When Inverse Data Outperforms: Exploring the Pitfalls of Mixed Data in Multi-Stage Fine-Tuning

Anonymous ACL submission

Abstract

Existing work has shown that o1-level performance can be achieved with limited data distillation, but most existing methods focus on unidirectional supervised fine-tuning (SFT), overlooking the intricate interplay between diverse reasoning patterns. In this paper, we construct **r1k**, a high-quality reverse reasoning dataset derived by inverting 1,000 forward examples from s1k (Muennighoff et al., 2025), and examine how SFT and Direct Preference Optimization (DPO) affect alignment under bidirectional reasoning objectives. SFT on r1k yields a **5.4%** accuracy improvement over s1k across evaluated benchmarks. However, naively mixing forward and reverse data during SFT weakens the directional distinction. Although DPO can partially recover this distinction, it also suppresses less preferred reasoning paths by shifting the probability mass toward irrelevant outputs. These findings suggest that mixed reasoning data introduce conflicting supervision signals, underscoring the need for robust and direction-aware alignment strategies.

1 Introduction

Recent studies show that Large Language Models (LLMs) can achieve strong reasoning performance by distilling knowledge from a small set of high-quality examples. Methods like s1 (Muennighoff et al., 2025) and LIMO (Ye et al., 2025) demonstrate that with just 817 to 1,000 curated samples, a 32B model can match or surpass larger systems. However, these approaches focus mainly on single-direction reasoning—solving problems step by step from question to answer. As shown in Figure 1, a model may learn to compute the kinetic energy of a gas molecule from its temperature, but not the reverse: inferring temperature from energy.

This narrow focus overlooks a core aspect of human cognition: **the inherently bidirectional nature** of reasoning. Humans commonly engage in backward reasoning, particularly in goal-directed problem solving (Newell and Simon, 1972; Hawes et al., 2012). Rather than reasoning solely from premises to conclusions, people often begin with a desired outcome and work backward through intermediate steps to reach known facts (Senn and Sacramento, 2015). Motivated by this cognitive insight, recent studies have begun to explore reverse or backward reasoning in LLM. MathGenie (Li et al., 2024) utilizes reverse derivation paths to improve robustness on math word problems. Iterative Question Composing (Cobbe et al., 2024) constructs intermediate subquestions that align with goal-driven, backward-style planning. In optimization modeling, OptiBench (Yang et al., 2024; Chang et al., 2024) promotes reflective and Socratic-style reformulations that partially embody reverse reasoning principles. While promising, these approaches remain constrained to short-context reasoning or domain-specific tasks. This leaves open the broader question of whether backward supervision can enhance long chain-of-thought (CoT) reasoning and generalize across diverse scenarios.

To investigate this, we construct a high-quality dataset, **r1k**, by systematically inverting 1,000 forward reasoning examples from s1k (Muennighoff et al., 2025). Reverse questions and reasoning paths are generated with cost-efficient DeepSeek-R1 model (Team, 2024), without the need for expensive data collection, cleaning, or selection procedures. Fine-tuning on **r1k** yields an approximate 5.4% improvement over s1k.

To further study the interaction between mixed data, we conducted extensive experiments on their combined effects. We observe that SFT on reverse data improves performance, whereas mixing forward and reverse examples leads to degradation. Mechanistic analysis shows that this reduces the model's ability to distinguish reasoning paths. While Direct Preference Optimization (DPO) partly alleviates this, it still suffers from suboptimal initialization and tends to shift reverse reasoning prob-

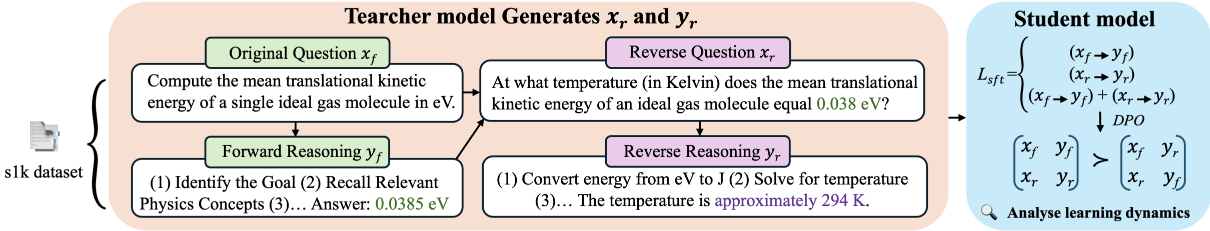

Figure 1: We begin with the s1k dataset (x_f, y_f) and generate reverse questions x_r, along with their corresponding reverse CoTs and answers y_r. We then fine-tune student models using cross-entropy loss under three settings as comparison: forward-only data, reverse-only data, and a mixture of both. To enhance directional consistency, we apply DPO to encourage directionally aligned responses while suppressing misaligned ones. Concurrently, we track the log probability of y_f and y_r across multiple fine-tuning stages to investigate the models' learning dynamics.

ability toward irrelevant outputs. These findings underscore the need for improved alignment strategies to support robust reasoning.

2 Related work

Data-Efficient Reasoning in LLMs: s1 (Muennighoff et al., 2025), LIMO (Ye et al., 2025) and LIMA (Zhou et al., 2023) demonstrate that training on a small set of high-quality examples enables more effective performance, suggesting that massive datasets may not always be necessary to achieve competitive results. This perspective is further supported by methods such as iterative refinement (Madaan et al., 2023) and self-rewarding feedback (Huang et al., 2023), which demonstrate that reusing or distilling informative examples can improve model performance without relying on large-scale data. Complementary findings from data pruning and selection studies (Ivison et al., 2025; Deng et al., 2025; Agarwal et al., 2024) reveal that indiscriminate scaling often yields diminishing returns, highlighting the value of targeted data curation in reasoning-intensive tasks.

Learning Dynamics of LLM Fine-Tuning: Neural Tangent Kernel (NTK) theory (Jacot et al., 2018; Arora et al., 2019) provides a framework for analyzing the influence of individual training examples during LLM fine-tuning. A gradient-based decomposition was later proposed (Ren and Sutherland, 2024), approximating the change in model confidence for an output y on input x_o after training on a single example (x_u, y_u) as:

$$\Delta \log \pi_t(y \mid x_o) \approx -\eta\, A_t(x_o)\, K_t(x_o, x_u)\, G_t(x_u, y_u)$$

where K_t is the empirical NTK, G_t the gradient, and A_t a scaling factor tied to model certainty. This perspective helps explain interference, hallucination, and memorization (Pruthi et al., 2020). It also explains the diversity collapse (Dang et al., 2025), where correctness optimization concentrates the probability mass on a single reasoning path, limiting diversity. These insights inspire a learning-dynamics perspective on how mixed reasoning data shapes model behavior in multi-stage fine-tuning.

3 Methodology

3.1 Reverse Data Construction and Alignment

We begin with a forward reasoning dataset $\mathcal{D}_{s1k} = \{(x_f^{(i)}, y_f^{(i)})\}_{i=1}^{1000}$, consisting of 1,000 high-quality examples from the s1k dataset. Each example includes a question x_f and its corresponding CoT and answer y_f generated by Deepseek-R1 (R1). Based on each s1k's question and final answer, we leverage R1 to construct the reverse dataset $\mathcal{D}_{r1k} = \{(x_r^{(i)}, y_r^{(i)})\}_{i=1}^{1000}$. For each forward example (x_f, y_f), we prompt the model to generate a reverse question x_r that naturally elicits the original reasoning in reverse. Conditioned on x_r, we prompt R1 to generate the corresponding reverse reasoning chain and answer y_r. We merge the above two datasets to obtain $\mathcal{D} = \mathcal{D}_{s1k} \cup \mathcal{D}_{r1k}$, which serves to investigate how bidirectional supervision influences model behavior and alignment.

We fine-tune the Qwen-2.5 (A et al., 2024) 7B and 14B models using the standard cross-entropy objective, where the input is the question x and the target output y is the concatenation of the CoT and the final answer with special separation tokens.

Although SFT introduces forward and reverse reasoning, it does not equip LLMs with the ability to switch between two directions. To better align model responses with question directionality, we apply DPO following SFT. For this, we construct preference pairs of the form (x, y^+, y^-), where x is the question, y^+ is the preferred response, and y^- is the response of reverse question. Specifically, for each example $(x_f, y_f) \in \mathcal{D}_{s1k}$, we treat the forward output as the preferred response, i.e., $y^+ = y_f$, and the corresponding reverse output as the rejected response, $y^- = y_r$. In contrast, for each example $(x_r, y_r) \in \mathcal{D}_{r1k}$, we assign $y^+ = y_r$ as the preferred response and $y^- = y_f$ as the rejected one. Each pair of preferences (x, y^+, y^-) is used to optimize the DPO objective(Rafailov et al., 2023), which encourages the model to prefer y^+ over y^-.

3.2 Analysis of the Pitfalls of Mixed Data

To investigate the fine-tuning behavior during both the SFT and DPO stages, we construct a small *probe training set* consisting of 100 examples: 50 forward instances $\mathcal{D}_f$ and their corresponding 50 reverse counterparts $\mathcal{D}_r$. The union of the two forms a mixed test dataset $\mathcal{D}_m = \mathcal{D}_f \cup \mathcal{D}_r$.

Throughout both the SFT and DPO stages, we monitor model behavior by recording intermediate checkpoints and evaluating the average log-probability (ALP) per token for both y^+ and y^-:

$$\begin{cases} \mathrm{ALP}(y^+) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{1}{|y_i^+|}\sum_{t=1}^{|y_i^+|}\log p(y_{i,t}^+ \mid x_i, y_{i,<t}^+), \\ \mathrm{ALP}(y^-) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{1}{|y_i^-|}\sum_{t=1}^{|y_i^-|}\log p(y_{i,t}^- \mid x_i, y_{i,<t}^-). \end{cases}$$

here, N denotes the number of examples in the *probe testing set*, and $|y_i^{\pm}|$ the length of each evaluated output, capped at 1000 tokens. This evaluation window is typically sufficient to capture the divergence between y^+ and y^-. Since the responses are long-form sequences, we normalize log-probability by sequence length to ensure fair comparison. Motivated by recent theoretical analyses of learning dynamics in LLMs (Ren and Sutherland, 2024), we track the margin:

$$\Delta = \mathrm{ALP}(y^+) - \mathrm{ALP}(y^-),$$

which serves as an empirical proxy for $\Delta \log \pi_t(y \mid x_o)$ in the NTK formulation, with ALP reflecting model certainty $A_t(x_o)$, forward–reverse pairs indicating input similarity $K_t(x_o, x_u)$, and training supervision contributing gradient signals $G_t(x_u, y_u)$.

4 Experiments

We conducted experiments on the DeepSeek-R1's s1k dataset, which consistently outperformed the Gemini-based variant. As the s1k has already been curated with quantity, diversity, and difficulty, we did not apply an additional filtering process. We fine-tune Qwen2.5-Instruct models (7B and 14B) in two stages on 8 A800-80GB GPUs. First, we applied SFT with LoRA (rank = 256, α = 512). Then, we performed DPO using the `trl` library (von Werra et al., 2022), with DeepSpeed ZeRO-3 (Jacobs et al., 2023) and Flash Attention (Shah et al., 2024) to reduce memory usage. We employ open-source lm-eval-harness (Gao et al., 2021), with gpt-4o-mini to evaluate the accuracy.

4.1 Impact of Reverse and Mixed Data

To evaluate the effect of reverse data construction and bidirectional supervision, we conducted fine-tuning experiments on different training datasets. As shown in Table 1, we compare the distillation performance of models trained on the $\mathcal{D}_{s1k}$, $\mathcal{D}_{r1k}$, and $\mathcal{D}$ across three challenging benchmarks: AIME24-NoFigures (Mathematical Association of America, 2024), Math 500 (Lightman et al., 2023), and GPQA (Clark et al., 2022) benchmarks.

Table 1: Effect of Reverse Data ($\mathcal{D}_{r1k}$) and Mixed Training Sets ($\mathcal{D}$ and $\mathcal{D}_{0.5k}$) on downstream performance. Here, $\mathcal{D}_{0.5k}$ consists of 500 forward examples paired with their corresponding reverse data. In the same setting experiment, the best results are shown in bold.

Data	Model	AIME	Math	GPQA	Average
$\mathcal{D}_{s1k}$	7B	16.7%	77.0%	34.0%	42.6%
$\mathcal{D}_{r1k}$(Ours)	7B	**20.0%**	**77.4%**	**42.4%**	**46.6%**
$\mathcal{D}_{0.5k}$	7B	13.3%	71.8%	35.8%	40.3%
$\mathcal{D}$	7B	6.7%	56.0%	31.8%	31.5%
$\mathcal{D}_{s1k}$	14B	20.0%	83.2%	48.4%	50.6%
$\mathcal{D}_{r1k}$(Ours)	14B	**33.3%**	**86.0%**	**53.0%**	**57.4%**
$\mathcal{D}$	14B	30.0%	81.6%	49.1%	53.6%

Under the same distillation pipeline, models trained on our reverse dataset $\mathcal{D}_{r1k}$ achieve an average improvement of 5.4% compared to those trained on the original $\mathcal{D}_{s1k}$. However, combining forward and reverse examples leads to a significant drop in performance. As the size of the mixed dataset increases from $\mathcal{D}_{0.5k}$ to $\mathcal{D}$, the degradation becomes more pronounced, suggesting that mixed-direction reasoning data introduce interference between reasoning modes and hinder effective learning.

4.2 Impact of Directional Preference

Our DPO experiments use a temperature-weighting hyperparameter of $\beta = 0.6$, with the SFT-trained model fixed as the reference model. We apply DPO fine-tuning to four SFT-based models: three 7B

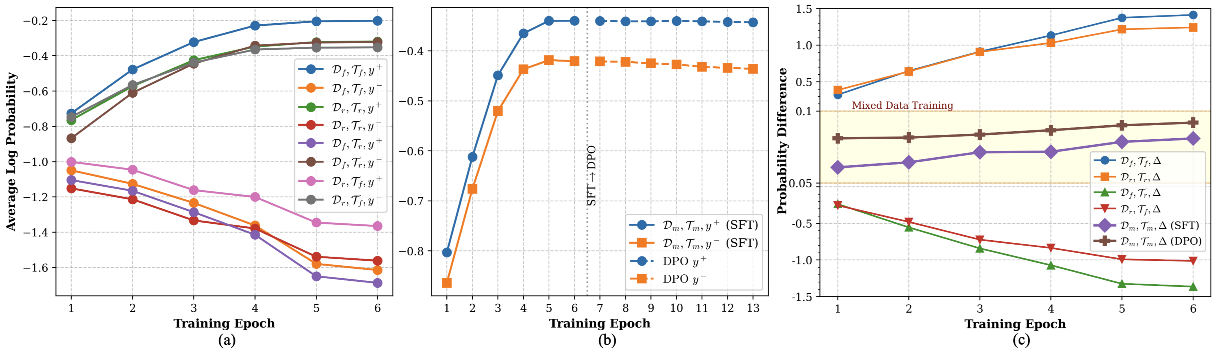

Figure 2: $\mathcal{D}$ denotes the training dataset, and $\mathcal{T}$ denotes the testing dataset. We report the Average Log Probability (ALP) for both the preferred responses (y^+) and the less preferred responses (y^-) in Figures (a) and (b), respectively. Figure (c) shows the difference $\mathrm{ALP}(y^+) - \mathrm{ALP}(y^-)$.

models individually trained on $\mathcal{D}_{s1k}$, $\mathcal{D}_{r1k}$, and $\mathcal{D}$, and a 14B model trained on $\mathcal{D}$. All DPO models are further fine-tuned using preference pairs from $\mathcal{D}$, where each pair consists of two responses y^+ and y^-, generated from the opposite question.

Table 2: Effect of DPO on SFT-Based Models. ↓ and ↑ indicate performance decrease and increase respectively; parentheses show relative change from the SFT baseline.

Based Model	**AIME**	**Math**	**GPQA**	**Average**
$\mathcal{D}_{s1k}$ (7B)	13.3%↓	71.8%↓	35.9%↓	40.3% (↓2.3%)
$\mathcal{D}_{r1k}$(7B)	16.7%↓	75.4%↓	39.4%↓	43.8% (↓2.8%)
$\mathcal{D}$ (7B)	16.7%↑	64.2%↑	34.8%↑	38.6% (↑ 7.1%)
$\mathcal{D}$ (14B)	40.0%↑	81.2%↓	46.4%↓	55.9% (↑ 2.3%)

Table 2 shows that applying DPO with mixed preference data on the $\mathcal{D}_{s1k}$ (7B) and $\mathcal{D}_{r1k}$ (7B) reference models leads to a performance decline. For the model initially fine-tuned on the mixed dataset $\mathcal{D}$, DPO achieves some performance improvements, but its overall performance remains inferior to SFT trained on $\mathcal{D}_{r1k}$.

4.3 Analysis of the Pitfalls of Mixed Data

We analyze the in-distribution pairs $(\mathcal{D}_f, \mathcal{T}_f)$, $(\mathcal{D}_r, \mathcal{T}_r)$, and $(\mathcal{D}_m, \mathcal{T}_m)$, where $\mathcal{D}$ and $\mathcal{T}$ denote the training and testing datasets respectively. SFT is run for 12 epochs with evaluation every 2 epochs, and DPO for 7 epochs with evaluation after each. Changes in the ALP of y^+ reflect the learned strategy, while y^- indicates hallucination. We also consider the out-of-distribution pairs $(\mathcal{D}_f, \mathcal{T}_r)$ and $(\mathcal{D}_r, \mathcal{T}_f)$, where variations in the ALP of y^+ measure the generalization capability to handle reverse question, whereas y^- indicate the likelihood of generating irrelevant or off-target responses.

Figure 2(a) shows that under out-of-distribution scenarios, models trained on the $\mathcal{D}_r$ exhibit lower hallucination rates y^- and better generalization. For in distribution settings, (a) demonstrates that the likelihood of y^+ increase significantly, but this improvement is accompanied by a corresponding rise in hallucinations y^-. Figure 2(b) reveals that mixed-data training $\mathcal{D}_m$ induces a stronger increase in hallucinations, while the likelihood of preferred responses fails to reach the levels achieved by training solely on $\mathcal{D}_r$ and $\mathcal{D}_f$. Even though the subsequent DPO improves preference alignment by suppressing the probability of y^- to irrelevant responses, this suppression is limited.

Figure 2(c) shows that models trained on $\mathcal{D}_f$ and $\mathcal{D}_r$ maintain a gap between y^+ and y^-, whereas the model trained on mixed data $\mathcal{D}_m$ produces only a narrow margin (0.05–0.1). This suggests that SFT on mixed data weakens LLM's ability to discriminate the learned strategies and hallucination. Although DPO slightly separates y^+ and y^-, the effect remains limited. We hypothesize that the conflicting signals from $\mathcal{D}_m$ lead the model to optimize in competing directions, hindering the formation of coherent preferences. This phenomenon also helps explain why models trained on smaller but higher-quality datasets, such as LIMO or s1k, can outperform larger ones: consistent supervision leads to more effective optimization.

5 Conclusion

We constructed a high-quality reverse reasoning dataset r1k and demonstrated its effectiveness in improving reasoning ability. We further investigate the effects of mixed data during multi-stage fine-tuning, underscoring the need for improved alignment strategies to support robust reasoning.

6 Limitations

This work explores the integration of reverse reasoning data in multi-stage fine-tuning, but several limitations remain. Our reverse dataset $\mathcal{D}_{r1k}$ is constructed by automated prompting without human validation, which may introduce subtle errors or inconsistencies in reasoning quality. Additionally, the DPO formulation assumes a strict directional preference between forward and reverse outputs, potentially oversimplifying cases where both reasoning directions offer complementary insights. Furthermore, while this study adopts a standard SFT + DPO pipeline, alternative alignment strategies may offer more robust solutions to conflicting supervision in mixed data settings.

References

Author A, Author B, and Author C. 2024. Qwen-2.5: Large-scale language models with enhanced comprehension and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Accessed: 2025-05-16.

Ishika Agarwal, Krishnateja Killamsetty, Lucian Popa, and Marina Danilevsky. 2024. Delift: Data efficient language model instruction fine-tuning. *arXiv preprint arXiv:2411.04425*.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, and Ruosong Wang. 2019. Harnessing the power of infinitely wide deep nets on small-data tasks. *arXiv preprint arXiv:1910.01663*.

Bowen Chang and 1 others. 2024. Optibench meets resocratic: Measure and improve llms for optimization modeling. *arXiv preprint arXiv:2404.xxxxx*.

Elizabeth Clark, Sameer Singh, Matt Gardner, Tushar Khot, Oyvind Tafjord, Antoine Bosselut, and Oren Etzioni. 2022. Teaching physics to large language models. *arXiv preprint arXiv:2207.05221*.

Karl Cobbe and 1 others. 2024. Iterative question composing for complex reasoning. *arXiv preprint arXiv:2401.09003*.

Jiawei Dang, John Doe, and Jane Smith. 2025. Weight averaging mitigates diversity collapse in llm fine-tuning. *arXiv preprint arXiv:2501.12345*.

Yue Deng and 1 others. 2025. Efficient preference optimization for language model alignment. *arXiv preprint arXiv:2403.12385*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Lewis Tunstall, and Jan Mueller. 2021. Eleutherai/lm-eval-harness: A framework for few-shot evaluation of autoregressive language models. https://github.com/EleutherAI/lm-eval-harness.

Daniel R. Hawes, Alexander Vostroknutov, and Aldo Rustichini. 2012. Experience and abstract reasoning in learning backward induction. *Frontiers in Neuroscience*, 6:23.

Xinyang Huang, Yuxiang Deng, Jing Xu, Sainbayar Sukhbaatar, and Jason Weston. 2023. Self-rag: Learning to retrieve with feedback. *Preprint*, arXiv:2312.04385. ArXiv:2312.04385.

Alex Ivison and 1 others. 2025. Data quality matters: Curation over scale in language modeling. *arXiv preprint arXiv:2402.15732*.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*.

Arthur Jacot, Franck Gabriel, and Clément Hongler. 2018. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 8580–8589.

Yukun Li and 1 others. 2024. Mathgenie: Automated reasoning with llms for math word problems. *arXiv preprint arXiv:2402.16352*.

Samuel Lightman, Adam Roberts, Colin Raffel, Rachel Rudinger, Sarah Wiegreffe, and Samuel R Bowman. 2023. Let's verify step by step. *arXiv preprint arXiv:2303.11366*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mathematical Association of America. 2024. American invitational mathematics examination (aime) problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2024-05-15.

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

Allen Newell and Herbert A Simon. 1972. *Human problem solving*. Prentice-Hall.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in*

Neural Information Processing Systems (NeurIPS), volume 33, pages 19920–19930.

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

Yi Ren and Danica J Sutherland. 2024. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*.

Walter Senn and João Sacramento. 2015. Backward reasoning the formation rules. *Nature Neuroscience*, 18:1705–1706.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685.

DeepSeek Team. 2024. Deepseek r1: Advancing open-source language models with reinforced training and multi-stage reasoning. *arXiv preprint arXiv:2404.12345*.

Jonas von Werra, Younes Belkada, Nathan Lambert, Suraj Patil, Lewis Tunstall, and Thomas Wolf. 2022. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Zhicheng Yang, Yiwei Wang, Yinya Huang, Zhijiang Guo, Wei Shi, Xiongwei Han, Liang Feng, Linqi Song, Xiaodan Liang, and Jing Tang. 2024. Optibench meets resocratic: Measure and improve llms for optimization modeling. *arXiv preprint arXiv:2407.09887*.

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

Chunting Zhou and 1 others. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Appendix

Distributed Optimization. We apply DeepSpeed ZeRO-3 for memory efficiency and optimize multi-GPU communication through NCCL tuning.