# The Emperor's New Reasoning: Format Imitation Overshadows Genuine Mathematical Understanding in SFT

**Anonymous ACL submission**

## Abstract

Recent advances in large language models (LLMs) have yielded impressive gains on mathematical reasoning benchmarks via supervised fine-tuning (SFT). However, the brittleness of these models under input perturbations has cast doubt on whether such improvements reflect genuine reasoning abilities or merely superficial alignment with expected output formats. We investigate the mechanisms behind SFT improvements in small-scale LLMs, addressing four key questions: (1) Are performance gains primarily due to format alignment rather than reasoning? (2) Can high-quality supervision encourage genuine reasoning? (3) Does scaling data shift learning from format alignment to deeper reasoning? (4) Are format alignment gains consistent across model sizes and architectures? Through controlled experiments, we find that most performance improvements arise from format alignment rather than genuine reasoning enhancement. Moreover, SFT's effectiveness is strongly influenced by the alignment between the base model's inductive biases and the teacher model's output distribution, rather than the teacher's raw strength. Finally, scaling up training data offers diminishing returns and does not fundamentally alter the model's reasoning behavior. These findings suggest that current SFT practices may overestimate the reasoning abilities of LLMs and underscore the need for more rigorous evaluation methods.

## 1 Introduction

Mathematical reasoning has emerged as a critical benchmark for evaluating the logical thinking capabilities of large language models (LLMs). With the growing scale of models and advances in training techniques, LLMs have demonstrated impressive performance on challenging mathematical benchmarks such as OlympiadBench and Omni-Math, leading to the development of powerful models like OpenAI-o1(Jaech et al., 2024), OpenAI-o3 (OpenAI, 2025), Kimi1.5 (Team et al., 2025) and DeepSeek-R1(Guo et al., 2025). Interestingly, recent work (Yang et al., 2024; Guan et al., 2025; Team, 2025; Abdin et al., 2025; Luo et al., 2025) has shown that even small-scale models, when equipped with carefully designed training strategies, can match or even surpass the reasoning performance of some proprietary LLMs.

However, a central question remains unresolved: **Do these models truly acquire mathematical reasoning abilities, or are they merely learning to reproduce the superficial patterns of reasoning-like outputs?** Several studies (Li et al., 2024; Gulati et al., 2024; Mirzadeh et al., 2024) have highlighted the fragility of current models under adversarial or perturbed inputs. For example, (Huang et al., 2025) found that introducing hard perturbations to math problems led to substantial drops in accuracy for top-tier models (e.g., -16.49% for O1-mini, -12.9% for Gemini-2.0 flash-thinking), suggesting a reliance on surface-level cues rather than genuine understanding. This paradox gives rise to our core research question: *What underlies the apparent reasoning performance gains in small-scale open-source LLMs?*

Recent research has significantly advanced the reasoning capabilities of **small-scale LLMs** by building upon strong base models, particularly the Qwen2.5-Math-7B model (Ye et al., 2025; Zeng et al., 2025; Li et al., 2025; Ma et al., 2025; Liu et al., 2025; Cui et al., 2025). Leveraging its robust mathematical foundation, these studies have applied a variety of enhancement techniques—including supervised fine-tuning (SFT), reinforcement learning (RL) (Xie et al., 2025; Li et al., 2025), and integration with methods such as Monte Carlo Tree Search (MCTS) and process reward model (PRM) (Guan et al., 2025)—to substantially improve reasoning performance. Among these approaches, supervised fine-tuning (SFT) has proven particularly effective, enabling student models (typically smaller-scale LLMs) to achieve sub-

Example of step-by-step reasoning

**Problem:** If $5x - 3 = 12$, what is the value of $5x + 3$?

**Solution:** To solve for the value of (5x + 3) given the equation (5x – 3 = 12), we can follow these steps:

1. Start with the given equation: 5x – 3 = 12
2. Add 6 to both sides of the equation to isolate the term 5x: 5x – 3 + 6 = 12 + 6
3. Simplifying both sides, we get: 5x + 3 = 18

Therefore, the value of (5x+3) is (\boxed{18}).

Figure 1: This figure illustrates the meaning of "format" as used in this paper. The term refers to the step-by-step solution procedure provided for a given problem. The demonstration samples used in SFT follow this format. The answer only refers to the final output enclosed in the "`\boxed{}`".

stantial improvements through fine-tuning on high-quality data distilled from more capable teacher models (generally larger LLMs). These SFT-based enhancements not only yield significant performance gains over base models but, in some cases, allow compact student models to surpass the capabilities of substantially larger counterparts.

A common characteristic of these methods is their structured output requirement—the generation of explicit, step-by-step solutions designed to mimic human reasoning patterns (as illustrated in Figure 1). We formalize this output *format* as syntactically constrained generation that requires intermediate reasoning steps. Although this approach yields outputs that superficially resemble human reasoning, it raises a crucial question: **Are the observed performance gains indicative of genuine reasoning capabilities, or do they primarily result from alignment with expected reasoning formats?**

We hypothesize that much of the improvement stems not from enhanced reasoning ability, but from format alignment, i.e., models learn to mimic the structural patterns of human derivations without truly internalizing their semantic content. To test this hypothesis, we conduct a systematic empirical study using lightweight SFT, aiming to disentangle performance gains driven by reasoning ability from those driven by superficial pattern imitation.

**RQ1**: Do recent performance improvements in small-scale LLMs during SFT primarily arise from format alignment rather than genuine reasoning enhancement?

**RQ2**: If format alignment dominates, can high-quality supervision (e.g., data distilled from stronger teacher models) help models move beyond format imitation toward genuine reasoning ability?

**RQ3**: As the amount of training data increases, does the primary driver of performance shift from format alignment to deeper reasoning?

**RQ4**: Are format alignment gains consistent across models of varying sizes and architectures?

To address these questions, we design a set of controlled experiments across different model sizes, data sources, and data scales. Our key findings are as follows:

**Result 1** (Section 3.2): Even using a minimal SFT setup, just 10 random format-correct samples (as shown in Figure 1), leads to near or surpassing state-of-the-art improvements (e.g., +49.8% vs SOTA's +44.6% on MATH500; +81.6% vs SOTA's +75.2% on GaokaoEn 2023). This suggests that the improvement is unlikely due to the acquisition of deep reasoning ability under such limited supervision. Instead, the model appears to benefit from learning to imitate the surface structure of step-by-step solutions. Further experiments show that even samples are being disturbed in content (e.g., incorrect answers, garbled text, mismatched solutions), only using 10 samples with correct format can still increase the performance. This further demonstrates the importance of format than that of content.

**Result 2** (Section 3.3): The match between the teacher and student models—in terms of reasoning style and output distribution—is more critical than teacher strength alone. In low-resource SFT settings, compatibility between the teacher's inductive bias and the student's capacity significantly influences knowledge transfer efficiency.

**Result 3** (Section 3.4): While format imitation is highly sample-efficient, its benefits saturate quickly. Achieving true reasoning ability likely requires deeper abstraction and generalization, which cannot be attained through format learning alone. Simply scaling up the number of formatted examples is insufficient to drive continued progress.

**Result 4** (Section 3.5): Competent small models can effectively utilize formatted demonstrations, while weaker models fail to generalize from them, and stronger models show diminishing marginal returns.

## 2 Related Work

**Enhancing Mathematical Reasoning in LLMs**. Recent years have witnessed significant progress in improving the mathematical reasoning capabilities of LLMs. A prominent line of research

leverages Chain-of-Thought (CoT)-based methods (Ling et al., 2023; Magister et al., 2022; Li et al., 2023; Yuan et al., 2024), where models are fine-tuned on specific math QA datasets containing step-by-step reasoning processes to guide coherent derivations. Further advancements extend this paradigm by formalizing reasoning as graph-structured processes (Lei et al., 2023), where nodes represent intermediate steps and edges denote logical dependencies. Techniques like Tree-of-Thought (ToT) (Yao et al., 2023) and Monte Carlo Tree Search (MCTS) (Feng et al., 2023; Gao et al., 2024; Xu, 2023; Xin et al., 2024) exemplify this approach, with (Guan et al., 2025) demonstrating that even smaller LLMs (e.g., 7B) can achieve strong mathematical reasoning through self-evolution within this framework. Additionally, recent RL-based methods (Guo et al., 2025; Xie et al., 2025; Zeng et al., 2025) combine format rewards and answer accuracy rewards to push the state-of-the-art (SOTA) performance further. Concurrently, another strand of research explores low-resource fine-tuning strategies to enhance reasoning efficiency. (Zhou et al., 2023; Li et al., 2025; Muennighoff et al., 2025) reveal that a carefully curated, small dataset (e.g., 1,000 high-quality samples) suffices to elicit high-quality outputs, while (Chen et al., 2025)) demonstrate that small-scale SFT (e.g., 0.072B data) can significantly improve instruction-following capabilities.

**Questioning the True Reasoning Capabilities of LLMs**. The nature of LLM reasoning capabilities remains controversial (Huang et al., 2025; Jiang et al., 2024; Li et al., 2024; Gulati et al., 2024; Srivastava et al., 2024). As shown in (Jiang et al., 2024), LLM reasoning appears to operate through probabilistic pattern matching rather than formal logical reasoning. (Mirzadeh et al., 2024) constructed the GSM-Symbolic dataset using symbolic templates, revealing that numerical variations cause performance degradation across all models. MATH-Perturb (Huang et al., 2025) further demonstrated significant performance drops in state-of-the-art proprietary models (e.g., o1-mini) under hard perturbations. While these studies exposed reasoning fragility through benchmark perturbations, they did not investigate the mechanisms behind observed performance improvements.

Departing from previous approaches, we analyze the fundamental drivers behind state-of-the-art performance improvements in leading small-scale LLMs within extreme low-resource SFT regimes. Our simple SFT experiments demonstrate that structural imitation alone suffices for competitive performance: using merely 10 randomly selected samples with correct formatting (regardless of answer accuracy), the model achieves near-state-of-the-art results across multiple mathematical reasoning benchmarks. This compelling evidence suggests that prevailing reasoning improvements may primarily arise from pattern matching of step-by-step solution formats, rather than the acquisition of genuine mathematical reasoning capabilities.

## 3 Experiments

To investigate whether small-scale LLMs can genuinely acquire reasoning capabilities during the SFT stage, we conduct a series of controlled experiments using a distilled dataset. Specifically, we fine-tune several commonly used base LLMs using a distilled dataset consisting of only 10 step-by-step reasoning examples per dataset, sourced from advanced reasoning models. By limiting the amount of training data and focusing exclusively on reasoning demonstrations, our goal is to examine whether such sparse fine-tuning can lead to meaningful improvements in mathematical reasoning tasks, thereby shedding light on whether the models are truly learning to reason or simply aligning with the expected output structure.

### 3.1 Experiment Setup

**Evaluation Datasets.** We compare different models on diverse commonly-used mathematical benchmarks. Beyond the widely adopted GSM8K dataset (Cobbe et al., 2021), our evaluation also incorporates diverse challenging benchmarks spanning multiple mathematical domains. These include: (i) advanced problem sets targeting competition and Olympiad-level reasoning, such as MATH-500 (Lightman et al., 2023), Olympiad Bench (He et al., 2024). (ii) undergraduate-level mathematics challenges from the College Math dataset (Tang et al., 2024); and (iii) a cross-lingual and culturally distinct math benchmark, GaoKao En 2023 (Liao et al., 2024), drawn from China's national college entrance examination. We exclude benchmarks such as AIME 2024 and AMC due to their limited number of available problems, which introduces high variance and reduces evaluation reliability.

**Base Models and Setup.** To verify the effects of SFT on LLMs, we use LLMs of different sizes as the base models, covering both general LLMs and

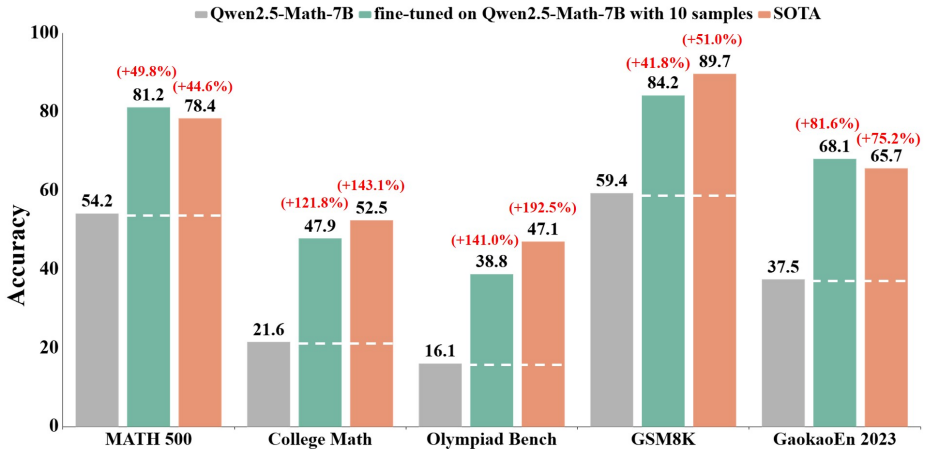


Figure 2: Performance comparison: base model vs. 10-sample fine-tuned model vs. SOTA

| Base Model (Qwen2.5-Math-7B) | Datasets | | | | | Avg. |
|---|---|---|---|---|---|---|
| | MATH 500 | College Math | Olympiad Bench | GSM8K | GaokaoEn 2023 | |
| Base model | 54.2 | 21.6 | 16.1 | 59.4 | 37.5 | 37.76 |
| +10 samples with correct answer | **81.2** | **47.9** | **38.8** | 84.2 | **68.1** | **64.04** |
| **Without format** | | | | | | |
| +10 samples with only answers | 16.8 | 8.4 | 2.5 | 2.6 | 11.2 | 8.3 |
| **Content-disturbed** | | | | | | |
| +10 samples with incorrect answer | 79.2 | 44.6 | 37.8 | **88.2** | 63.1 | 62.58 |
| +10 samples with garbled characters | 64.4 | 28.2 | 25.9 | 80.1 | 59.5 | 51.62 |
| +10 samples with unmatched solutions | 48.4 | 45.1 | 24.3 | 67.6 | 48.3 | 46.74 |
| +10 samples with disrupted steps | 69.8 | 21.8 | 29.5 | 83.6 | 57.1 | 52.36 |
| **SOTA (Guan et al., 2025)** | 78.4* | 52.5* | 47.1* | 89.7* | 65.7* | 66.68* |

Table 1: Performance of models trained on different SFT data. The highest results are highlighted in bold and the second-best results are marked with underline except SOTA. For SOTA, we use the results from their original reports (Guan et al., 2025), denoted by *.

mathematical reasoning models. Specifically, the selected base models include Qwen2.5-Math-1.5B, Mistral-7B-v0.3, DeepSeek-Math-7B, Qwen2.5-Math-7B, Qwen2.5-7B, Llama-3.1-8B, Qwen2.5-14B, Qwen2.5-32B, Llama-3.1-70B, and Qwen2.5-Math-72B.

For the training data, we construct a set of minimal SFT datasets derived from the BigMath dataset (Albalak et al., 2025). For each target benchmark, we select a corresponding subset from BigMath that aligns with the problem distribution and reasoning requirements of the evaluation set. For example, to evaluate performance on the MATH-500 benchmark, we extract a representative subset from the MATH portion of BigMath, ensuring that the training data reflect similar topic coverage and difficulty. For each such subset, we distill a small number of high-quality reasoning trajectories using different LLMs. These distilled examples are then used to perform SFT on the selected base models.

The SFT experiments were conducted using a consistent training configuration. Each model was fine-tuned for 20 epochs, in order to ensure sufficient exposure to the small training set and to allow the model to fully internalize the reasoning patterns. A low learning rate of 7e-6 was used to prevent overfitting and to ensure stable optimization, especially given the small dataset size. For models larger than 7B parameters, due to GPU memory constraints, we adopted parameter-efficient fine-tuning via LoRA instead of full fine-tuning. The experiments were conducted on a server with 8 H20 GPUs, each has 80GB memeory.

**Evaluation Metric.** We evaluate the performance of various LLMs using the evaluation toolkit provided by Qwen2.5-Math (Yang et al., 2024) and report the Pass@1 accuracy for all models.

### 3.2 R1: Do recent performance improvements in small-scale LLMs during SFT primarily arise from format alignment rather than true reasoning enhancement?

SFT has been widely recognized for its effectiveness in improving the performance of LLMs across a variety of tasks. In the domain of mathematical reasoning, prior work has shown that even small instruction-following datasets can significantly improve model performance, especially when the base model already possesses strong reasoning capabilities (Yuan et al., 2023). This has led to the hypothesis that SFT may not fundamentally improve reasoning ability, but rather activate pre-existing capabilities by aligning model outputs with a preferred reasoning format.

To examine whether the performance gains from SFT arise primarily from format alignment rather than genuine reasoning improvement, we conduct a minimal-data experiment that isolates the effect of learning reasoning format. By evaluating the model's performance before and after sparse fine-tuning, we aim to assess whether exposure to a small number of formatted demonstrations can substantially improve reasoning performance—suggesting that SFT may function more as a format imitation mechanism than as a means of endowing LLMs with new reasoning capabilities.

**Experimental Design.** We construct 6 distinct fine-tuning datasets based on Qwen2.5-Math-7B-Instruct, each containing only 10 examples distilled from the original training corpus. These include (1) well-structured examples with correct answers, (2) examples that include only the final answer, (3) well-structured examples with incorrect answers, (4) well-structured reasoning steps paired with mismatched solutions, (5) examples with shuffled reasoning steps, and (6) examples with garbled text. We then fine-tune Qwen2.5-Math-7B on these datasets and evaluate all models across five math benchmarks to measure the impact of each data condition on reasoning performance.

**Results and Analysis.** We first compare the models trained on the well-structured dataset with correct answers with the base and instruct models, as shown in Figure 2. Remarkably, fine-tuning the base model with only 10 well-structured and correctly answered examples yields substantial performance gains across all benchmark datasets, approaching the performance of the fully instruction-tuned model. This result suggests that the improvement does not stem from a fundamental enhancement in the model's reasoning ability—an unlikely outcome given the extremely limited supervision—but rather from aligning the model's output to the expected reasoning format. The model appears to benefit primarily from exposure to the step-by-step reasoning structure, allowing it to imitate the format of correct solutions without truly acquiring deeper reasoning skills.

We further compare fine-tuning with step-by-step reasoning data against data that with only the final answer. As shown in Table 1, when fine-tuned on 10 examples containing only the final answers, the model's performance drastically drops to an average of 8.3, far below even the base model. This confirms that simply exposing the model to correct answers provides negligible benefit. In contrast, even examples with incorrect answers or mismatched solutions, as long as they retain a coherent step-by-step format, significantly improve performance—achieving average scores of 62.58 and 46.74 respectively. Similarly, models trained on examples with disrupted steps or garbled content still outperform the base model. These results collectively reinforce the conclusion that the observed performance gain stems primarily from format alignment: **LLMs learn to mimic the structure of step-by-step reasoning rather than internalizing the reasoning process themselves.**

To further probe the role of format quality in alignment-driven performance gains, we conduct a fine-grained comparison of how different types of formatting perturbations affect model outcomes. As shown in Table 1, models fine-tuned on well-structured but incorrect examples still achieve impressive gains (62.58 avg.), indicating that answer correctness is not the dominant factor. More surprisingly, models trained on mismatched solutions—where the reasoning process does not match the final answer—still outperform the base model (46.74 avg.), suggesting that as long as the reasoning format remains intact, the model can extract useful patterns. On the other hand, when the reasoning steps are disrupted (52.36 avg.), or when the inputs include garbled characters (51.62 avg.), performance degrades more noticeably, though still remains above the base level. This suggests that preserving the logical order of reasoning steps is more critical than answer correctness, and even partially corrupted input retains value if the formatting skeleton is intact. These findings collectively underscore that **format alignment is highly sensitive to**

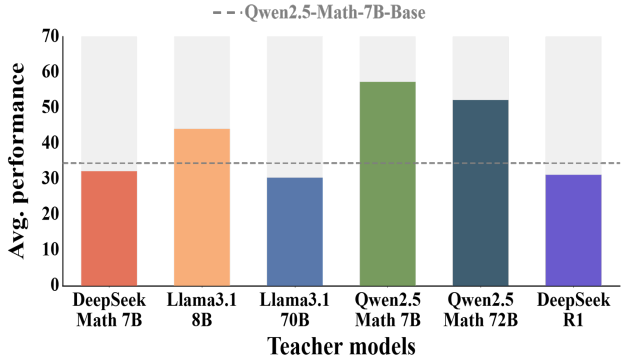


Figure 3: Performance of Qwen2.5-Math-7B trained on data distilled from different teacher models that are of the instruct version.

**both surface-level coherence and deeper structural regularities.** Format elements such as step order, logical flow, and format consistency appear to act as strong priors that guide the model toward generating well-structured outputs, even in the absence of valid reasoning. This highlights the importance of high-quality formatting in SFT datasets and provides further evidence that performance improvements in few-shot fine-tuning scenarios often reflect format mimicry rather than genuine reasoning acquisition.

### 3.3 RQ 2: If format alignment dominates, can high-quality supervision (e.g., data distilled from stronger teacher models) help models move beyond format imitation toward genuine reasoning ability?

To further understand how LLMs learn to reason from format-aligned data during SFT, we examine not only their sensitivity to format preference but also the effect of supervision strength. While earlier experiments show that even a small number of exemplars can steer models toward preferred reasoning formats, it remains unclear whether stronger supervision signals—such as those derived from more capable teacher models—provide benefits beyond superficial imitation. In this section, we investigate whether training student models on demonstrations produced by models with different levels of reasoning ability leads to variations in performance. This analysis helps assess whether high-quality supervision enables models to internalize deeper reasoning patterns.

**Experimental Design.** To examine how the quality of supervision affects reasoning performance, we distill SFT data from teacher models of varying capabilities, including the instruct versions of DeepSeek-Math-7B, Llama-3.1-8B, Llama-3.1-70B, Qwen2.5-Math-7B, Qwen2.5-Math-72B, and DeepSeek-R1. From these outputs, we randomly select 10 samples that are both correctly formatted and contain the correct answers. These samples are then used to fine-tune three base models with different initial capabilities, with their detailed experimental results presented in Appendix. All other training configurations, such as the number of epochs, batch size, and learning rate, remain consistent with previous SFT experiments.

**Results and Analysis.** Figure 3 presents the performance of various base models fine-tuned with data distilled from different teacher models. Contrary to expectations, we find that stronger teacher models do not consistently yield better SFT outcomes. For example, although DeepSeek-R1 achieves strong results on full-task benchmarks, it consistently underperforms as a data generator for 10-shot SFT. In contrast, smaller or mid-sized teacher models—such as Qwen2.5-Math-7B-Instruct—often produce supervision signals that lead to more effective downstream learning.

This counterintuitive trend appears to arise from differences in reasoning style and output distribution across teacher models. Specifically, DeepSeek-R1 tends to produce long, verbose chains of thought, which, while effective in isolation, may be suboptimal when used in few-shot SFT. When only a few examples are available, such stylistically complex outputs may overwhelm smaller base models, leading to underfitting or poor generalization—especially if the base model lacks the capacity or prior alignment to internalize such stylistic nuances.

More generally, our results suggest that raw accuracy or scale of the teacher is not the sole determinant of downstream SFT effectiveness. Instead, the compatibility between the teacher's output format and the base model's inductive biases plays a more crucial role. Instruction-tuned teachers like Qwen2.5-Math-7B-Instruct tend to produce concise, well-structured outputs, which are easier for base models to imitate and generalize from, particularly under limited supervision. In contrast, outputs from larger models like Llama-3.1-70B-Instruct—despite being high-quality in isolation—may contain stylistic or structural patterns that are less transferable to less capable models.

Taken together, these findings highlight a key insight: **In low-resource SFT settings, compati-**

**bility between the teacher's inductive bias and the student's capacity significantly influences knowledge transfer efficiency.** Careful selection of teacher models whose reasoning style aligns with the learning capacity of the base model is essential for efficient and effective knowledge transfer under data-scarce conditions.

### 3.4 RQ3: As the amount of training data increases, does the primary driver of performance shift from format alignment to deeper reasoning?

To better understand the relationship between supervision scale and reasoning performance, we next investigate how the amount of SFT data influences model behavior. While previous results have shown that even limited exposure to well-formatted demonstrations can improve reasoning performance, it remains unclear whether increasing the quantity of training data results in proportionally stronger reasoning capabilities, or whether it primarily reinforces superficial format adherence. In this section, we vary the size of the fine-tuning dataset to compare the performance of finetuned models with different SFT data sizes.

**Experimental Design.** To examine how the scale of supervision affects reasoning performance, we conduct a controlled experiment using varying amounts of training data distilled from a teacher model. Specifically, we sample different quantities of high-quality data from each subsets generated by Qwen2.5-Math-7B-Instruct. These subsets are then mixed and used to fine-tune the base model, Qwen2.5-Math-7B, allowing us to assess whether increased data volume leads to proportional improvements in reasoning ability, or primarily reinforces stylistic conformity.

**Results and Analysis.** The results in Figure 4 reveal a clear yet nuanced relationship between the data scale of SFT and reasoning performance. In general, increasing the number of training samples distilled from Qwen2.5-Math-7B-Instruct leads to consistent performance improvements, particularly when moving from extremely low supervision (2 samples from each subset) to moderate-scale SFT (e.g., 10 or 25 samples from each subset). This trend is evident across all evaluated datasets, with average performance rising from 62.4 to 65.3 when increasing from 2 to 10 samples.

However, increasing the number of training examples from each subset beyond 10 does not lead to consistent performance gains. Instead, we observe fluctuations in performance as the dataset size scales up to 2,500 examples, with the highest average score reaching only 64.2—lower than the performance achieved with just 10 examples. This trend suggests that simply scaling the amount of SFT data does not guarantee improved reasoning ability. One possible explanation is that the initial gains primarily result from the model learning to mimic the surface-level structure and format of the supervision data. Once this alignment is achieved, additional examples offer diminishing or even adverse returns, especially if they reinforce formatting regularities without introducing novel reasoning strategies.

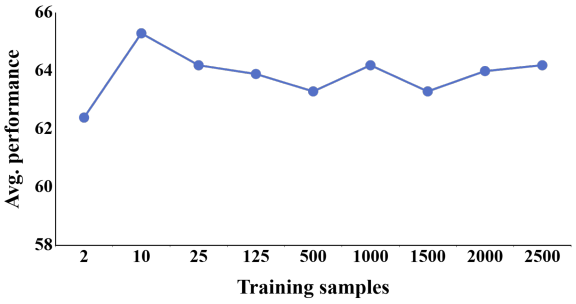


Figure 4: Performance of Qwen2.5-Math-7B trained on different data sizes. Training samples are derived from a subset of BigMath dataset and distilled from Qwen2.5-Math-7B-Instruct. The average performance is evaluated across four benchmarks: MATH500, College Math, Olympiad Bench, and GSM8K.

These findings highlight a potential bottleneck in current SFT strategies for reasoning tasks: **while format imitation is highly sample-efficient, its contribution may quickly saturate. Since understanding and applying reasoning strategies likely requires deeper generalization than format learning alone, merely scaling up formatted demonstrations may not be sufficient to drive further progress.** This underscores the need for complementary approaches—such as higher-quality demonstrations, curriculum design, or targeted reasoning feedback—that go beyond stylistic conformity to facilitate deeper reasoning skill acquisition.

### 3.5 RQ 4: Are format alignment gains consistent across models of varying sizes and architectures?

Building on the results from RQ1, we find that Qwen2.5-Math-7B achieves substantial performance gains during SFT primarily through format alignment, we are now interested in extend-

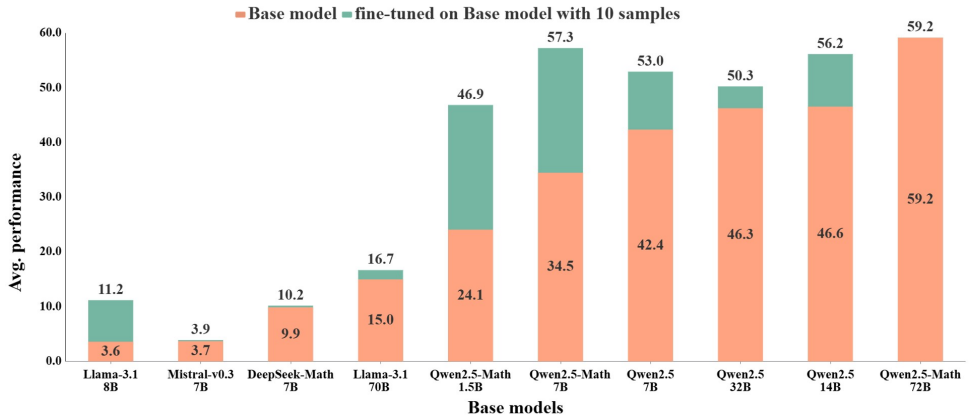


Figure 5: A comparison of SFT performance across multiple base models, all trained on the same 10 samples distilled from Qwen2.5-Math-7B-Instruct. Model performance is evaluated by averaging results across six mathematical reasoning benchmarks: MATH500, College Math, Olympiad Bench, GSM8K, GaokaoEn 2023, and Omni-math.

ing this investigation to other base models. The purpose of this section is to examine whether similar improvements in reasoning performance can be achieved across different models when fine-tuned with a small number of well-structured, correct-answer data samples. By doing so, we aim to explore how the inherent reasoning capabilities of the base model influence its ability to learn the correct reasoning format through SFT. This analysis will provide valuable insights into whether the effectiveness of SFT is model-dependent, and how the baseline reasoning ability of a model interacts with the fine-tuning process to shape its overall performance on reasoning tasks.

**Experimental Design.** We use the 10 samples generated from Qwen2.5-Math-7B-Instruct that are both well-structured and contain the correct answers to train different base models and evaluate the impact of SFT on their reasoning capabilities.

**Results and Analysis.** The experimental results presented in Figure 5 reveal a striking trend: among various base models, only the stronger small-scale models—such as Qwen2.5-Math-1.5B and Qwen2.5-Math-7B—exhibit substantial performance gains from fine-tuning with 10 step-by-step reasoning samples. This shows that this mechanism is only effective for models with sufficient reasoning capacity and alignment receptiveness.

In contrast, weaker models such as Mistral-7B-v0.3 fail to benefit significantly from the same fine-tuning. This suggests that when a model's inherent reasoning and instruction-following ability is too weak, it cannot effectively absorb or generalize the reasoning format from a small number of examples. Similarly, for very large models like Qwen2.5-32B and Qwen2.5-Math-72B, the observed gains are limited. One possible explanation is that these models have already been heavily exposed to instruction-like patterns during pretraining or prior alignment stages, such that the marginal benefit from additional formatted data is minimal.

In summary, the impact of format alignment in SFT depends strongly on the capability of the base model. **Small but capable models are best positioned to leverage format-aligned examples, while weaker models lack the inductive bias to learn from them, and strong models offer diminishing marginal returns.**

## 4 Conclusion

We investigate the underlying mechanisms driving apparent reasoning improvements in small-scale LLMs during SFT. Our experiments show that state-of-the-art small LLMs can approach or surpass SOTA performance on five math benchmarks using only 10 random format-correct samples. This suggests that gains may stem from surface-level format alignment rather than true reasoning advances. Moreover, we identify three key factors influencing SFT efficacy: (i) distributional alignment between base and teacher model outputs matters most, while teacher quality has limited effect; (ii) data scaling yields diminishing returns without qualitative reasoning improvements; (iii) base model capability determines format utilization: small but capable models benefit most from aligned examples. These findings indicate that current methods may overestimate reasoning ability, highlighting the need for frameworks that distinguish format imitation from genuine understanding.

## Limitations

Our investigation is limited in two key aspects: (1) it focuses exclusively on SFT-based approaches, leaving open whether reinforcement learning or other training paradigms exhibit similar format bias susceptibility; and (2) it examines conventional step-by-step formats, omitting alternative structures like self-correction mechanisms. These limitations motivate future research directions exploring the effects of format alignment in different training methodologies and the comparative efficacy of various formats.

## References


Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and 1 others. 2025. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*.

Zui Chen, Tianqiao Liu, Mi Tian, Weiqi Luo, Zitao Liu, and 1 others. 2025. Advancing mathematical reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. In *The Thirteenth International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2024. Interpretable contrastive monte carlo tree search reasoning. *arXiv preprint arXiv:2410.01707*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno de Moraes Dumont, and Sanmi Koyejo. 2024. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, and 1 others. 2025. Math-perturb: Benchmarking llms' math reasoning abilities against hard perturbations. *arXiv preprint arXiv:2502.06453*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*.

Bin Lei, Chunhua Liao, Caiwen Ding, and 1 others. 2023. Boosting logical reasoning in large language models through a new framework: The graph of thought. *arXiv preprint arXiv:2308.08614*.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2023. Query and response augmentation cannot help out-of-domain math reasoning generalization.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.

Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. 2024. Mario: Math reasoning with code interpreter output–a reproducible pipeline. *arXiv preprint arXiv:2401.08190*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36:36407–36433.

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. 2025. There may not be aha moment in r1-zero-like training—a pilot study.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, and 1 others. 2025. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S2r: Teaching llms to self-verify and self-correct via reinforcement learning. *arXiv preprint arXiv:2502.12853*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI. 2025. Openai o3-mini system card. *OpenAI o3-mini System Card, January 2025*.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, and 1 others. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. Mathscale: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. *QwQ-32B: Embracing the Power of Reinforcement Learning*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, and 1 others. 2024. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*.

Haotian Xu. 2023. No train still gain. unleash mathematical reasoning of large language models with monte carlo tree search guided by energy function. *arXiv preprint arXiv:2309.03224*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, and 1 others. 2024. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more

for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

# A Detailed experimental results

| Model | Datasets | | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | MATH 500 | College Math | Olympiad Bench | GSM8K | GaokaoEn 2023 | Omni-math | |
| **1.5B Models** | | | | | | | |
| Qwen2.5-Math-1.5B | 35.4 | 7.0 | 22.7 | 38.3 | 27.0 | 14.5 | 24.1 |
| Qwen2.5-Math-1.5B +10 samples | 71.8 | 43.7 | 35.3 | 48.1 | 61.3 | 21.1 | 46.9 |
| Qwen2.5-Math-1.5B-Instruct | 75.4 | 48.3 | 37.9 | 85.1 | 67.5 | 29.0 | 57.2 |
| **7∼8B Models** | | | | | | | |
| Qwen2.5-Math-7B | 54.2 | 21.6 | 16.1 | 59.4 | 37.5 | 18.3 | 34.5 |
| Qwen2.5-Math-7B+10 samples | 81.2 | 47.9 | 38.8 | 84.2 | 68.1 | 23.6 | 57.3 |
| Qwen2.5-Math-7B-Instruct | 83.2 | 47.0 | 41.2 | 95.7 | 68.6 | 30.2 | 61.0 |
| DeepSeek-Math-7B | 12.4 | 7.6 | 2.5 | 21.9 | 11.7 | 3.0 | 9.9 |
| DeepSeek-Math-7B + 10 samples | 12.4 | 7.6 | 2.5 | 21.8 | 13.8 | 3.0 | 10.2 |
| DeepSeek-Math-7B-Instruct | 45.6 | 31.0 | 13.5 | 82.2 | 42.1 | 12.1 | 37.8 |
| Qwen2.5-7B | 60.4 | 30.8 | 28.9 | 65.8 | 47.0 | 21.5 | 42.4 |
| Qwen2.5-7B+10 samples | 69.8 | 43.6 | 34.7 | 88.6 | 57.1 | 24.2 | 53.0 |
| Qwen2.5-7B-Instruct | 76.8 | 46.8 | 39.6 | 92.5 | 65.2 | 26.6 | 57.9 |
| Llama-3.1-8B | 4.4 | 2.9 | 1.9 | 5.5 | 4.9 | 2.0 | 3.6 |
| Llama-3.1-8B + 10 samples | 16.4 | 8.4 | 4.3 | 17.3 | 17.1 | 3.9 | 11.2 |
| Llama-3.1-8B-Instruct | 50.8 | 30.9 | 14.4 | 83.3 | 40.8 | 13.2 | 38.9 |
| Mathstral-7B-v0.3 | 3.4 | 2.2 | 1.5 | 8.0 | 5.7 | 1.4 | 3.7 |
| Mathstral-7B-v0.3 + 10 samples | 3.8 | 2.2 | 1.5 | 8.9 | 5.7 | 1.6 | 3.9 |
| Mathstral-7B-v0.3-Instruct | 13.2 | 6.4 | 2.8 | 50.8 | 16.9 | 4.5 | 15.8 |
| **14B Models** | | | | | | | |
| Qwen2.5-14B | 61.0 | 35.3 | 27.3 | 85.8 | 49.6 | 20.5 | 46.6 |
| Qwen2.5-14B +10 samples | 78.0 | 46.6 | 39.4 | 83.4 | 64.2 | 25.5 | 56.2 |
| Qwen2.5-14B-Instruct | 80.0 | 47.9 | 43.3 | 94.9 | 66.8 | 28.9 | 60.3 |
| **32B Models** | | | | | | | |
| Qwen2.5-32B | 59.2 | 37.8 | 29.0 | 82.0 | 50.1 | 19.9 | 46.3 |
| Qwen2.5-32B + 10 samples | 69.8 | 41.4 | 32.4 | 68.6 | 67.0 | 22.9 | 50.3 |
| Qwen2.5-32B-Instruct | 83.8 | 48.7 | 44.4 | 95.9 | 70.1 | 31.9 | 62.5 |
| **∼ 70B Models** | | | | | | | |
| Qwen2.5-Math-72B | 79.4 | 44.7 | 43.7 | 86.9 | 64.9 | 35.4 | 59.2 |
| Qwen2.5-Math-72B + 10 samples | 81.0 | 44.1 | 44.4 | 87.0 | 63.9 | 35.0 | 59.2 |
| Qwen2.5-Math-72B-Instruct | 85.6 | 49.6 | 48.7 | 95.9 | 70.9 | 32.9 | 63.9 |
| Llama-3.1-70B | 17.4 | 7.2 | 2.2 | 40.2 | 18.4 | 4.5 | 15.0 |
| Llama-3.1-70B +10 samples | 18.8 | 8.6 | 2.2 | 43.1 | 22.9 | 4.7 | 16.7 |
| Llama-3.1-70B-Instruct | 50.8 | 32.0 | 25.6 | 84.2 | 39.0 | 15.0 | 41.1 |

Table 2: SFT results trained on different base models

Table 2 reports the detailed evaluation results of different base models fine-tuned on 10 examples distilled from Qwen2.5-Math-7B-Instruct. We evaluate the resulting models on 6 mathematical reasoning benchmarks to assess the effectiveness of this high-quality supervision. The average performance across the five benchmarks, as shown in the final column of Table 2, corresponds directly to the data points plotted in Figure 5.

| Base Model | Teacher Models | Datasets | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | | MATH 500 | College Math | Olympiad Bench | GSM8K | GaokaoEn 2023 | Omni-math | |
| Qwen2.5-Math-7B | Qwen2.5-Math-7B-Instruct | 81.2 | 47.9 | 38.8 | 84.2 | 68.1 | 23.6 | 57.3 |
| | Llama-3.1-8B-Instruct | 61.2 | 40.2 | 12.6 | 81.5 | 48.6 | 20.8 | 44.1 |
| | DeepSeek-Math-7B-Instruct | 46.2 | 34.5 | 11.1 | 51.9 | 29.1 | 20.1 | 32.2 |
| | Qwen2.5-Math-72B-Instruct | 83.4 | 44.0 | 40.9 | 71.9 | 60.8 | 12.0 | 52.2 |
| | Llama-3.1-70B-Instruct | 51.8 | 21.4 | 13.2 | 40.7 | 46.2 | 9.2 | 30.4 |
| | DeepSeek-R1 | 42.6 | 18.0 | 12.0 | 52.7 | 42.9 | 18.8 | 31.2 |
| Qwen2.5-Math-7B-Instruct | | 83.2 | 47.0 | 41.2 | 95.7 | 68.6 | 30.2 | 61.0 |
| Llama-3.1-8B | Qwen2.5-Math-7B-Instruct | 16.4 | 8.4 | 4.3 | 17.3 | 17.1 | 3.9 | 11.2 |
| | Llama-3.1-8B-Instruct | 8.0 | 8.4 | 3.1 | 19.6 | 6.8 | 3.0 | 8.2 |
| | DeepSeek-Math-7B-Instruct | 15.0 | 13.9 | 3.7 | 25.2 | 13.2 | 4.2 | 12.5 |
| | Qwen2.5-Math-72B-Instruct | 11.2 | 8.1 | 2.8 | 41.1 | 15.3 | 3.5 | 13.7 |
| | Llama-3.1-70B-Instruct | 10.0 | 9.2 | 3.4 | 30.8 | 12.2 | 3.3 | 11.5 |
| | DeepSeek-R1 | 9.2 | 6.4 | 3.4 | 36.0 | 14.8 | 4.0 | 12.3 |
| Llama-3.1-8B-Instruct | | 50.8 | 30.9 | 14.4 | 83.3 | 40.8 | 13.2 | 38.9 |
| Qwen2.5-Math-72B | Qwen2.5-Math-7B-Instruct | 81.0 | 44.1 | 44.4 | 87.0 | 63.9 | 35.0 | 59.2 |
| | Llama-3.1-8B-Instruct | 78.2 | 44.5 | 43.3 | 86.8 | 63.9 | 35.4 | 58.7 |
| | DeepSeek-Math-7B-Instruct | 78.4 | 44.5 | 44.6 | 87.3 | 65.7 | 35.1 | 59.3 |
| | Qwen2.5-Math-72B-Instruct | 80.0 | 44.6 | 44.4 | 87.3 | 62.6 | 35.2 | 59.0 |
| | Llama-3.1-70B-Instruct | 79.6 | 44.4 | 43.0 | 86.5 | 62.6 | 35.0 | 58.5 |
| | DeepSeek-R1 | 80.8 | 44.8 | 43.0 | 86.7 | 62.1 | 34.9 | 58.7 |
| Qwen2.5-Math-72B-Instruct | | 86.0 | 49.7 | 48.4 | 95.8 | 73.2 | 32.9 | 64.3 |

Table 3: SFT results trained on various teacher-generated distilled data

Table 3 presents the full evaluation results of different base models fine-tuned on 10 examples distilled from different teacher models. Each distilled example is both correctly formatted and contains the correct final answer. The table reports the performance of the resulting SFT models on 6 mathematical reasoning benchmarks. The average accuracy across these benchmarks, shown in the rightmost column of Table 3, corresponds to the values plotted in Figure 3, enabling a direct comparison of teacher-specific effects on downstream reasoning performance.

| Base Model | Average Training Samples from Each Subset | Datasets | | | | |
|---|---|---|---|---|---|---|
| | | MATH 500 | College Math | Olympiad Bench | GSM8K | Avg. |
| Qwen2.5-Math-7B | 2 | 80.4 | 44.0 | 37.4 | 87.6 | 62.4 |
| | 10 | 82.2 | 47.1 | 41.3 | 90.5 | 65.3 |
| | 25 | 80.0 | 47.2 | 39.3 | 90.1 | 64.2 |
| | 125 | 77.6 | 47.5 | 40.0 | 90.4 | 63.9 |
| | 500 | 78.2 | 47.1 | 38.5 | 89.2 | 63.3 |
| | 1000 | 80.0 | 47.4 | 39.4 | 90.1 | 64.2 |
| | 1500 | 78.0 | 47.4 | 37.8 | 89.8 | 63.3 |
| | 2000 | 78.2 | 47.5 | 40.3 | 89.9 | 64.0 |
| | 2500 | 79.8 | 47.4 | 41.0 | 88.7 | 64.2 |
| Qwen2.5-Math-7B-Instruct | - | 83.2 | 47.0 | 41.2 | 95.7 | 66.8 |

Table 4: Performance of models trained on different data sizes

Table 4 reports the evaluation results of the Qwen2.5-Math-7B model fine-tuned on mixed SFT data of varying scales. Each row corresponds to a different data volume, and the model is evaluated on 4 mathematical reasoning benchmarks. The average performance across these benchmarks, shown in the rightmost column, is used to generate the trend shown in Figure 4. Due to the limited availability of correctly formatted and answer-correct samples for GaokaoEn 2023 and Omni-math, these datasets were not included in the mixed SFT training data. As a result, we also exclude these two datasets from the evaluation in this setting.