persist under parameter updates. This distinction motivates our direct comparison of topic- and approach-matched transfer under source-specific fine-tuning. Our factors are the mathematical topic and solution approach, and our primary endpoint is the paired accuracy difference after source-specific fine-tuning.

More broadly, a growing body of work uses skill structure to guide training-data selection: Skill-it (Chen et al., 2023) models inter-skill dependencies to design training mixtures, LESS (Xia et al., 2024) selects influential examples via gradient-based signals, STAT (He et al., 2026) diagnoses missing skills and synthesizes targeted supervision, and skill-aware reasoning distillation (Zhang et al., 2026) combines selection with explicit skill decomposition. These methods ask *how* to select training data given a skill taxonomy. Our work addresses a complementary question: *which* axis of similarity—topic or reasoning approach—produces greater transfer when the other is held out.

The form and content of mathematical supervision also affect downstream performance more generally: chain-of-thought demonstrations improve prompting (Wei et al., 2022), and synthetically generated questions support supervised adaptation (Yu et al., 2024). On the evaluation side, benchmarks have become increasingly precise about the kind of generalization they test. MATH (Hendrycks et al., 2021) provides competition-level problems with worked solutions; procedural mathematics benchmarks (Saxton et al., 2019) permit controlled compositional splits; GSM-Symbolic (Mirzadeh et al., 2024) probes robustness to variation within fixed problem templates; and OMEGA (Sun et al., 2025) distinguishes exploratory, compositional, and transformative generalization. Our evaluation paradigm differs from these: rather than measuring accuracy on an external benchmark, we hold out one cell of a shared 2×2 construction pipeline and compare how training on each remaining cell transfers to the held-out target.

Finally, our design depends on distinguishing three properties that need not coincide: final-answer correctness, solution-step validity, and approach classification. Lightman et al. (2024) establish that outcome-based and process-based verification target different aspects of mathematical reasoning; this motivates our use of final-answer accuracy as the primary transfer endpoint rather than intermediate-step correctness. Lee et al. (2026) show that lexical and embedding variation in solutions need not track variation in underlying mathematical strategies; this motivates our treatment of similarity measures as descriptive diagnostics rather than substitutes for the approach rubric. The transfer endpoint accepts any valid solution method; adherence to the reference approach is not required.

3 A ROTATIONAL TEST OF MATHEMATICAL TRANSFER

3.1 CELLS AND SOURCE–TARGET RELATIONS

Let $\mathcal{T} = \{P, C\}$ denote probability and combinatorics, and $\mathcal{A} = \{I, D\}$ denote invariant reasoning and double-counting approaches. The four cells are $\mathcal{C} = \mathcal{T} \times \mathcal{A}$. Each cell contains problem statements, worked reference solutions, and final answers. A cell label describes the intended organization of its reference solutions; it does not preclude alternative valid methods for the same problem.

At the conceptual level, invariant reasoning exploits a quantity or relation preserved under specified transformations; probabilistic formulations may use expectation-preserving quantities. Double counting relates two enumerations of the same collection or incidence structure. Boundary cases—including martingale arguments, potential functions, mixed-method solutions, and sums of indicators—can blur the distinction between the two approach categories.

For target $\tau = (t, a)$, write $\bar{t}$ and $\bar{a}$ for the complementary topic and approach. The two comparison sources are

$$s_{\mathrm{SA}}(\tau) = (\bar{t}, a), \qquad s_{\mathrm{ST}}(\tau) = (t, \bar{a}). \tag{1}$$

A separate checkpoint is trained on each source cell. For a given target, neither the SA nor the ST checkpoint has been trained on that cell. Source and target roles rotate across the experiment; no checkpoint is trained on the union of the four cells for the primary comparison.

3.2 Outcome and estimand

Let $A_{m,s}(\sigma \to \tau)$ be final-answer accuracy, in percent, on target τ for model m and seed s after training on source σ. The target-specific contrast is

$$\delta_{m,s,\tau} = A_{m,s}(s_{\mathrm{SA}}(\tau) \to \tau) - A_{m,s}(s_{\mathrm{ST}}(\tau) \to \tau). \tag{2}$$

Our model-level estimand equally averages four targets and three seeds:

$$\bar{\delta}_m = \frac{1}{12} \sum_{s \in \{7,17,27\}} \sum_{\tau \in \mathcal{C}} \delta_{m,s,\tau}. \tag{3}$$

For brevity, we refer to the first and second terms in (2) as A_{SA} and A_{ST}, respectively. We also report $\Delta_{\mathrm{SA}} = A_{\mathrm{SA}} - A_{\mathrm{base}}$ and $\Delta_{\mathrm{ST}} = A_{\mathrm{ST}} - A_{\mathrm{base}}$. The target-specific contrast can equivalently be written as $\delta = \Delta_{\mathrm{SA}} - \Delta_{\mathrm{ST}}$. A positive contrast $\Delta_{\mathrm{SA}} - \Delta_{\mathrm{ST}}$ does not imply that both conditions improve over the base model; SA can exceed ST even when one or both degrade absolute accuracy. The five-model mean summarizes the evaluated suite and is not a population-level estimate.

3.3 What rotation controls

Suppose a transfer score admits the decomposition

$$A(\sigma \to \tau) = \mu + u_\sigma + v_\tau + w_{\sigma,\tau}, \tag{4}$$

where u_σ and v_τ are additive source and target effects, respectively. The target term cancels within each SA–ST comparison. Each source appears once under each relation, so the u_σ terms cancel after averaging over targets. Under the additive model, a source cell that is uniformly better, easier, or noisier cannot, on its own, produce the equally weighted aggregate contrast. What rotation does not eliminate are source–target interactions $w_{\sigma,\tau}$, including statement-level resemblance under embeddings or TF–IDF. Section 6 reports those checks. Rotation controls for fixed source identity, not for every leftover.

4 Dataset and experimental protocol

4.1 Construction, answer auditing, and labels

The dataset contains 2,000 problems, 500 in each cell. Candidates were generated using gpt-oss-120b at high reasoning effort (OpenAI, 2025), conditioned on topic and approach and accompanied by explicit reference-solution steps. Each candidate was then re-solved eight times without an approach instruction by the generator, and three large LLM judges, namely GLM-5.3, Deepseek-V4-Pro, and GPT5.6-terra, independently labeled the approaches of the resulting solutions. A problem was retained only if all three judges agreed on its approach label and a token-similarity criterion was satisfied. This procedure selected problems compatible with the generator's unconstrained behavior—problems whose natural solution route matches the intended approach—rather than solutions forced to follow a prescribed method.

To verify the correctness of the final answers, GLM-5.3 and DeepSeek-V4-Pro were each asked to solve every problem twice without access to the reference solution. A problem was retained only when the solvers agreed on the answer. During further auditing, 83 of the 2,000 entries (4.15%) were found to have inconsistent answers across the solvers; these entries were discarded and replaced using the same procedure. Finally, 50 randomly selected problems were manually reviewed, and all were found to be correct. All numerical results reported here use the final validated dataset.

4.2 Models, training, and evaluation

The primary suite comprises five base or pretrained checkpoints: Qwen3-4B-Base (Qwen Team, 2025), Meta-Llama-3-8B (Meta AI, 2024), DeepSeek-Math-7B-Base (Shao et al., 2024), Mistral-7B-v0.3 (Mistral AI, 2024), and Gemma-3-4B-pt (Gemma Team, 2025). This selection combines general-purpose and math-specialized pretraining while remaining at the base/pretrained level. Exact checkpoint identifiers are listed in Appendix E.

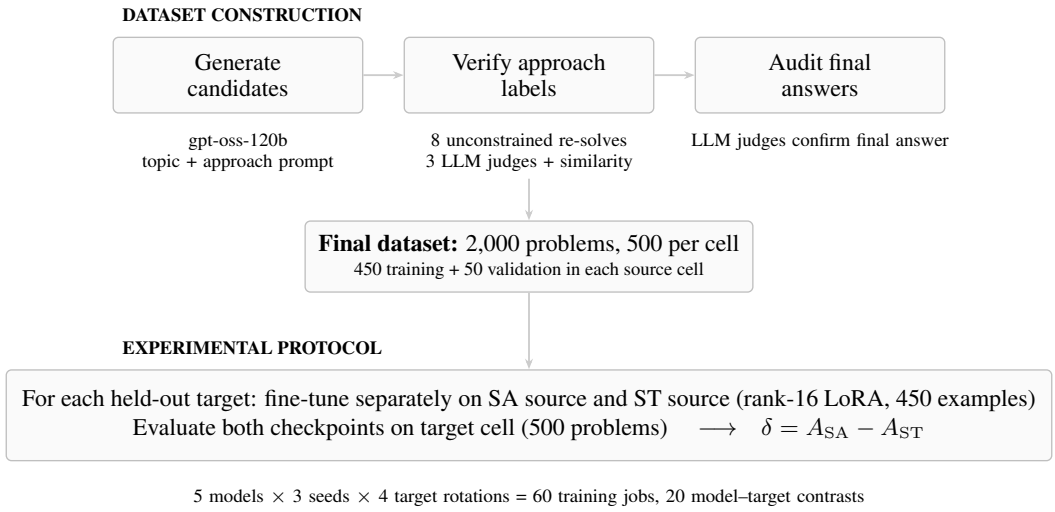


Figure 2: **Dataset construction and experimental protocol.** *Top:* problems are generated with topic-and-approach conditioning, verified via unconstrained re-solution and multi-judge agreement, and corrected through final-answer auditing. *Middle:* the final dataset contains 500 problems per cell, split 450/50 for training and validation. *Bottom:* for each of four target rotations, a separate checkpoint is fine-tuned on the SA source and the ST source and evaluated on the held-out target; the primary outcome is their accuracy difference δ.

Each model is adapted with rank-16 LoRA (Hu et al., 2021) on reference solutions, using 450 examples from one source cell and seeds 7, 17, and 27. The remaining 50 source-cell items serve as validation. Cross-cell evaluation uses all 500 items of each held-out target. A given item may appear in training for one rotational run and in evaluation for another, but never in the checkpoint being evaluated on its own cell.

This yields $5 \times 3 \times 4 = 60$ source-training jobs, all completed. Each model–seed aggregate pools four target contrasts, producing 15 aggregates. Final-answer accuracy is the outcome; the evaluator does not require the student to reproduce the reference wording or approach.

4.3 Uncertainty and reporting

Accuracy differences are reported in percentage points. We compute 95% confidence intervals using paired, cell-stratified percentile bootstraps over evaluation problems. For multi-seed estimates, per-problem outcomes are first averaged across seeds before resampling. The resulting intervals quantify uncertainty due to the finite evaluation set conditional on the trained checkpoints; they do not incorporate training-seed variability. Seed-to-seed variation is therefore reported separately through the per-seed SA–ST contrasts. Full details of the interval procedure are provided in Appendix F.

5 Results

5.1 SA exceeds ST across all five models

The model-level SA–ST advantages are $+16.2$ points for Qwen, $+10.5$ for Llama, $+10.2$ for DeepSeek, $+9.1$ for Mistral, and $+8.2$ for Gemma (Figure 3). Every reported 95% confidence interval lies above zero, and the equally weighted mean across models is $+10.8$ points. All 15 model–seed aggregate contrasts are also positive (Appendix A).

The result is not driven by a few favorable comparisons. All 20 seed-pooled model–target contrasts favor SA, and each reported target-level interval excludes zero. The smallest contrast is Gemma on combinatorics/double counting at $+2.8$ points $[1.6, 4.1]$; the largest is Qwen on combinatorics/invariant at $+27.9$ points $[24.2, 31.8]$. The SA advantage therefore holds at every level of disaggregation—across models, seeds, and individual target cells.

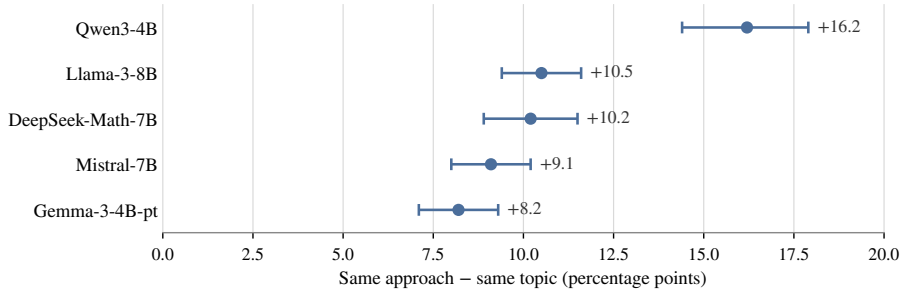


Figure 3: **Approach-matched supervision transfers better in all five models.** Each point averages four held-out targets and seeds 7, 17, and 27. Error bars are 95% CIs; right labels show point estimates. The mean across models is $+10.8$ percentage points.

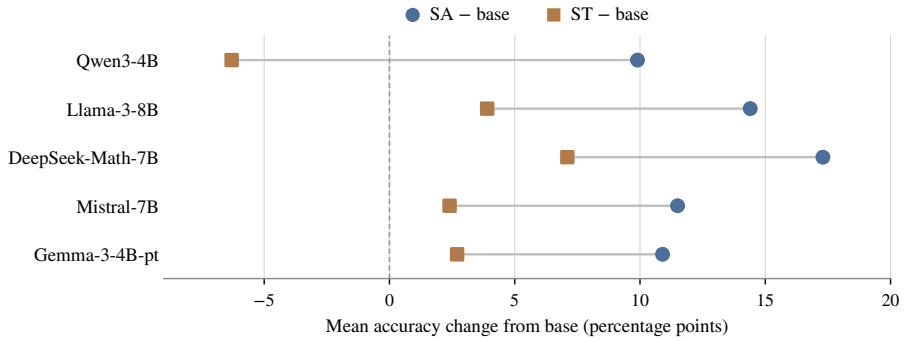


Figure 4: **Positive and negative transfer coexist with the same ordering.** Points are descriptive means of the four target-wise changes, pooled over three seeds. SA improves average performance in every model; ST improves it in four and lowers it in Qwen. Full target-level CIs appear in Appendix A.

5.2 The advantage is not only less negative transfer

A natural objection is that SA might outperform ST simply because topic-matched training is harmful. Figure 4 separates the SA–ST contrast from absolute changes relative to the unadapted base model.

For Llama, DeepSeek, Mistral, and Gemma, both SA and ST improve accuracy over the base in all four targets—but SA improves it substantially more. Across these four models, SA gains average roughly 11–17 percentage points while ST gains average 2–7 points. The SA advantage therefore persists where adaptation is beneficial under both relations, not only where ST degrades performance.

Qwen provides a complementary regime: SA averages $+9.9$ points over base, while ST averages -6.3 points, with negative ST transfer in every target. This demonstrates that topic matching does not guarantee beneficial cross-approach adaptation. Across all 20 model–target combinations, SA–base is positive in 19; the single exception is Qwen on probability/double counting (-0.3 points $[-4.5, 3.8]$), where SA still exceeds ST by $+6.3$ points.

Low base accuracies in some models should not be interpreted as absence of pretrained mathematical knowledge, since base performance depends on prompting format, generation strategy, and answer extraction. The finding concerns improved target accuracy after adaptation, not learning from a zero-knowledge starting point.

5.3 Target-wise effects are consistent but heterogeneous

Figure 5 displays all 20 contrasts. Averaged across models, the SA–ST gaps are 12.0 points for probability/invariant, 5.6 for probability/double counting, 19.5 for combinatorics/invariant, and 6.2 for combinatorics/double counting. Combinatorics/invariant is the strongest target for every model, but both double-counting targets retain positive contrasts throughout the suite.

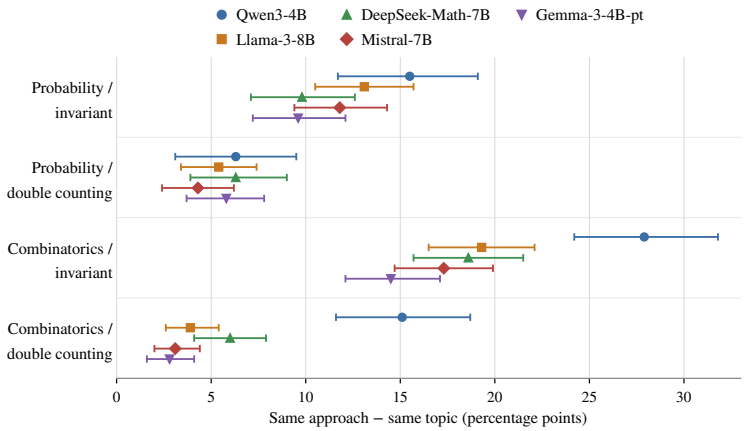


Figure 5: **Every target favors SA.** All points pool three seeds. Error bars are 95% CIs. Invariant targets show larger contrasts, particularly in combinatorics, but the ordering is positive for both approaches in both topics. These are target-wise effects, not estimates of an intrinsic transferability ranking among approaches.

This heterogeneity does not undermine the aggregate finding but does require careful interpretation. A globally stronger invariant source would increase the invariant-target contrast (where it serves as SA) and reduce the double-counting-target contrast (where it serves as ST). Such a source effect cancels in the equally weighted aggregate but not in each target subgroup. The target-wise contrasts therefore do not distinguish source quality from a genuine target–approach interaction, and we do not interpret the pattern as evidence that invariant reasoning is intrinsically more transferable.

5.4 The ordering replicates with different topics and approaches

To test whether the observed ordering is specific to the topic and approach choices in the primary design, we constructed a second 2×2 dataset containing 800 problems. This design crosses number theory and geometry with complement and pigeonhole reasoning. The dataset was constructed and validated using the same generation, approach-verification, answer-auditing, and filtering pipeline as the primary dataset. Each cell contains 200 problems, with 180 used for training and 20 for validation; evaluation on a held-out target uses all 200 target problems. Training uses the same rank-16 LoRA protocol as the primary experiment. We repeated the rotational comparison across the same five base models and training seeds 7, 17, and 27.

The model-level SA–ST advantages are $+16.0$ points for Llama, $+15.7$ for Qwen, $+14.6$ for DeepSeek, $+13.0$ for Gemma, and $+12.0$ for Mistral, averaging $+14.3$ percentage points. All five model-level 95% confidence intervals exclude zero. All 15 model–seed aggregate contrasts and all 20 seed-pooled model–target point estimates favor SA. SA also improves over the base model on average in every model, with gains ranging from $+4.0$ to $+18.1$ points. ST improves average accuracy in four models but reduces it in Qwen.

Statement-similarity diagnostics again favor ST: embedding nearest-neighbor similarity is 0.727 for ST versus 0.544 for SA, and TF–IDF nearest-neighbor similarity is 0.306 versus 0.116. Thus, the SA-over-ST ordering replicates when both topics and approaches change, despite greater measured statement resemblance under ST. Appendix B reports the model-level, per-seed, and target-level results.

5.5 The ordering persists under full fine-tuning

The primary experiments use rank-16 LoRA, raising the possibility that the SA advantage depends on the adaptation method. We therefore repeated the full rotation using full-parameter fine-tuning on all five models at seed 17.

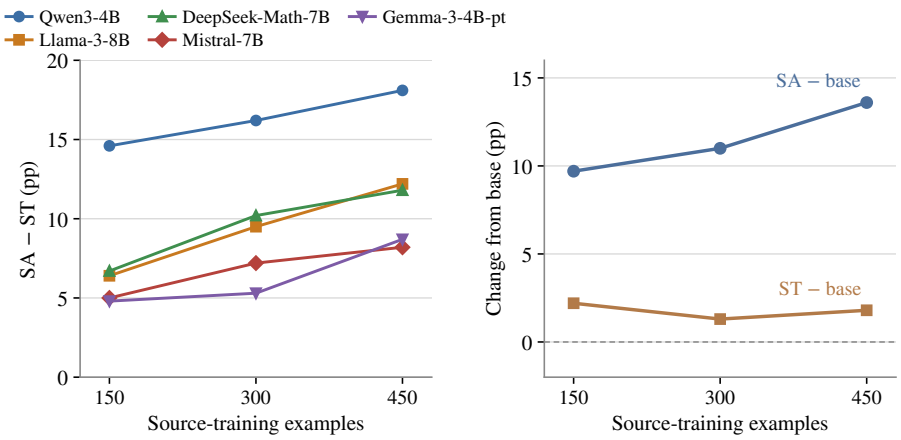


Figure 6: **The ordering holds at every training budget and widens with data.** Left: SA–ST for each model at 150, 300, and 450 source-training examples. Right: suite-mean change from base for each relation. All five models; single seed; epochs are fixed, so optimizer steps grow with the data.

Mistral did not converge under full fine-tuning, so we exclude it from the summaries below and report it separately in Appendix C. For the remaining four models, the SA–ST contrast is positive in all 16 model–target comparisons, with a mean advantage of $+11.1$ points versus $+12.0$ in the matched LoRA runs. The SA-over-ST ordering is therefore not specific to low-rank adaptation.

5.6 The advantage appears at smaller training budgets

The contrasts above use 450 source-training examples. A budget ablation subsamples each source cell into nested, level-balanced subsets of 150, 300, and 450 problems and repeats the full rotation on all five models. It rests on weaker evidence than Section 5 because it uses a single seed. We therefore read it primarily for directional robustness rather than precise effect-size estimation.

Every model aggregate favors SA at 150, 300, and 450 exmaples. All five models increase monotonically, lifting the suite mean from $+7.5$ to $+9.7$ to $+11.4$ points across the three budgets (Figure 6, left).

Separating the two relations shows where the widening originates (Figure 6, right). Averaged over the suite, SA–base grows from $+9.7$ to $+13.3$ points, while ST–base stays small and relatively flat across budgets ($+2.2$, $+1.3$, $+1.9$). Additional same-topic supervision does not convert into target accuracy at these budgets, whereas additional same-approach supervision does. Appendix D reports the per-model values.

6 Statement similarity and the transfer ordering

If the SA advantage were driven by greater surface resemblance between source and target problems, we would expect SA sources to be more similar to targets than ST sources. The opposite holds. Using Qwen3-Embedding representations (Zhang et al., 2025) and TF–IDF, we compare source–target statement similarity under both all-pairs averages and nearest-neighbor comparisons. For normalized embeddings $e(x)$, the directional nearest-neighbor similarity is

$$\mathrm{NN}(T, S) = \frac{1}{|T|} \sum_{x \in T} \max_{z \in S} \langle e(x), e(z) \rangle. \tag{5}$$

Across cells, statement embedding similarity favors ST over SA for all pairs (0.458 versus 0.437) and nearest neighbors (0.739 versus 0.705). Statement TF–IDF nearest-neighbor similarity likewise favors ST (0.260 versus 0.232). For probability/invariant, the nearest out-of-cell statement is in the ST cell for 63.0% of items and in the SA cell for 35.4%. The model is evaluated on statements, so these are the relevant resemblance cues. They favor ST; transfer favors SA.

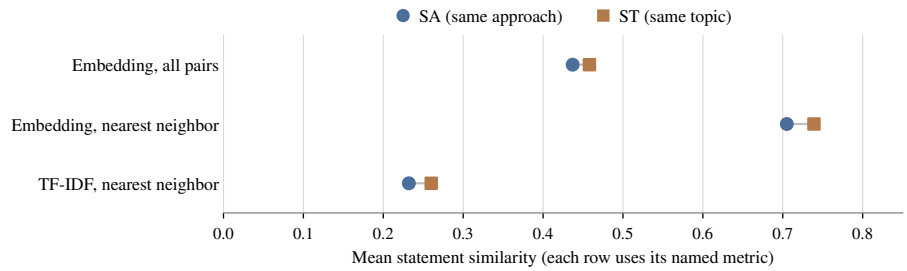


Figure 7: **Greater statement resemblance does not track greater transfer.** ST exceeds SA on every reported statement measure. NN denotes nearest neighbor.

7 Discussion and Limitations

The central finding is that approach-matched supervision consistently outperforms topic-matched supervision in cross-cell mathematical transfer. The result holds across five base models, three training seeds, and all target rotations, yielding positive SA–ST contrasts in all 40 seed-pooled model–target comparisons. Because the second design changes both the mathematical topics and reasoning approaches, the ordering is not tied to a single factorial construction.

The designs were built as controlled tests in which topic sharing and approach sharing can be varied independently while remaining mathematically meaningful. This makes the repeated SA-over-ST ordering across four topics and four approaches stronger evidence than a result from one topic pair alone. The similarity diagnostics reinforce this interpretation: in both designs, ST sources are more similar to targets under lexical and embedding measures, yet they transfer less effectively.

The practical implication is that mathematical fine-tuning data should not be organized by topic alone; reasoning approach should be considered explicitly when selecting supervision for transfer. The study is still a controlled investigation rather than an exhaustive survey of mathematics, and broader domains and solution structures would provide additional tests of generality. The synthetic construction also does not eliminate every possible source–target interaction, and final-answer accuracy does not identify the internal mechanism of transfer.

8 Conclusion

Across two independently constructed 2×2 designs, five base models, and three training seeds per design, approach-matched supervision outperforms topic-matched supervision in every one of the 40 seed-pooled model–target comparisons. The advantage is large and replicable: it averages 10.8 percentage points in the first design and 14.3 points in the second, despite changing both the mathematical topics and the reasoning approaches. The ordering also holds across all model–seed aggregates and every target cell in both designs.

This advantage is not explained by greater surface resemblance. Across both datasets, same-topic sources are more similar to their targets under embedding and lexical measures, yet they transfer less effectively. Thus, mathematical transfer follows shared reasoning approach more consistently than shared topic in our experiments.

Taken together, these results provide evidence that reasoning approach is a stronger organizing principle than topic for selecting transfer-oriented mathematical supervision in LLMs. Rather than treating mathematical domain as the primary criterion for choosing training examples, our findings suggest that the structure of the solution method should play a central role in how mathematical fine-tuning data are organized and selected.

AI USE STATEMENT

AI systems were used to generate mathematical problems and reference solutions, produce unconstrained re-solutions, judge approach labels, and audit final answers. AI assisted with manuscript rewriting, literature research, reference checking, and preparation of LaTeX and figure code.

REPRODUCIBILITY STATEMENT

Section 3 defines the estimand and rotation; Appendices A and B provide the complete summary statistics for the two designs.

REFERENCES

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36:36000–36040, 2023.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152, 1981.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2318d75a06437eaa257737a5cf3ab83c-Abstract-Conference.html.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

Mary L Gick and Keith J Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983.

Siyuan Guo, Aniket Didolkar, Nan Rosemary Ke, Anirudh Goyal, Ferenc Huszár, and Bernhard Schölkopf. Learning beyond pattern matching? assaying mathematical understanding in llms. *arXiv preprint arXiv:2405.15485*, 2024.

Yinghui He, Abhishek Panigrahi, Yong Lin, and Sanjeev Arora. STAT: Skill-targeted adaptive training. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/17d60fef592086d1a5cb136f1946df59-Abstract-Conference.html.

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 Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Sangmook Lee, Minbeom Kim, Jeonghye Kim, Dohyung Kim, Sojeong Rhee, and Kyomin Jung. Are we measuring strategy or phrasing? the gap between surface- and approach-level diversity in llm math reasoning. *arXiv preprint arXiv:2606.29985*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, volume 2024, pp. 39578–39601, 2024.

Meta AI. Meta-Llama-3-8B model card, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-8B.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. URL https://arxiv.org/abs/2410.05229.

Mistral AI. Mistral-7B-v0.3 model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-v0.3.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://openai.com/index/gpt-oss-model-card/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019. URL https://arxiv.org/abs/1904.01557.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. OMEGA: Can LLMs reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *arXiv preprint arXiv:2506.18880*, 2025. URL https://arxiv.org/abs/2506.18880.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024. URL https://proceedings.mlr.press/v235/xia24c.html.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c400474e8a36d0812fdee52739288b12-Abstract-Conference.html.

Lechen Zhang, Yunxiang Zhang, Wei Hu, and Lu Wang. Skill-aware data selection and fine-tuning for data-efficient reasoning distillation. *arXiv preprint arXiv:2601.10109*, 2026. URL https://arxiv.org/abs/2601.10109.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A COMPLETE RESULTS

Tables use seeds 7, 17, and 27. Intervals are those supplied with the summaries. Point estimates and differences were rounded separately, so a displayed subtraction can differ by 0.1.

Table 1: Model-level SA–ST contrasts, pooled over four targets and three seeds.

Model	SA–ST (pp)	95% CI
Qwen3-4B-Base	+16.2	$[+14.4, +17.9]$
Meta-Llama-3-8B	+10.5	$[+9.4, +11.6]$
DeepSeek-Math-7B	+10.2	$[+8.9, +11.5]$
Mistral-7B-v0.3	+9.1	$[+8.0, +10.2]$
Gemma-3-4B-pt	+8.2	$[+7.1, +9.3]$
Equal-weight model mean	+10.8	—

Table 2: Per-seed SA–ST, averaged over the four targets.

Model	Seed 7	Seed 17	Seed 27
Qwen3-4B-Base	+15.8	+16.5	+16.2
Meta-Llama-3-8B	+10.1	+11.4	+9.9
DeepSeek-Math-7B	+11.4	+10.4	+8.8
Mistral-7B-v0.3	+9.5	+9.1	+8.8
Gemma-3-4B-pt	+7.4	+9.6	+7.5

Table 3: Absolute accuracy (%), pooled over the three seeds. P = probability; C = combinatorics; I = invariant; D = double counting.

Model	Target	Base	SA	ST
Qwen3-4B	P / I	28.4	37.3	21.9
	P / D	45.2	44.9	38.5
	C / I	28.8	50.8	22.9
	C / D	28.0	37.1	22.0
Llama-3-8B	P / I	2.2	17.9	4.8
	P / D	0.4	12.2	6.8
	C / I	0.8	25.7	6.4
	C / D	0.4	5.5	1.6
DeepSeek-Math-7B	P / I	3.8	18.3	8.5
	P / D	4.0	22.1	15.8
	C / I	0.0	26.9	8.3
	C / D	0.0	9.7	3.7
Mistral-7B	P / I	3.4	16.1	4.3
	P / D	1.4	9.7	5.4
	C / I	1.4	23.0	5.7
	C / D	0.6	4.0	0.9
Gemma-3-4B-pt	P / I	1.8	14.1	4.5
	P / D	0.8	10.3	4.5
	C / I	0.4	18.8	4.3
	C / D	0.2	3.7	0.9

Table 4: Target-level accuracy differences (percentage points) with 95% CIs, pooled over seeds 7, 17, and 27.

Target	SA–base [95% CI]	ST–base [95% CI]	SA–ST [95% CI]
Qwen3-4B-Base			
P / I	$+8.9$ $[+4.9, +13.0]$	-6.5 $[-10.5, -2.5]$	$+15.5$ $[+11.7, +19.1]$
P / D	-0.3 $[-4.5, +3.8]$	-6.7 $[-10.5, -2.7]$	$+6.3$ $[+3.1, +9.5]$
C / I	$+22.0$ $[+17.1, +27.0]$	-5.9 $[-10.0, -1.9]$	$+27.9$ $[+24.2, +31.8]$
C / D	$+9.1$ $[+4.7, +13.2]$	-6.0 $[-9.8, -2.3]$	$+15.1$ $[+11.6, +18.7]$
Meta-Llama-3-8B			
P / I	$+15.7$ $[+13.1, +18.3]$	$+2.6$ $[+0.7, +4.3]$	$+13.1$ $[+10.5, +15.7]$
P / D	$+11.8$ $[+9.8, +13.9]$	$+6.4$ $[+4.9, +8.0]$	$+5.4$ $[+3.4, +7.4]$
C / I	$+24.9$ $[+22.0, +27.8]$	$+5.6$ $[+4.1, +7.2]$	$+19.3$ $[+16.5, +22.1]$
C / D	$+5.1$ $[+3.7, +6.7]$	$+1.2$ $[+0.3, +2.0]$	$+3.9$ $[+2.6, +5.4]$
DeepSeek-Math-7B			
P / I	$+14.5$ $[+11.5, +17.6]$	$+4.7$ $[+2.3, +7.1]$	$+9.8$ $[+7.1, +12.6]$
P / D	$+18.1$ $[+14.9, +21.5]$	$+11.8$ $[+9.1, +14.5]$	$+6.3$ $[+3.9, +9.0]$
C / I	$+26.9$ $[+24.0, +29.7]$	$+8.3$ $[+6.6, +9.9]$	$+18.6$ $[+15.7, +21.5]$
C / D	$+9.7$ $[+7.9, +11.5]$	$+3.7$ $[+2.7, +4.7]$	$+6.0$ $[+4.1, +7.9]$
Mistral-7B-v0.3			
P / I	$+12.7$ $[+10.1, +15.3]$	$+0.9$ $[-1.1, +2.7]$	$+11.8$ $[+9.4, +14.3]$
P / D	$+8.3$ $[+6.2, +10.3]$	$+4.0$ $[+2.4, +5.6]$	$+4.3$ $[+2.4, +6.2]$
C / I	$+21.6$ $[+18.8, +24.5]$	$+4.3$ $[+2.7, +5.9]$	$+17.3$ $[+14.7, +19.9]$
C / D	$+3.4$ $[+2.0, +4.8]$	$+0.3$ $[-0.5, +1.0]$	$+3.1$ $[+2.0, +4.4]$
Gemma-3-4B-pt			
P / I	$+12.3$ $[+9.9, +14.7]$	$+2.7$ $[+1.1, +4.1]$	$+9.6$ $[+7.2, +12.1]$
P / D	$+9.5$ $[+7.4, +11.5]$	$+3.7$ $[+2.4, +4.9]$	$+5.8$ $[+3.7, +7.8]$
C / I	$+18.4$ $[+15.9, +21.0]$	$+3.9$ $[+2.7, +5.1]$	$+14.5$ $[+12.1, +17.1]$
C / D	$+3.5$ $[+2.3, +4.8]$	$+0.7$ $[+0.1, +1.4]$	$+2.8$ $[+1.6, +4.1]$

B SECOND-DESIGN REPLICATION

The second dataset contains 800 problems and crosses number theory and geometry with complement and pigeonhole reasoning. It was constructed using the same generation, unconstrained re-solution, approach-label verification, filtering, and final-answer auditing framework as the primary dataset. The only changes are the topic–approach definitions and dataset size.

Each of the four cells contains 200 problems. For a source cell, 180 problems are used for training and 20 for validation. Evaluation on a held-out target cell uses all 200 target problems. Training uses the same rank-16 LoRA configuration as the primary experiment, with seeds 7, 17, and 27. The same rotational source–target assignment defined in Section 3 is applied to all four cells.

Table 5 reports model-level contrasts averaged equally over the four target cells and three seeds. Table 6 reports the individual seed-level contrasts, and Table 7 reports seed-pooled target accuracies and contrasts. All 15 model–seed aggregate contrasts and all 20 seed-pooled model–target point estimates favor SA. Confidence intervals in Table 5 are model-level intervals; target-level intervals are not reported in these tables. Point estimates and differences are rounded separately.

Table 5: Second-design model-level accuracy differences, in percentage points, averaged over four target cells and three seeds. Confidence intervals are for SA–ST.

Model	SA–ST	95% CI	SA–base	ST–base
Meta-Llama-3-8B	$+16.0$	$[+14.0, +17.8]$	$+18.1$	$+2.1$
Qwen3-4B-Base	$+15.7$	$[+13.0, +18.3]$	$+4.0$	-11.7
DeepSeek-Math-7B-Base	$+14.6$	$[+12.6, +16.7]$	$+17.7$	$+3.1$
Gemma-3-4B-pt	$+13.0$	$[+11.2, +14.8]$	$+14.9$	$+1.9$
Mistral-7B-v0.3	$+12.0$	$[+10.2, +13.7]$	$+13.1$	$+1.1$

Table 6: Second-design SA–ST contrasts by training seed, averaged over the four target cells. Values are in percentage points.

Model	Seed 7	Seed 17	Seed 27
Meta-Llama-3-8B	$+17.1$	$+16.1$	$+14.6$
Qwen3-4B-Base	$+14.2$	$+16.2$	$+16.5$
DeepSeek-Math-7B-Base	$+15.1$	$+13.4$	$+15.2$
Gemma-3-4B-pt	$+15.1$	$+10.6$	$+13.1$
Mistral-7B-v0.3	$+10.0$	$+11.9$	$+14.1$

Table 7: Second-design target-level accuracy (%) and SA–ST contrasts (percentage points), pooled over seeds 7, 17, and 27. NT = number theory; Geo = geometry; Comp = complement; Pig = pigeonhole. Differences are computed before rounding.

Model	Target	Base	SA	ST	SA–ST
Meta-Llama-3-8B	NT / Comp	1.0	5.7	0.7	+5.0
	NT / Pig	0.0	25.5	4.5	+21.0
	Geo / Comp	0.0	3.8	1.5	+2.3
	Geo / Pig	0.5	38.8	3.3	+35.5
Qwen3-4B-Base	NT / Comp	37.5	41.3	38.0	+3.3
	NT / Pig	49.0	53.2	29.2	+24.0
	Geo / Comp	27.0	38.5	28.3	+10.2
	Geo / Pig	65.5	61.8	36.7	+25.2
DeepSeek-Math-7B-Base	NT / Comp	0.0	7.3	2.7	+4.7
	NT / Pig	1.0	25.2	4.2	+21.0
	Geo / Comp	1.0	6.3	4.0	+2.3
	Geo / Pig	0.5	34.3	4.0	+30.3
Gemma-3-4B-pt	NT / Comp	0.0	4.3	1.5	+2.8
	NT / Pig	0.0	23.3	1.5	+21.8
	Geo / Comp	0.5	4.2	1.5	+2.7
	Geo / Pig	0.0	28.2	3.7	+24.5
Mistral-7B-v0.3	NT / Comp	0.5	2.3	1.5	+0.8
	NT / Pig	1.5	18.7	3.3	+15.3
	Geo / Comp	0.5	3.7	0.5	+3.2
	Geo / Pig	1.0	31.2	2.5	+28.7

C FULL FINE-TUNING ABLATION

This check updates all language-model weights instead of rank-16 LoRA adapters, on the same five models, at seed 17. Settings otherwise follow the LoRA runs (three epochs, effective batch 16, validation-loss checkpointing) at lr 10^{-5}. Models up to 4B use AdamW; the 7–8B models use Adafactor.

Table 8: Full fine-tuning, averaged over the four target cells. The LoRA column is the matched low-rank run. †Mistral did not converge and is excluded from the mean.

Model	SA–ST	SA–base	ST–base	LoRA SA–ST
Qwen3-4B-Base	+19.8	+11.0	-8.8	+16.5
Meta-Llama-3-8B	+6.9	+7.5	+0.6	+11.4
DeepSeek-Math-7B-Base	+12.2	+19.8	+7.6	+10.4
Gemma-3-4B-pt	+5.4	+8.0	+2.6	+9.6
Mean of the four above	+11.1	—	—	+12
Mistral-7B-v0.3†	+0.5	+0.4	-0.1	+9.1

Table 9: Target-level SA–ST contrasts under full fine-tuning (percentage points).

Model	P / I	P / D	C / I	C / D
Qwen3-4B-Base	+28.0	+13.0	+25.0	+13.0
Meta-Llama-3-8B	+8.2	+1.0	+16.0	+2.4
DeepSeek-Math-7B-Base	+10.6	+8.8	+23.6	+5.6
Gemma-3-4B-pt	+10.4	+0.4	+10.4	+0.4
Mistral-7B-v0.3†	+0.4	+0.8	+0.4	+0.4

Mistral's four contrasts are $+0.4$ to $+0.8$ points and its SA accuracies fall below base on two targets; the row is reported for completeness. Excluding it, the mean is $+11.1$ against $+12.0$ for the matched LoRA runs.

D TRAINING-SET SIZE ABLATION

Each source cell is subsampled into nested, level-balanced subsets ($150 \subset 300 \subset 450$), with the same 50 validation items and evaluation on all 500 target items. Every size trains for three epochs, so the step budget grows with the data (about 30, 57, and 87 steps).

Table 10: Accuracy differences (percentage points) by number of source-training examples, averaged over the four target cells.

	SA–ST			SA–base			ST–base		
Model	150	300	450	150	300	450	150	300	450
Qwen3-4B-Base	+14.6	+16.2	+16.5	+11.1	+9.5	+10.2	−3.5	−6.7	−6.3
Meta-Llama-3-8B	+6.4	+9.5	+11.4	+10.5	+12.7	+14.9	+4.1	+3.1	+3.5
DeepSeek-Math-7B-Base	+6.7	+10.2	+10.4	+12.9	+15.7	+17.7	+6.2	+5.5	+7.3
Mistral-7B-v0.3	+5.0	+7.2	+9.1	+7.0	+8.9	+11.7	+2.0	+1.8	+2.6
Gemma-3-4B-pt	+4.8	+5.3	+9.6	+7.1	+8.0	+11.9	+2.2	+2.7	+2.3
Suite mean	+7.5	+9.7	+11.4	+9.7	+11.0	+13.3	+2.2	+1.3	+1.9

E CHECKPOINTS AND PROTOCOL

Table 11: Public checkpoints used in the primary suite.

Model	Checkpoint
Qwen3-4B-Base	`Qwen/Qwen3-4B-Base`
Meta-Llama-3-8B	`meta-llama/Meta-Llama-3-8B`
DeepSeek-Math-7B-Base	`deepseek-ai/deepseek-math-7b-base`
Mistral-7B-v0.3	`mistralai/Mistral-7B-v0.3`
Gemma-3-4B-pt	`google/gemma-3-4b-pt`

Each cell contains 500 problems. For a source cell, 450 problems are used for training and the remaining 50 for validation. Evaluation on a held-out target cell uses all 500 target problems. The primary setting uses rank-16 LoRA with training seeds 7, 17, and 27, yielding $5 \times 3 \times 4 = 60$ source-training jobs.

Approach labels correspond to the two construction categories defined in Section 3. Invariant reasoning uses a quantity or relation preserved under the transformations in the problem, whereas double counting derives a result by counting the same collection or incidence structure in two ways. A problem was retained only when all three approach judges agreed on its label.

F STATISTICAL ANALYSIS

Confidence intervals. We use a paired, cell-stratified nonparametric percentile bootstrap over evaluation problems. Within each target cell, evaluation problems are resampled with replacement, preserving the original number of evaluation problems in that cell. For the SA–ST contrast, let

$$d_i = y_i^{\mathrm{SA}} - y_i^{\mathrm{ST}},$$

where y_i^{SA} and y_i^{ST} are the binary strict-correct outcomes for problem i. For each bootstrap replicate, we resample the evaluation problems within that target cell with replacement and compute the mean of d_i. The 95% confidence interval is given by the 2.5th and 97.5th percentiles of the resulting bootstrap distribution. SA–base and ST–base intervals are computed analogously. Because SA, ST, and base are evaluated on the same target problems, outcomes remain paired within each resampled problem.

For quantities averaged across the four target cells, each target cell is resampled independently within its own evaluation set. The four cell-level means are then equally averaged within each bootstrap replicate, and the confidence interval is obtained from the percentiles of this replicate-level distribution.

For results pooled across seeds 7, 17, and 27, each problem's binary correctness outcome is first averaged across seeds. The resulting per-problem outcome therefore takes a value in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Bootstrap resampling is then performed over these seed-averaged problem-level outcomes using the same paired, cell-stratified procedure.

The primary analyses use 4,000 bootstrap replicates. Bootstrap sampling is initialized with `np.random.default_rng(7)` for reproducibility. Reported confidence intervals quantify uncertainty arising from the finite set of evaluation problems conditional on the trained checkpoints. Training runs themselves are not resampled, and seeds are averaged rather than resampled; consequently, the intervals do not represent training-seed uncertainty. We report the individual seed-level SA–ST contrasts separately to characterize variation across training seeds.

Reported 95% confidence intervals are pointwise and are not adjusted for multiple comparisons.

Base predictions across seeds. Base-model predictions are reused across training seeds because base evaluation uses deterministic greedy decoding and does not depend on the fine-tuning seed.

G SIMILARITY DIAGNOSTICS

Table 12: Statement similarity by relation. Centred embeddings subtract the dataset-mean vector. Metrics are not interchangeable.

Metric	Within	SA	ST
TF–IDF, all pairs	0.054	0.023	0.026
TF–IDF, NN	0.538	0.232	0.260
Embedding, all pairs	0.537	0.437	0.458
Embedding, NN	0.898	0.705	0.739
Centred emb., all pairs	0.153	−0.031	0.004
Centred emb., NN	0.811	0.445	0.518

Table 13: Share of problems whose nearest out-of-cell statement neighbor lies in the SA or ST cell (Qwen3 embeddings). The leftover share matches neither topic nor approach.

Cell	SA	ST
P / I	35.4%	63.0%
C / I	52.8%	47.0%
P / D	45.8%	54.2%
C / D	33.2%	64.4%