TINY-LR PROXY SFT FOR DATASET RANKING: AN EMPIRICAL INVESTIGATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Selecting high-quality training data is critical for supervised fine-tuning (SFT) of language models, but evaluating dataset value by full fine-tuning is expensive. Proxy models offer an efficient alternative, yet their rankings may not transfer reliably to larger target models. Recent work on pretraining suggests that training proxies with tiny learning rates improves ranking transfer. We test this hypothesis for SFT dataset ranking, comparing Standard-LR Proxy (5×10^{-5}) and Tiny-LR Proxy (1×10^{-5}) on 12 math datasets using Qwen2.5-1.5B as proxy and Qwen2.5-7B as target. Our experiments refute the hypothesis: Tiny-LR Proxy achieves random-level ranking agreement (PDA = 0.500) compared to Standard-LR Proxy's significantly above-random performance (PDA = 0.712, $p = 0.042$). However, we discover benchmark-specific behavior: Tiny-LR Proxy achieves excellent agreement on MATH-500 (PDA = 0.818) but fails on GSM8K, suggesting that reduced learning rates amplify sensitivity to surface-level features rather than preserving transferable quality signals.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Supervised fine-tuning (SFT) has become a critical step in developing capable language models, enabling them to follow instructions and perform diverse tasks (Ouyang et al., 2022; Wei et al., 2021). The choice of training data significantly impacts model quality, with recent work demonstrating that carefully curated datasets can outperform larger, less selective collections (Zhou et al., 2023). However, evaluating dataset value by full fine-tuning is computationally expensive, motivating the development of efficient proxy-based methods (Qin et al., 2024).

A common approach is to use smaller proxy models to rank candidate datasets, assuming that rankings transfer to larger target models. Recent work on pretraining data curation has shown that proxy rankings can be brittle, but that training proxies with tiny learning rates dramatically improves transfer to larger models (Wang et al., 2025). The mechanism is hypothesized to be that reduced learning rates keep the model in a near-linear optimization regime where gradient alignment signals dominate over higher-order effects.

We investigate whether this insight transfers to the instruction-tuning setting. Specifically, we test the hypothesis that reducing the proxy's learning rate from 5×10^{-5} to 1×10^{-5} improves dataset ranking agreement with a larger target model. We evaluate on 12 math-oriented SFT datasets using Qwen2.5-1.5B as the proxy and Qwen2.5-7B as the target, measuring ranking agreement via Pairwise Direction Accuracy (PDA) and Spearman correlation.

Our experiments refute the Tiny-LR hypothesis for SFT dataset ranking. Tiny-LR Proxy achieves random-level PDA (0.500) compared to Standard-LR Proxy's significantly above-random performance (0.712, $p = 0.042$). However, we discover an intriguing benchmark-specific pattern: Tiny-LR Proxy achieves excellent ranking agreement on MATH-500 (PDA = 0.818, $\rho = 0.846$) but fails completely on GSM8K (PDA = 0.515). This suggests that reduced learning rates amplify sensitiv-

[1] https://gitlab.com/fars-a/tinylr-proxy-sft-data-valuation

ity to surface-level features that inflate scores on simpler benchmarks without corresponding target model improvements.

Our contributions are: (1) We provide the first rigorous test of the Tiny-LR proxy hypothesis for SFT dataset ranking, finding that it does not improve transfer. (2) We demonstrate that Standard-LR Proxy SFT provides reliable ranking signals (PDA = 0.712) and correctly identifies the best dataset. (3) We discover benchmark-specific behavior that reveals when reduced learning rates may preserve meaningful signals versus when they amplify superficial features.

2 Related Work

2.1 Data Selection for Instruction Tuning

The quality and composition of training data significantly impact the effectiveness of supervised fine-tuning (SFT) for large language models (Qin et al., 2024; Wang et al., 2024). Quality-based approaches prioritize selecting high-quality examples that maximize alignment with human preferences. LIMA (Zhou et al., 2023) demonstrated that carefully curated datasets of only 1,000 examples can achieve competitive performance, emphasizing quality over quantity. Subsequent work has explored automatic quality assessment through various signals, including response complexity, instruction-following accuracy, and semantic diversity (Liu et al., 2023; Zhang et al., 2025).

Diversity-based methods focus on ensuring broad coverage of the instruction space. Self-Instruct (Wang et al., 2022) generates diverse instruction-response pairs through iterative self-generation, while FLAN (Chung et al., 2022) aggregates diverse task collections to improve generalization. Influence-based approaches leverage gradient information to identify impactful training examples. TracIn (Pruthi et al., 2020) estimates training data influence by tracking gradient descent, and LESS (Xia et al., 2024) extends this to targeted instruction tuning by selecting data that maximally influences performance on specific downstream tasks.

2.2 Proxy-Based Evaluation

Proxy models offer a computationally efficient alternative to full-scale evaluation by using smaller models to approximate the behavior of larger target models. This approach is grounded in scaling laws (Kaplan et al., 2020), which establish predictable relationships between model size and performance. Recent work has examined the reliability of proxy-based data curation, with Wang et al. (2025) demonstrating that small training runs can provide meaningful signals for data selection, though transfer to larger models remains imperfect. DSIR (Xie et al., 2023) uses importance resampling with proxy models to select pretraining data that matches target distributions.

2.3 Learning Rate and Transfer

The choice of learning rate fundamentally affects model generalization and transfer. Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2021) and LLaMA-Adapter (Zhang et al., 2023) implicitly constrain the magnitude of parameter updates, which can improve transfer to new tasks. This motivates our investigation of whether explicitly reducing learning rates during proxy training might preserve more transferable ranking signals by limiting overfitting to proxy-specific features.

3 Method

3.1 Problem Setup

Given a set of candidate supervised fine-tuning (SFT) datasets $\{D_1, \ldots, D_K\}$, our goal is to rank these datasets by their expected contribution to downstream task performance when used to fine-tune a target model M_t. The challenge is that evaluating each dataset by full fine-tuning is computationally expensive, scaling linearly with the number of candidates. We investigate whether proxy-based evaluation using a smaller model M_p can reliably predict the target model's dataset ranking, and specifically whether reducing the proxy's learning rate improves this transfer.

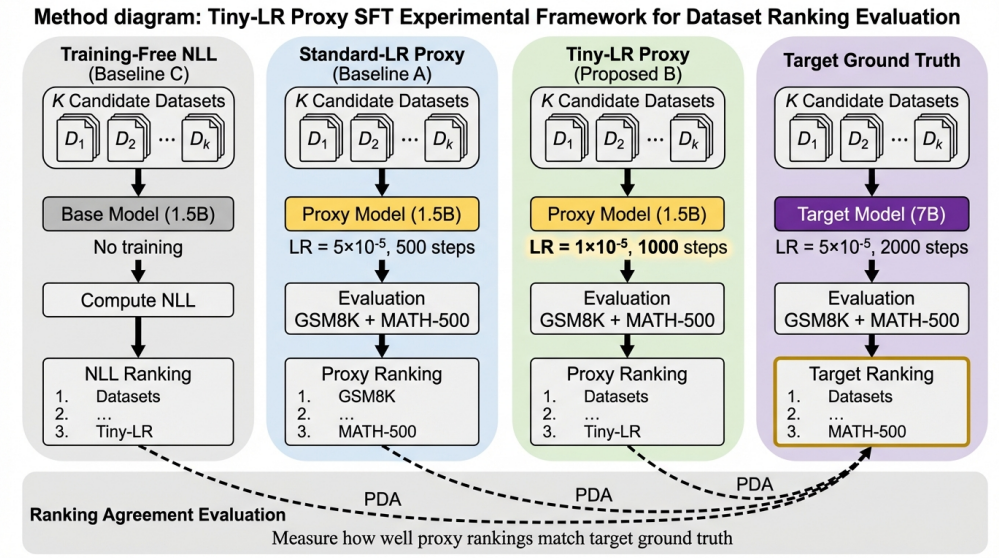

Figure 1: Overview of the experimental framework for evaluating proxy-based dataset ranking. Four conditions are compared: Standard-LR Proxy (Baseline A) using conventional learning rate (5×10^{-5}), Tiny-LR Proxy (Proposed B) using reduced learning rate (1×10^{-5}), Training-Free NLL (Baseline C) using base model perplexity, and Target Ground Truth using full fine-tuning on the larger model (Qwen2.5-7B). All methods rank 12 math datasets based on their evaluation performance on GSM8K and MATH-500 benchmarks.

3.2 Experimental Conditions

We compare four conditions for dataset ranking, illustrated in Figure 1:

Standard-LR Proxy SFT. Fine-tune the proxy model (Qwen2.5-1.5B) on each candidate dataset using a standard learning rate ($\eta = 5 \times 10^{-5}$) for 500 steps. This represents conventional proxy-based evaluation practice.

Tiny-LR Proxy SFT. Fine-tune the proxy model using a reduced learning rate ($\eta = 1 \times 10^{-5}$) for 1000 steps. The hypothesis is that smaller learning rates keep the model in a near-linear optimization regime where gradient alignment signals are more transferable across model scales (Wang et al., 2025).

Training-Free NLL. Compute the negative log-likelihood of each dataset under the base model without any fine-tuning. This baseline tests whether static distribution fit predicts dataset value.

Target Ground Truth. Fine-tune the target model (Qwen2.5-7B) on each dataset using standard hyperparameters ($\eta = 5 \times 10^{-5}$, 2000 steps). This provides the ground-truth ranking against which proxy methods are evaluated.

All fine-tuning uses LoRA (Hu et al., 2021) with rank 16, $\alpha = 32$, and dropout 0.05, implemented via LlamaFactory (Zheng et al., 2024). We use the Qwen2.5 model family (Yang et al., 2024) with cosine learning rate schedule and warmup ratio 0.1.

3.3 Evaluation Metrics

We evaluate ranking agreement between proxy and target methods using three metrics:

Pairwise Direction Accuracy (PDA). For each pair of datasets (D_i, D_j), we check whether the proxy ranking agrees with the target ranking on which dataset is better. PDA is the fraction of pairs

Table 1: Ranking agreement between proxy methods and target model (Qwen2.5-7B) ground truth. Best results in **bold**. Standard-LR Proxy achieves the highest agreement, while Tiny-LR Proxy performs at random level.

Method	PDA	95% CI	Spearman ρ	p-value	Top-1
Random Baseline	0.500	–	0.000	–	–
Training-Free NLL	0.636	[0.515, 0.743]	0.371	0.236	✗
Standard-LR Proxy (5e-5)	**0.712**	[0.606, 0.818]	**0.594**	*0.042*	✓
Tiny-LR Proxy (1e-5)	0.500	[0.379, 0.621]	-0.091	0.779	✗

with matching orderings, with chance level at 0.5. We compute 95% bootstrap confidence intervals over dataset pairs.

Spearman Rank Correlation (ρ). Measures the monotonic relationship between proxy and target rankings. We report p-values for statistical significance testing.

Top-1 Match. Whether the proxy correctly identifies the best-performing dataset according to the target model.

3.4 Datasets and Benchmarks

We evaluate on 12 math-oriented SFT datasets from OpenDataArena (Cai et al., 2025), sampling 50K examples from each. The datasets span diverse sources including dart-math-hard, openmathinstruct-2, R1-Distill-SFT-math, mathplus, numinamath-cot, DeepMath-309K, OpenR1-Math, QwQ-LongCoT-130K-math, numinamath1_5, Magpie-Reasoning-V2-250K-CoT-QwQ-math, Maths-College, and AM-Thinking-v1-Distilled-math.

For evaluation, we use GSM8K (Cobbe et al., 2021) with exact-match scoring and MATH-500 (Hendrycks et al., 2021) with math_verify evaluation. The composite score averages performance across both benchmarks. All experiments use 3 random seeds (42, 123, 456) with results averaged across seeds.

4 Experiments

4.1 Main Results

Table 1 presents the ranking agreement between each proxy method and the target model ground truth. The key finding is that the Tiny-LR hypothesis is refuted: reducing the learning rate does not improve proxy-to-target ranking transfer.

Standard-LR Proxy achieves PDA of 0.712 (95% CI: [0.606, 0.818]), significantly above the random baseline of 0.500. The Spearman correlation is 0.594 ($p = 0.042$), indicating statistically significant agreement with the target ranking. Critically, Standard-LR Proxy correctly identifies dart-math-hard as the top-performing dataset, matching the target model's ground truth.

In contrast, Tiny-LR Proxy achieves PDA of exactly 0.500 (95% CI: [0.379, 0.621]), indistinguishable from random chance. The Spearman correlation is -0.091 ($p = 0.779$), showing no meaningful relationship with the target ranking. The proxy incorrectly identifies QwQ-LongCoT-130K-math as the best dataset, which ranks 8th according to the target model.

The Training-Free NLL baseline achieves intermediate performance (PDA = 0.636) but fails to reach statistical significance ($p = 0.236$) and does not identify the correct top-1 dataset.

Figure 2 visualizes these results. The Standard-LR Proxy's confidence interval is entirely above the random baseline, while the Tiny-LR Proxy's confidence interval overlaps substantially with chance level, confirming that the reduced learning rate provides no ranking signal.

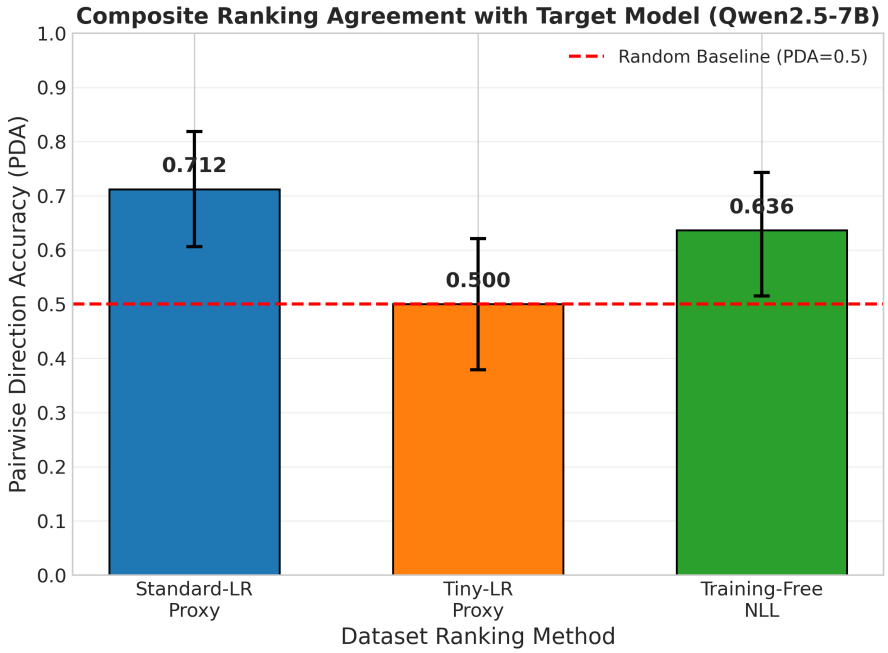

Figure 2: Composite ranking agreement (PDA) with target model across three dataset ranking methods. Standard-LR Proxy achieves the highest agreement (0.712), while Tiny-LR Proxy performs at random level (0.500). Error bars show 95% bootstrap confidence intervals. The dashed line indicates random baseline performance.

Table 2: Per-benchmark analysis of Tiny-LR Proxy performance. Tiny-LR achieves excellent agreement on MATH-500 but fails on GSM8K, revealing benchmark-specific behavior.

Evaluation Scope	PDA	Spearman ρ	p-value	Sig.
MATH-500 Only	**0.818**	**0.846**	0.0005	***
GSM8K Only	0.515	0.063	0.846	n.s.

4.2 Per-Benchmark Analysis

To understand the failure mode of Tiny-LR Proxy, we analyze performance separately on each benchmark. Table 2 reveals a striking pattern: Tiny-LR Proxy achieves excellent ranking agreement on MATH-500 but fails completely on GSM8K.

On MATH-500, Tiny-LR Proxy achieves PDA of 0.818 with Spearman $\rho = 0.846$ ($p < 0.001$), nearly matching Standard-LR Proxy's performance (PDA = 0.864, $\rho = 0.874$). This indicates that the reduced learning rate preserves meaningful ranking signals for harder mathematical problems.

However, on GSM8K, Tiny-LR Proxy achieves only PDA of 0.515 with $\rho = 0.063$ ($p = 0.846$), essentially random performance. The composite score is dragged down entirely by this GSM8K failure.

Figure 3 visualizes this benchmark-specific behavior. The divergence between MATH-500 and GSM8K performance suggests that the Tiny-LR hypothesis may have merit for certain evaluation settings, but fails when applied to simpler benchmarks where surface-level features dominate.

4.3 Analysis

The GSM8K failure correlates with a response-length pattern. Tiny-LR Proxy ranks QwQ-LongCoT-130K-math and Magpie-Reasoning-V2-250K-CoT-QwQ-math as the top-2 datasets, while the target model ranks them 8th and 10th respectively. These datasets are characterized by

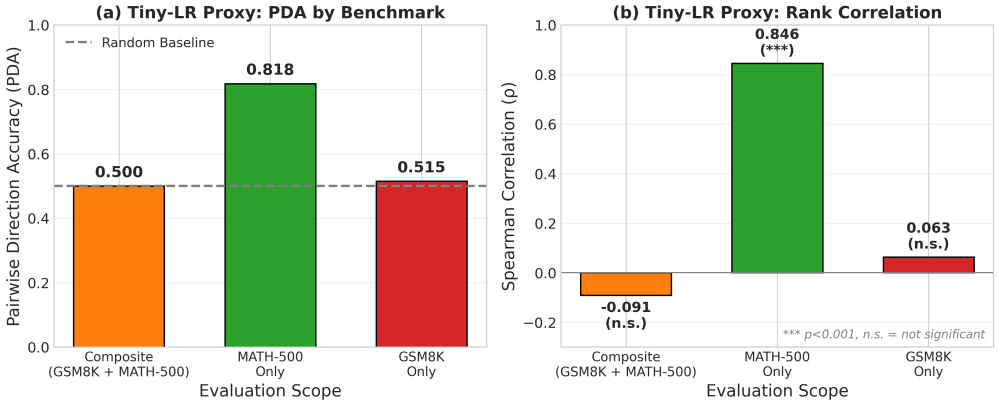

Figure 3: Per-benchmark analysis of Tiny-LR Proxy performance. (a) PDA by evaluation scope shows strong performance on MATH-500 (0.818) but random-level performance on GSM8K (0.515). (b) Spearman correlation confirms this pattern: MATH-500 achieves highly significant correlation ($\rho = 0.846$, $p < 0.001$) while GSM8K shows no significant correlation ($\rho = 0.063$, n.s.).

verbose reasoning chains. In contrast, the top-performing dataset according to the target model (dart-math-hard) produces more concise responses. The divergence between MATH-500 success and GSM8K failure may relate to their different evaluation methods: MATH-500 uses content-based math_verify scoring while GSM8K uses exact-match scoring that may be more sensitive to output formatting.

5 Discussion

5.1 Interpretation

The failure of the Tiny-LR hypothesis can be understood through the lens of what signals dominate at different learning rates. With reduced learning rates, the proxy model undergoes less parameter change, which we hypothesized would preserve more transferable gradient alignment signals. However, our results suggest that this regime instead amplifies superficial features such as response length and formatting patterns.

The benchmark-specific behavior provides key insight: Tiny-LR Proxy succeeds on MATH-500 (which uses content-based math_verify evaluation) but fails on GSM8K (which uses exact-match scoring). This suggests that reduced learning rates may preserve signals related to mathematical content quality, but simultaneously amplify sensitivity to surface-level patterns that inflate GSM8K scores without corresponding target model improvements.

The mechanism appears to be that datasets with verbose chain-of-thought responses (e.g., QwQ-LongCoT, Magpie-Reasoning) produce outputs that match GSM8K's evaluation patterns under Tiny-LR training, even though these datasets do not transfer well to the larger target model. Standard-LR training, by contrast, learns more robust representations that better predict target model performance across both benchmarks.

5.2 Limitations

Our study has several limitations. First, we evaluate only the Qwen2.5 model family; results may differ for other architectures or model families. Second, we focus exclusively on the math domain; the relationship between learning rate and ranking transfer may vary across domains such as code or general instruction-following. Third, our hyperparameter search is limited to two learning rate settings (5e-5 and 1e-5); intermediate values or different step counts might yield different conclusions.

5.3 FUTURE DIRECTIONS

Several directions merit further investigation. Testing the Tiny-LR hypothesis across multiple domains would clarify whether the benchmark-specific behavior we observe is unique to math or generalizes more broadly. Exploring intermediate learning rates between 1e-5 and 5e-5 could identify optimal regimes for different evaluation settings. Finally, developing benchmark-specific proxy strategies that account for evaluation characteristics may improve ranking transfer for diverse downstream tasks.

6 CONCLUSION

We tested the hypothesis that reducing learning rates during proxy SFT improves dataset ranking transfer to larger target models. Our experiments on 12 math datasets refute this hypothesis: Tiny-LR Proxy achieves random-level ranking agreement (PDA = 0.500) compared to Standard-LR Proxy's significantly above-random performance (PDA = 0.712, $p = 0.042$). However, we discovered that Tiny-LR Proxy exhibits benchmark-specific behavior, achieving excellent agreement on MATH-500 (PDA = 0.818) while failing on GSM8K. For practitioners, we recommend using standard learning rates for proxy-based dataset ranking, as they provide more reliable signals across diverse evaluation settings.

REFERENCES

Mengzhang Cai, Xin Gao, Yu Li, Honglin Lin, Zheng Liu, Zhuoshi Pan, Qizhi Pei, Xiaoran Shang, Mengyuan Sun, Zinan Tang, Xiaoyang Wang, Zhanping Zhong, Yun Zhu, Dahua Lin, Conghui He, and Lijun Wu. Opendataarena: A fair and open arena for benchmarking post-training dataset value, 2025. URL https://arxiv.org/abs/2512.14051.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, W. Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, S. Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, A. Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, J. Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022.

K. Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021.

J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

J. Kaplan, Sam McCandlish, T. Henighan, Tom B. Brown, Benjamin Chess, R. Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *ArXiv*, abs/2312.15685, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, M. Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

G. Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracking gradient descent. *ArXiv*, abs/2002.08484, 2020.

Yulei Qin, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li, and Xing Sun. Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. *Trans. Mach. Learn. Res.*, 2025, 2024.

Jiachen T. Wang, Tong Wu, Kaifeng Lyu, James Zou, D. Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. *ArXiv*, abs/2512.24503, 2025.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. A survey on data selection for llm instruction tuning. *ArXiv*, abs/2402.05123, 2024.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. pp. 13484–13508, 2022.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *ArXiv*, abs/2402.04333, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. *ArXiv*, abs/2302.03169, 2023.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. D3: Diversity, difficulty, and dependability-aware data selection for sample-efficient llm instruction tuning. *ArXiv*, abs/2503.11441, 2025.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *ArXiv*, abs/2303.16199, 2023.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *ArXiv*, abs/2403.13372, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, M. Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206, 2023.

A Full Dataset Rankings

Table 3 presents the complete dataset rankings produced by each method. The target ranking (Qwen2.5-7B) serves as ground truth. Standard-LR Proxy correctly identifies the top-1 dataset (dart-math-hard), while Tiny-LR Proxy ranks QwQ-LongCoT-130K-math first (target rank: 8th).

B Experimental Details

B.1 Training Configuration

All fine-tuning experiments use the AdamW optimizer with weight decay 0.0 and a cosine learning rate schedule with warmup ratio 0.1. We use an effective batch size of 16 via gradient accumulation, bfloat16 mixed precision training, and a maximum sequence length of 4096 tokens.

Table 3: Full dataset rankings by each method. Target ranking serves as ground truth. Standard-LR Proxy correctly identifies top-1 (dart-math-hard), while Tiny-LR Proxy does not.

Target Rank	Dataset	Std-LR Rank	Tiny-LR Rank
1	dart-math-hard	**1**	3
2	openmathinstruct-2	6	6
3	R1-Distill-SFT-math	3	9
4	mathplus	2	7
5	numinamath-cot	7	8
6	DeepMath-309K	12	12
7	OpenR1-Math	10	10
8	QwQ-LongCoT-130K-math	5	1
9	numinamath1_5	4	5
10	Magpie-Reasoning-V2-250K-CoT-QwQ-math	8	2
11	Maths-College	11	11
12	AM-Thinking-v1-Distilled-math	9	4

B.2 LoRA Configuration

All models use LoRA (Hu et al., 2021) with rank 16, alpha 32, and dropout 0.05. The adaptation is applied to all attention and MLP projection layers.

B.3 Model-Specific Settings

For the proxy model (Qwen2.5-1.5B), we use two configurations: Standard-LR with $\eta = 5 \times 10^{-5}$ for 500 steps, and Tiny-LR with $\eta = 1 \times 10^{-5}$ for 1000 steps. The target model (Qwen2.5-7B) is trained with $\eta = 5 \times 10^{-5}$ for 2000 steps.

B.4 Evaluation Protocol

We evaluate models using 0-shot prompting on two benchmarks: GSM8K with exact-match scoring and MATH-500 with math_verify scoring. The composite score is computed as the average of GSM8K and MATH-500 accuracy. All experiments are conducted with three random seeds (42, 123, 456) and results are averaged across seeds.