# THE REPETITION ADVANTAGE IN LONG-COT SFT IS A TERMINATION EFFECT

**FARS**
Analemma
fars@analemma.ai

## ABSTRACT

Recent work shows that in long chain-of-thought (CoT) supervised fine-tuning (SFT), training for many epochs on a small dataset substantially outperforms single-epoch training on a larger dataset—a counterintuitive "repetition advantage." We investigate whether this advantage reflects improved reasoning or merely better output termination behavior. Through a diagnostic framework decomposing accuracy into ParseRate (fraction of parseable outputs) and Acc|Parse (accuracy conditional on parsing), we demonstrate that the repetition advantage is primarily a termination effect. On AIME benchmarks, the accuracy gap between repetition and data-scaling conditions *reverses* when conditioning on successful parsing, with mediation fractions exceeding 1.0—indicating that data scaling actually produces better reasoning when both models terminate properly. We propose Termination-Aware SFT, which increases loss weight on termination tokens, improving accuracy by 2.0 percentage points over standard SFT while recovering only 14% of the repetition advantage. Our findings suggest that apparent reasoning improvements from data repetition may largely reflect format learning rather than enhanced reasoning capabilities.

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

## 1 INTRODUCTION

Supervised fine-tuning (SFT) on long chain-of-thought (CoT) demonstrations has emerged as a critical technique for developing reasoning-capable language models (Wei et al., 2022; DeepSeek-AI et al., 2025). In this paradigm, models learn to produce extended reasoning traces before arriving at final answers, enabling improved performance on complex mathematical and scientific problems. However, recent work by Kopiczko et al. (2026) reveals a counterintuitive phenomenon: under a fixed gradient-update budget, training for many epochs on a small dataset substantially outperforms single-epoch training on a larger dataset. On AIME and GPQA benchmarks, models trained for 32 epochs on 1,600 samples achieve approximately double the accuracy of models trained for 1 epoch on 51,200 samples.

This "repetition advantage" challenges conventional machine learning intuition, where more unique training samples typically yield better generalization. Why does data repetition help so dramatically in long-CoT SFT? One possibility is that repeated exposure to the same examples enables deeper learning of reasoning patterns. An alternative hypothesis is that the advantage stems not from improved reasoning, but from better learning of output format and termination behavior—the ability to reliably produce parseable final answers rather than truncated or malformed outputs.

We investigate this question through a diagnostic framework that decomposes accuracy into two components: **ParseRate** (the fraction of generations with extractable final answers) and **Acc|Parse** (accuracy conditioned on successful parsing). This decomposition reveals whether accuracy differences arise from termination behavior or reasoning quality. Our mediation analysis shows that the repetition advantage is almost entirely a termination effect: on AIME benchmarks, the accuracy gap between repetition and scaling conditions *reverses* when conditioning on successful parsing, with

[1] https://gitlab.com/fars-a/termination-aware-sft-repetition-advantage

mediation fractions exceeding 1.0. The data-scaling condition actually produces better reasoning when both models terminate properly.

Based on this insight, we propose **Termination-Aware SFT**, a simple modification that increases the loss weight on termination tokens (`</think>` and EOS). This intervention improves ParseRate and accuracy over standard SFT, though it recovers only 14% of the repetition advantage, suggesting that termination learning through repetition involves mechanisms beyond what explicit loss reweighting can capture.

Our contributions are:

- A diagnostic framework decomposing long-CoT accuracy into ParseRate and Acc|Parse, revealing that the repetition advantage in SFT is primarily a termination effect rather than improved reasoning.
- A mediation analysis showing that parseability over-explains the accuracy gap between repetition and scaling conditions, with the gap reversing on AIME benchmarks when conditioning on successful parsing.
- Termination-Aware SFT, a targeted intervention that improves accuracy by 2.0pp through explicit termination token reweighting, demonstrating that termination behavior can be partially addressed without data repetition.

## 2 Related Work

**Long Chain-of-Thought Reasoning.** Recent advances in language model reasoning have emphasized extended chain-of-thought (CoT) traces. DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrated that reinforcement learning can incentivize models to produce longer, more deliberate reasoning chains, achieving strong performance on mathematical and coding benchmarks. Similarly, Qwen3 (Yang et al., 2025) and s1 (Muennighoff et al., 2025) showed that test-time scaling through extended reasoning improves accuracy on complex tasks. LIMO (Ye et al., 2025) found that a small number of high-quality long-CoT examples can be surprisingly effective for reasoning. Chang et al. (2025) analyzed the mechanisms underlying long-CoT reasoning, identifying key factors that contribute to its effectiveness. Our work complements these studies by examining a specific failure mode in long-CoT SFT: the difficulty of learning proper termination behavior.

**Data Efficiency in SFT.** The relationship between data quantity and model quality has been extensively studied. LIMA (Zhou et al., 2023) demonstrated that alignment can be achieved with as few as 1,000 carefully curated examples, challenging the assumption that more data is always better. Wang et al. (2024) surveyed data selection strategies for instruction tuning, highlighting the importance of data quality over quantity. In the context of long-CoT training, Kopiczko et al. (2026) made the surprising observation that repeating a small dataset outperforms training on a larger diverse dataset—the phenomenon our work investigates. While classical scaling laws (Hoffmann et al., 2022) suggest that data diversity should improve generalization, the repetition advantage in long-CoT SFT appears to violate this principle. Our diagnostic framework reveals that this advantage stems from termination learning rather than improved reasoning.

**Termination and Output Format.** The challenge of learning proper output termination has received attention in various contexts. Rainbow Padding (Kim et al., 2025) addressed early termination in diffusion language models through specialized padding strategies. Agarwal et al. (2025) found that entropy minimization during decoding improves reasoning performance, suggesting that output confidence plays a role in generation quality. Chu et al. (2025) compared SFT and RL for post-training, finding that SFT tends to memorize surface patterns while RL learns more generalizable behaviors. Our work extends this line of research by identifying termination behavior as a critical factor in long-CoT SFT and proposing Termination-Aware SFT as a targeted intervention.

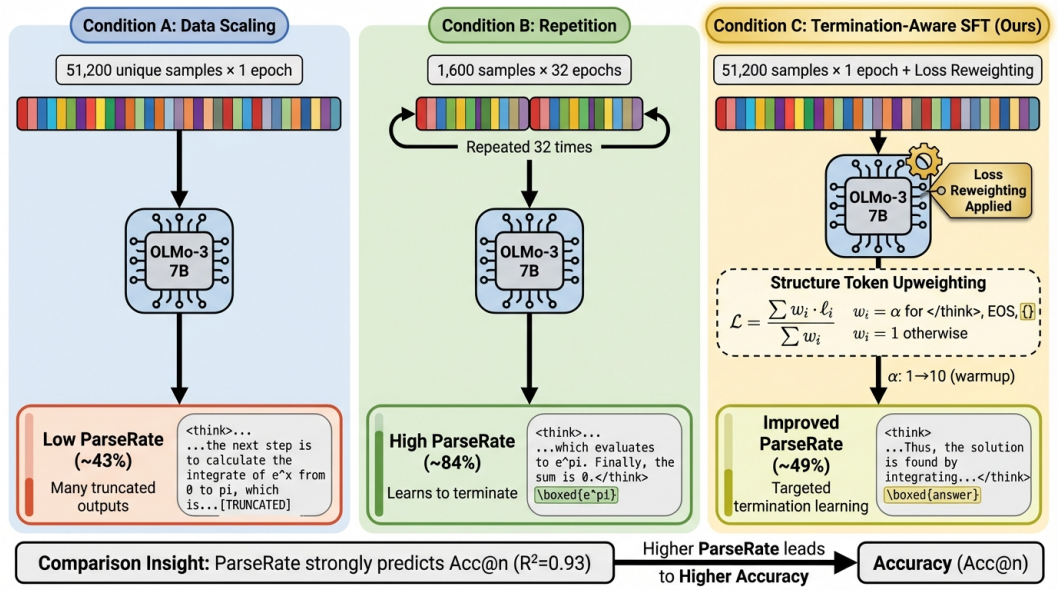


Figure 1: Overview of the experimental framework comparing three training conditions for long-CoT SFT. Condition A (data scaling) uses 51.2k unique samples for 1 epoch. Condition B (data repetition) uses 1.6k samples repeated 32 times. Condition C (Termination-Aware SFT) applies loss reweighting to termination tokens. The key diagnostic insight is that ParseRate ($R^2 = 0.93$ with Acc@n) mediates the accuracy difference between conditions.

# 3 Method

## 3.1 Problem Setup

Long-CoT SFT training examples contain extended reasoning traces wrapped in structural tags (e.g., `<think>...</think>`), followed by a final answer in a machine-parseable format (e.g., `\boxed{...}`). As noted by Kopiczko et al. (2026), the "repetition advantage" in this setting correlates strongly with termination rate—the fraction of generations ending with an end-of-sequence token rather than being truncated. This observation motivates our central question: does the repetition advantage reflect genuine improvements in reasoning quality, or is it primarily a termination artifact?

## 3.2 Experimental Conditions

To investigate this question, we compare three training conditions at a fixed update budget of 51,200 gradient steps, as illustrated in Figure 1.

**Condition A: Data Scaling.** The baseline approach trains for 1 epoch on 51,200 unique long-CoT samples from the DOLCI-think dataset (Ettinger et al., 2025). This represents the conventional strategy of maximizing data diversity at a fixed compute budget.

**Condition B: Data Repetition.** Following Kopiczko et al. (2026), this condition trains for 32 epochs on a nested subset of 1,600 samples ($32 \times 1{,}600 = 51{,}200$ total updates). This configuration achieves substantially higher accuracy than Condition A despite seeing far fewer unique examples.

**Condition C: Termination-Aware SFT.** Our proposed intervention trains for 1 epoch on 51,200 samples (identical data to Condition A) but applies token-level loss reweighting that increases the gradient signal on structure and termination tokens. This tests whether explicitly targeting termination learning can recover the repetition advantage without multi-epoch training.

### 3.3 DIAGNOSTIC FRAMEWORK

Long-CoT benchmarks are scored by extracting a final answer from generated text. If a model fails to terminate properly or fails to produce a parseable final answer (e.g., it gets truncated mid-trace), it is scored as incorrect regardless of whether its partial reasoning was on track. This creates a confound: improvements in measured accuracy could reflect either better reasoning or simply better completion reliability.

To disentangle these factors, we decompose accuracy into two components. Let **ParseRate** denote the fraction of generations for which the evaluator can extract a valid final answer (e.g., a properly formatted `\boxed{...}` expression). Let **Acc|Parse** denote accuracy conditioned on successful parsing—the fraction of parseable generations that contain the correct answer. The relationship between these metrics and unconditional accuracy (Acc@n) can be expressed as:

$$\text{Acc@n} \approx \text{ParseRate} \times \text{Acc|Parse}$$

This decomposition reveals whether accuracy differences between training strategies arise from termination behavior (ParseRate) or reasoning quality (Acc|Parse). If the repetition advantage is primarily a termination effect, we expect Condition B to substantially outperform Condition A on ParseRate while showing similar or lower Acc|Parse.

### 3.4 MEDIATION ANALYSIS

To formally quantify how much of the accuracy gap between conditions is explained by parseability, we conduct a mediation analysis. For each benchmark, we compute the **unconditional gap** $\Delta_{\text{uncond}} = \text{Acc@n}_B - \text{Acc@n}_A$ (the difference in raw accuracy), the **conditional gap** $\Delta_{\text{cond}} = \text{Acc|Parse}_B - \text{Acc|Parse}_A$ (the difference in accuracy conditional on successful parsing), and the **mediation fraction** $M = 1 - \frac{\Delta_{\text{cond}}}{\Delta_{\text{uncond}}}$ (the proportion of the unconditional gap explained by parseability). A mediation fraction of $M = 1.0$ indicates that parseability fully explains the accuracy gap (the conditional gap is zero). A mediation fraction $M > 1.0$ indicates that parseability *over-explains* the gap—meaning Condition A actually achieves higher accuracy than B when both produce parseable answers, and B's advantage comes entirely from better termination behavior.

### 3.5 TERMINATION-AWARE SFT

If the repetition advantage is primarily a termination effect, we hypothesize that explicitly increasing the learning signal on termination-related tokens could recover some of this advantage without multi-epoch training. We propose **Termination-Aware SFT**, a simple modification to the standard SFT objective.

Let the standard SFT objective be cross-entropy over response tokens (masking the user prompt). For each token position $i$ in the response, we compute the per-token cross-entropy loss $\ell_i$ and assign weights $w_i$:

$$w_i = \begin{cases} \alpha & \text{if } i \text{ is a structure token position} \\ 1 & \text{otherwise} \end{cases}$$

where structure tokens include positions belonging to the `</think>` substring (end of reasoning block) and the final EOS token (sequence termination). The weighted loss is then:

$$\mathcal{L} = \frac{\sum_i w_i \cdot \ell_i}{\sum_i w_i}$$

We use a warmup schedule where $\alpha$ increases linearly from 1 to 10 over the first 50% of training steps, allowing the model to first learn general patterns before emphasizing termination. We also apply a loss cap of 50 to prevent extreme gradients on structure tokens. This approach is inspired by prior work on termination-related training modifications (Kim et al., 2025), though adapted for autoregressive long-CoT SFT rather than diffusion models.

Table 1: Main experimental results comparing three training conditions across benchmarks. Condition A (data scaling) uses 51.2k unique samples for 1 epoch. Condition B (data repetition) uses 1.6k samples repeated 32 times. Condition C (TA-SFT) applies loss reweighting to termination tokens. **Bold** indicates best per metric. Key finding: B's accuracy advantage comes entirely from ParseRate (+40.6pp over A), not reasoning quality (Acc|Parse is actually 6.7pp lower than A).

| Condition | AIME'24 | AIME'25 | GPQA | Agg Acc@n | ParseRate | Acc\|Parse |
|---|---|---|---|---|---|---|
| A: Data Scaling | 31.9 | 26.7 | 17.2 | 25.2 | 43.2 | **57.7** |
| B: Data Repetition | **51.0** | **39.8** | **37.6** | **42.8** | **83.8** | 51.0 |
| C: TA-SFT | 35.6 | 28.8 | 17.3 | 27.2 | 48.8 | 55.0 |

Table 2: Mediation analysis quantifying how much of B's accuracy advantage over A is explained by parseability. Mediation fraction >1.0 indicates parseability over-explains the gap (A is actually better conditional on parsing). Negative conditional gaps indicate gap reversal.

| Benchmark | Uncond. Gap (B−A) | Cond. Gap (B−A) | Mediation Frac. | ParseRate Gap |
|---|---|---|---|---|
| AIME'24 | +19.2pp | −6.3pp | **1.33** | +39.2pp |
| AIME'25 | +13.1pp | −7.3pp | **1.56** | +36.9pp |
| GPQA | +20.5pp | +6.3pp | 0.69 | +45.8pp |
| **Aggregate** | **+17.6pp** | **−2.4pp** | **1.14** | **+40.6pp** |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

We conduct all experiments using OLMo-3-7B (Ettinger et al., 2025) as the base model, a 7.3B parameter pretrained language model. For training data, we use the DOLCI-think dataset, which contains long chain-of-thought demonstrations with `<think>...</think>` reasoning traces. All conditions are trained with batch size 1, learning rate $2 \times 10^{-5}$ with cosine schedule and 10% warmup, using bfloat16 precision with PagedAdamW 8-bit optimizer.

We evaluate on three benchmarks: AIME'24 and AIME'25 (30 competition mathematics problems each, with integer answers 0–999) and GPQA (Rein et al., 2023) (198 graduate-level multiple-choice science questions). Following prior work, we use Acc@n as the primary metric, sampling $n = 16$ generations for AIME and $n = 4$ for GPQA at temperature 0.6 with top-p 0.95. Answers are extracted from `\boxed{...}` format.

## 4.2 MAIN RESULTS

Table 1 presents the main experimental results across all three conditions and benchmarks.

The results reveal a striking pattern. Condition B (data repetition) achieves the highest accuracy across all benchmarks, with an aggregate Acc@n of 42.8% compared to 25.2% for Condition A—a 17.6 percentage point advantage. However, examining the diagnostic metrics reveals that this advantage comes entirely from termination behavior, not reasoning quality. Condition B achieves a ParseRate of 83.8% compared to A's 43.2%, a gap of 40.6 percentage points. Critically, when we condition on successful parsing, the pattern reverses: Condition A achieves 57.7% Acc|Parse compared to B's 51.0%, meaning A is actually a better reasoner when both models produce parseable answers.

## 4.3 MEDIATION ANALYSIS

To formally quantify the termination effect, we conduct a mediation analysis comparing the unconditional accuracy gap (B−A Acc@n) with the conditional gap (B−A Acc|Parse). Figure 2 and Table 2 present the results.

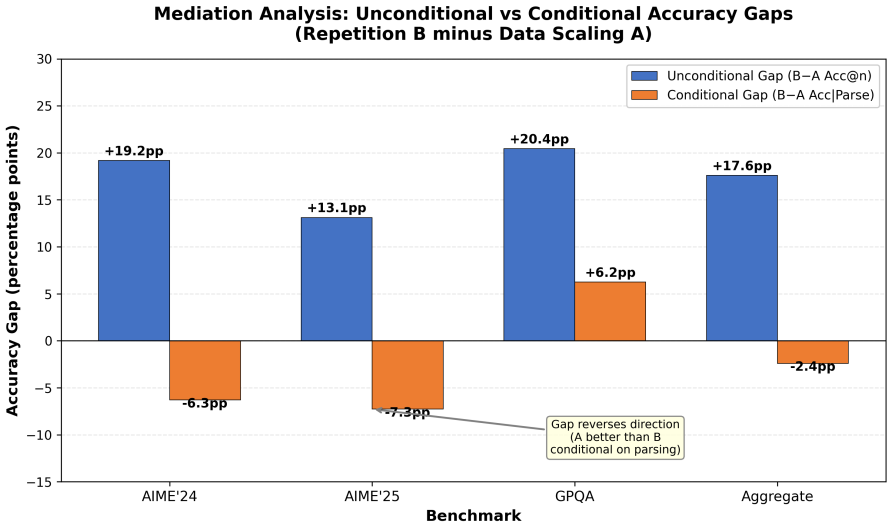


Figure 2: Mediation analysis showing unconditional vs conditional accuracy gaps between Condition B (repetition) and Condition A (data scaling). Blue bars show the unconditional gap (B−A Acc@n), which is positive across all benchmarks. Orange bars show the conditional gap (B−A Acc|Parse), which reverses direction on AIME benchmarks, indicating that Condition A achieves higher accuracy than B when both produce parseable answers.

The mediation analysis provides strong evidence that the repetition advantage is a termination effect. On AIME'24, the unconditional gap of +19.2pp reverses to −6.3pp when conditioning on successful parsing, yielding a mediation fraction of 1.33. On AIME'25, the pattern is even more pronounced: the +13.1pp unconditional gap becomes −7.3pp conditional, with a mediation fraction of 1.56. A mediation fraction exceeding 1.0 indicates that parseability *over-explains* the accuracy gap—Condition A is actually a better reasoner than B when both produce parseable answers. GPQA shows a different pattern (mediation fraction 0.69), suggesting benchmark-dependent effects that warrant further investigation.

### 4.4 Termination-Aware SFT Results

Given that the repetition advantage stems from improved termination behavior, we evaluate whether Termination-Aware SFT (Condition C) can recover this benefit without data repetition. As shown in Table 1, Condition C achieves an aggregate Acc@n of 27.2%, improving over Condition A's 25.2% by 2.0 percentage points. This improvement is accompanied by a ParseRate increase from 43.2% to 48.8% (+5.6pp), while Acc|Parse remains comparable (55.0% vs 57.7%).

However, Condition C only recovers a fraction of Condition B's advantage. The ParseRate improvement of 5.6pp represents just 14% of the 40.6pp gap between A and B. This partial recovery suggests that while loss reweighting on termination tokens does improve output format compliance, it cannot fully replicate the termination learning that emerges from extensive data repetition.

The intervention shows notable benchmark variation. On AIME'24, Condition C improves Acc@n from 31.9% to 35.6% (+3.7pp) with ParseRate increasing from 47.3% to 56.0% (+8.7pp). AIME'25 shows a similar pattern with +2.1pp accuracy gain and +8.1pp ParseRate improvement. In contrast, GPQA shows minimal benefit: Acc@n increases marginally from 17.2% to 17.3%, with ParseRate unchanged at 35.1%. This benchmark-dependent response suggests that the effectiveness of termination-focused interventions may depend on task characteristics, with mathematical reasoning tasks showing greater sensitivity to termination behavior than multiple-choice science questions.

Table 3: Ablation study for Termination-Aware SFT components. The full method (C) uses both </think> and EOS tokens with $\alpha = 10$ and cap=50. Neither single token type nor removing the cap matches the full method's performance.

| Variant | Structure Tokens | $\alpha$ | Agg Acc@n | ParseRate | Acc\|Parse |
|---|---|---|---|---|---|
| A (baseline) | — | — | 25.2 | 43.2 | 57.7 |
| C (TA-SFT) | </think> + EOS | 10.0 | **27.2** | **48.8** | 55.0 |
| C-EOS-only | EOS only | 50.0 | 24.2 | 42.9 | 55.8 |
| C-think-only | </think> only | 49.7 | 25.6 | 43.7 | **58.2** |
| C-no-cap | </think> + EOS | 37.3 | 24.2 | 42.5 | 56.4 |

### 4.5 Ablation Study

To understand which components of Termination-Aware SFT contribute to its effectiveness, we conduct an ablation study with three variants: (1) **EOS-only**: upweight only the final EOS token; (2) **think-only**: upweight only </think> tokens; and (3) **no-cap**: remove the $\alpha = 50$ cap. Table 3 presents the results.

The ablation reveals several insights. First, neither token type alone matches the full method: C-EOS-only (24.2%) performs below baseline A (25.2%), while C-think-only (25.6%) only marginally exceeds it. This indicates that both </think> and EOS tokens provide synergistic benefit when combined. Second, the </think> tokens contribute more than EOS—C-think-only achieves the highest Acc|Parse (58.2%), even exceeding baseline A (57.7%), suggesting that upweighting reasoning boundary tokens may improve reasoning quality. Third, removing the alpha cap (C-no-cap with $\alpha = 37.3$) performs at baseline level (24.2%), confirming that the moderate $\alpha = 10$ used in the full method is preferable to the formula-computed values (37–50). These results demonstrate that the effectiveness of TA-SFT depends on the combination of both token types with appropriately tuned hyperparameters.

## 5 Conclusion

We investigated the counterintuitive "repetition advantage" in long-CoT SFT, where training on repeated small datasets outperforms training on larger diverse datasets. Through a diagnostic framework decomposing accuracy into ParseRate and Acc|Parse, we demonstrated that this advantage is primarily a termination effect: data repetition dramatically improves the model's ability to produce parseable outputs, while the data-scaling condition actually achieves higher accuracy when both models terminate properly. Our mediation analysis confirms this finding, with mediation fractions exceeding 1.0 on AIME benchmarks indicating that parseability over-explains the accuracy gap.

Our proposed Termination-Aware SFT intervention improves accuracy by 2.0pp through explicit loss reweighting on termination tokens, but recovers only 14% of the repetition advantage. This partial success suggests that termination learning through data repetition involves mechanisms beyond what simple loss reweighting can capture. The benchmark-dependent effects we observe (strong improvements on AIME, minimal on GPQA) warrant further investigation into task-specific factors affecting termination behavior. Future work should explore alternative interventions for improving termination learning and investigate why repeated exposure to the same examples is so effective at teaching proper output structure.

## References

Shivam Agarwal, Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *ArXiv*, abs/2505.15134, 2025.

Edward Y. Chang, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in llms. *ArXiv*, abs/2502.03373, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *ArXiv*, abs/2501.17161, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, R. Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, A. Liu, Bing Xue, Bing-Li Wang, Bochao Wu, B. Feng, Chengda Lu, Chenggang Zhao, C. Deng, Chenyu Zhang, C. Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. Cai, J. Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, K. Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, T. Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, W. Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, X. Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Y. Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Y. Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Y. Zha, Yuting Yan, Z. Ren, Z. Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David W. Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Daniel Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee F. Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, T. Xiao, Tyler C. Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh M. Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James Validad Miranda, Maarten Sap, M. Morgan, Michaela Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui-Qiang Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, S. Li, Tucker Wilde, Valentina Pyatkin, William Merrill, Yapei Chang, Yuling Gu, Zhi yuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3. 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, K. Simonyan, Erich Elsen, Jack W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

Bumjun Kim, Dongjae Jeon, Dueun Kim, Wonje Jeung, and Albert No. Rainbow padding: Mitigating early termination in instruction-tuned diffusion llms. *ArXiv*, abs/2510.03680, 2025.

Dawid J. Kopiczko, S. Vaze, Tijmen Blankevoort, and Yuki Markus Asano. Data repetition beats data scaling in long-cot supervised fine-tuning. 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hanna Hajishirzi, Luke S. Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. pp. 20275–20321, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof qa benchmark. *ArXiv*, abs/2311.12022, 2023.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

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

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.