FinLCR: Scaling Financial Long-Context Reasoning via Curriculum Reinforcement Learning

Wenjie Qiu[1,†,*], **Yong Xie**[2,‡,*], **Karan Aggarwal**[2], **Aitzaz Ahmad**[2] and **Stephen Lau**[2]

[1]Rutgers University, [2]Amazon

Abstract

Long-context financial document understanding and reasoning pose significant challenges for small language models (SLMs). In this paper, we scale the financial long-context reasoning capability of SLMs through reinforcement learning. Specifically, we propose an efficient curriculum reinforcement learning recipe that features staged training across context lengths and difficulty-aware data sampling. To support our curriculum recipe, we curate `Fin-Ladder`, a difficulty-graded financial corpus spanning diverse datasets and context lengths. We apply the proposed corpus and recipe to train **Fin**ancial **L**ong-**C**ontext **R**easoner (`FinLCR`), a 4-billion parameter model with enhanced financial reasoning capabilities. Experiments demonstrate that our curriculum learning recipe yields superior performance on financial numerical reasoning benchmarks.

1 Introduction

Financial document understanding is critical for investment analysis, corporate governance, and regulatory compliance. However, financial texts, such as SEC filings and analyst reports, typically span hundreds of pages and require simultaneous long-context retrieval and multi-step numerical reasoning. While small language models offer significant efficiency gains over frontier models, they empirically struggle with long-context reasoning, presenting a fundamental cost-performance trade-off in financial applications.

Existing approaches to enhancing financial reasoning have primarily focused on short-context tasks (Qian et al., 2025; Liu et al., 2025). However, reasoning in a long context poses significant challenges that cannot be generalized from short-context reasoning training; we observe at most a 34% performance gap from mainstream LMs between FinQA (Chen et al., 2021) and DocFinQA (Reddy et al., 2024) despite that the latter consists of the same question sets. Besides, the gap widens from small language models compared with their large counterparts. Enhancement of reasoning capability on long-context tasks becomes critical for the adoption of small language models in the financial domain.

In this paper, we propose a curriculum reinforcement learning recipe to scale the reasoning capability of small language models on long-context financial tasks. Our recipe features a progressive two-stage short-to-long training framework reinforced by difficulty-aware sampling. We structure the curriculum to advance from short-context examples, establishing core reasoning patterns, to long-context scenarios that require complex information retrieval and integration. Unlike unstable online filtering methods, our approach utilizes an offline difficulty-aware strategy to identify and prioritize "learnable" data points—those that are challenging yet achievable—thereby optimizing training stability and efficiency.

To facilitate the recipe, we curate `Fin-Ladder`, a composite dataset characterized by graded difficulty and varying context lengths derived from diverse financial benchmarks. We instantiate the proposed corpus and recipe to train a **Fin**ancial **L**ong-**C**ontext **R**easoner (`FinLCR-4B`), a 4-billion parameter model based on Qwen3 (Zhao et al., 2025) with enhanced financial reasoning capabilities.

We evaluate `FinLCR-4B` on both short- and long-context financial reasoning tasks. On DocFinQA (128K context), it achieves 47.29% accuracy, outperforming its base model (42.52%), Qwen3-30B (42.41%), DeepSeek-R1 (31.02%), and Claude Haiku 4.5 (42.30%). We further validate the recipe against alternative RL approaches including DAPO (Yu et al., 2025) and SEC (Chen et al., 2025), demonstrating superior performance. Ablation

[†]Work done during author's internship at Amazon.
[*]Equal contribution.
[‡]Corresponding author: yonxie@amazon.com

studies confirm that neither staged training nor difficulty-aware sampling suffices alone. Further analysis shows that the two stages target complementary capabilities, i.e., reasoning and retrieval, respectively.

2 Method

2.1 Reinforcement Learning

We scale financial long-context reasoning through reinforcement learning, where the policy π_θ is optimized to maximize the expected correctness of generated responses. We employ Grouped Sequence Policy Optimization (GSPO) as our primary algorithm, a variant of Group Relative Policy Optimization (GRPO) designed to mitigate training instability (Zheng et al., 2025; Shao et al., 2024). Besides, we adopt an outcome reward model based on rules or LLM-as-a-judge depending on the task nature, assigning a binary reward $r(y, \hat{y}) = 1$ if the response $\hat{y}$ is correct and 0 otherwise.

2.2 Curriculum Learning Recipe

Two-Stage Short-to-Long Curriculum. Reasoning in long financial documents requires two distinct capabilities: numerical reasoning and information retrieval. To disentangle these challenges, we adopt a two-stage training strategy. Stage 1 trains on short-context data to establish robust numerical reasoning and answer formatting skills. Stage 2 extends training to long-context documents, initializing from the best Stage 1 checkpoint. This progression is motivated by the intuition that reasoning patterns, e.g., formula application and table interpretation, transfer directly from short to long contexts. By mastering these skills in Stage 1, the model can focus on the skill of long-context retrieval in Stage 2, stabilizing the learning process on complex long-context questions.

Difficulty-Aware Offline Sampling. To ensure data efficiency, we employ an offline sampling strategy that prioritizes samples within the model's "learnable zone." Standard random sampling often includes "over-easy" examples (which provide minimal policy gradient signal) or "over-hard" examples (which yield zero rewards, destabilizing RL). Inspired by curriculum learning principles (Bengio et al., 2009), we estimate the difficulty of each sample using the pass rate of the base model: $PassRate(x) = \frac{1}{k}\sum_{i=1}^{k} r(y, \hat{y}_i)$. We steer the training by retaining and sampling only the data with PassRate in the intermediate learnable range $[\tau_l,\tau_h]$. The offline approach eliminates repeated inference overhead and prevents distribution shifts during training, ensuring a consistent and high-quality signal for policy optimization while reducing the sample size and training cost.

2.3 Dataset Curation: `Fin-Ladder`

To facilitate our curriculum recipe, we construct `Fin-Ladder`, a composite dataset with graded difficulty and varying context lengths. We aggregate high-quality financial benchmarks and stratify them into two categories based on context length.

Short-Context Datasets (Stage 1). We source data from FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), BizBench (Krumdick et al., 2024), and FinCoT (Qian et al., 2025). These datasets primarily feature context lengths of 2K-8K tokens and focus on multi-step numerical reasoning over financial tables and text. They serve as the foundation for learning arithmetic operations, table lookups, and financial logic.

Long-Context Datasets (Stage 2). For the second stage, we utilize DocFinQA (Reddy et al., 2024), which comprises real-world annual reports and SEC filings. These documents range from 32K to 128K tokens, requiring the model to retrieve evidence across widely separated pages before performing reasoning. This split ensures a diverse distribution of reasoning types while rigorously testing long-context capabilities.

We apply the difficulty-aware sampling strategy to the raw data sources to maximize training efficiency and stabilize the training signal. Using the base model as the actor and the reward model for verification, we compute pass rates for all candidate samples. We strictly retain only those falling within the $[0.2, 0.8]$ pass rate interval, effectively pruning samples that are either already mastered ($PassRate > 0.8$) or currently provide sparse reward signal ($PassRate < 0.2$). Furthermore, to account for varying complexity levels and distinct difficulty distributions across datasets, we employ targeted down-sampling or up-sampling to ensure a balanced distribution. This curation results in a high-quality subset concentrated in the optimal learning zone with average pass rates of 58% for the short stage and 57% for the long stage. We defer more analysis on the dataset to Appendix B.

	Size	FinQA	DocFinQA	DocMathEval				Avg^L	Avg
				CL	CS	SL	SS		
(Curriculum) Reinforcement Learning									
`Qwen3-4B-Thinking-2507`	4B	67.79	42.52	28.82	75.50	50.00	68.50	40.45	55.52
`GSPO`	4B	71.26	44.36	31.60	74.50	42.00	68.50	39.32	55.37
`DAPO`	4B	69.31	43.71	30.56	75.50	47.00	67.00	40.42	55.51
`SEC (GSPO)`	4B	70.93	44.47	30.93	78.50	49.00	69.00	41.46	57.13
`FinLCR-4B`	4B	68.87	47.29	30.56	77.50	50.00	72.00	42.62	57.70
Frontier Models									
`Qwen3-30B-A3B-Thinking-2507`	30B	67.57	42.41	31.94	77.00	52.00	73.00	42.12	57.32
`Qwen3-235B-A22B-Instruct`	235B	63.02	35.45	37.11	79.00	52.00	73.50	41.52	56.68
`DeepSeek-R1`	671B	65.29	31.02	37.80	75.50	51.00	73.00	39.94	55.60
`Claude Haiku 4.5`	-	63.99	42.30	34.71	72.00	54.00	76.00	43.67	57.17
`Claude Sonnet 4.5`	-	72.67	55.75	40.55	81.50	61.00	80.00	52.43	65.24
`Claude Opus 4.5`	-	73.86	56.29	39.18	85.50	62.00	80.00	52.49	66.14

Table 1: Accuracy on financial reasoning benchmarks (in percentage). CL/CS/SL/SS denote *complex-long, complex-short, simple-long, simple-short* splits of DocMathEval benchmark. Avg^L denotes average accuracy over long-context benchmark. Avg denotes overall average accuracy. SEC is applied to GSPO for fair comparison.

3 Experiments

We conduct intensive experiments on applying our curriculum reinforcement learning recipe and `Fin-Ladder` to train SLMs. Models are evaluated on diverse financial benchmarks to answer 3 key research questions: (RQ1) Does RL upon our recipe and dataset improve SLM's long-context reasoning capability? (RQ2) How is our recipe compared with other curriculum learning approaches? (RQ3) What are the benefits of staged training and difficulty-aware sampling in our recipe?

Backbone. We utilize Qwen3-4B-Thinking-2507 from the Qwen series (Zhao et al., 2025) as our backbone. It is a SOTA small reasoning model with native support for the context length of 262,144. We use it due to its strong inherent reasoning capabilities and long-context performance, serving as an ideal testbed for enhancing domain-specific long-context reasoning. We name the resulting model as `Financial Long-Context Reasoner` (`FinLCR-4B`).

Benchmarks. We use three financial benchmarks covering diverse context lengths and task types. FinQA (Chen et al., 2021) contains financial questions requiring multi-step numerical reasoning over tables and text. DocFinQA (Reddy et al., 2024) extends FinQA to long documents representing real financial reports, allowing us to assess length generalization capabilities. Specifically, we test on a 128K context length. DocMathEval (Zhao et al., 2024), consisting of 4 subsets, i.e., *simple-short, simple-long, complex-short, complex-long*, tests mathematical reasoning on documents with varying lengths and complexity. We report accuracy calculated upon rule-based numerical equivalence.

Baselines. We compare `FinLCR-4B` against two distinct categories of baselines to validate both its absolute performance and the efficacy of our training recipe. First, we evaluate against frontier models, including SOTA proprietary and open-weight models such as the Claude 4.5 series, Qwen3-30B-A3B, Qwen3-235B-A22B (Zhao et al., 2025), and DeepSeek-R1 (DeepSeek-AI, 2025).

Second, to demonstrate the specific efficacy of the curriculum learning recipe, we compare it against alternative approaches. These includes native GSPO, Decoupled Clip Dynamic Sampling Policy Optimization (DAPO), which filters training samples dynamically based on real-time pass rates (Yu et al., 2025), and Self-Evolving Curriculum (SEC), which constructs curriculum through a non-stationary multi-armed bandit (Chen et al., 2025). We deploy the baseline approaches on the mixture of raw data sources without staged training and difficulty-aware sampling.

3.1 Main Results

Table 1 presents the main results. Specifically, `FinLCR-4B` achieves a significant performance leap over its backbone model. On the critical long-context benchmark DocFinQA, our model improves accuracy by 4.77%. Furthermore, for the average accuracy over long-context benchmarks (Avg^L), which includes DocFinQA and the long-context splits of DocMathEval, `FinLCR-4B` scores 42.62%, markedly outperforming the base model. This confirms that our specialized RL recipe suc-

	Avg^S	Avg^L	Avg
`Short Only`	70.44	41.81	56.13
`Long Only`	70.29	40.06	55.18
`Naive GSPO`	71.42	39.32	55.37
+ `S2L`	68.57	40.48	54.52
+ `DAS`	67.97	38.71	53.33
+ `Both`	72.79	42.62	57.70
τ=[0.2,0.7]	73.57	43.90	58.74
τ=[0.3,0.8]	73.01	44.39	58.70

Table 2: Ablation on curriculum components and DAS threshold sensitivity. Short/Long train on short- or long-context data only. Bottom rows vary τ=[τ_l,τ_h] from the default [0.2,0.8].

cessfully scales long-context reasoning capabilities that general-purpose training fails to elicit (RQ1).

Remarkably, `FinLCR-4B` delivers competitive performance against generalist models orders of magnitude larger. On DocFinQA, it outperforms Qwen3-30B (47.29% vs 42.41%) and surpasses the very large DeepSeek-R1 (31.02%), suggesting that general-purpose scaling does not guarantee domain-specific long-context proficiency. While the proprietary Claude Sonnet and Opus 4.5 remain the topline, `FinLCR-4B` significantly narrows the gap and performs on par with Haiku 4.5.

Compared to other training recipes, `FinLCR-4B` achieves 47.29% on DocFinQA, outperforming SEC (44.47%) and DAPO (43.71%). We attribute this to DAPO's focus on reward stabilization rather than skill progression, and SEC's inability to construct an effective context-length curriculum despite accounting for difficulty. In contrast, our recipe ensures progressive skill acquisition throughout training, resulting in enhanced long-context reasoning (RQ2).

3.2 Ablation Study

To answer RQ3, we investigate the individual and combined contributions of our recipe's components, attribute gains to retrieval vs. reasoning, and validate robustness to hyperparameter choices. Table 2 presents the setups and results.

Stage Synergy. The component analysis reveals a crucial synergy between the short-to-long (S2L) curriculum and difficulty-aware sampling (DAS). Results in Table 2 imply that the components in isolation produce trade-offs; for instance, S2L helps arrest the decline in long-context tasks compared with Naive GSPO (Avg^L recovers to 40.48%) but causes a regression in short performance (68.57%). The DAS ensures the curriculum focuses on high-signal examples, preventing the noisy rewards often seen during training. Consequently, the combined approach surpasses all baselines in both regimes, validating that progressive complexity and data quality should be optimized jointly.

Retrieval vs. Reasoning Attribution. FinQA and DocFinQA sharing the same QA pairs but different evidence length provides a natural ablation to disentangle retrieval and reasoning contributions. A comparison of accuracy on these matched pairs isolates the retrieval bottleneck. On FinQA (with oracle evidence), accuracy is proxy for reasoning capability, It improves from 68.4% (backbone) to 69.2% (Stage 1) to 69.3% (`FinLCR-4B`), indicating Stage 1 captures most reasoning improvement. The accuracy on DocFinQA (with full documents) rises from 43.4% to 44.4% to 47.0%, with the larger jump in Stage 2 confirming it specifically targets retrieval. Besides, we leverage LLM-as-a-judge to classify whether models' thinking references the gold evidence; we find that retrieval success rate increases from 75.1% (backbone) to 76.0% after Stage 1 (+0.9%) and 77.9% after Stage 2 (+1.9%), confirming that Stage 2 contributes twice the retrieval gain of Stage 1. Experiment setup and further analysis are deferred to Appendix C.2.

Threshold Robustness. We validate the robustness of DAS by varying the threshold window. We vary the pass rate threshold by 0.1 as its distribution is clustered. As seen in the bottom panel of Table 2, shifting the lower boundary up (τ=[0.3,0.8]) or the upper boundary down (τ=[0.2,0.7]) yields Avg^L of 44.39% and 43.9% respectively, surpassing all baselines. The threshold variants also yield leading results on short-context benchmarks. These variants retain different subsets of the training data yet achieve consistent gains, confirming that performance stems from the curriculum structure rather than threshold choices. Additional threshold variants are deferred to Appendix C.1.

4 Conclusion

In this paper, we propose a novel curriculum reinforcement learning recipe for financial long-context reasoning by using difficulty-aware sampling with a two-stage short-to-long context progression. We curate `Fin-Ladder`, a comprehensive dataset, and instantiate the recipe to train `Financial Long-Context Reasoner` (`FinLCR-4B`), a

specialized 4B small language model for financial long-context reasoning. Our experiments demonstrate that `FinLCR-4B` not only substantially outperforms its base model but also achieves parity with or surpasses generalist models orders of magnitude larger on challenging benchmarks. These findings validate that structured curriculum learning and rigorous data curation are critical pathways for unlocking high-performance reasoning in small language models.

Limitations. Our work has limitations. First, our difficulty-aware sampling relies on a single base model's performance for difficulty estimation, which may not perfectly align with other models' learning curves. Second, our curriculum is coarse-grained with only two stages; finer-grained progression might yield further improvements. Third, our evaluation focuses primarily on numerical reasoning tasks; the effectiveness of our approach on other types of long-context financial understanding (e.g., summarization) remains to be explored.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 41–48.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 2025. Self-evolving curriculum for LLM reasoning. *CoRR*, abs/2505.14970.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Arzu Burcu Güven, Anna Rogers, and Rob Van Der Goot. 2025. Do syntactic categories help in developmentally motivated curriculum learning for language models? In *Proceedings of the First BabyLM Workshop*, pages 288–300, Suzhou, China. Association for Computational Linguistics.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Greg Kamradt. 2023. Needle In A Haystack - Pressure Testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332, Bangkok, Thailand. Association for Computational Linguistics.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, Chao Li, Sheng Xu, Dezhi Chen, Yun Chen, Zuo Bai, and Liwen Zhang. 2025. Fin-r1: A large language model for financial reasoning through reinforcement learning. *Preprint*, arXiv:2503.16252.

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 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

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

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. 2025. Curriculum reinforcement learning from easy

to hard tasks improves LLM reasoning. *CoRR*, abs/2506.06632.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Yilun Zhao, Jimin Huang, Qianqian Xie, and Jian yun Nie. 2025. Fino1: On the transferability of reasoning-enhanced llms and reinforcement learning to finance. *Preprint*, arXiv:2502.08127.

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. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. DocFinQA: A long-context financial reasoning dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Ning Shang, Li Lyna Zhang, Siyuan Wang, Gaokai Zhang, Gilsinia Lopez, Fan Yang, Weizhu Chen, and Mao Yang. 2025. Longrope2: Near-lossless llm context window scaling. *Preprint*, arXiv:2502.20082.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *Preprint*, arXiv:2101.10382.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *Preprint*, arXiv:2409.12640.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *Preprint*, arXiv:2010.13166.

Zhenting Wang, Guofeng Cui, Kun Wan, and Wentian Zhao. 2025. DUMP: automated distribution-level curriculum learning for rl-based LLM post-training. *CoRR*, abs/2504.09710.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10184–10201. Association for Computational Linguistics.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models. *Preprint*, arXiv:2309.16039.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jixiao Zhang and Chunsheng Zuo. 2025. Grpo-lead: A difficulty-aware reinforcement learning approach for

concise mathematical reasoning in language models. *Preprint*, arXiv:2504.09696.

Shijie Zhang, Guohao Sun, Kevin Zhang, Xiang Guo, and Rujun Guo. 2025. CLPO: curriculum learning meets policy optimization for LLM reasoning. *CoRR*, abs/2509.25004.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. 2025. Qwen3guard technical report. *CoRR*, abs/2510.14276.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. 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 3277–3287, Online. Association for Computational Linguistics.

A Related Work

Financial Reasoning. Financial document understanding requires models to perform complex numerical reasoning over tables, charts, and text. FinQA (Chen et al., 2021) introduced the first large-scale dataset for numerical reasoning over financial data, focusing on answering questions that require multi-step calculations. TAT-QA (Zhu et al., 2021) extended this to more complex table-and-text hybrid scenarios. More recently, DocFinQA (Reddy et al., 2024) and FinanceBench (Islam et al., 2023) scaled financial QA to long documents (128K tokens and more), representing real-world annual reports and SEC filings.

Several works have explored specialized models for financial reasoning. Fin-o1 (Qian et al., 2025) introduced chain-of-thought reasoning dataset called FinCoT for financial problems. Fin-R1 (Liu et al., 2025) applied reinforcement learning specifically to financial question answering. However, these approaches primarily focus on short-context documents (typically <10K tokens) and have not systematically explored long-context (>100K tokens) scenarios with small models (<10B parameters).

Post-training for LLMs. Post-training techniques have become essential for aligning LLMs with human preferences and improving their reasoning capabilities. Supervised Fine-Tuning (SFT) (Ouyang et al., 2022b) remains the foundation, where models learn from human demonstrations. RLHF (Ouyang et al., 2022a) introduced reinforcement learning from human feedback, using a learned reward model to guide policy optimization through PPO (Schulman et al., 2017). More recent works have explored outcome-based RL approaches that eliminate the need for separate critic models. GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025) directly optimize policies using group-relative advantages, making training more stable and efficient. DPO (Rafailov et al., 2023) reframes RL as supervised learning on preference pairs. DAPO (Yu et al., 2025) introduces dynamic filtering strategies to select training data based on model performance. Our work builds on GSPO but introduces a difficulty-aware sampling strategy that provides stability benefits similar to DAPO while avoiding the overhead of online policy evaluation during training.

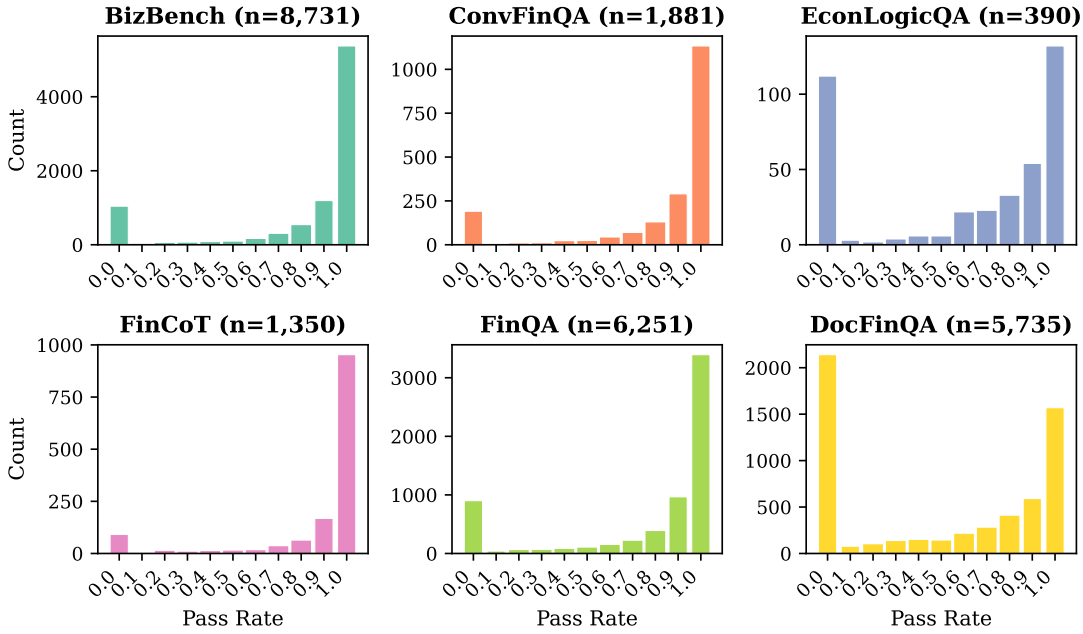

Figure 1: Pass-rate distribution by data source. Most samples cluster at the extremes (0.0 or 1.0), confirming the bimodal structure that motivates difficulty-aware filtering. The retained [0.2, 0.8] interval targets the informative middle where group-relative advantages are non-trivial.

Long-Context Reasoning. Extending context windows in LLMs has been an active research area. Early position encoding schemes like RoPE (Su et al., 2023) enabled context extension through interpolation and extrapolation. YaRN (Peng et al., 2024), ABF (Xiong et al., 2023), and other methods further improved position encoding for longer sequences. Evaluation of long-context understanding has also evolved. The "Needle in a Haystack" (NIAH) test (Kamradt, 2023) measures retrieval capabilities across varying context lengths. RULER (Hsieh et al., 2024) and MRCR (Vodrahalli et al., 2024) provide more comprehensive evaluations. However, most long-context work focuses on general-domain tasks. Our work specifically addresses long-context numerical reasoning in the financial domain, which poses unique challenges combining information retrieval and multi-step calculation.

Curriculum Learning. Curriculum Learning (CL) (Bengio et al., 2009; Soviany et al., 2022) proposes training models on easy examples before progressively increasing difficulty. Curriculum learning has been widely studied across various domains, including computer vision and NLP tasks (Wang et al., 2021; Güven et al., 2025). In the context of LLMs, curriculum strategies have been explored for data ordering (Liu et al., 2024; Xie et al., 2024), difficulty-based sampling (Platanios et al., 2019), and task progression (Xu et al., 2020; Zhang and Zuo, 2025). Prior work on context extension has explored position encoding modifications (Press et al., 2022; Zhu et al., 2024; Shang et al., 2025) and training strategies. Recently, curriculum reinforcement learning has been explored to increase training efficiency and improve performance (Chen et al., 2025; Zhang et al., 2025; Wang et al., 2025; Parashar et al., 2025), but none of them are investigated on long-context reasoning tasks.

B Data Statistics

Figure 1 shows the pass rate distribution across all data sources used to construct `Fin-Ladder`. The distributions are heavily bimodal: the majority of samples are either trivially solved (pass rate = 1.0) or impossible for the base model (pass rate = 0.0). This confirms that standard random sampling would be dominated by uninformative extremes that yield zero policy gradient signal under group-relative optimization.

The retained [0.2, 0.8] band captures the informative middle where rollout groups produce diverse outcomes, enabling meaningful advantage estimation. Short-context sources (BizBench, FinQA, ConvFinQA, FinCoT) exhibit a strong right skew resulting in retention rates of 10–15%. In contrast, DocFinQA shows a more balanced distribution with substantial mass at both extremes, yielding

Thresholds	Train Size	FinQA	DocFinQA	DocMathEval				Avg^L	Avg
				CL	CS	SL	SS		
$[0.2, 0.8]$	3,890	68.87	**47.29**	30.56	**77.50**	50.00	72.00	42.62	57.70
$[0.3, 0.7]$	2,222	69.96	43.93	**30.90**	76.50	48.00	72.50	40.94	56.96
$[0.1, 0.9]$	7,310	**70.39**	44.14	29.20	73.00	51.00	71.50	41.44	56.53
$[0.2, 0.7]$	2,400	69.20	46.20	29.50	77.00	**56.00**	**74.50**	43.90	**58.74**
$[0.3, 0.8]$	3,712	69.52	46.96	30.20	77.00	**56.00**	72.50	**44.39**	58.70

Table 3: Threshold sensitivity analysis. All models trained with the full two-stage recipe (S2L + DAS), varying only the pass-rate threshold. Train Size = total retained samples (short + long).

a higher retention rate of 23.7%. This difference reflects the greater difficulty variance in long-context documents, where retrieval complexity introduces a wider range of challenge levels.

C Ablation Analysis

C.1 Threshold Sensitivity Analysis

To validate that our results are not an artifact of a specific threshold choice, we train four additional models using the full two-stage recipe while varying the DAS pass-rate window $[\tau_l, \tau_h]$. We test two symmetric shifts (narrowing to [0.3, 0.7] and widening to [0.1, 0.9]) and two asymmetric boundary changes ([0.2, 0.7] and [0.3, 0.8]). Table 3 reports the full results.

Narrowing the window to [0.3, 0.7] reduces the training set by 43% (from 3,890 to 2,222 samples), while widening to [0.1, 0.9] nearly doubles it (7,310 samples). Despite this large variation in data volume, both variants remain within 1–2% of the default on FinQA and DocMathEval. However, DocFinQA drops by $\sim$3% under both symmetric shifts, indicating that the full [0.2, 0.8] band provides the richest signal for long-context document QA.

The asymmetric experiments isolate the contribution of each boundary band. Removing only the upper band ([0.2, 0.7]) or only the lower band ([0.3, 0.8]) yields DocFinQA scores of 46.20% and 46.96% respectively. They are close to the default (47.29%) and substantially better than removing both bands simultaneously ([0.3, 0.7] at 43.93%). This confirms that both the 0.2–0.3 (near-hard) and 0.7–0.8 (medium-easy) samples contribute useful training signal, and their combined presence is optimal.

FinQA accuracy is stable across all variants (68.9–70.4%), well within confidence intervals. Short-context numerical reasoning is insensitive to threshold selection, as expected given the abundance of short-context training data across all configurations. Overall, performance is robust to threshold perturbations of ± 0.1, with no variant significantly degrading results. The default [0.2, 0.8] achieves the best DocFinQA accuracy.

C.2 Retrieval vs. Reasoning Attribution

A key question is whether `FinLCR-4B`'s gains on long-context benchmarks stem from improved evidence retrieval, improved numerical reasoning, or both. We present two complementary analyses that disentangle these contributions across training stages.

Oracle-Context Ablation Analysis FinQA and DocFinQA share 922 identical questions: FinQA provides only the gold evidence paragraphs and tables, while DocFinQA embeds the same evidence in the full annual report. Comparing model accuracy on these matched pairs directly isolates the retrieval bottleneck—FinQA accuracy reflects reasoning-only performance, and the gap to DocFinQA accuracy measures retrieval difficulty.

Table 4 reveals a clear division of labor between training stages. The retrieval gap shrinks from 25.05% (backbone) to 24.84% after Stage 1, and then to 22.34% after Stage 2; a 2.5% reduction concentrated entirely in the long-context training phase. Meanwhile, oracle accuracy based on FinQA, a proxy for reasoning, improves primarily in Stage 1 (+0.76%) with negligible further gain in Stage 2 (+0.11%).

The model agreement distribution reinforces this interpretation. *Oracle-Only Correct*—cases where the model reasons correctly given gold evidence but fails on the full document—drops from 28.0% to 24.4%, with the majority of the reduction in Stage 2 (27.0% $\rightarrow$ 24.4%). Conversely, *Both Wrong* (reasoning failures regardless of context) remains fixed at 28.6% across all stages, confirming that reasoning gains are modest and front-loaded in Stage 1.

	Backbone	Stage 1	FinLCR-4B
(a) Benchmark Performance			
Oracle Acc. (FinQA)	68.44	69.20	69.31
Full Doc Acc. (DocFinQA)	43.38	44.36	46.96
Retrieval Gap (Δ)	25.05	24.84	22.34
(b) Model Agreement			
Both Correct	40.5%	42.2%	44.9%
Oracle-Only Correct	28.0%	27.0%	24.4%
Full-Only Correct	2.9%	2.2%	2.1%
Both Wrong	28.6%	28.6%	28.6%
(c) Retrieval Analysis			
Retrieval Rate	75.1%	76.0%	77.9%
Reasoning | Retrieval	57.5%	58.2%	60.0%

Table 4: Retrieval vs. reasoning attribution. (a) Accuracy on matched FinQA (oracle evidence) and DocFinQA (full document) pairs; the retrieval gap measures performance lost to imperfect evidence selection. (b) Model agreement between oracle and full-document settings; *Oracle-Only Correct* indicates retrieval failure. (c) LLM-as-a-judge analysis of whether the thinking trace references gold evidence; $Reasoning|Retrieval$ is accuracy conditioned on successful retrieval.

Figure 2: Thinking pattern comparison. (a) Qwen3-4B-Thinking-2507 uses bottom-up reasoning for exploration. (b) After SFT on GPT-4o-generated CoT data, the model adopts top-down decomposition, disrupting natural thinking patterns and degrading performance.

```
Let me search through the text to find information
about Brazilian paper sales as a percentage of
printing papers in 2006.

Looking at the text, I see a section on "SALES VOLUMES
BY PRODUCT" with a table showing sales data for 2006,
2005, and 2004.

In the table, there\'s a line item for "BRAZILIAN
PAPER" with sales of $496 million in 2006.
```

(a) Bottom-Up Reasoning.

```
Alright, let's figure out what percentage of the total
sales in 2006 was due to Brazilian paper sales. First,
I need to find the total sales for the Printing Papers
segment. According to the data, it's $6,930 million.
Now, I need to see how much of that was specifically
from Brazilian paper. The data says Brazilian paper
sales were $496 million. Okay, so now I need to
calculate what portion of the total sales this
represents. To do that, ...
```

(b) Top-Down Reasoning

Retrieval Analysis To corroborate the analysis with an independent signal, we use Claude Opus 4.5 as a judge to classify whether each model's thinking trace references the correct evidence from the oracle context. For each of the 922 DocFinQA predictions per model, the judge compares the full thinking trace against the FinQA gold evidence and classifies retrieval as *retrieved*, *partially retrieved*, or *not retrieved*. Combined with answer correctness, this yields a retrieval $\times$ reasoning decomposition without relying on paired accuracy comparisons.

Table 4 confirms the same pattern: retrieval rate improves by +0.9% in Stage 1 and +1.9% in Stage 2, while reasoning conditioned on retrieval improves more uniformly (+0.7% and +1.8% respectively). The larger retrieval gain in Stage 2 aligns with our findings and validates that long-context training specifically enhances evidence localization.

Together, both analyses converge on the same conclusion: the two-stage curriculum targets complementary capabilities, with Stage 1 establishing numerical reasoning and Stage 2 addressing the retrieval bottleneck in long documents.

D Qualitative Analysis

D.1 Thinking Patterns

To understand how different training approaches affect model reasoning, we analyze thinking patterns of the base model, SFT-trained models, and FinLCR-4B.

Bottom-Up vs Top-Down Reasoning. We observe two distinct reasoning patterns in financial QA. **Bottom-up** reasoning starts by exploring available data (e.g., "Let me search through the text...") before formulating solution steps. **Top-down** reasoning begins with problem decomposition (e.g., "First, I need to find...Then, I need to calculate...") before data retrieval. Figure 2 illustrates representative examples of both patterns.

Why SFT Disrupts Thinking Patterns. Initial attempts to improve the base model through supervised fine-tuning (SFT) on FinCoT (generated by GPT-4o) (Qian et al., 2025) resulted in performance degradation. Analysis reveals that this failure stems from fundamental pattern mismatches: Qwen3-4B-Thinking-2507 naturally employs bottom-up reasoning with extensive exploration (avg. 2.5k tokens), while GPT-4o uses concise top-down decomposition (avg. 300 tokens) on the FinQA benchmark. SFT on GPT-4o-generated demonstrations forced the model to adopt top-down patterns, disrupting its natural reasoning structure and significantly reducing thinking length. This mismatch led to incomplete reasoning and lower accuracy.

RL Preserves Thinking Patterns. In contrast, our GSPO-based approach optimizes for outcome correctness without prescribing specific reasoning paths. Analysis of FinLCR-4B's outputs shows that

the model retains its natural bottom-up pattern with appropriate thinking length while improving accuracy. This preservation of natural thinking structure explains why RL succeeds where SFT fails: RL enhances what the model already does well rather than replacing it with incompatible patterns.