Structure Trumps Size: Rethinking Data Quality for LLM Reasoning

Hu Xu **Zeyan Li** **Rui Wang***
School of Computer Science
Shanghai Jiao Tong University
{xuhu6736, wangrui12}@sjtu.edu.cn

Jianfeng Xu*
Koguan School of Law
China Institute for Smart Justice
School of Computer Science
Shanghai Jiao Tong University
xujf@sjtu.edu.cn

Abstract

As domain-specific datasets continue to expand, Large Language Models (LLMs) have achieved significant improvements across various fields through supervised fine-tuning (SFT). However, is more data always better for model fine-tuning? Through a series of controlled experiments, we discover that dataset **structure—rather than mere size—plays a decisive role in enhancing LLM reasoning capabilities**. While existing methods acknowledge that good data quality can make training more efficient, they primarily rely on simple heuristic strategies and lack systematic, quantitative frameworks for evaluating data quality. To address this gap, we introduce **MCSQ**—the first multi-dimensional quantitative framework for reasoning data management. MCSQ rigorously evaluates and optimizes datasets along six orthogonal dimensions. Through comprehensive controlled experiments, we find that selectively incorporating "distorted" (model-disagreed) or "mismatched" (low-relevance) samples—which are typically discarded in traditional approaches—can outperform conventional "clean" data on certain advanced reasoning benchmarks. Our findings challenge traditional assumptions about data "quality" in LLM fine-tuning and provide actionable, quantitative guidance for efficient, structure-aware dataset management. The datasets and codes are both available at https://github.com/xuhu0115/MCSQ.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive reasoning capabilities across diverse domains, yet complex multi-step reasoning remains a significant challenge (Wei et al., 2022b; Ouyang et al., 2022). Supervised fine-tuning (SFT) on reasoning traces has emerged as a promising approach to enhance these abilities by directly encoding reasoning patterns into model parameters (Zelikman et al., 2022; Muennighoff et al., 2025; Zhang et al., 2024). By training on structured Question, Reasoning Trace, Answer triplets, this method enables models to internalize step-by-step problem-solving strategies, potentially offering more efficient reasoning compared to inference-time techniques like Chain-of-Thought prompting (Wei et al., 2022b) or tree search (Yao et al., 2023; Besta et al., 2024).

Reasoning-focused SFT has gained significant traction in recent years, with numerous studies demonstrating its effectiveness across mathematical (Lewkowycz et al., 2022; Hendrycks et al., 2021), scientific (Huang et al., 2023), and logical reasoning tasks (Ye et al., 2025). This approach has evolved from early work on step-by-step reasoning distillation (Zelikman et al., 2022) to more sophisticated methods incorporating diverse reasoning patterns (Muennighoff et al., 2025), planning algorithms (Zhang et al., 2024), and multi-stage reasoning frameworks (Wang et al., 2023; Yao et al., 2023). Recent research has shown that models fine-tuned on reasoning traces can achieve significant performance improvements on complex problem-solving benchmarks (Li et al., 2025b; Deng et al., 2025), often matching or exceeding the capabilities of much larger models. Despite these advances, the field still lacks a systematic understanding of what makes reasoning datasets effective, with most work relying on intuition-based curation strategies rather than principled analysis of dataset characteristics.

This knowledge gap manifests in several critical, unresolved questions that directly impact reasoning SFT effectiveness. First, the relationship between data volume and reasoning performance remains unclear—while scaling laws suggest "more is better," recent work hints at diminishing returns or even performance degradation with excessive data (Ye et al., 2025; Li et al., 2025b). Second, the optimal composition of reasoning domains and types is

* Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11489–11513
November 4-9, 2025

poorly understood; practitioners lack guidance on whether to prioritize breadth (covering many domains) or depth (focusing on specific areas). Third, the impact of data quality variations—including imperfect reasoning traces or examples of marginal relevance—has not been systematically investigated, leaving dataset filtering decisions to rely on unvalidated heuristics rather than empirical evidence.

These uncertainties point to a fundamental limitation in current approaches: the lack of a principled framework for analyzing and optimizing reasoning datasets beyond simple volume or binary quality metrics. While extensive research has explored data selection for general LLM pretraining (Kudugunta et al., 2023; Elazar et al., 2024) and instruction tuning (Wei et al., 2022a; Raffel et al., 2020), the unique requirements of reasoning-centric SFT data demand specialized analysis tools that can capture the multi-faceted nature of reasoning quality.

To address these challenges, we introduce the **M**ulti-dimensional Quantitative Framework for Evaluating Distilled **C**hain-of-Thought **S**FT Data **Q**uality (MCSQ). Grounded in information theory principles (Xu et al., 2025b), MCSQ decomposes reasoning dataset quality into six concrete, measurable dimensions: Volume, Scope, Granularity, Variety, Distortion, and Mismatch. Through controlled experiments where we systematically vary one dimension while holding others constant, we uncover several counter-intuitive findings that challenge conventional wisdom in reasoning dataset construction. Our key discoveries include: (1) data structure consistently outweighs volume in determining reasoning performance, with scaling beyond moderate thresholds yielding diminishing returns; (2) "imperfect" data can significantly boost performance on certain advanced reasoning tasks; and (3) the optimal balance between domain specialization and diversity depends critically on the target task. Based on these findings, we make three major contributions to the field:

1. We propose the MCSQ framework, the first systematic, multi-dimensional approach for reasoning dataset analysis and optimization, providing a principled foundation for understanding how dataset characteristics shape LLM reasoning capabilities.

2. We empirically establish that dataset structure—not just size or binary quality—fundamentally shapes reasoning capabilities, with different dimensions affecting different aspects of performance in ways that challenge conventional wisdom about data curation.

3. We demonstrate that conventional data curation heuristics can be counterproductive, offering instead quantitative guidelines for structure-aware dataset design that can maximize reasoning performance while managing trade-offs with general knowledge retention.

2 Related Work

Research on LLM reasoning has progressed from inference-time prompting—such as Chain-of-Thought (Wei et al., 2022b; Kojima et al., 2022), self-consistency (Wang et al., 2023), and search-based methods (Yao et al., 2023; Zhang et al., 2024)—to approaches that embed reasoning capabilities directly into model parameters via targeted data and fine-tuning. We briefly review three relevant strands: (1) general data selection methods for LLM training, (2) reasoning trace generation and fine-tuning, and (3) reasoning-centric data selection strategies. These areas collectively motivate the need for a principled, multi-dimensional approach to reasoning dataset construction.

Data Selection for Language Model Training. Effective data selection is critical at every stage of LLM development. During pretraining, large-scale corpora are carefully filtered, deduplicated, and balanced to optimize coverage and quality (Kudugunta et al., 2023; Elazar et al., 2024; Axelrod, 2017; Du et al., 2022). Instruction-tuning further refines model capabilities through curated (Instruction, Output) pairs (Wei et al., 2022a; Raffel et al., 2020), while preference fine-tuning (e.g., RLHF, DPO) aligns models with human values and user preferences (Ouyang et al., 2022; Rafailov et al., 2023). However, these stages typically focus on general text quality or alignment signals, rather than the nuanced properties necessary for advanced reasoning.

Reasoning Trace Generation and Fine-tuning. A growing body of work has explored the supervised fine-tuning of LLMs on long-form reasoning traces, often generated by strong teacher models or advanced planning algorithms (Muennighoff et al., 2025; Li et al., 2025a; Zelikman et al., 2022; Zhang et al., 2024). These traces, typically in the form of Question, Reasoning Trace, Answer triplets, are

designed to instill step-by-step problem-solving skills in student models, particularly for mathematics and scientific domains(Xu et al., 2025a; Hou et al., 2025). While this strategy has shown promise, the curation of optimal reasoning datasets remains largely heuristic—most prior works select data based on perceived difficulty, teacher model confidence, or coarse quality filters, without a systematic understanding of how specific data properties influence downstream reasoning.

Reasoning-Centric Data Selection. Beyond "More is Better." Recent studies have begun to question the assumption that more data or cleaner data always leads to better reasoning. The "Less-Is-More" hypothesis (Ye et al., 2025) and related works (Li et al., 2025b; Deng et al., 2025) demonstrate that targeted data selection—based on sample complexity, learning impact, or preference signals—can yield superior or comparable reasoning performance with less data(Wang et al., 2025). However, these methods typically focus on sample-level or single-metric selection, such as uncertainty or teacher-student agreement, and lack a holistic, multi-dimensional view of dataset composition.

In contrast to prior work, our approach is the first to introduce a comprehensive, quantitative framework (MCSQ) for reasoning dataset analysis and adapting it to the LLM reasoning context. By decomposing dataset quality into distinct, measurable dimensions—Volume, Scope, Granularity, Variety, Distortion, and Mismatch—our work enables systematic investigation of how structural dataset properties interact and influence model reasoning. This advances the field beyond piecemeal or heuristic data selection, providing actionable guidance for structure-aware, efficient reasoning SFT.

3 Methodology

In this section, we present our approach for quantitatively characterizing and manipulating reasoning-centric fine-tuning datasets.

3.1 The MCSQ Framework

MCSQ is a principled framework for analyzing and curating reasoning datasets in LLM fine-tuning. Rather than relying on heuristic selection, MCSQ provides a theoretically grounded and mathematically justified approach to dataset evaluation. The framework is inspired by Objective Information Theory (OIT) (Xu et al., 2015, 2023; Xu, 2024), which offers a rigorous information-theoretic foundation for assessing data quality. While OIT defines 11 general information metrics for characterizing pretraining datasets—operationalized in the GIME framework (Xu et al., 2025b)—not all are directly applicable to the highly structured and targeted nature of supervised fine-tuning data. MCSQ distills this foundation into a low-dimensional, interpretable mapping, tailored to the empirical needs of reasoning-centric LLM fine-tuning (see Figure 1).

Pretraining data typically exhibits statistical distribution properties such as large-scale, weakly-structured (Zhang et al., 2021), and high-diversity characteristics, whereas supervised fine-tuning (SFT) data emphasizes high-quality, strongly-annotated, and target-relevant attributes (Wolfe, 2023; Dong et al., 2024). Consequently, certain OIT measures (e.g., temporal and coverage metrics) lack operational feasibility and semantic foundation when applied to SFT data, rendering them inapplicable. By adapting OIT to the characteristics of the s1-54k dataset, we distill six interpretable, orthogonal dimensions central to reasoning-centric LLM fine-tuning: Volume, Scope, Granularity, Variety, Distortion, and Mismatch. The adaptation process and detailed formal definitions are provided in Appendix A.

3.2 Dataset Construction

As shown in Figure 2, all experiments are based on the s1-59k dataset (Muennighoff et al., 2025), a large-scale, high-quality collection of 59k (question, reasoning trace, answer) triplets. After preprocessing and filtering for annotation completeness, we obtain 54,046 samples(s1-54k). Each sample is annotated with domain, reasoning type, and model-based correctness/relevance metadata, enabling precise calculation of all MCSQ dimensions. Unlike conventional reasoning datasets (e.g., MATH, GSM8K, or AQuA) that focus on a single domain or reasoning type, s1-59k is constructed to maximize domain diversity and reasoning style coverage, with each sample carefully annotated for multi-dimensional analysis. This breadth, alongside rich metadata and model-based scoring, enables fine-grained measurement and manipulation of all MCSQ axes—facilitating controlled, quantitative studies not feasible with most public datasets.

For each experiment, we construct data subsets $\mathcal{D}_{\text{sub}}$ by systematically varying a single MCSQ dimension, while holding all others approximately

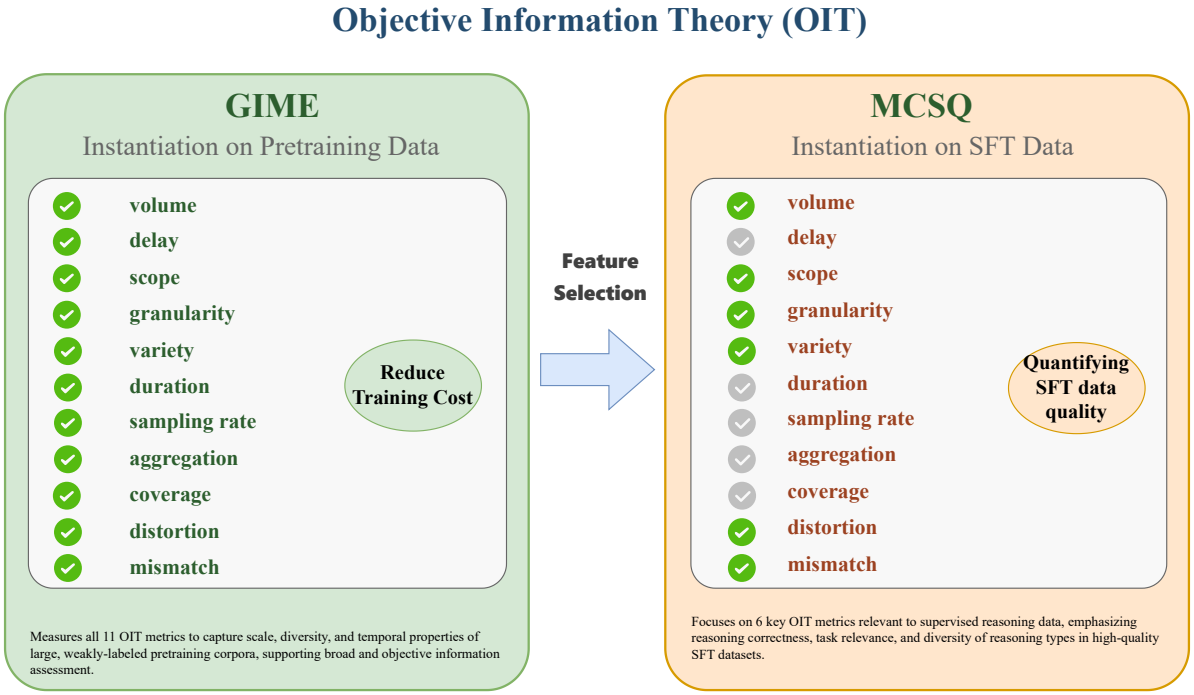

Figure 1: Comparison of OIT-based Metric Instantiation: GIME for Pretraining Data vs. MCSQ for SFT Data. OIT provides a unified theoretical foundation for information quality assessment across data types. GIME instantiates all 11 metrics for large-scale, weakly-labeled pretraining data, while MCSQ selectively adapts 6 core metrics relevant and measurable for SFT reasoning data.

constant. See the Appendix B for detailed sampling strategy. This design enables isolation of each dimension's effect, supporting causal inference about data structure and reasoning performance.

4 Experiments

The core objective of this work is to systematically investigate how distinct structural properties of reasoning datasets affect LLM fine-tuning and performance. To this end, we conduct extensive experiments aimed at answering the following research questions:

- **Volume:** Does increasing the amount of reasoning data always improve LLM performance?
- **Scope:** How does expanding the breadth of subject domains impact model's generalization?
- **Granularity:** What kind of initial data distribution across domains is most beneficial for reasoning performance?
- **Variety:** Is it advantageous to maximize the diversity of reasoning types, or does specialization in a single type offer superior results?
- **Distortion:** Does incorporating "imperfect" or model-disagreed samples necessarily harm reasoning ability?
- **Mismatch:** How much does including non-reasoning or marginally relevant data hurt—or even help—reasoning performance?

By systematically investigating these questions, our results(Figure 3,Appendix D) provide actionable insights into how each dataset property shapes LLM reasoning, enabling principled and effective data curation strategies for advanced model development.

4.1 Experimental Setup

Training Details. All experiments are conducted using Qwen2.5-3B-Instruct as the base model and fine-tuned with LlamaFactory on two NVIDIA 2*A800 80GB GPUs. We adopt full-parameter supervised fine-tuning, enabled by DeepSpeed ZeRO Stage 3 for large-batch optimization. The training uses a learning rate of 1×10^{-5} (cosine schedule, 10% warmup), runs for 5 epochs with BF16 precision, and an effective batch size of 16 (gradient accumulation: 16, per-device batch: 1).

For evaluation, models are deployed using the

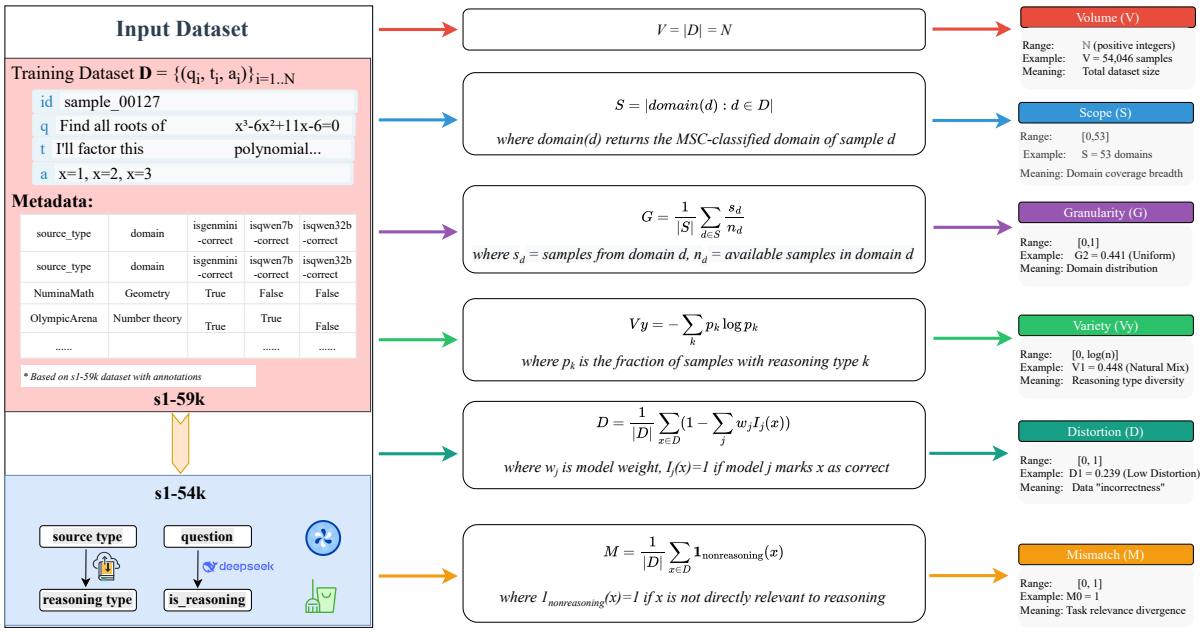

Figure 2: Computation methodology for the six MCSQ dimensions. The left section shows the input dataset structure with sample entries and metadata fields. The center section outlines the multi-step calculation process for each dimension, including the mathematical formulations. The right section displays the resulting dimensions with their value ranges, examples from our experiments, and interpretations. Each dimension captures a distinct aspect of dataset properties: Volume measures dataset size, Scope quantifies domain coverage breadth, Granularity reflects sampling uniformity across domains, Variety captures reasoning type diversity through entropy, Distortion assesses data quality via model disagreement, and Mismatch measures task relevance through automated scoring.

`vllm` framework for efficient large-scale inference. For **reasoning-intensive tasks**, we adopt the zero-shot setting; for **general capability tasks**, we use five-shot prompting. All evaluations use a unified inference pipeline to ensure fair and consistent comparison across experiments.

- **Reasoning-Intensive Tasks:** *AMC (2023)*, *GPQA-diamond* (Rein et al., 2024), *MATH500* (Hendrycks et al., 2021), *Minerva* (Lewkowycz et al., 2022), and *OlympiadBench* (He et al., 2024). For these tasks, we report both **Pass@1** (accuracy of the first generated response) and **Pass@5** (accuracy if any of the top five generated responses is correct). Pass@1 reflects direct generation quality, while Pass@5 is a more robust metric, especially for small benchmarks..
- **General Capability Tasks:** *MMLU* (Hendrycks et al., 2021), *CMMLU* (Li et al., 2023), and *C-Eval* (Huang et al., 2023)(5-shot). We report standard accuracy for these tasks.

Inference and Sampling Settings. TTo minimize evaluation randomness and ensure reproducibility, we repeat each configuration three times using the following sampling protocols and report the averaged results.

- For **Pass@1**, $k = 1$, temperature=0, n_sampling=1, top_p=1.
- For **Pass@5**, $k = 5$, temperature=0.7, n_sampling=5, top_p=0.95.

Beyond Qwen2.5-3B-Instruct, we validated the framework across multiple model scales (Qwen2.5-0.5B and 7B) as well as the challenging DeepMath-103K corpus. Results consistently confirm the same qualitative trends: performance improves with increased data volume but saturates quickly, and structure-first effects (e.g., Scope balance, controlled Distortion/Mismatch) persist regardless of model size or dataset family. For example, on MATH500, Qwen2.5-0.5B, 3B, and 7B all exhibit diminishing returns beyond 10k samples, with relative improvements preserved across scales (see Appendix C). These findings strengthen the generality of our conclusions.

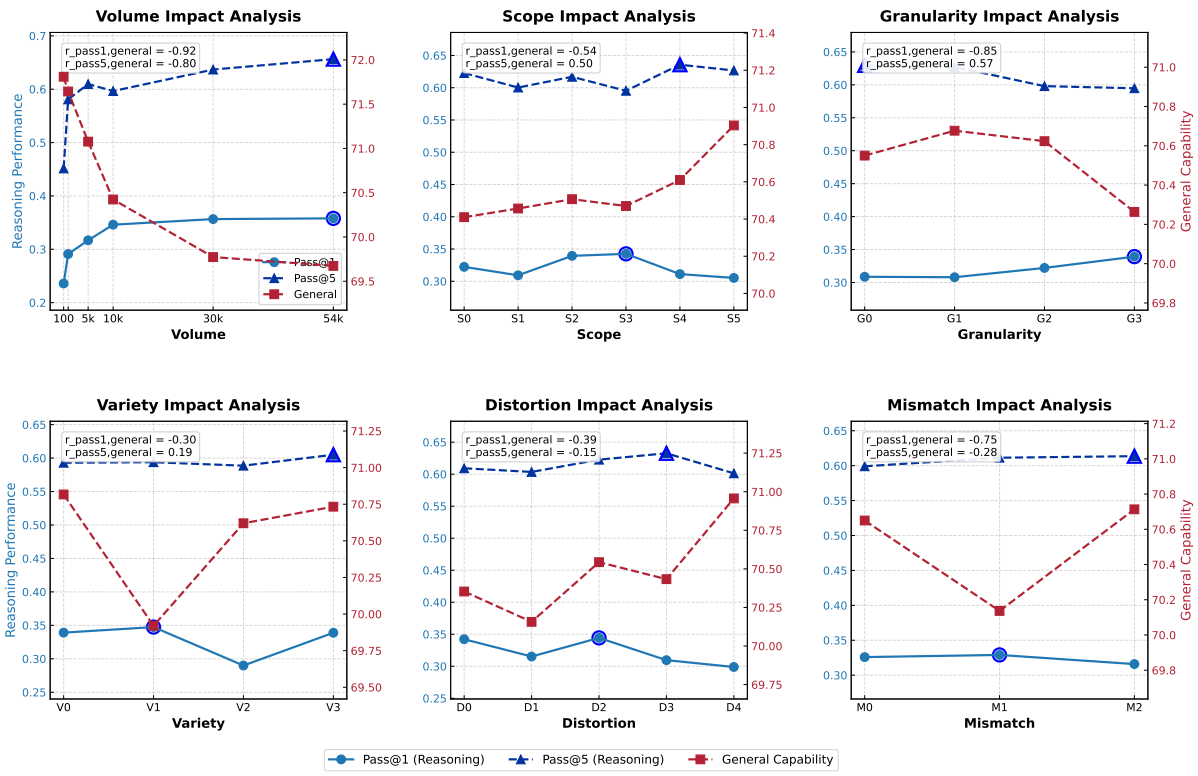

Figure 3: Each subplot illustrates the effect of varying a single MCSQ dimension on reasoning performance (Pass@1 and Pass@5) and general capability. The solid blue line and dashed triangular line represent Pass@1 and Pass@5 reasoning performance (left vertical axis), respectively, while the dashed red square line denotes the general capability score (right vertical axis). The correlation coefficient displayed in the upper-left corner quantifies the strength and direction of the relationship between reasoning performance and general capability. Blue hollow markers indicate the highest performance points in each measurement.

4.2 Experimental Results

RQ 1: Does increasing the amount of reasoning data always improve LLM performance?

We first revisit the most fundamental variable—dataset volume. As shown in Figure 3 and Tables 7-8, a substantial performance leap when the data scale increases from 100 to 1,000 samples, followed by a relatively significant yet diminishing marginal gain as the dataset further expands from 1,000 to 10,000 instances.

A striking trade-off emerges—larger reasoning datasets consistently degrade general ability scores (Table 9), indicative of catastrophic forgetting or over-specialization. Fine-tuning on 54k samples produces the highest reasoning scores, yet general benchmarks drop 2–4 points relative to the base model. This underscores a central tension: maximizing reasoning skills via SFT can erode general knowledge, highlighting the need for data-efficient, structure-aware curation.

Key insight: Naive volume scaling alone is insufficient. Beyond a moderate threshold, marginal gains become small and side effects on generality become pronounced. This motivates the need to look beyond data size and optimize dataset structure.

RQ 2: How does expanding the breadth of subject domains impact model's generalization?

We vary the subject domain scope from a narrow expert selection (5 domains) to all 53 available domains (Table 2). Results (Tables 11, 12) reveal that intermediate scope settings (20 and 30 domains) often outperform both narrower and broader configurations on Pass@1 across several tasks. For Pass@5, including more domains generally helps, but the benefit saturates or reverses beyond a certain breadth(S4). The general capability exhibits gradual improvement with the expansion of scope breadth, while demonstrating progressively increasing marginal returns.

Key insight: A broader domain scope is generally preferable. However, significant disparities in sample sizes across different domains may ad-

versely affect the model's reasoning capabilities. Consequently, the optimal scope should be adaptively determined based on the inherent characteristics of the data.

RQ 3: What kind of initial data distribution across domains is most beneficial for reasoning performance?

Granularity measures the sample distribution across domains. We compare strategies that uniformly sample, over-sample minor domains, or concentrate on dominant domains (Table 3). Results (Tables 15-16) show a clear divergence: A finer granularity demonstrates the most substantial enhancement in reasoning performance. However, such extreme sampling typically incurs significant degradation in the model's generalizability. As illustrated in Figure 3, the model exhibits a 1.63% performance decline compared to its no-fine-tuned state. Moreover, this approach yields marginal gains under pass@1 evaluation while resulting in a 2.90% deterioration in reasoning capability under pass@5 relative to the baseline. In contrast, both G2 (uniform sampling) and G1 (original data distribution) achieve a more balanced trade-off between reasoning proficiency and generalizability. Notably, the original data distribution in this sampling regime maintains relative uniformity, being drawn from the top 10 domains by data volume. Consequently, uniform sampling emerges as a suboptimal yet robust choice when the underlying data distribution remains unknown.

Key insight: Data distribution across domains is a powerful lever—optimizing granularity per task and metric can unlock significant gains unattainable by volume scaling. The optimal granularity is highly task-dependent. Increasing the granularity of domain-relevant data sampling demonstrates the most pronounced performance gains on task-specific metrics. However, such extreme sampling strategies typically incur significant degradation in the model's cross-domain generalization capability.

RQ 4: Is it advantageous to maximize the diversity of reasoning types, or does specialization in a single type offer superior results?

Varying the mix of reasoning types (Math, Science, General) demonstrates that no single composition is universally optimal (Tables 19, 20). There was little difference in reasoning performance and general capability between the group trained predominantly on mathematical reasoning data mixed with other types(V3) and the group trained solely on mathematical reasoning data(V0). This suggests that the existing mathematical reasoning dataset may already contain a considerable amount of general knowledge. It is worth noting that V3 is 1.2% higher than V0 in pass@5 accuracy, which means that V3 brings a larger solution space.

Key insight: Expanding reasoning type diversity beyond a strong core (such as math) offers only marginal overall gains, but can improve solution diversity and model robustness for challenging benchmarks. Thus, moderate diversity is beneficial when targeting broader or more complex reasoning tasks.

RQ 5: Does incorporating "imperfect" or model-disagreed samples necessarily harm reasoning ability?

Perhaps most counter-intuitive, we find that introducing distorted data—samples that judge models disagree on or mark as incorrect—can improve performance for advanced reasoning tasks (Tables 23-24). For example, high-distortion data(D3) yields the best Pass@5 on GPQA-Diamond and AMC23, outperforming even "pure" data. Meanwhile, for MATH500 and Minerva, strictly clean data remains optimal. On the other hand, including moderately difficult problems can help improve reasoning skills while maintaining solid general abilities, but too much difficulty will only hurt these abilities. Past studies (Muennighoff et al., 2025; Ye et al., 2025) have reflected this point, but their difficulty was only controlled within the range acceptable to the base model, and no phenomenon of decreased ability was found when the difficulty was too high.

Intriguingly, the most "distorted" data (D4) causes the least general ability degradation, possibly because the model struggles to fit extreme noise and thus preserves pre-trained knowledge. This challenges the conventional wisdom that only high-quality, teacher-agreed data should be used.

Key insight: Selectively including controversial or hard-to-judge data can inject useful diversity and unlock advanced reasoning potential, especially for tasks with out-of-distribution challenges.

RQ 6: How much does including non-reasoning or marginally relevant data hurt—or even help—reasoning performance?

We test the effect of incorporating data with varying relevance scores (Table 27). Surprisingly, data labeled "low relevance" by automated scoring can significantly boost Pass@5 on GPQA-Diamond (Table 28), outperforming high-relevance data and even the base model. This suggests that automated relevance filters can be overly conservative, discarding valuable edge-case or long-tail samples.

Key insight: Automated relevance metrics are imperfect. "Irrelevant" data may contain essential signals for advanced benchmarks, and blanket filtering can be counterproductive.

4.3 Best Practice Recommendations

Based on our comprehensive experimental analysis of the six MCSQ dimensions, we summarize actionable best practices for curating and optimizing reasoning-centric fine-tuning datasets for large language models:

Volume: Do not blindly increase the dataset size. Substantial gains are achieved by scaling up from very small datasets (e.g., 100 to 1,000 samples), but performance improvements diminish rapidly beyond moderate volumes (e.g., 10,000 samples). Excessive data may even degrade general abilities due to overfitting or catastrophic forgetting. **Recommendation:** Focus on moderate, high-quality volumes rather than maximizing raw data size.

Scope: Expanding the breadth of subject domains generally improves generalization, but too broad a scope can dilute domain expertise and harm reasoning performance. Intermediate coverage (e.g., scope rete=0.556) often strikes the optimal balance. **Recommendation:** Select a domain scope that matches your target application and data distribution, prioritizing moderate breadth.

Granularity: Uniform or near-uniform domain sampling often yields robust performance, especially when the underlying data distribution is unknown. Over-concentration on major domains boosts task-specific performance but harms cross-domain generalization. **Recommendation:** Adopt balanced sampling across domains unless clear task-driven priorities exist.

Variety: Maximizing the diversity of reasoning types does not always lead to significant improvements over specialization, especially when the core dataset (such as mathematical reasoning) already contains substantial general knowledge. However, incorporating a balanced mix of reasoning types can modestly expand the solution space and enhance robustness on more complex or out-of-distribution tasks. **Recommendation:** Prioritize reasoning type diversity when targeting challenging or highly diverse downstream tasks, but for datasets with a strong core (e.g., mathematics), moderate diversity is sufficient in most cases.

Distortion: Selectively incorporating "distorted" or model-disagreed samples can enhance reasoning on challenging tasks, while purely "clean" data may not always be optimal. However, excessive noise can harm the model. **Recommendation:** Do not discard all noisy or controversial samples—retain a controlled proportion to promote robustness and out-of-distribution reasoning.

Mismatch: Including a limited amount of marginally relevant or low-relevance data can inject valuable diversity and improve benchmark performance, but over-inclusion may dilute focus. **Recommendation:** Avoid overly aggressive automated filtering; consider retaining edge-case samples that may seem irrelevant but could benefit advanced reasoning.

In summary: Prioritize data structure and diversity over volume; tailor dataset curation to the specific reasoning tasks and benchmarks of interest. Combining moderate data volume, balanced domain and type diversity, and controlled inclusion of "imperfect" or "noisy" samples leads to the most robust improvements in LLM reasoning.

5 Discussion

One concern is the reliability of Distortion and Mismatch metrics, as they rely on model-based judgments. To mitigate bias, we employed ensemble scoring (Gemini, Qwen-7B, Qwen-32B) and discretized outputs into bins. Furthermore, a human annotation study on 500 samples showed over 90Another concern is the independence of MCSQ axes. While Objective Information Theory (OIT) provides theoretical justification for their orthogonality, we acknowledge that residual correlations may exist in practice. Our controlled sampling design minimizes confounding, and future work will quantify these correlations explicitly with dependency analysis. Finally, we recognize the trade-off between reasoning specialization and general capability observed in our experiments. We propose a preliminary "three-stage filtering funnel"—Macro

(domain scope), Core (challenge and relevance), and Micro (granularity/diversity)—to balance reasoning gains with general knowledge retention. Early results indicate this strategy achieves competitive reasoning improvements while maintaining general benchmarks (see Appendix D).

In summary, MCSQ provides the first systematic, quantitative framework for reasoning-centric SFT data. Our expanded validation across multiple model scales and datasets confirms that data structure trumps data volume, with consistent trends under varying optimization hyperparameters. Moreover, metrics based on automated judgments are strongly aligned with human evaluation, and preliminary multi-stage curation strategies show promise in mitigating the specialization–generalization trade-off. While our current experiments focus on English math and science reasoning, the MCSQ framework is domain-agnostic by design and can be extended to multilingual, coding, and open-domain tasks. We will release code, data, and annotation pipelines to support such extensions and encourage community contributions.

6 Conclusion

In this work, we present MCSQ—a systematic, quantitative framework for dissecting and optimizing the data characteristics that underpin LLM reasoning. By decomposing dataset composition into six interpretable, measurable dimensions—Volume, Scope, Granularity, Variety, Distortion, and Mismatch—we move beyond conventional, volume-centric heuristics and provide the first large-scale evidence that dataset structure decisively shapes reasoning performance. Through controlled fine-tuning experiments, we reveal that targeted adjustments to data structure consistently yield greater improvements in LLM reasoning than naive data scaling.Moreover, even certain data containing 'noise' or exhibiting 'low relevance' can contribute to the enhancement of reasoning capabilities. Importantly, our findings expose a robust specialization–generalization trade-off, highlighting the need for data-efficient and structure-aware curation strategies.

Our results challenge prevailing assumptions about data "quality" in LLM fine-tuning and provide actionable guidance for dataset curation, with broad implications for model distillation, instruction tuning, and automated data selection. We believe that the MCSQ framework paves the way for a new generation of reasoning-centric LLMs, inspiring further research into automated, principled, and task-aware dataset optimization.

Limitations

This study has several limitations. Firstly, Our experiments were only conducted on the Qwen family, and whether they are valid for other architecture models remains to be verified. Secondly, isolating the effect of one MCSQ dimension while keeping others perfectly constant is challenging due to inherent correlations in real-world data; our control is approximate. Thirdly, the definition and measurement of some MCSQ dimensions (e.g., Distortion based on model disagreement, Mismatch based on automated scoring) are operationalizations that could be refined or replaced with alternative metrics. Fourthly, the observed trade-off between reasoning and general capabilities, while consistent in our setup, requires further investigation to understand its underlying mechanisms fully and explore mitigation strategies. Finally, computational constraints limited the number of configurations and hyperparameter tuning explored for each dimension.

References

Amittai Axelrod. 2017. Cynical selection of language model training data. *Preprint*, arXiv:1709.02279.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving llm alignment via preference data selection. *Preprint*, arXiv:2502.14560.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, and 1 others. 2022. Glam: Efficient scaling of language

models with mixture-of-experts. In *International conference on machine learning*, pages 5547–5569. PMLR.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. 2025. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:4189–4209.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, and 1 others. 2025a. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

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

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. *Preprint*, arXiv:2501.19393.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, and Willie Neiswanger. 2025. Tina: Tiny reasoning models via lora. *arXiv preprint arXiv:2504.15777*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Cameron R. Wolfe. 2023. Understanding and using supervised fine-tuning (sft) for language models.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025a. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.

Jianfeng Xu. 2024. Research and application of general information measures based on a unified model. *IEEE Transactions on Computers*, 73(3):915–927.

Jianfeng Xu, Congcong Liu, Xiaoying Tan, Xiaojie Zhu, Anpeng Wu, Huan Wan, Weijun Kong, Chun Li, Hu Xu, Kun Kuang, and Fei Wu. 2025b. General information metrics for improving ai model training efficiency. *Preprint*, arXiv:2501.02004.

Jianfeng Xu, Zhenyu Liu, Shuliang Wang, Tao Zheng, Yashi Wang, Yingfei Wang, and Yingxu Dang. 2023. Foundations and applications of information systems dynamics. *Engineering*, 27:254–265.

Jianfeng Xu, Jun Tang, Xuefeng Ma, Bin Xu, Yanli Shen, and Yongjie Qiao. 2015. Modeling and measurement of objective information. *Scientia Sinica Informationis*, 45(3):336–353. (in Chinese).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025. What, how, where, and how well? a survey on test-time scaling in large language models. *arXiv preprint arXiv:2503.24235*.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

A Formation of the MCSQ framework

A.1 Objective Information Theory (OIT) and GIME Metrics

OIT is a rigorous framework for quantifying information from an objective, mathematically grounded perspective. OIT conceptualizes information as a mapping from the state of a noumenon (object) over a time interval to the state of its carrier at another time. Based on OIT, the General Information Metrics Evaluation (GIME) (Xu et al., 2025b) method introduces eleven universal information metrics, each with a formal definition, allowing for comprehensive and systematic evaluation of datasets and information resources in machine learning. The eleven GIME metrics are:

- **Volume**: $\text{volume}_\sigma(I) = \sigma(g(c, T_m))$
- **Delay**: $\text{delay}(I) = \sup T_m - \sup T_h$
- **Scope**: $\text{scope}_\sigma(I) = \sigma(o)$
- **Granularity**: $\text{granularity}_\sigma(I) = \frac{\int_\Lambda \sigma(o_\lambda) d\mu}{\mu(\Lambda)}$
- **Variety**: $\text{variety}_R(I) = [f(o, T_h)]_R$
- **Duration**: $\text{duration}(I) = \sup T_h - \inf T_h$
- **Sampling rate**: $\text{sampling rate}(I) = \frac{|\Lambda|}{\text{duration}(I)}$
- **Aggregation**: $\text{aggregation}(I) = \frac{|R|}{|f(o,T_h)|}$
- **Coverage**: $\text{coverage}_\sigma(I) = \int_\Lambda \sigma(c_\lambda) d\mu$
- **Distortion**: $\text{distortion}_J(I) = d(f, f_e)$
- **Mismatch**: $\text{mismatch}_{I_0}(I) = d(I, I_0)$

These metrics provide a mathematical foundation for describing the size, diversity, structure, temporal dynamics, and reliability of information, and have been successfully applied to pretraining data selection and evaluation in large-scale AI systems.

A.2 Mapping GIME's 11 Metrics to MCSQ's 6 Dimensions

Difference between Pretraining Data and SFT Data Pretraining data for large language models (LLMs) is typically massive, heterogeneous, and weakly supervised, comprising diverse sources (e.g., web pages, books, code) with an emphasis on coverage, scale, and diversity. In contrast, supervised fine-tuning (SFT) data is usually smaller in scale, highly curated, and strongly annotated, often taking the form of (instruction, reasoning traces, answer) triplets that target specific capabilities such as stepwise reasoning. This fundamental difference leads to a shift in the relevance and interpretability of certain information metrics when moving from pretraining to SFT settings.

Challenges in Adapting Metrics Several GIME metrics, especially those related to temporal properties (*delay, duration, sampling rate*) and certain structural properties (*aggregation, coverage*), are difficult to meaningfully instantiate in the context of static, single-turn, and highly structured SFT data. For example, SFT samples do not typically encode natural timestamps, nor are they sampled from a temporal process. Similarly, the notions of aggregation or coverage, which are meaningful for large-scale, unstructured corpora, may not translate directly to smaller, highly focused SFT datasets. Furthermore, SFT data puts more emphasis on properties such as task relevance, stepwise correctness, and reasoning diversity, which requires reinterpreting and reweighting the original metrics.

Theoretical Basis and Formal Mapping The MCSQ framework distills six core dimensions from the GIME/OIT metrics, tailored for the reasoning-centric SFT scenario. Each MCSQ dimension is a principled instance or aggregation of GIME metrics, as shown below:

1. Volume (V): OIT/GIME:

$$\text{volume}_\sigma(I) = \sigma(g(c, T_m))$$

where $g(c, T_m)$ is the reflecting set of information I and σ is a measure function; for discrete data, this is simply cardinality.

In SFT: Each sample $x_i = (q_i, t_i, a_i)$ is a reflected information unit. Thus, for an SFT dataset $D = \{x_1, x_2, \ldots, x_N\}$,

$$V = |D| = N = \sigma(g(c, T_m))$$

2. Scope (S): OIT/GIME:

$$\text{scope}_\sigma(I) = \sigma(o)$$

where o is the ontology of information.

In SFT: The ontology is mapped as the set of unique subject domains $\mathcal{S}$ (e.g., mathematics subfields). Thus,

$$S = |\mathcal{S}| = \sigma(o)$$

Here, $o = \bigcup_{x \in D} \text{domain}(x)$, and σ is the set cardinality.

3. Granularity (G): OIT/GIME:

$$\text{granularity}_\sigma(I) = \frac{\int_\Lambda \sigma(o_\lambda) d\mu}{\mu(\Lambda)}$$

where o_λ is an atomic ontology, Λ is the index set, and μ is a measure on Λ.

In SFT: Each domain $d \in \mathcal{S}$ is an atomic ontology. Let n_d be the number of samples from domain d in D and N_d be the available pool size for d. We instantiate:

$$G = \frac{1}{|\mathcal{S}|} \sum_{d \in \mathcal{S}} \frac{n_d}{N_d}$$

This is the mean relative coverage per domain, reflecting the sampling uniformity across ontologies.

4. Variety (Vy): OIT/GIME:

$$\text{variety}_R(I) = [f(o, T_h)]_R$$

where $[\cdot]_R$ denotes the number of equivalence classes of states under relation R.

In SFT: We define R as the equivalence relation induced by reasoning type (e.g., Math, Science, General). Let p_k be the proportion of samples of type k. Instead of only counting classes, we use Shannon entropy to measure both richness and balance:

$$Vy = H(\text{type}) = -\sum_k p_k \log p_k$$

This is a strict refinement, as $H(\cdot)$ achieves its maximum when all types are equally present.

5. Distortion (D): OIT/GIME:

$$\text{distortion}_J(I) = d(f, f_e)$$

where d is a distance function between the reflected state f and the restored state f_e.

In SFT: For each sample x, teacher/judge models $\mathcal{M} = \{M_j\}$ provide binary correctness $I_j(x)$

and weights w_j (with $\sum_j w_j = 1$). The "distance" is $1 - \sum_j w_j I_j(x)$, i.e., the probability the ensemble disagrees with x. Then:

$$D = \frac{1}{|D|} \sum_{x \in D} \left[1 - \sum_j w_j I_j(x) \right]$$

This is the empirical mean distortion across all samples.

6. Mismatch (M): OIT/GIME:

$$\text{mismatch}_{I_0}(I) = d(I, I_0)$$

where I_0 is the target information and d is a distance function.

In SFT: Let I_0 be the subset of "reasoning-relevant" information. For each $x \in D$, define the indicator $1_{\text{nonreasoning}(x)}$ (1 if x is not reasoning-relevant, 0 otherwise). Then:

$$M = \frac{1}{|D|} \sum_{x \in D} 1_{\text{nonreasoning}(x)}$$

This is the average mismatch rate, i.e., the fraction of samples not matching the intended reasoning target.

Summary The above mappings show that each MCSQ metric is a concrete instantiation or refinement of its OIT/GIME ancestor, with all formulae explicitly derived from the OIT definitions under the semantics of SFT data. Metrics such as *delay, duration, sampling rate, aggregation, coverage* are omitted since SFT data is typically static, single-turn, and non-temporal, making these metrics either constant or uninformative in this context. However, in future, when SFT scenarios become more dynamic or multi-turn, these metrics could be appropriately instantiated.

In conclusion, MCSQ is not a heuristic selection but a theoretically grounded, mathematically justified reduction of the OIT/GIME system, ensuring both interpretability and empirical relevance for reasoning-centric LLM fine-tuning data.

B Dataset construction details

B.1 The Construction of s1-54k Dataset

Our experiments are based on the s1-54k dataset, which is derived from the s1 project(We follow the Apache license 2.0 for its data use agreement.). The original collection consists of 59,029 questions sourced from 16 diverse sources, covering a broad range of domains and reasoning styles. To ensure high data quality, we performed rigorous filtering to remove any samples with missing fields, resulting in a final dataset of 54,046 valid instances.

The dataset integrates questions from various sources, each annotated with its corresponding reasoning type and detailed source information. Table 1 summarizes the breakdown of samples by source and reasoning type.

B.2 Methods for Dividing Data Subsets under 6 Metrics

To systematically analyze how different dataset properties affect LLM reasoning, we construct data subsets along six quantitative dimensions defined in the MCSQ framework. Below are the detailed partitioning methods for each dimension.

1. Volume We sample subsets of different sizes from the full s1-54k dataset to study the effect of data scale. Specifically, we construct sets with volumes of 100, 1k, 5k, 10k, 30k, and 54k samples using stratified sampling to preserve the original data distribution.

2. Scope Domain classification: We use Claude 3.5 Sonnet to classify each question into a domain based on the Mathematics Subject Classification (MSC) system from the American Mathematical Society, covering both mathematical and scientific topics (such as geometry, combinatorics, biology, physics, economics, etc.). Experimental groups and sampling strategies(Table 2)

3. Granularity This dimension studies the effect of domain-level sample distribution. Volume fixed at 10k samples, Scope fixed to the top 10 domains (as in S1), ensuring moderate diversity. Granularity is controlled by adjusting intra-domain sampling ratios(Table 3).

4. Variety This dimension controls reasoning type diversity(Table 4). The original s1-59k lacks the reasoning type label, and we divide it according to the classification strategy of diverse real-world scenarios (Zhang et al., 2025). Science_Focus and General_Focus are not discussed here due to too little data.

5. Distortion Distortion quantifies model disagreement on sample correctness, calculated as: $Distortion = 1 - (w_1 Gemini\_acc + w_2 Qwen7B\_acc + w_3 Qwen32B\_acc)$, where w_1, w_2, w_3 are weights reflecting validation sample

sizes for each model. We partition samples into five distortion levels(Table 5).

6. Mismatch Mismatch is measured using Deepseek-V3, which labels each question with an `isreasoning` boolean and an `isreasoning_score` (0–100) via the following prompt:

Evaluate the suitability of the following question for reasoning training. Respond ONLY with an integer score between 0 and 100. Do not include any other text, explanation, or labels.
Question: {question_text}
Subsets are constructed by score ranges(Table 6).

Summary For each dimension, we vary only the target property while keeping other variables as constant as possible, enabling controlled experiments to isolate the causal effect of each data characteristic on LLM reasoning and generalization performance.

C Detailed Evaluation Results

This section provides the detailed tables referenced in Section 4.2.

Due to resource constraints, our primary experiments focused on Qwen2.5-3B-Instruct and the s1-54k dataset to enable tightly controlled, causally interpretable studies. In response, we have conducted additional experiments using both different parameter sizes within the Qwen2.5 family. The results on the MATH500 dataset are summarized in Table 31:

In addition to volume, we also evaluated the effect of mismatch on multiple model scales (Qwen2.5-0.5B/3B/7B-Instruct) by averaging Pass@1 and Pass@5 across AMC23, GPQA-Diamond, MATH500, Minerva, and Olympiad-Bench. The results are shown in Table 32:

To further validate generality, we also report results of Qwen2.5-3B-Instruct on the challenging DeepMath-103K reasoning corpus, evaluating the effect of training volume. The following Table 33 shows Pass@1 and Pass@5 results across five math benchmarks:

These results further confirm that, even on a significantly different and more challenging dataset, increasing fine-tuning data volume yields consistent improvements, though with diminishing returns. The structure-first conclusions are also preserved.

D Additional Experimental Details and Results

We acknowledge that exclusive reliance on automated model scoring may introduce bias, but large-scale human annotation is impractical for our dataset size. To reduce this risk:

- Distortion: We adopt an ensemble of teacher models (Gemini-2.0-flash-thinking-exp-1219, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-32B), weighted by reliability, to compute distortion. This ensemble approach helps mitigate bias from any single model and yields more robust labels.
- Mismatch: Given that LLM judges such as DeepSeek-V3-0324 may produce clustered scores, we discretize the outputs into three bins (M0–M2), reframing the task as a classification problem to enhance reliability.

To further validate our approach, we conducted a human annotation study on 500 samples. As shown in Table 34, the agreement between automated and human labels is high, particularly for reasoning tasks.

The inherent trade-off between enhancing reasoning specialization and preserving general capabilities during supervised fine-tuning (SFT) is a core challenge for the development and deployment of large language models. Based on our experimental findings and recent developments in LLM training, we recommend several concrete strategies:

1. Structural Regularization with "Imperfect" Data. Injecting a controlled proportion of higher-distortion or mismatched data during SFT can serve as regularization, discouraging overfitting and helping preserve general capabilities while boosting out-of-domain reasoning.
2. Multi-Objective Data Curation (MCSQ-guided). By leveraging the quantitative axes of MCSQ, dataset construction can be posed as a multi-objective optimization: maximize reasoning accuracy and minimize general ability degradation. This enables principled balancing (e.g., via Pareto frontier search or active data selection).
3. Curriculum/Staged Fine-Tuning. First fine-tune on broad data for generalization, then on

reasoning-optimized data (following MCSQ principles) to incrementally enhance specialization. Our and prior work both support this staged approach.

4. Architectural and Regularization Techniques. Model-level strategies (e.g., mixture-of-experts, multi-task learning) and explicit regularization (e.g., elastic weight consolidation, distillation) can further mitigate capability loss.

Preliminary Validation: Three-Stage Filtering Funnel

We implemented a simple three-stage filtering funnel:

- Macro: Select data domains (Scope)
- Core: Optimize challenge and relevance (Mismatch/Distortion)
- Micro: Balance diversity and quantity (Granularity/Variety/Volume)

Compared to s1 and LIMO baselines (Qwen2.5-3B-Instruct), our MCSQ-based funnel achieves competitive or better results on reasoning and general benchmarks(Table 35, Table 36).

Table 1: Statistics of the s1-54k dataset by source and reasoning type.

Source	Reasoning Type	Samples	Subset Details
NuminaMATH	Math	30,655	AI-MO/NuminaMath-CoT/aops_forum: 30,171; AI-MO/NuminaMath-CoT/cn_k12: 311; AI-MO/NuminaMath-CoT/olympiads: 173
MATH	Math	11,953	qfq/openaimath/Algebra: 2,807; qfq/openaimath/Intermediate Algebra: 2,095; qfq/openaimath/Prealgebra: 1,991; qfq/openaimath/Number Theory: 1,334; qfq/openaimath/Geometry: 1,298; qfq/openaimath/Precalculus: 1,232; qfq/openaimath/Counting & Probability: 1,196
OlympicArena	Science	226	GAIR/OlympicArena/Math: 169; GAIR/OlympicArena/Physics: 29; GAIR/OlympicArena/Chemistry: 14; GAIR/OlympicArena/Astronomy: 8; GAIR/OlympicArena/Biology: 6
OmniMath	Math	4,237	KbsdJames/Omni-MATH: 4,237
AGIEval	General-purpose (basic, not reasoning-intensive)	2,312	baber/agieval/logiqa: 654; baber/agieval/lsat_lr: 512; baber/agieval/lsat_rc: 271; baber/agieval/aqua_rat: 259; baber/agieval/lsat_ar: 233; baber/agieval/sat_math: 225; baber/agieval/sat_en: 137; baber/agieval/math_agieval: 21
OlympiadBench	Science	896	Hothan/OlympiadBench/Theorem proof/Math: 503; Hothan/OlympiadBench/Open-ended/Physics: 236; Hothan/OlympiadBench/Open-ended/Math: 132; Hothan/OlympiadBench/Theorem proof/Physics: 25
AIME (1983–2021)	Math	769	qq8933/AIME_1983_2024: 769
TheoremQA	Science	747	TIGER-Lab/TheoremQA/float: 360; TIGER-Lab/TheoremQA/integer: 200; TIGER-Lab/TheoremQA/bool: 112; TIGER-Lab/TheoremQA/list of integer: 51; TIGER-Lab/TheoremQA/option: 16; TIGER-Lab/TheoremQA/list of float: 8
JEEBench	Science	514	daman1209arora/jeebench/math: 236; daman1209arora/jeebench/chem: 155; daman1209arora/jeebench/phy: 123
GPQA (eval only)	Science	307	Idavidrein/gpqa: 307
SciEval	Science	227	OpenDFM/SciEval/chemistry/multiple-choice/SocraticQA: 52; OpenDFM/SciEval/biology/multiple-choice/MedQA: 43; OpenDFM/SciEval/chemistry/filling/reagent selection: 43; OpenDFM/SciEval/biology/judge/PubMedQA: 35; OpenDFM/SciEval/biology/multiple-choice/SocraticQA: 29; OpenDFM/SciEval/physics/multiple-choice/SocraticQA: 25
s1-prob	Math	205	qfq/quant: 23; qfq/stats_qual: 182
s1-teasers	Math	998	0xharib/xword1: 998
All 54K questions		**54,046**	

Table 2: Domain coverage and sampling strategies for Scope dimension

Group	Scope Rate	Sampling Strategy
S0 (Scope-Expert)	5/53 = 0.094	Select the five core domains recommended by mathematics education experts: Geometry, Number Theory, Combinatorics, Real Functions, and Probability Theory and Stochastic Processes; oversample these domains.
S1	10/53 = 0.189	Select samples from the top 10 domains by sample size.
S2	20/53 = 0.377	Select samples from the top 20 domains by sample size.
S3	30/53 = 0.566	Select samples from the top 30 domains by sample size.
S4	40/53 = 0.755	Select samples from the top 40 domains by sample size.
S5	53/53 = 1.0	Use all 53 domains.

Table 3: Sampling strategies for Granularity dimension

Group	Granularity Value	Sampling Strategy
G0 (Lean)	0.154	Bipolar: Oversample the top 2 domains (Geometry and Number Theory), each at 35% (3.5k samples), with the remaining 30% (3k) drawn from the other 8 domains according to their original proportions.
G1 (Preserved)	0.204	Natural: Draw 10k samples from the top 10 domains in proportion to their original distribution.
G2 (Uniform)	0.441	Uniform: Sample 1,000 from each of the top 10 domains (10k/10 = 1k). If a domain has fewer than 1k, take all its samples and reallocate the remainder to other domains to keep the distribution as balanced as possible.
G3 (Extreme)	0.564	Extreme: Sample only from the smallest domains; take 100% of the available samples from as many small domains as needed to reach 10k (up to 6/10 = 0.6 ratio).

Table 4: Sampling strategies for Variety dimension

Group	Composition	Description	Variety Ratio
V0	Math(100%)	Only Math questions, minimal diversity	0.000
V1	Math(92.7%), Science(3.8%), General(3.5%)	Natural mix reflecting the original dataset	0.448
V2	Math(81.2%), Science(18.8%)	Remove General type for moderate diversity	0.697
V3	Math(60.5%), Science(18.8%), General(20.7%)	Maximize diversity within 10k budget	1.363
Science_Focus	Science(100%)	Science only; Volume = 1,877	–
General_Focus	General(100%)	General only; Volume = 2,066	–

Table 5: Sampling strategies for Distortion dimension

Group	Distortion Ratio	Description	Sampling Details
D1	0.0	No distortion: all models (Gemini, Qwen-32B, Qwen-7B) agree and are correct.	10k samples, all models correct.
D2	0.239	Low distortion: 31.1% all models correct, 53% two models correct, 15.9% one model correct. Gemini acc 100%, Qwen-32B acc 80.6%.	10k samples.
D3	0.604	Medium distortion: only one model (mostly Gemini) is correct; Gemini acc 87%, Qwen-32B acc 7.9%, Qwen-7B acc 5.1%.	10k samples.
D4	0.981	High distortion: 92.9% all models incorrect, 7.1% only Qwen-7B correct; Gemini and Qwen-32B acc 0%.	10k samples.
D5	1.0	Maximum distortion: all models incorrect.	10k samples, all models incorrect.

Table 6: Sampling strategies for Mismatch dimension

Group	Description	Sampling Criteria	Count	Avg Score
M0	Low mismatch	isreasoning_score in [90,100], volume=10k	25,104	90.04
M1	Medium mismatch	isreasoning_score in (80,90), volume=10k	17,528	85.00
M2	High mismatch	isreasoning_score in [0,80], volume=10k	11,307	70.44

Table 7: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Fine-tuning Data Volume. Scores represent accuracy. Best performance among fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Dataset Volume	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
Volume=100	0.2250	0.1061	0.4580	0.2243	0.1659
Volume=1,000	0.3000	0.1667	0.5460	0.2096	0.2326
Volume=5,000	0.3250	0.1869	0.5800	0.2463	0.2459
Volume=10,000	0.4000	0.1919	0.6020	**0.2610**	0.2756
Volume=30,000	**0.4500**	**0.2020**	**0.6140**	0.2206	**0.2963**
Volume=54,046	0.5000	0.2170	0.6260	0.2853	0.3115

Table 8: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Fine-tuning Data Volume. Scores represent accuracy. Best performance among fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Volume	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
Volume=100	0.3750	0.4242	0.7140	0.4375	0.3067
Volume=1,000	0.6000	0.6263	0.8120	0.4485	0.4178
Volume=5,000	0.6500	0.6919	**0.8160**	0.4375	0.4533
Volume=10,000	0.6250	0.6414	0.8040	0.4485	0.4637
Volume=30,000	**0.7000**	**0.7071**	0.8020	**0.4596**	**0.5156**
Volume=54,046	0.7750	0.7269	0.8280	0.4659	0.5963

Table 9: Average Performance on General Capability Benchmarks with Varying Fine-tuning Data Volume. Scores represent accuracy (MMLU is 5-shot avg). Best performance among fine-tuned models is in **bold**.

Model / Dataset Volume	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
Volume=100)	**66.98**	**74.68**	73.77
Volume=1,000	66.15	74.49	**74.29**
Volume=5,000	65.91	73.84	73.48
Volume=10,000	65.33	73.58	72.36
Volume=30,000	64.93	72.84	71.55
Volume=54,046	64.48	72.52	70.98

Table 10: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Fine-tuning Data Volume.

Volume	MMLU					CMMLU					C-Eval				
	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
100	66.98	61.56	77.48	60.98	70.79	74.68	67.88	74.75	77.78	77.87	73.77	66.28	84.36	77.04	72.40
1k	66.15	61.53	76.96	58.92	70.70	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
5k	65.91	60.34	76.47	59.55	70.30	73.84	66.61	73.99	77.18	77.08	73.48	66.74	84.00	77.04	71.09
10k	65.33	59.74	76.37	58.62	69.77	73.58	66.46	74.10	76.66	76.49	72.36	65.58	83.27	75.10	70.31
30k	64.93	58.18	75.82	58.85	69.68	72.84	65.07	73.17	75.69	76.74	71.55	62.79	81.82	75.49	71.35
54k	64.48	58.40	75.46	57.93	69.28	72.52	64.41	72.77	75.57	76.67	70.98	62.32	81.45	75.49	70.05

Table 11: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Domain Scope (S) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among S0-S5 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Scope Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
S0 - Expert 5	0.3000	0.2222	0.6000	0.2243	0.2652
S1 - Top 10	0.2500	0.2020	0.6080	0.2243	0.2622
S2 - Top 20	0.3750	**0.2424**	0.6020	0.2169	0.2607
S3 - Top 30	**0.4000**	0.1869	**0.6200**	**0.2316**	**0.2741**
S4 - Top 40	0.2500	0.2020	0.5900	0.2721	0.2415
S5 - All 53	0.3000	0.2121	0.5660	0.2169	0.2311

Table 12: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Domain Scope (S) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among S0-S5 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Scope Setting	**AMC23**	**GPQA-Diamond**	**MATH500**	**Minerva**	**OlympiadBench**
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
S0 - Expert 5	0.6500	0.7273	0.8080	0.4412	0.4874
S1 - Top 10	0.6250	0.6616	0.8020	**0.4596**	0.4533
S2 - Top 20	0.6750	0.6869	0.8100	0.4522	0.4607
S3 - Top 30	0.6750	0.5606	**0.8180**	0.4338	**0.4889**
S4 - Top 40	**0.7250**	**0.7475**	**0.8180**	0.4301	0.4578
S5 - All 53	0.6750	0.7273	0.8080	0.4449	0.4785

Table 13: Average Performance on General Capability Benchmarks with Varying Domain Scope (S) at Fixed Volume (10k samples). Scores represent accuracy (MMLU is 5-shot avg).

Model / Scope Setting	**MMLU (Avg)**	**CMMLU (Avg)**	**C-Eval (Avg)**
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
S0 - Expert 5)	65.25	73.84	72.14
S1 - Top 10)	65.10	73.91	72.36
S2 - Top 20)	65.18	73.46	72.88
S3 - Top 30)	65.25	73.57	72.59
S4 - Top 40)	65.24	73.71	72.88
S5 - All 53)	**65.69**	**73.99**	**73.03**

Table 14: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Domain Scope (S) at Fixed Volume (10k samples).

Scope	**MMLU**					**CMMLU**					**C-Eval**				
Setting	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
S0	65.25	59.71	76.21	58.55	69.71	73.84	66.73	74.01	76.98	77.11	72.14	63.72	82.91	77.82	70.05
S1	65.10	59.91	76.67	58.17	69.03	73.91	66.89	74.04	77.06	77.15	72.36	64.65	83.64	77.04	69.79
S2	65.18	60.17	76.63	57.87	69.56	73.46	65.63	74.10	77.02	76.43	72.88	66.05	84.00	76.26	70.31
S3	65.25	58.88	76.93	58.30	70.20	73.57	65.78	73.99	77.06	76.84	72.59	64.19	84.36	78.21	69.79
S4	65.24	59.31	76.57	58.53	69.74	73.71	66.53	74.32	77.02	76.36	72.88	65.12	84.36	77.04	70.57
S5	65.69	60.17	76.96	58.53	70.51	73.99	66.50	74.12	77.54	77.29	73.03	65.81	84.36	76.26	70.83

Table 15: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Domain Granularity (G) at Fixed Volume (10k samples) and Scope (Top 10 Domains). Scores represent accuracy. Best performance among the four settings for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Granularity Setting	**AMC23**	**GPQA-Diamond**	**MATH500**	**Minerva**	**OlympiadBench**
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
G0=0.154	0.2750	0.1869	0.5840	0.2500	0.2474
G1=0.204	0.2500	0.1768	0.6040	**0.2647**	0.2444
G2=0.441	0.3000	0.1818	**0.6180**	0.2426	**0.2681**
G3=0.564	**0.3750**	**0.2323**	0.5840	0.2574	0.2474

Table 16: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Domain Granularity (G) at Fixed Volume (10k samples) and Scope (Top 10 Domains). Scores represent accuracy. Best performance among the four settings for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Granularity Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
G0=0.154	**0.7750**	0.6414	**0.8220**	0.4485	0.4607
G1=0.204	0.7500	0.6414	0.7980	**0.4706**	**0.4756**
G2=0.441	0.5500	**0.6919**	0.8160	0.4596	0.4726
G3=0.564	0.5750	0.6717	0.8180	0.4412	0.4681

Table 17: Average Performance on General Capability Benchmarks with Varying Domain Granularity (G) at Fixed Volume (10k samples) and Scope (Top 10 Domains). Scores represent accuracy (MMLU is 5-shot avg). Best performance among the four settings is in **bold**.

Model / Granularity Setting	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
G0=0.154	65.37	**74.21**	72.07
G1=0.204	**65.55**	73.52	**72.96**
G2=0.441	65.41	73.50	**72.96**
G3=0.564	65.52	73.43	71.84

Table 18: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Domain Granularity (G) at Fixed Volume (10k samples) and Scope (Top 10 Domains).

Granularity	MMLU					CMMLU					C-Eval				
Setting	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
Extreme	65.52	60.24	76.67	58.19	70.51	73.43	65.67	73.85	76.70	76.87	71.84	63.02	83.27	77.04	70.05
Uniform	65.41	60.30	76.11	58.28	70.36	73.50	65.63	74.01	76.58	77.08	72.96	65.35	83.64	78.21	70.31
Preserved	65.55	59.64	76.57	58.96	70.14	73.52	65.51	73.77	77.10	77.11	72.96	65.12	84.00	76.65	71.35
Lean	65.37	59.87	76.54	58.36	70.05	74.21	67.17	74.23	77.50	77.49	72.07	64.42	83.27	75.88	70.05

Table 19: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Reasoning Type Variety (V). Scores represent accuracy. Best performance among V1-V2 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference. ($Science/General_Focus$ results are pending/omitted for brevity).

Model / Variety Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
V0 - $Math_Focus$	0.4250	0.1919	0.5960	0.2279	0.2548
V1 - Preserved	**0.4750**	0.1667	**0.6000**	0.2390	0.2563
V2 - $Reasoning_Focus$	0.2250	0.1717	0.5880	0.2206	0.2444
V3 - Balanced	0.3750	**0.2222**	0.5880	**0.2500**	**0.2593**

Table 20: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Reasoning Type Variety (V). Scores represent accuracy. Best performance among V1-V4 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Variety Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
V0 - $Math_Focus$)	0.6250	0.5960	**0.8200**	**0.4485**	0.4756
V1 - Preserved)	0.6000	0.6515	0.8080	0.4265	0.4830
V2 - $Reasoning_Focus$)	0.5750	0.6414	0.8180	0.4154	**0.4933**
V3 - Balanced)	**0.7000**	**0.6566**	0.8020	0.4191	0.4474

Table 21: Average Performance on General Capability Benchmarks with Varying Reasoning Type Variety (V). Scores represent accuracy (MMLU is 5-shot avg). Best performance among V1-V4 fine-tuned models is in **bold**.

Model / Variety Setting	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
V0 - $Math_Focus$)	65.65	**73.99**	72.81
V1 - Preserved)	64.36	73.56	71.84
V2 - $Reasoning_Focus$)	65.62	73.65	72.59
V3 - Balanced)	**65.77**	73.32	**73.11**

Table 22: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Reasoning Type Variety (V).

Variety	MMLU					CMMLU					C-Eval				
Setting	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
V0 (MathFoc)	65.65	60.11	76.54	59.23	69.80	73.99	66.69	74.15	77.58	77.08	72.81	64.42	85.09	77.43	70.31
V1 (Pres.)	64.36	59.24	76.11	56.26	69.74	73.56	65.82	73.99	77.10	76.74	71.84	63.72	84.36	76.26	69.01
V2 (ReasFoc)	65.62	59.81	76.54	59.06	70.20	73.65	65.90	74.01	77.18	76.91	72.59	64.19	84.73	76.65	70.57
V3 (Balanced)	65.77	60.14	76.86	59.11	70.14	73.32	65.98	73.80	76.62	76.29	73.11	65.12	83.27	78.99	70.83

Table 23: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Data Distortion (D) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among D1-D5 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Distortion Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
D0 = 0.0	0.3250	**0.2677**	0.6060	0.2353	**0.2770**
D1 = 0.24	0.2750	0.1768	**0.6260**	0.2316	0.2667
D2 = 0.60	**0.3750**	0.2525	0.5980	0.2279	0.2681
D3 = 0.98	0.3250	0.1768	0.5500	**0.2500**	0.2459
D4 = 1.0	0.3000	0.1869	0.5500	0.2279	0.2296

Table 24: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Data Distortion (D) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among D1-D5 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Distortion Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
D0 = 0.0	0.5750	0.7071	**0.8320**	**0.4743**	0.4578
D1 = 0.24	0.6250	0.6667	0.8180	0.4265	0.4815
D2 = 0.60	0.6250	0.7525	0.8220	0.4265	**0.4874**
D3 = 0.98	**0.6750**	**0.7677**	0.8180	0.4375	0.4652
D4 = 1.0	0.6000	0.6869	0.8060	0.4412	0.4726

Table 25: Average Performance on General Capability Benchmarks with Varying Data Distortion (D) at Fixed Volume (10k samples). Scores represent accuracy (MMLU is 5-shot avg). Note the counter-intuitive trend.

Model / Distortion Setting	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
D0 = 0.0)	65.13	73.42	72.51
D1 = 0.24)	64.73	73.53	72.21
D2 = 0.60)	65.23	73.74	72.66
D3 = 0.98)	65.28	73.66	72.36
D4 = 1.0)	**65.53**	**73.79**	**73.55**

Table 26: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Data Distortion (D) at Fixed Volume (10k samples).

Distortion	MMLU					CMMLU					C-Eval				
Setting (D value)	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
D0 (0.00)	65.13	59.24	76.80	58.04	69.80	73.42	65.67	73.93	76.46	76.94	72.51	66.28	82.55	77.43	69.01
D1 (0.24)	64.73	58.91	76.63	57.43	69.43	73.53	66.10	73.80	77.34	76.39	72.21	64.42	84.36	75.88	69.79
D2 (0.60)	65.23	59.41	76.76	58.19	69.90	73.74	66.61	73.71	76.90	77.29	72.66	63.49	85.09	77.82	70.57
D3 (0.98)	65.28	59.15	76.54	58.62	69.99	73.66	66.02	73.66	77.30	77.18	72.36	64.65	82.91	77.04	70.31
D4 (1.00)	**65.53**	59.44	76.60	58.98	70.20	**73.79**	66.46	73.82	77.46	76.98	**73.55**	66.51	83.64	78.21	71.09

Table 27: Performance on Reasoning Benchmarks (Pass@1 Accuracy) with Varying Data Mismatch (M) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among M0-M2 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Mismatch Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.4500	0.3182	0.6340	0.3088	0.2770
M0	**0.3250**	0.2020	0.5820	**0.2537**	0.2667
M1	**0.3250**	0.2121	**0.6160**	0.2132	**0.2785**
M2	0.2750	**0.2424**	0.5940	0.2206	0.2474

Table 28: Performance on Reasoning Benchmarks (Pass@5 Accuracy) with Varying Data Mismatch (M) at Fixed Volume (10k samples). Scores represent accuracy. Best performance among M0-M2 fine-tuned models for each benchmark is in **bold**. Base model performance is provided for reference.

Model / Mismatch Setting	AMC23	GPQA-Diamond	MATH500	Minerva	OlympiadBench
Base (Qwen2.5-3B-Instruct)	0.7500	0.6414	0.8180	0.4522	0.4444
M0	0.6250	0.6566	**0.8040**	**0.4522**	0.4578
M1	**0.7000**	0.6414	**0.8040**	0.4301	**0.4815**
M2	0.6000	**0.7576**	**0.8040**	0.4338	0.4726

Table 29: Average Performance on General Capability Benchmarks with Varying Data Mismatch (M) at Fixed Volume (10k samples). Scores represent accuracy (MMLU is 5-shot avg).

Model / Mismatch Setting	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)
Base (Qwen2.5-3B-Instruct)	66.69	74.49	74.29
M0 - Low Mismatch)	65.66	73.63	72.66
M1 - Medium Mismatch)	65.21	73.21	71.99
M2 - High Mismatch)	65.63	73.63	72.88

Table 30: Detailed Performance on General Capability Benchmarks (MMLU, CMMLU, C-Eval) by Subject Category with Varying Data Mismatch (M) at Fixed Volume (10k samples).

Mismatch	MMLU					CMMLU					C-Eval				
Setting	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other	Avg	STEM	Soc Sci	Humanities	Other
Base	66.69	61.76	77.77	59.64	70.97	74.49	67.13	74.75	78.02	77.56	74.29	66.98	85.09	75.88	73.70
M0 (Low)	65.66	59.91	76.44	59.21	70.20	73.63	65.63	74.12	77.42	76.84	72.66	64.88	83.64	76.65	71.09
M1 (Med)	65.21	59.15	76.34	58.62	69.83	73.21	65.43	73.47	76.86	76.56	71.99	63.95	82.91	76.65	70.05
M2 (High)	65.63	59.58	76.67	58.85	70.45	73.63	66.10	73.85	77.10	77.05	72.88	65.12	84.00	77.43	70.31

Table 31: Performance on MATH500 across different Qwen2.5 model scales.

Fine-tune Setting	Qwen2.5-0.5B	Qwen2.5-3B	Qwen2.5-7B
base+volume(100)	0.1780	0.4580	0.7340
base+volume(1000)	0.1240	0.5460	0.6820
base+volume(5000)	0.1280	0.5800	0.7360
base+volume(10000)	0.1540	0.6020	0.7540
base+volume(30000)	0.1900	0.6140	0.7720
base+volume(54046)	0.1920	0.6200	0.7840

Table 32: Effect of Mismatch dimension across Qwen2.5 models (Pass@1 and Pass@5 averaged over reasoning benchmarks).

Fine-tune\base	pass@k	Qwen2.5-0.5B	Qwen2.5-3B	Qwen2.5-7B
base+M0	pass@1	0.0689	0.3459	0.3698
base+M1	pass@1	0.0688	0.3290	0.3549
base+M2	pass@1	0.0615	0.3159	0.3501
base+M0	pass@5	0.2826	0.6258	0.6904
base+M1	pass@5	0.3159	0.6114	0.6984
base+M2	pass@5	0.3098	0.6336	0.7198

Table 33: Validation on DeepMath-103K: reasoning performance across benchmarks with increasing training volume.

Fine-tune	**Metric**	**AMC**	**GPQA**	**MATH**	**Minerva**	**OlympiadBench**
base+volume(100)	Pass@1	0.0000	0.0758	0.0300	0.0331	0.0059
base+volume(1000)	Pass@1	0.0500	0.1162	0.0700	0.0551	0.0178
base+volume(5000)	Pass@1	0.1000	0.1768	0.1500	0.0919	0.0415
base+volume(10000)	Pass@1	0.1500	0.2121	0.2000	0.1287	0.0593
base+volume(100)	Pass@5	0.0750	0.4141	0.2820	0.1213	0.0889
base+volume(1000)	Pass@5	0.1250	0.4848	0.3700	0.1691	0.1304
base+volume(5000)	Pass@5	0.2250	0.5808	0.5000	0.2500	0.2000
base+volume(10000)	Pass@5	0.3000	0.6313	0.5700	0.3015	0.2504

Table 34: Agreement between automated judges and human annotations on 500 samples.

Judge	**#Samples**	**Agreement with Human Annotation (Mismatch)**
Human (reference)	500	100%
DeepSeek-V3-0324	500	94.4%
Qwen2.5-32B-Instruct	500	92%

Table 35: Comparison of different curation methods on reasoning benchmarks.

Methods	Volume	GPQA-Diamond	MATH500	Minerva	OlympiadBench	Avg
MCSQ	1000/54k	**0.3283**	0.5720	0.2059	0.2267	0.3316
S1	1000/59k	0.1846	**0.5880**	0.2279	**0.2593**	0.3070
LIMO	817/100k	0.2879	0.5800	**0.2537**	0.2193	**0.3332**

Table 36: Comparison of different curation methods on general capability benchmarks.

Methods	Volume	MMLU (Avg)	CMMLU (Avg)	C-Eval (Avg)	Total (Avg)
MCSQ	1000/54k	64.65	**74.81**	72.30	**70.587**
S1	1000/59k	64.89	74.18	**72.44**	70.503
LIMO	817/100k	**65.00**	73.73	71.92	70.217