Rethinking "RL Generalizes, SFT Memorizes": The Role of SFT Data

Yunlong Hou [*1] **Fengzhuo Zhang** [†*2] **Yuan Cheng** [*1] **Jiachun Pan** [1] **Xingyao Li** [1] **Zhuoran Yang** [2]

Abstract

Large language models trained with Reinforcement Learning (RL) with verifiable rewards exhibit strong reasoning ability and broad generalization, whereas models trained with Supervised Fine-Tuning (SFT) are often viewed as more prone to memorization and limited transfer. This paper rethinks this distinction through the lens of SFT training data. First, we study the role of data source and show that it is critical: a carefully mixed SFT dataset substantially outperforms data generated solely by a larger model. Second, we study the role of data scale and show that matching the number of correct rollouts between SFT and RL greatly improves SFT generalization, while matching the total rollout budget enables SFT to generalize as well as RL. Combining these two factors further enables SFT to generalize even better than RL. Third, by using LLM annotations to characterize the solution methods in training rollouts, we show that larger datasets cover more tail methods and that these tail methods provide generalizable reasoning signals. Finally, we support these empirical findings theoretically by analyzing the training dynamics of shallow transformers under both RL and SFT.

1. Introduction

Large Language Models (LLMs) trained via Reinforcement Learning (RL) or Supervised Fine-Tuning (SFT) have achieved remarkable improvements in reasoning ability, as exemplified by the success of DeepSeek-R1 (Guo and Others, 2025) and s1 (Muennighoff et al., 2025). A line of work compares the strengths and limitations of RL and SFT for reasoning (Chen et al., 2025a; Shenfeld et al., 2025; Huan et al., 2025). These studies suggest that "RL generalizes, whereas SFT memorizes" (Chu et al., 2025; Huan et al., 2026), indicating that SFT often struggles to match RL on reasoning tasks beyond the training domain.

However, it remains unclear whether the conclusion that "RL generalizes, SFT memorizes" is drawn in a setting where the full potential of SFT is realized and SFT is given comparable data resources to RL. In this work, we revisit this conclusion through the lens of the training dataset. Specifically, we study how data source and scale affect SFT generalization, and ask:

1. *What kind of data source best unlocks the generalization potential of SFT?*

2. *How should we match the data consumed by SFT and RL to compare their generalization in a fairer setting?*

To answer the first question, we perform SFT using training data from two sources and their mixtures, and evaluate the resulting models on general reasoning tasks to assess generalization. Specifically, we construct SFT datasets from rollouts of two models: the strong Qwen3-14B model and the RL-tuned base model, pairing each question prompt with one correct rollout. Although both SFT-trained models generalize worse than the RL-trained model, the RL-tuned dataset yields much better generalization than the Qwen3-14B dataset. Moreover, interpolating between the two sources shows that adding an appropriate amount of more diverse data further improves general reasoning. These findings highlight the importance of data source for SFT generalization.

To answer the second question, we match the SFT training dataset size to that of RL under different criteria. Rather than matching SFT and RL training data prompt-wise, we scale the SFT dataset according to three criteria: (1) the number of solvable prompts, (2) the number of correct rollouts, and (3) the total number of rollouts. After training and evaluating models on these datasets, we find that matching the number of correct rollouts used in RL greatly improves SFT generalization, while further matching the total number of rollouts enables SFT to achieve generalization comparable to RL. Moreover, data source and data scale have a compounding effect: SFT with a mixture of the two data

[*]Equal contribution [†]Project lead [1]National University of Singapore, Singapore [2]Yale University, United States. Correspondence to: Fengzhuo Zhang <fzzhang@u.nus.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

sources and a matched number of total rollouts achieves better generalization than RL.

To further demystify the role of data scale, we study how scaling the SFT dataset affects the coverage of LLM-annotated solution methods in training rollouts and, in turn, general reasoning performance. Our analysis shows that larger datasets cover more tail methods in the long-tailed method distribution. By further annotating evaluation rollouts on general reasoning tasks, we find that these tail methods are used more often and solve more questions as the dataset grows. These results suggest that data scaling improves the coverage of long-tailed reasoning patterns, thereby enhancing generalization.

Finally, we theoretically justify our insights through a training-dynamics analysis of shallow transformers. For long-tailed SFT data, we prove that both SFT and RL achieve high accuracy on in-distribution problems. However, with insufficient method coverage, SFT generalizes worse than RL on out-of-distribution (OOD) problems. As more methods are covered, potentially through a larger data scale, SFT can match RL generalization. These results justify the importance of data scale and tail-method coverage for SFT generalization.

In conclusion, our findings suggest that "RL generalizes, whereas SFT memorizes" requires prerequisites on SFT training datasets. In particular, we show that both the SFT training data source and data scale are essential for unlocking the full potential of SFT. When trained with the same total number of samples as RL, SFT with a appropriate mixture of data sources can achieve generalization comparable to, or even better than, that of RL.

2. Preliminaries

Reinforcement Learning with Verifiable Rewards (RLVR) fine-tunes LLMs by maximizing the expected reward of model responses, where the reward is automatically computed from verifiable outcomes (Guo and Others, 2025). Given a set of question prompts $\mathcal{D}_{\mathrm{RL}} \subseteq \mathcal{V}^*$, where $\mathcal{V}^*$ denotes the set of all finite sequences over the alphabet $\mathcal{V}$, the regularized RLVR objective for model parameter $\theta \in \Theta$ is

$$J_{\mathrm{RL}}(\theta) = \mathbb{E}\big[r(X,Y)\big] - \beta\mathbb{E}\Big[\frac{\log\big(\mathbb{P}_\theta(Y \mid X)}{\mathbb{P}_{\text{ref}}(Y \mid X)\big)}\Big],$$

where the expectations are taken for $X \sim \mathcal{D}_{\mathrm{RL}}, Y \sim \mathbb{P}_\theta(\cdot \mid X)$ The first term is the expected reward of model rollouts $Y \sim \mathbb{P}_\theta(\cdot \mid X)$ on the question-prompt set, where $r(X,Y)$ evaluates the correctness of the response Y to the question X, and $X \sim \mathcal{D}_{\mathrm{RL}}$ means that X is sampled uniformly from $\mathcal{D}_{\mathrm{RL}}$. The second term regularizes the model $\mathbb{P}_\theta$ to remain close to the reference model $\mathbb{P}_{\mathrm{ref}}$, typically the pre-RLVR model. Directly optimizing the model parameters with $J_{\mathrm{RL}}(\theta)$ often leads to poor performance due to large gradient variance and coverage issues (Schulman et al., 2015; 2017). Instead of maximizing J_{RL} directly, Proximal Policy Optimization (PPO) (Schulman et al., 2017) and its recent powerful variant Group Relative Policy Optimization (GRPO) (Guo and Others, 2025) maximize the following surrogate objective at each step:

$$J_{\mathrm{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|Y_i|}\sum_{t=1}^{|Y_i|}J_{i,t}\Bigg] \quad \text{where}$$

$$J_{i,t} = \min\big(w_{i,t}(\theta)A_i,\ \mathrm{clip}_\varepsilon(w_{i,t}(\theta)A_i\big) - \beta\mathrm{KL}(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\mathrm{ref}}).$$

where the expectation is taken over questions $X \sim \mathcal{D}_{\mathrm{RL}}$ and G rollouts $\{Y_i\}_{i=1}^G \sim \mathbb{P}_{\theta_{\mathrm{old}}}(\cdot \mid X)$. Here, $w_{i,t}(\theta) = \mathbb{P}_\theta(Y_{i,t} \mid X, Y_{i,<t})/\mathbb{P}_{\theta_{\mathrm{old}}}(Y_{i,t} \mid X, Y_{i,<t})$ is the importance-sampling ratio between the current model and the previous-step model, $\mathrm{clip}_\varepsilon(x)$ clips x to $[1-\varepsilon, 1+\varepsilon]$, and $A_i = (r_i - \mathrm{mean}(\{r_j\}_{j\in[G]}))/\mathrm{std}(\{r_j\}_{j\in[G]})$ is the group-normalized advantage of the i-th rollout. GRPO maximizes J_{GRPO} to update θ, typically using Adam. Empirically, Guo and Others (2025) show that GRPO substantially improves multi-step reasoning performance on mathematics, logical deduction, and scientific reasoning tasks. Subsequent studies further show that RL-induced gains on the training domain, such as mathematics, can generalize to other reasoning domains, including physics and broader scientific topics (Chu et al., 2025; Huan et al., 2026). See more related works in Appendix A.

SFT equips pretrained LLMs with long-form reasoning abilities by directly minimizing the cross-entropy loss on curated long-reasoning data (Muennighoff et al., 2025; Wang et al., 2026a). Let $\mathcal{D}_{\mathrm{SFT}}$ denote a long-reasoning dataset consisting of pairs (X, Y), where X is a question prompt and Y is the corresponding long-reasoning answer. The objective of long-reasoning SFT is $L_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\mathrm{SFT}}}[\sum_{t=1}^{T}\log\mathbb{P}_\theta(Y_t \mid X, Y_{<t})]$. In practice, this optimization is typically solved with first-order optimizers such as Adam (Kingma and Ba, 2017) or AdamW (Loshchilov and Hutter, 2019). Because dense supervision is available at every token position, SFT is often sample-efficient (Zhao et al., 2026a). However, prior work (Chu et al., 2025; Huan et al., 2026) suggests that SFT mainly improves model reasoning ability on tasks that are in distribution with respect to the training data, and that the resulting models exhibit weaker generalization than RLVR-trained models.

3. Experiments

In this section, we conduct extensive experiments to compare the generalization ability of RL and SFT. Following Huan et al. (2026), we train models with SFT and RL on

mathematical reasoning datasets and evaluate them on other general reasoning tasks to assess out-of-domain generalization. To further understand what affects the generalization of SFT models, we study two critical properties of the training data: (1) its source and (2) the dataset scale. We first describe the basic training setups used throughout the experiments.

Experimental Setup. We adopt Qwen2.5-7B (Qwen et al., 2025) as the base model and train it on mathematical problems from the DAPO dataset (Yu et al., 2026). For RL, we use the questions in the DAPO dataset with a reasoning prompt prefix, and optimize the model with GRPO (Guo and Others, 2025) implemented in VeRL (Sheng et al., 2025) under the standard training setup. For SFT, we use rollout models, Qwen3-14B (Yang et al., 2025) by default, to generate training responses for the same DAPO prompts, and optimize the model with LlamaFactory (Zheng et al., 2024); detailed hyperparameters are provided in Appendix B.2. We evaluate the trained models on both in-distribution mathematical reasoning tasks and general reasoning tasks. For mathematical reasoning, we primarily use the SimpleRL evaluation pipeline (Zeng et al., 2025), covering AIME24 (Jia, 2024), Minerva Math (Lewkowycz et al., 2022), GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024). For general reasoning, we use twelve tasks from the Language Model Evaluation Harness (Gao et al., 2024), including GPQA-Diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024), with the full list deferred to Appendix B.3. These tasks span science, action planning, deep semantic understanding, multilingual reasoning, and related domains beyond mathematics. For each task, we normalize the metric to $[0, 100]$ and report the average score across tasks.

3.1. Training Data Source Matters for SFT Generalization

In this section, we investigate how the training data source influences the generalization ability of SFT. We first introduce baselines and two data sources.

Baselines and Data Sources. Throughout the paper, we use the base model, i.e., the model without RL or SFT tuning, and the model trained by RL as baselines. We consider two data sources for SFT. The first is constructed by rejection sampling from a strong teacher model, Qwen3-14B (Dong et al., 2023; Yuan et al., 2023; Xiong et al., 2025a): for each question, we generate at most 128 rollouts and retain one correct rollout if available. We refer to the resulting dataset, of size 15,607, as the *Qwen3-14B dataset*. The second data source uses the model trained by RL for rejection sampling under the same protocol. Since the RL-tuned model is obtained from the base model, its trajectories are expected to be more aligned with the base model parameters. We refer to this dataset, of size 14,512, as the *RL-tuned dataset*. We highlight that both the Qwen3-14B and RL-tuned datasets match the RL dataset in the number of prompts. To ablate the effect of data source more finely, we further interpolate between these two datasets with 5 mixing ratios. The k-th mixed dataset, denoted by Mix-k, contains $(11-2k) \times 10\%$ data from the RL-tuned dataset and $(2k-1) \times 10\%$ data from the Qwen3-14B dataset, where $k \in \{1, \ldots, 5\}$.

Table 1 reports results on all mathematical reasoning datasets and a representative subset of general reasoning datasets, with the full results deferred to Appendix B.4. Figure 1 shows the average performance over the mathematical and general reasoning datasets.

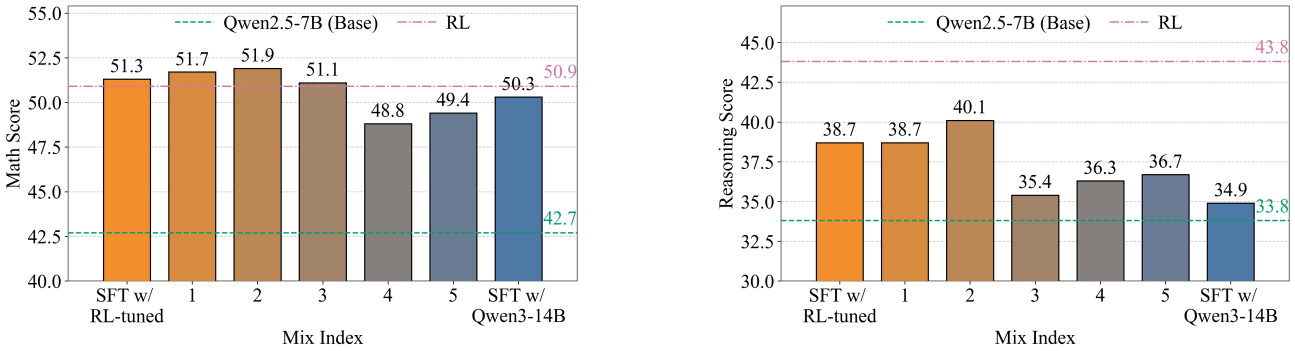
First, for the two baselines, our results show that RL tuning improves the base model on all mathematical reasoning tasks: the average math score increases from 42.7 to 50.9 after RL training on DAPO. Notably, this improvement also transfers to general reasoning tasks, where the average score increases from 33.8 to 43.8. Second, comparing SFT with Qwen3-14B and RL-tuned data, Figure 1 shows that both data sources improve the base model's mathematical reasoning ability, reaching performance comparable to or even higher than the model trained by RL. However, both SFT variants perform worse than the RL-tuned model on general reasoning tasks, while SFT with the RL-tuned dataset still outperforms SFT with the Qwen3-14B dataset. Finally, comparing different mixtures of Qwen3-14B and RL-tuned data, Table 1 and Figure 1 show that pure RL-tuned data or strong model-generated data does not always yield the best general-reasoning performance. Instead, Mix-2, which consists of 70% data from the RL-tuned model and 30% from Qwen3-14B, achieves the highest average score on general reasoning tasks. These lead to the following observation.

Observation 1: The training data source is crucial for generalization: data from a model closer to the one being fine-tuned can be more effective than data from a stronger model, while data from a single model is not necessarily optimal; incorporating *an appropriate amount of more diverse data* can further improve general reasoning ability.

3.2. Matched Rollout Budgets Enable SFT to Generalize Like RL

While the previous section shows that a suitable SFT data source improves general reasoning, the resulting model still lags behind the model trained by RL. We next study an orthogonal factor: how SFT training-set size affects generalization.

In Section 3.1, RL and SFT are matched at the question-prompt level, as both use questions from the DAPO dataset.



(a) Math benchmarks

(b) General reasoning benchmarks

Figure 1. Performance comparison of the SFT-trained models with training datasets of different mixing ratios: Mix-k consists of $(11-2k)\times 10\%$ RL-tuned dataset and $(2k-1)\times 10\%$ Qwen3-14B dataset. 70% from RL-tuned model and 30% from Qwen3-14B yields the best performance.

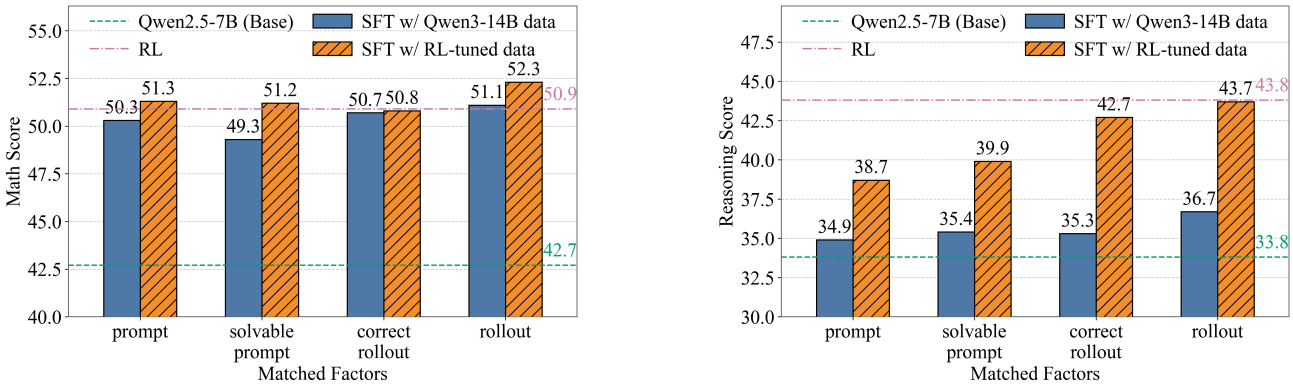
	Math					General Reasoning				
Model	AIME24 (Avg@32)	Minerva Math	GSM8k	MATH500	Olympiad Bench	ACP Bench	BBH	C-Eval	Ground Cocoa	MMLU -Pro
Qwen2.5-7B	4.7	26.5	88.6	63.6	30.2	23.3	10.9	40.1	34.7	9.2
RL-tuned	14.2	34.9	90.8	74.8	39.7	50.1	42.9	43.6	34.8	45.7
SFT w/ RL-tuned	13.8	37.1	91.4	74.8	**39.3**	**58.0**	33.8	42.4	35.6	19.4
SFT w/ Mix-1	**13.9**	37.5	91.3	**76.6**	39.0	56.6	34.4	36.4	36.8	17.9
SFT w/ Mix-2	13.4	**41.2**	**91.9**	76.2	36.9	53.4	37.3	47.8	**37.0**	**21.3**
SFT w/ Mix-3	12.4	38.6	91.6	75.0	37.8	50.7	35.6	28.2	36.7	3.8
SFT w/ Mix-4	10.7	31.2	90.8	73.8	37.6	43.7	34.1	52.9	34.6	0.7
SFT w/ Mix-5	11.0	34.2	90.2	73.2	38.2	46.6	37.7	**53.8**	34.2	1.2
SFT w/ Qwen3-14B	**13.9**	33.5	90.5	75.0	38.5	45.3	**42.1**	36.5	33.2	1.1

Table 1. Evaluation results on math and selected general reasoning tasks of models trained from mixed data sources. The highest and second highest number within the SFT-tuned models are highlighted in **bold** and underlined, respectively. SFT w/ Mix-2 achieves the best performance among the SFT-tuned models.

However, RL generates 8 rollouts per prompt, whereas SFT observes only one correct rollout per prompt. As a result, RL trains on $8\times$ more rollouts than SFT. This motivates us to increase the SFT dataset size under different rollout-budget matching criteria.

- We first match *the number of solvable prompts*, defined as prompts for which at least one correct rollout is obtained during RL training. Under this criterion, the corresponding Qwen3-14B and RL-tuned datasets each contain one correct rollout for every solvable prompt. This yields a training set of about 10,000 prompt–rollout pairs, fewer than the one used in Section 3.1.
- We then match *the number of correct rollouts* per prompt in the SFT training dataset to that observed during RL training. Specifically, for each prompt, we record the number of correct rollouts generated during RL and construct the SFT dataset with the same number of correct rollouts for that prompt. This yields a dataset with about 39,000 prompt–rollout pairs.
- Lastly, we match *the total number of rollouts* for each prompt. Since RL training uses both correct and incorrect rollouts, we augment the 39,000 correct training samples with about 98,000 incorrect-rollouts. This criterion also matches the gradient computation cost of SFT to that of RL. The resulting SFT datasets contain about 138,000 prompt–rollout pairs.

More details of constructing these SFT datasets are defered to Appendix B.2. Evaluation results are presented in Figure 2. First, comparing the prompt-matched and solvable-prompt-matched SFT datasets shows that the solvable-prompt-matched dataset achieves comparable mathematical reasoning ability but better general reasoning ability. One possible explanation is that forcing the model to learn from prompts that are not solvable during RL training can hurt generalization. Second, comparing the prompt-matched and correct-rollout-matched datasets shows that matching the number of correct rollouts substantially improves SFT generalization, yielding general reasoning performance only slightly below that of the RL-trained model. Finally, comparing the correct-rollout-matched and rollout-matched datasets shows that including rollouts with incorrect final answers can further improve SFT generalization. This may be because such rollouts still contain correct intermediate reasoning steps that can benefit SFT (Tian et al., 2026). Moreover,



(a) Math benchmarks

(b) General reasoning benchmarks

Figure 2. Performance comparisons of the SFT-trained model on math and general reasoning benchmarks when different factors during training are aligned with RL training. SFT with the same amount of training data as RL substantially improves general reasoning performance.

SFT with the rollout-matched dataset achieves comparable or better mathematical and general reasoning ability than the RL-trained model. These results lead to the following observation.

Observation 2: The amount of training data substantially influences SFT generalization: matching the number of correct rollouts greatly improves SFT generalization, while further matching the total rollout budget enables SFT to achieve comparable generalization to RL.

We also conducted an ablation study to exclude the effect of the compute in Appendix B.5, which further supports our observations on the training-set size.

Compounding Effects of Data Source and Scale. Observations 1 and 2 show that choosing an appropriate data source and increasing the training-set size can each improve the generalization of SFT. We next investigate whether combining these two factors enables SFT to generalize even better than RL. Specifically, we mix the rollout-matched training dataset across different data sources. Since Section 3.1 suggests that the best mixture ratio is around 70% from the RL-tuned model and 30% from Qwen3-14B, we focus on ratios near this setting to reduce computation. We consider mixture ratios $(11-2k)\times 10\%$ of the RL-tuned dataset and $(2k-1)\times 10\%$ of the Qwen3-14B dataset for $k \in \{1,2,3\}$, yielding three training datasets, each with about 138,000 prompt–rollout pairs.

The experimental results are reported in Figure 3, It shows that the best SFT performance is achieved by the dataset mixing 90% RL-tuned dataset with 10% Qwen3-14B dataset, which yields better generalization than the RL-trained model. We reach the following observation.

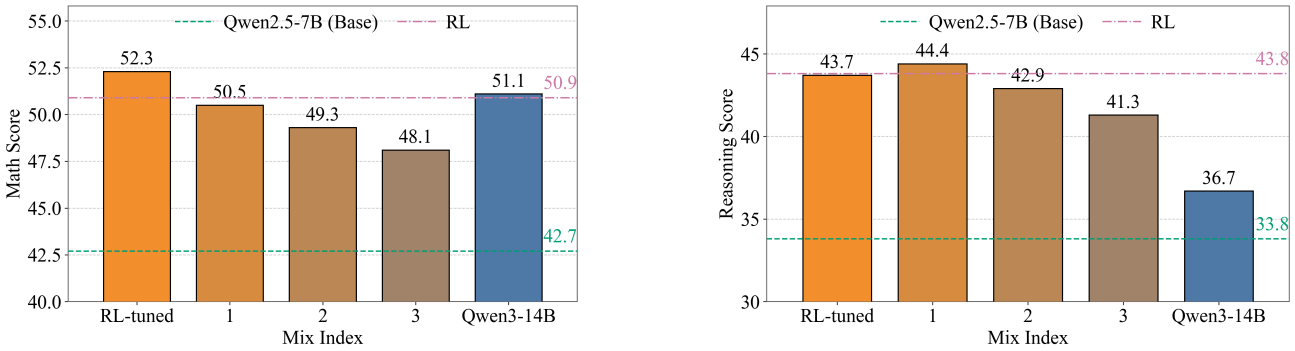
Observation 3: With a carefully mixed dataset and a rollout budget matched to RL, SFT achieves better general reasoning ability than RL, qualifying "RL generalizes, SFT memorizes."

3.3. Long-Tailed Rollouts Provide Generalizable Reasoning Signals

Our experiments in Section 3.2 show that training-set size plays a critical role in enabling SFT to match RL generalization. In this section, we investigate what additional information in larger datasets drives this effect. As the training set scales, it naturally includes more diverse rollouts, including rare tail rollouts that are less likely to appear in smaller datasets. We conjecture that these rare but informative rollouts are essential for SFT generalization, and we verify this hypothesis below.

We first characterize the rollout distribution by the *methods* used to solve each math problem. To study how dataset size changes the contained information, we construct training datasets with $k \in \{1,2,4,8\}$ rollouts per prompt, where $k = 8$ matches the total rollout budget of RL. We then prompt Qwen3-32B to annotate the methods used in each rollout, allowing multiple labels per rollout. After consolidating the annotations, we identify 138 unique methods for solving math problems. More experimental details are in Appendix B.6. In the following, for robustness, we say that a method is *covered* by a dataset if it appears at least 100 times in that dataset.

Figure 4(a) plots the top-50 method frequencies in the datasets with $k = 8$, with methods ranked by their frequencies in the Qwen3-14B dataset. The figure reveals a clear long tail: many methods appear only infrequently. For each dataset, we define *tail methods* as those in the bottom



(a) Math benchmarks

(b) General reasoning benchmarks

Figure 3. Performance comparison of the SFT-trained model with scaled training datasets of different mixing ratios. When the training data is scaled, 90% from RL-tuned model and 10% from Qwen3-14B yields the best performance.

20% by frequency among all identified methods. Figure 4(b) further plots the number of covered tail methods as k varies. For both the Qwen3-14B and RL-tuned datasets, larger datasets cover more tail methods, exposing SFT models to a broader set of rare but potentially generalizable reasoning patterns.

Tail Method Helps. To investigate how tail methods in the training dataset influence generalization to general reasoning tasks, we collect the reasoning rollouts during evaluation from models SFT-tuned on the four datasets. We then prompt Qwen3-32B to identify the methods employed in each rollout and compute the usage statistics on these tasks. Figure 5 shows the usage frequency of the tail methods on two representative tasks, MMLU-Pro (Wang et al., 2024) and MMLU-Pro+ (excluding math) (Asgari et al., 2024), for the models trained on datasets with different k. We report the usage counts both over all evaluation rollouts and over correct rollouts only.

We observe that tail methods are used more frequently as the training-set size scales. Furthermore, these tail methods are capable of solving more questions as the tail-method coverage increases. We arrive at the following observation.

Observation 4: One benefit of increasing the training samples is improved coverage of long-tailed patterns. Training on these rare but informative samples can enhance performance on downstream general reasoning tasks.

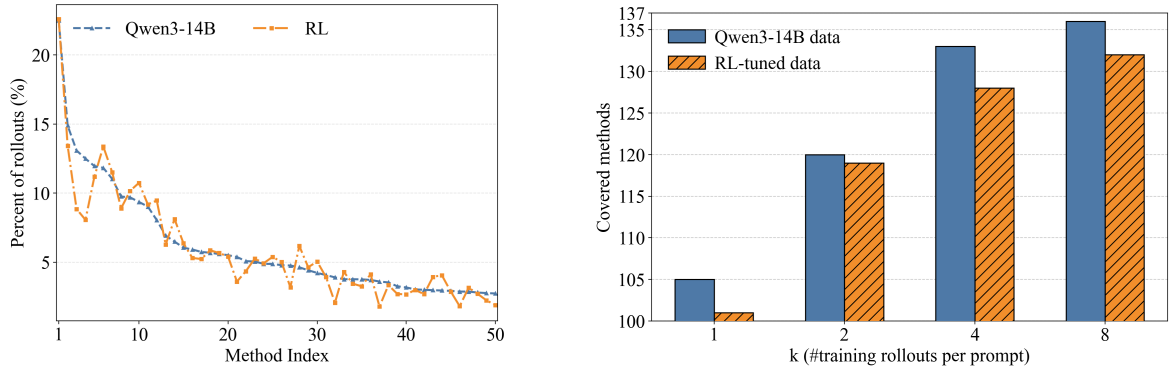
4. A Case Study of Shallow Transformers

Section 3 shows that SFT generalization depends critically on training-data scale, especially the coverage of long-tailed methods. We now provide theoretical grounding for these findings by analyzing the training dynamics of shallow transformers with a simplified data model.

Data model. We first introduce our simplified data model. The question dataset contains L questions. For the l-th question, the prompt consists of $N+1$ token embeddings, $P = (X_1^\top, \ldots, X_N^\top, X_q^\top)^\top \in \mathbb{R}^{(N+1)\times d}$, with the first N token denoted as $P_{\text{in}} = (X_1^\top, \ldots, X_N^\top)$. The row-wise last token embedding $X_q = u_l$ represents the query, where $\mathcal{U} = \{u_1, \ldots, u_L\} \subset \mathbb{R}^{1\times d}$ denotes the set of query embeddings and d is the embedding dimension. The first N tokens are sampled from two types of embeddings: informative embeddings $\mathcal{I}_l = \{i_{l,1}, \ldots, i_{l,K}\} \subset \mathbb{R}^{1\times d}$ and context embeddings $\mathcal{C} = \{c_1, \ldots, c_J\} \subset \mathbb{R}^{1\times d}$. The informative embeddings represent solution cues in the question, while the context embeddings represent connecting words. For the l-th question, each X_i is sampled independently by first choosing $\mathcal{I}_l$ or $\mathcal{C}$ with equal probability, and then sampling uniformly from the selected set. Thus, prompts across all questions contain tokens from $\mathcal{I} = \cup_{l=1}^{L} \mathcal{I}_l$, $\mathcal{C}$, and $\mathcal{U}$. For simplicity, we assume that all embeddings in $\mathcal{I} \cup \mathcal{C} \cup \mathcal{U}$ are orthonormal (Huang et al., 2023a; Nichani et al., 2024). We denote by $\mathbb{P}_{\text{pt}}$ the distribution over sampled questions and their corresponding prompt embeddings, induced by the uniform distribution over questions. Each question has K correct answers, $\mathcal{A}_l = \{a_{l,1}, \ldots, a_{l,K}\} \subset \mathbb{R}^{1\times(LK+1)}$, where $a_{l,k}$ corresponds to the informative token $i_{l,k}$. Here, $a_{l,k}$ represents both the final answer and the solving steps implied by the solution cue in $i_{l,k}$. Let $\mathcal{A} = \cup_{l=1}^{L} \mathcal{A}_l$ denote the set of all correct answers. The full answer space is $\bar{\mathcal{A}} = \mathcal{A} \cup \{\tilde{a}\}$, where $\tilde{a}$ represents all possible wrong answers. Similar to prompt embeddings, we assume that the embeddings in $\bar{\mathcal{A}}$ are orthonormal.

Further, we assume that answers presented in the SFT training dataset is long-tailedly distributed.

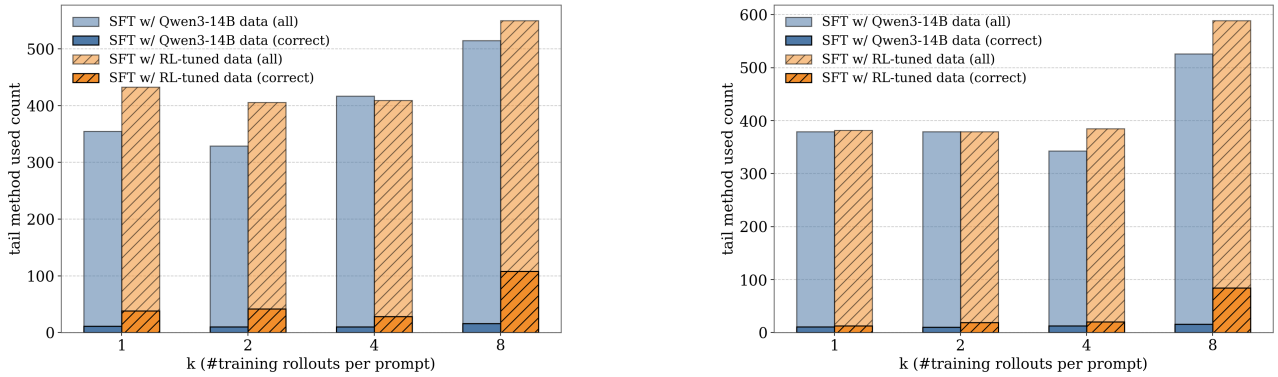
Assumption 4.1 (Long-Tailed Distribution in SFT Data)**.** For each $l \in [L]$, we index the candidate answers so that $a_{l,1}$ is the unique head answer and $\{a_{l,2}, \ldots, a_{l,K}\}$ are tail answers. Let $\mathbb{P}_{\text{SFT}}(A = a_{l,k} \mid X_q = u_l) = p_{l,k}$ be the conditional answer distribution of the SFT data. We assume



(a) Distribution of method usage (top-50).

(b) Method coverage

Figure 4. Characterization of the method distribution and coverage. (a) The method distribution is long-tailed; (b) Method coverage increases as training data scales.



(a) MMLU-Pro

(b) MMLU-Pro+

Figure 5. Tail method usage statistics. Tail methods are used more frequently (light bars) and can correctly solve more questions (dark bars) in general reasoning tasks (e.g., MMLU-Pro, MMLU-Pro+), as the training dataset encompasses a broader range of methods.

for every $l \in [L]$, $p_{l,1} = \Theta(1)$ and $p_{l,k} = o(1)$ for $k \neq 1$.

This assumption captures long-tailedness under mild requirements. Specifically, we only require the head probability to be $O(1)$ and the tail probabilities to be $o(1)$, mimicking the empirical pattern in Figure 4(a).

Model architecture. We next introduce the model structure. Specifically, we consider the model with a one-layer attention module: $F(P_{\text{in}}, X_q; W) = \texttt{soft}(\texttt{soft}(X_q W P_{\text{in}}^\top) P_{\text{in}} W_{\text{lm}}) \in \mathbb{R}^{1\times(LK+1)}$, where $\texttt{soft}(\cdot)$ denotes the softmax operator applied to a row vector. The inner softmax gives the attention weights, with the query embedding X_q as the query and the prompt embeddings P as both keys and values. The parameter $W \in \mathbb{R}^{d\times d}$ is optimized to capture the correlation between the query and the keys. The attention output is then mapped by the language-model head $W_{\text{lm}} \in \mathbb{R}^{d\times(LK+1)}$ to answer-token logits, and the outer softmax converts these logits into answer probabilities. We assume that W_{lm} is fixed and inherited from prior knowledge in the pretrained model. Specifically, W_{lm} stores associative memories between informative tokens and their corresponding answers, as commonly observed in pretrained models (Fang et al., 2024; Meng et al., 2022a;b). We model this as

$$W_{\text{lm}} = c_{\text{lm}} \log(K) \Big(\sum_{l=1}^{L} \sum_{k=1}^{K} i_{l,k}^\top (a_{l,k} - \sum_{l' \neq l} \sum_{k'=1}^{K} a_{l',k'} - \widetilde{a}) + \sum_{j=1}^{J} c_j^\top \widetilde{a} \Big) \in \mathbb{R}^{d\times(LK+1)}, \quad \text{where } c_{\text{lm}} > 0.$$

This construction encodes one-to-one associations between informative tokens and correct answers, while associating context tokens with the incorrect answer $\widetilde{a}$. We index the k-th answer of the l-th question in $\bar{\mathcal{A}}$ by the global index $(l-1)K + k$. For simplicity, when there is no ambiguity, we omit the dependence on P, X_q, and W, and write $F_{l,k}$, or equivalently $F_{(l-1)K+k}$, for the predicted probability of this answer in $\bar{\mathcal{A}}$.

Training recipes. SFT and RL update the attention parameter W according to their respective training recipes. Both start from zero initialization, i.e., $W_{\mathrm{SFT}}^{(0)} = W_{\mathrm{RL}}^{(0)} = 0_{d\times d}$. SFT optimizes W with the following loss:

$$L_{\mathrm{SFT}}(W) = -\sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{E}\big[\mathbb{1}\{X_q = u_l, A = a_{l,k}\}\log(F(P_{\mathrm{in}}, X_q; W)_{(l-1)K+k})\big].$$

The expectation $\mathbb{E} = \mathbb{E}_{P_{\mathrm{in}},X_q,A\sim\mathcal{D}_{\mathrm{SFT}}}$ is taken with respect to the joint distribution induced by the prompt-embedding distribution $\mathbb{P}_{\mathrm{pt}}$ and the conditional answer distribution in Assumption 4.1. With learning rate $\eta_{\mathrm{SFT}} > 0$, SFT updates the parameter by $W_{\mathrm{SFT}}^{(t+1)} = W_{\mathrm{SFT}}^{(t)} - \eta_{\mathrm{SFT}}\nabla_W L_{\mathrm{SFT}}(W_{\mathrm{SFT}}^{(t)})$. As introduced in Section 2, RL updates the model by maximizing the expected reward of model rollouts. For theoretical simplicity, we let RL directly optimize the unregularized expected reward, i.e.,

$$J_{\mathrm{RL}}(W) = \sum_{l=1}^{L}\sum_{k=1}^{K}\mathbb{E}_{P_{\mathrm{in}},X_q}\left[\mathbb{1}\{X_q = u_l\}F(P_{\mathrm{in}}, X_q; W)_{(l-1)K+k}\right].$$

For each question prompt P, the reward is defined as the total probability mass assigned to all correct answers, i.e., $\sum_{k=1}^{K} F(P_{\mathrm{in}}, X_q; W)_k$. The parameter is then optimized by gradient ascent: $W_{\mathrm{RL}}^{(t+1)} = W_{\mathrm{RL}}^{(t)} + \eta_{\mathrm{RL}}\nabla_W L_{\mathrm{RL}}(W_{\mathrm{RL}}^{(t)})$. For simplicity, we consider only a one-step RL update, which suffices for the shallow model to reach optimality. This simplification is standard in the learning-dynamics literature (Huang et al., 2026; Wang et al., 2025; Ba et al., 2022).

Theorem 4.2 (In-Distribution Convergence of SFT and RL)**.** *Consider the regime $N \gg \mathrm{poly}(K) \gg \mathrm{polylog}(N)$, $N \gg d$, $K \gg L$. Let $\epsilon = \log(K)/K > 0$ denote the error threshold. Then the following results hold with probability at least $1 - \exp(-\Omega(\mathrm{poly}(K)))$.*

- *Under Assumption 4.1, when SFT tunes the parameter with a learning rate large enough $\eta_{\mathrm{SFT}} = \Theta(LK/p_{l,1})$, after one step gradient update, the learned model achieves high accuracy on the training dataset. Specifically, for the l-th question, the wrong answer probability of the learned model is upper-bounded as follows: $1 - \sum_{k=1}^{K} F_{(l-1)K+k} \le \epsilon$.*
- *When RL tunes the parameter with learning rates $\eta_{\mathrm{RL}} = \Theta(\mathrm{poly}(K, \log(1/\epsilon)))$, the learned model achieves high accuracy on the training dataset, i.e., for the l-th question, the model learns all the correct answers equally, i.e., $|F_{(l-1)K+k} - 1/K| \le \epsilon/K$, $\forall k \in [K]$. The wrong answer probability is also upper bounded by ϵ.*

These results theoretically show that both SFT and RL can achieve high accuracy on in-distribution problems, i.e., problems covered by the training dataset. This is consistent with our experimental results in Section 3, where SFT achieves math reasoning ability comparable to RL when the two methods are matched by the number of question prompts. This finding is also consistent with Chu et al. (2025); Lu et al. (2026). Additionally, we also show that RL-trained models explore and learn a balanced distribution across answers, which supports their generalization ability.

To assess the generalization ability of models learned by SFT and RL, we evaluate them on OOD prompts. For each query $X_q = u_l$, an OOD prompt, denoted by $P_{l,k}^{\mathrm{OOD}}$, contains a single informative token type $i_{l,k}$ repeated $N/2$ times, with the remaining $N/2$ positions filled by context tokens, where $k \in [K]$. Thus, $a_{l,k}$ is the unique correct answer for this prompt. For each $l \in [L]$, the OOD test distribution mixes over $k \in [K]$ with weights $\{\rho_{l,k}\}_{k=1}^{K}$, where $\sum_{k=1}^{K}\rho_{l,k} = 1$ and $\rho_{l,k} \ge 0$. We evaluate a model by its probability of outputting the correct answer: $\mathrm{Eval}(W) = L^{-1}\sum_{l=1}^{L}\sum_{k=1}^{K}\rho_{l,k}F(P_{l,k}^{\mathrm{OOD}}, u_l; W)_{(l-1)K+k}$.

Theorem 4.3 (OOD Generalization)**.** *Denote the trained weights of SFT and RL as W_{SFT} and W_{RL}, respectively. Let $\mathrm{Eval}_{\mathrm{SFT}} = \mathrm{Eval}(W_{\mathrm{SFT}})$, $\mathrm{Eval}_{\mathrm{RL}} = \mathrm{Eval}(W_{\mathrm{RL}})$. We have*

- *(RL always generalizes.) For any OOD mixture $\{\rho_{l,k}\}_{k=1}^{K}$, we have $1 - \mathrm{Eval}_{\mathrm{RL}} = O(K^{-3})$.*
- *(SFT generalizes on covered methods.) If the OOD mixture distribution is supported on answers whose training probabilities satisfy $p_{l,k} \ge 1/\log K$, then $1 - \mathrm{Eval}_{\mathrm{SFT}} = O(K^{-3})$. In contrast, if it is supported on answers whose training probabilities satisfy $p_{l,k} \le 1/K$, then $\mathrm{Eval}_{\mathrm{SFT}} \le 1/2$.*

First, we show that RL generalizes to OOD prompts, which serve as a stylized analogue of the general reasoning tasks in our experiments. This is consistent with the strong generalization of RL observed in prior work (Chu et al., 2025). Second, we show that SFT generalizes to well-covered methods, characterized by $p_{l,k} \ge 1/\log K$. This explains Observations 2 and 4: increasing the data scale covers more methods and raises the empirical frequencies of tail methods in the dataset, as shown in Figure 5.

5. Conclusion

This work rethinks the claim that "RL generalizes, whereas SFT memorizes" from the perspective of SFT training data. Specifically, we show that both data source and data scale substantially influence SFT generalization. With a carefully mixed dataset and a rollout budget matched to RL, SFT can even generalize better than RL. We further show that this improvement is supported by increased coverage of tail methods, and provide theoretical analysis to justify these insights. A limitation of our work is that we focus on LLMs; extending our analysis to other generative models, such as diffusion models, is left for future work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Daya Guo and Others. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025a.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why on-policy reinforcement learning forgets less. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning, 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. In *Forty-second International Conference on Machine Learning*, 2025.

Maggie Ziyu Huan, Yuetai Li, Tianyu Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general LLM capabilities? understanding transferability of LLM reasoning, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, et al. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*, 58(8):1–35, 2026a.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026a.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

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. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, page 1279–1297. Association for Computing Machinery, 2025. doi: 10.1145/3689031.3696075.

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, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun MA, and Junxian He. SimpleRL-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*, 2025.

Maxwell Jia. Aime problem set 2024, 2024.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Advances in neural information processing systems*, 2021.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

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 q&a benchmark. In *First Conference on Language Modeling*, 2024.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and Hanze Dong. A minimalist approach to llm reasoning: from rejection sampling to reinforce, 2025a.

Xueyun Tian, Minghua Ma, Bingbing Xu, Nuoyan Lyu, Wei Li, Heng Dong, Zheng Chu, Yuanzhuo Wang, and Huawei Shen. Learning from mistakes: Negative reasoning samples enhance out-of-domain generalization, 2026.

Saeid Asgari, Aliasghar Khani, and Amir Hosein Khasahmadi. MMLU-pro+: Evaluating higher-order reasoning and shortcut learning in LLMs. In *Neurips Safe Generative AI Workshop 2024*, 2024.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023a.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Yu Huang, Zixin Wen, Yuejie Chi, Yuting Wei, Aarti Singh, Yingbin Liang, and Yuxin Chen. On the learning dynamics of rlvr at the edge of competence. *arXiv preprint arXiv:2602.14872*, 2026.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Aojun Lu, Tao Feng, Hangjie Yuan, Wei Li, and Yanan Sun. Why does rl generalize better than sft? a data-centric perspective on vlm post-training. *arXiv preprint arXiv:2602.10815*, 2026.

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

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025a. Notion Blog.

Bingxiang He, Zekai Qu, Zeyuan Liu, Yinghao Chen, Yuxin Zuo, Cheng Qian, Kaiyan Zhang, Weize Chen, Chaojun Xiao, Ganqu Cui, Ning Ding, and Zhiyuan Liu. Justrl: Scaling a 1.5b llm with a simple rl recipe, 2025.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your group-relative advantage is biased, 2026.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework under non-linear rl objectives, 2025b.

Jens Tuyls, Dylan J Foster, Akshay Krishnamurthy, and Jordan T. Ash. Representation-based exploration for language models: From test-time to post-training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms, 2025.

Yiyou Sun, Yuhan Cao, Pohao Huang, Haoyue Bai, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. RL grokking recipe: How does RL unlock and transfer new algorithms in LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025.

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, Philip Torr, and Lei Bai. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning, 2026.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr, 2026.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking RLVR through clipping, entropy, and spurious reward. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025a.

Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, et al. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10), 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024.

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

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

Haotian Luo, Li Shen, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Naiqiang Tan, Xiaochun Cao, and Dacheng Tao. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *arXiv preprint arXiv:2501.12570*, 2025b.

Bin Yu, Hang Yuan, Haotian Li, Xueyin Xu, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen. Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models. *arXiv preprint arXiv:2505.03469*, 2025.

Neel Rajani, Aryo Pradipta Gema, Seraphina Goldfarb-Tarrant, and Ivan Titov. Scalpel vs. hammer: Grpo amplifies existing capabilities, sft replaces them. *arXiv preprint arXiv:2507.10616*, 2025.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Doina Precup, and Mohammad Hamdaqa. RL fine-tuning heals the OOD forgetting in SFT. In *First Workshop on Foundations of Reasoning in Language Models*, 2025.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025b.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Rl squeezes, sft expands: A comparative study of reasoning llms. *arXiv preprint arXiv:2509.21128*, 2025.

Brian Lu, Hongyu Zhao, Shuo Sun, Hao Peng, Rui Ding, and Hongyuan Mei. Generalization of rlvr using causal reasoning as a testbed. *arXiv preprint arXiv:2512.20760*, 2025.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in sft-rl post-training: When high sft scores mislead and what to use instead. *arXiv preprint arXiv:2510.01624*, 2025.

Haitao Jiang, Wenbo Zhang, Jiarui Yao, Hengrui Cai, Sheng Wang, and Rui Song. Supervised fine-tuning versus reinforcement learning: A study of post-training methods for large language models. *arXiv preprint arXiv:2603.13985*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting, 2025c.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026b.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026b.

David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966. Association for Computational Linguistics, July 2019. doi: 10.18653/v1/P19-1092.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench: Reasoning about action, change, and planning. In *AAAI*. AAAI Press, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.findings-acl.824.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023b.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.naacl-long.77.

Harsh Kohli, Sachin Kumar, and Huan Sun. GroundCocoa: A benchmark for evaluating compositional & conditional reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages

8280–8295. Association for Computational Linguistics, April 2025. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.420.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391. Association for Computational Linguistics, October-November 2018. doi: 10.18653/v1/D18-1260.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of Transformers. In *International Conference on Machine Learning*, pages 19660–19722. PMLR, 2024.

Yuan Cheng, Fengzhuo Zhang, Yunlong Hou, Cunxiao Du, Chao Du, Tianyu Pang, Aixin Sun, and Zhuoran Yang. Demystifying the slash pattern in attention: The role of RoPE. *arXiv preprint arXiv:2601.08297*, 2026.

Luc Devroye. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pages 896–904, 1983.

A. Related Works

Reinforcement Learning with Verifiable Rewards has substantially improved the reasoning ability of LLMs, as exemplified by the success of DeepSeek-R1 (Guo and Others, 2025). A central component of this training pipeline is GRPO (Shao et al., 2024), a group-based policy optimization method that normalizes advantages by the within-group standard deviation. Beyond the original DeepSeek-R1 results (Guo and Others, 2025), the effectiveness of GRPO has been reproduced across models of different scales (Zeng et al., 2025; Luo et al., 2025a; He et al., 2025). Subsequent works improve GRPO from several perspectives, including bias correction (Liu et al., 2025; Yang et al., 2026; Yu et al., 2026), sample selection (Xiong et al., 2025b; Yu et al., 2026), reward design (Tuyls et al., 2026; Zhu et al., 2026), and entropy regularization (Wang et al., 2026b). A parallel line of work studies the mechanisms and limits of vanilla GRPO and RLVR. One central question is whether RL-tuned models acquire genuinely new skills or mainly refine existing reasoning trajectories. Yue et al. (2026) argues that RLVR sharpens correct reasoning trajectories on math and coding tasks without expanding the model's skill boundary. In contrast, Wen et al. (2025) show that, under CoT-Pass@K, which evaluates both final answers and intermediate reasoning steps, RLVR can extend the reasoning ability of the base model. Relatedly, Sun et al. (2026) find that RLVR-trained models can solve tasks requiring recombination of learned skills but remain weak on transformative tasks, while Liu et al. (2026) show that prolonged RL training can uncover reasoning strategies inaccessible to the base model. Other works investigate the generalization of RLVR beyond the training topics, with evidence that RL-tuned models can transfer learned abilities to unseen domains (Huan et al., 2026; Fu et al., 2025; Chu et al., 2025). Finally, Tan et al. (2026) studies the scaling behavior of GRPO on math tasks, while Cui et al. (2025); Shao et al. (2026); Chen et al. (2026a) show that reasoning performance can improve even without verifiable rewards, for example, through random rewards or entropy minimization under suitable conditions. See Wang et al. (2026a); Li et al. (2025a); Xu et al. (2025) for the comprehensive surveys.

Long-Reasoning SFT has recently emerged as a simple and effective paradigm for transferring explicit reasoning procedures to LLMs via supervised learning on long chain-of-thought trajectories. Early rationale-supervision methods such as STaR bootstrap reasoning by generating rationales, filtering correct ones, and fine-tuning on the retained traces (Zelikman et al., 2022). Instruction-tuning and explanation-distillation studies further show that chain-of-thought demonstrations or teacher-generated explanations improve reasoning, as in FLAN-CoT (Chung et al., 2024), Orca (Mukherjee et al., 2023), and Orca 2 (Mitra et al., 2023). In mathematical reasoning, specialized SFT datasets with intermediate derivations or program-aided traces have been developed by MAmmoTH (Yue et al., 2023), ToRA (Gou et al., 2023), and NuminaMath (Li et al., 2024). More recent o1-style studies show that long reasoning can be elicited from small but carefully curated SFT datasets: s1 uses 1,000 examples (Muennighoff et al., 2025), LIMO reports strong reasoning from 817 examples (Ye et al., 2025), and Li et al. (2025b) transfer long-chain reasoning through data-efficient SFT or LoRA on long-CoT demonstrations. Recent work also studies efficiency and length control, including O1-Pruner for reducing redundant reasoning (Luo et al., 2025b) and long-short mixture SFT for mitigating overthinking (Yu et al., 2025). Together, these works show that long-reasoning SFT effectively transfers reasoning behaviors from demonstrations, while its success depends on trace quality, reasoning structure, data curation, and inference-length control.

Understanding SFT and RL has been an active topic, as both SFT and RL can improve model reasoning but appear to do so through different mechanisms. A central finding is that SFT often imitates or memorizes supervised demonstrations, whereas RL can better elicit generalizable behavior (Chu et al., 2025; Chen et al., 2025a; Shenfeld et al., 2025; Huan et al., 2025). In particular, Huan et al. (2025) compare models trained on math tasks and evaluated on other reasoning tasks, showing that SFT is less effective than RL in improving general reasoning ability. Mechanistically, GRPO appears to amplify existing capabilities with milder parameter changes, while SFT can overwrite prior skills and degrade out-of-domain performance (Rajani et al., 2025). This distinction is refined by works showing that RL can recover out-of-distribution capability forgotten during SFT (Jin et al., 2025), mitigate forgetting through on-policy, mode-seeking data (Chen et al., 2025b), and reshape reasoning trajectories by compressing incorrect paths rather than homogenizing correct ones (Matsutani et al., 2025). Other studies qualify the advantage of RL, attributing it partly to implicit medium-difficulty data filtering (Lu et al., 2026), showing its dependence on model scale and task difficulty (Lu et al., 2025), and questioning whether RLVR creates new reasoning patterns beyond those already present in the base model (Yue et al., 2026). High SFT scores may also poorly predict post-RL gains (Kang et al., 2025), while broader surveys organize SFT, RL, and hybrid post-training methods under a unified comparison framework (Jiang et al., 2026). More recently, on-policy distillation has emerged as an SFT-style alternative that uses on-policy data to mitigate the forgetting issues of conventional SFT while improving sample and compute efficiency relative to RL (Lu and Thinking Machines Lab, 2025; Chen et al., 2025c). Subsequent self-distillation methods further allow the same model to act as both teacher and student (Zhao et al., 2026b; Hübotter et al., 2026).

Table 2. Hyperparameters for RL training with `verl.trainer.main_ppo`.

Hyperparameter	Value
Training batch size	128
PPO mini-batch size	64
PPO micro-batch size per GPU	1
Prompt length	1024
Response length	8192
Rollout sample number (n)	8
Learning rate	10^{-6}
Clip ratio high	0.28
Use KL loss	False
Use KL in reward	False
Entropy coefficient	0
Temperature	1.0
Top-p	1.0
Enable chunked prefill	False
Enforce eager execution	False
Free cache engine	False
Use remove padding	True
Enable gradient checkpointing	True
Ulysses sequence parallel size	1
Actor parameter offload	True
Actor optimizer offload	True
Reference parameter offload	True
Critic warmup	0

B. More Experimental Details

B.1. RL Training Details

As the original DAPO dataset duplicates each prompt 100 times due to reproduction concerns, we pre-process the dataset and obtain 17391 unique prompts. We add the following prompt suffix to the input question, as per the standard procedure in the literature (Liu et al., 2025; Chen et al., 2026b) to encourage reasoning:

"Please reason step by step, and put your final answer within \\boxed{}."

The parameter used in RL training is listed in Table 2.

B.2. SFT Training Details

B.2.1. General SFT Training Configurations

The parameter used in SFT training is listed in Table 3.

B.2.2. Details of the SFT Training Dataset Size

In Section 3.2, we match the SFT training dataset size with those during RL training. Specifically,

- When the number of solvable prompts is matched, the corresponding Qwen3-14B dataset contains 9963 prompt–rollout pairs; and the RL-tuned dataset contains 10020 prompt–rollout pairs.
- When the number of correct rollouts per prompt in the SFT training dataset is matched to that observed during RL training, both Qwen3-14B dataset and the RL-tuned dataset consist of 39623 prompt-rollout pairs, respectively.
- When the total number of rollouts for each prompt is matched, each of the two datasets contains 39623 correct training samples with 98641 prompt-wrong rollout pairs.

Table 3. Hyperparameters for supervised fine-tuning (SFT).

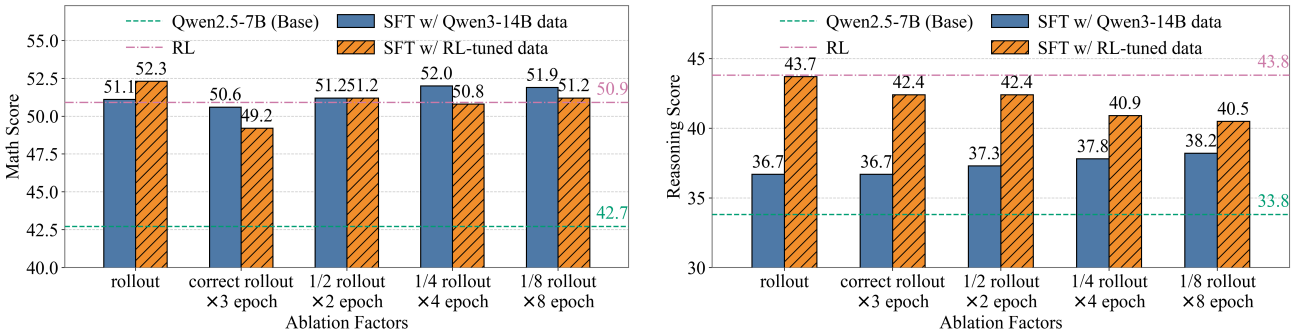
Hyperparameter	Value
DeepSpeed config	`deepspeed_z3`
Template	`qwen`
Cutoff length	16000
Per-device train batch size	1
Gradient accumulation steps	8
Learning rate	1.0×10^{-5}
Number of training epochs	1.0
Learning rate scheduler	`cosine`
Warmup ratio	0.1
BF16	True

B.3. Evaluation Setup Details

For the math ability evaluation, we adopt Avg@32 for AIME24 due to the small sample size, and Pass@1 for the rest of the tasks. To evaluate the generalization ability of the trained models, we evaluate on the following reasoning benchmarks, which are accessible via Gao et al. (2024):

- GPQA-Diamond (Rein et al., 2024): a challenging dataset of 448 high-quality and difficult multiple-choice questions written by domain experts in biology, physics, and chemistry.
- HEAD-QA (Vilares and Gómez-Rodríguez, 2019): a healthcare dataset for complex reasoning, which contains questions covering medicine, nursing, psychology, chemistry, pharmacology and biology.
- ACPBench (Kokel et al., 2025): a benchmark for evaluating the reasoning tasks in the field of planning, consisting of 7 reasoning tasks over 13 planning domains.
- ARC (Clark et al., 2018): it consists of 7787 science exam questions drawn from a variety of sources, including science questions provided under license by a research partner affiliated with AI2. The questions are sorted into a Challenge set of "hard" problems and a Easy set of "easy" problems. We adopt the challenge set.
- BIG-Bench Hard (BBH) (Suzgun et al., 2023): a suit of 23 challenging BIG-Bench tasks, where the language models usually perform worse than human-raters.
- C-Eval (Huang et al., 2023b): a comprehensive Chinese evaluation suite, targeting at the knowledge and reasoning abilities of LLMs within the context of Chinese language and culture.
- COPAL-ID (Wibowo et al., 2024): an Indonesian causal commonsense reasoning dataset that captures local nuances.
- GroundCocoa (Kohli et al., 2025): a lexically diverse benchmark connecting compositional and conditional reasoning skills to the real-world problem of flight booking.
- MMLU-Pro (Wang et al., 2024): an enhanced dataset which incorporates more challenging, reasoning-focused questions than the well-known MMLU dataset, and expands the choice set from four to ten options.
- MMLU-Pro+ (Asgari et al., 2024): an enchanced benchmark building upon MMLU-Pro, which is used to assess shortcut learning and higher-order reasoning in LLMs.
- OpenBookQA (Mihaylov et al., 2018): a question-answering dataset which models open book exams and requires external knowledge and reasoning.
- PIQA (Bisk et al., 2020): a physical commonsense reasoning and a corresponding benchmark dataset.

As post-training usually is done with templates, we apply the default chat template in Gao et al. (2024) during evaluation.



(a) Math benchmarks

(b) General reasoning benchmarks

Figure 6. Ablation study on the components of the training dataset with the training compute budget fixed. When the training dataset shrinks, the performance also downgrades.

B.4. More Evaluation Results for the Data Source Experiment

We provide more detailed evaluation results on the general reasoning tasks of the models trained via mixed dataset in Section 3.1 in Table.

Benchmark	Qwen2.5-7B	RL	SFT w/ RL-tuned	Mix model 1	Mix model 2	Mix model 3	Mix model 4	Mix model 5	SFT w/ Qwen3-14B
GPQA-Diamond	22.7	30.3	21.7	23.7	25.3	21.7	**28.8**	27.3	27.3
HEAD-QA	33.6	33.6	32.9	**33.3**	33.1	33.0	32.1	31.6	31.4
ARC	42.1	42.0	40.3	40.8	41.6	**42.9**	40.3	40.0	37.8
ACPBench	23.3	50.1	**58.0**	56.6	53.4	50.7	43.7	46.6	45.3
BIG-Bench Hard	10.9	42.9	33.8	34.4	37.3	35.6	34.1	37.7	**42.1**
C-Eval	40.1	43.6	42.4	36.4	47.8	28.2	52.9	**53.8**	36.5
COPAL-ID	61.7	60.8	60.1	59.7	60.3	**61.4**	59.0	59.4	59.0
GroundCocoa	34.7	34.8	35.6	36.8	**37.0**	36.7	34.6	34.2	33.2
MMLU-Pro	9.2	45.7	19.4	17.9	**21.3**	3.8	0.7	1.2	1.1
MMLU-Pro+	21.0	31.8	13.8	**15.1**	**15.1**	2.5	1.6	1.1	0.8
OpenBookQA	28.6	33.2	30.6	**32.8**	32.4	32.2	30.2	30.0	27.2
PIQA	77.1	76.1	76.2	76.3	76.7	76.5	77.3	77.1	**77.5**
Average	33.8	43.8	38.7	38.7	**40.1**	35.4	36.3	36.7	34.9

Table 4. Performance Comparison Across General Reasoning Benchmarks

B.5. Ablation Study on the Dataset Scale with Fixed Compute

In Section 3.2, we have studied the effect of the SFT training data scale on the generalization ability of the SFT-tuned model. The training compute also scales as more training data is introduced. To dig out the true workhorse behind data scale and compute, we carry out an ablation study.

Specifically, we fix the total compute used for SFT and vary the composition of the training data. In the first ablation study, we train the model with SFT using all 39623 prompt-correct rollout pairs for 3 epochs. In this setting, the total compute is approximately aligned (118869 vs. 138264), while the number of unique training samples differs. In the second set of experiments, we randomly sample $1/k$ fraction of the 8 rollouts for each prompt, and train the base models via SFT for k epochs for $k = 2, 4, 8$. In this case, the compute is aligned, but less training data is used.

The experimental results are presented in Figure 6. It indicates using fewer training samples but with the same compute as in RL training, the final performance on general reasoning tasks decreases compared to that of using the whole dataset.

We conclude that the training data scale, i.e., the number of unique rollouts, matters more than compute for SFT training given the same gradient compute budget as RL training.

B.6. Rollout Analysis

We list the classified 138 consolidated methods in Table 5.

C. Notations and Outline of Proofs

Notations	Descriptions
$P_{\text{in}} = (X_1^\top, X_2^\top, \ldots, X_N^\top)^\top$	The input informative and context tokens.
$S_n = \texttt{soft}(X_q W P_{\text{in}}^\top)_n$	The attention score from query X_q to token X_n at position n.
$S_{l,k} = \sum_{n=1}^N \mathbb{1}\{X_n = i_{l,k}\} S_n$	The attention score from query u_l to informative token $i_{l,k}$.
$\widetilde{S}_m = \sum_{n=1}^N \mathbb{1}\{X_n = c_j\} S_n$	The attention score from query X_q to context token c_j.
$\widetilde{S} = \sum_{j=1}^J \widetilde{S}_m$	The attention score from query X_q to all context token.
$B_{l,k} = u_l W i_{l,k}^\top$	The attention logit from query u_l to informative token $i_{l,k}$.
$\widetilde{B}_{l,j} = u_l W c_j^\top$	The attention logit from query u_l to context token c_j.
$\beta_{l,k} = -u_l \nabla_W L_{\text{SFT}}\, i_{l,k}^\top$	Update of logit from query u_l to informative token $i_{l,k}$.
$\widetilde{\beta}_{l,j} = -u_l \nabla_W L_{\text{SFT}}\, c_j^\top$	Update of attention logit from query u_l to context token c_j.
$\mathcal{E}_{\text{P}}^*$ (Lemma D.7)	High probability event for token concentration.
$A_l = \sum_{k=1}^K a_{(l-1)K+k}$	Class-specific correct-answer indicator vector.
$A_{-l} = \sum_{l' \neq l} \sum_{k'=1}^K a_{l',k'}$	Class-specific incorrect-answer indicator vector.
$\Phi_l = F A_l^\top$	Probability assigned to correct answers of class l
$\Phi_{-l} = F A_{-l}^\top$	Probability assigned to incorrect answers of class l

Table 6. Summary of frequently used notations in proof. We omit (t) here for simplicity.

Notations. We first summarize the frequently used notations in Table 6. In addition, the pre-trained knowledge

$$W_{\text{lm}} = c_{\text{lm}} \log(K) \Big(\sum_{l=1}^L \sum_{k=1}^K i_{l,k}^\top (a_{l,k} - A_{-l} - \widetilde{a}) + \sum_{j=1}^J c_j^\top \widetilde{a} \Big),$$

where $A_{-l} = \sum_{l' \neq l} \sum_{k'=1}^K a_{l',k'} \in \mathbb{R}^{LK+1}$. In addition, we have for the output,

$$F = \texttt{soft}(c_{\text{lm}} \log(K) (\sum_{k=1}^K S_{l,k} a_{l,k}^\top + (2\widetilde{S} - 1)\widetilde{a}^\top - (1 - \widetilde{S}) A_{-l}).$$

Roadmap of Proof Part. We prove the SFT part and RL part of Theorem 4.2 in Appendix D and Appendix E, respectively, and prove Theorem 4.3 in Appendix F. We next summarize the outline of the proof of the SFT part of Theorem 4.2, the proof of the RL part is similar.

D. Proofs of SFT

Proof Outline of SFT. Note that the output F is directly determined by attention scores S, and S itself depends on the attention logits B. Accordingly, we first formalize the key quantities, the attention logits and their update rules, in Definition D.2, and analyze their evolution throughout the optimization process. We first calculate the gradient in Lemma D.1 and characterize the update of logits in Lemma D.3. Building on these foundations, we characterize the update of one step gradient descent in Appendix D.2, which proves the SFT part of Theorem 4.2.

Table 5. Alphabetical Classification of Mathematical Methods

Method	Method	Method
3D geometry & volumes	equation solving	partial fractions
absolute value methods	estimation & bounding	partitions
algebraic manipulation	Euler's formula	pattern recognition
angle properties	expected value	percentages
area & perimeter	exponents & powers	perfect powers
arithmetic operations	extremal principle	periodicity
asymptotic analysis	factoring	permutations
averages	Fermat's & Euler's theorems	pigeonhole principle
bijection & double counting	Fibonacci & special sequences	polygons
binomial expansion	floor & ceiling functions	polynomial methods
calculation & computation	fractions & ratios	prime numbers
calculus (general)	function analysis	probability
case analysis	functional equations	problem reformulation
centroid & triangle centers	game theory & strategy	proof by contradiction
characteristic equation	GCD & LCM	proof by induction
Chinese Remainder Theorem	generating functions	Pythagorean theorem
circles & circle theorems	geometry (general)	quadratic methods
classical geometry theorems	graph theory	random walks & Markov chains
coefficient comparison	greedy algorithms	range & domain analysis
coloring & tiling arguments	grid & lattice methods	rational root theorem
combinations & binomial coefficients	group theory	rationalization
combinatorial identities	inclusion-exclusion	recurrence relations
combinatorics (general)	inequalities	rings & fields
completing the square	integer methods	root analysis
complex numbers	integration	search & traversal
conic sections	interval analysis	sequence analysis
constraint analysis	invariants	series & convergence
construction	iterative methods	set theory
coordinate geometry	known results & formulas	similarity & congruence
counting (general)	limits	simplification & expansion
counting principles	linear algebra	simulation & enumeration
critical points & optimization (calculus)	lines & slopes	sorting
degree analysis	logarithms	squaring both sides
derivatives	logical deduction	statistics
determinants	magnitude & norm	substitution
difference of squares	matrices	summation techniques
differential equations	midpoint methods	symmetry
digits & number representation	modular arithmetic	systems of equations
Diophantine equations	multivariable calculus	transformations
discriminant analysis	number theory (general)	triangle geometry
distance formulas	number-theoretic functions	trigonometric identities
distributions	optimization	trigonometry
divisibility & divisors	other	unit conversion
dot product & cross product	p-adic valuations	vectors
dynamic programming	parametric methods	verification & checking
eigenvalues & eigenvectors	parity arguments	Vieta's formulas

D.1. Gradient Computations

In this section, we first calculate the gradient with respect to W_{SFT} in Lemma D.1 and characterize the key variables: logits and their updates in Definition D.2 and Lemma D.3. We omit (t) and the subscript SFT in this section when there is no ambiguity and write $L_{\mathrm{SFT}}(W_{\mathrm{SFT}})$ as L_{SFT} here for simplicity.

Lemma D.1 (SFT Gradient)**.** *Recall that $P_{\mathrm{in}} = (X_1^\top, X_2^\top, \ldots, X_N^\top)^\top$, the gradient of the SFT loss function with respect to W is given by*

$$\nabla_W L_{\mathrm{SFT}} = \mathbb{E}_{P_{\mathrm{in}}, X_q, A \sim \mathcal{D}_{\mathrm{SFT}}} \Big[X_q \Big((F - A) W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \big(\mathrm{diag}(S) - S^\top S \big) \Big) P_{\mathrm{in}} \Big],$$

where $S = \mathtt{soft}(X_q^\top W P_{\mathrm{in}}^\top) \in \mathbb{R}^{N \times 1}$ is the attention score, and $F = \mathtt{soft}(S P_{\mathrm{in}} W_{\mathrm{lm}}) \in \mathbb{R}^{1 \times (LK+1)}$ is the output.

Proof. We first define the following intermediate variables for a single sample $(P_{\mathrm{in}}, X_q, A)$,

$$\begin{aligned} s &= X_q^\top W P_{\mathrm{in}}^\top \in \mathbb{R}^{1 \times N}, \\ S &= \mathtt{soft}(s) \in \mathbb{R}^{1 \times N}, \\ z &= S P W_{\mathrm{lm}} \in \mathbb{R}^{1 \times (LK+1)}, \\ F &= \mathtt{soft}(z) \in \mathbb{R}^{1 \times (LK+1)}. \end{aligned}$$

Here N is the prompt length, $P_{\mathrm{in}} \in \mathbb{R}^{N \times d}$, $X_q \in \mathbb{R}^{d \times 1}$, $W \in \mathbb{R}^{d \times d}$, and $W_{\mathrm{lm}} \in \mathbb{R}^{d \times (LK+1)}$.

And the single-sample loss is

$$\ell_{\mathrm{SFT}}(W; P_{\mathrm{in}}, X_q, A) = -\sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{1}\{A = a_{l,k}\} \log F_{(l-1)K+k}.$$

Then

$$L_{\mathrm{SFT}}(W) = \mathbb{E}_{P_{\mathrm{in}}, X_q, A \sim \mathcal{D}_{\mathrm{SFT}}} \big[\ell_{\mathrm{SFT}}(W; P_{\mathrm{in}}, X_q, A) \big].$$

We solve the gradient as follows. We first differentiate with respect to the output logits z. Using the standard cross-entropy gradient for softmax, we have

$$g_z = \nabla_z \ell_{\mathrm{SFT}} = \sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{1}\{A = a_{l,k}\} (F - e_{(l-1)K+k}) \in \mathbb{R}^{1 \times (LK+1)}.$$

Then we differentiate with respect to the inner softmax output S. Since $z = S P W_{\mathrm{lm}}$, we have

$$g_S = \nabla_S \ell_{\mathrm{SFT}} = g_z W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \in \mathbb{R}^{1 \times N}.$$

Then we derive through the inner softmax. Since $S = \mathtt{soft}(s)$, the Jacobian of the row-softmax is

$$J_{\mathtt{soft}}(s) = \mathrm{diag}(S) - S^\top S \in \mathbb{R}^{N \times N}.$$

Therefore,

$$\begin{aligned} g_s = \nabla_s \ell_{\mathrm{SFT}} &= g_S J_{\mathtt{soft}}(s) \\ &= g_S \big(\mathrm{diag}(S) - S^\top S \big) \\ &= \sum_{l=1}^{L} \sum_{k=1}^{K} \mathbb{1}\{A = a_{l,k}\} (F - e_{(l-1)K+k}) W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \big(\mathrm{diag}(S) - S^\top S \big) \in \mathbb{R}^{1 \times N}. \end{aligned}$$

We finally derive with respect to W. Recall $s = X_q^\top W P_{\mathrm{in}}^\top$, and its differential is $ds = X_q^\top \, dW \, P_{\mathrm{in}}^\top$. Hence

$$d\ell_{\mathrm{SFT}} = g_s \, ds$$

$$= g_s X_q^\top \, dW \, P_{\text{in}}^\top$$
$$= \operatorname{tr}\big(P_{\text{in}}^\top g_s X_q^\top \, dW\big) = \operatorname{tr}\big((X_q g_s P)^\top dW\big).$$

Therefore,

$$\nabla_W \ell_{\text{SFT}} = X_q \, g_s \, P \in \mathbb{R}^{d\times d}.$$

Substituting the expression for g_s, we obtain

$$\nabla_W \ell_{\text{SFT}}$$
$$= \sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{1}\{A = a_{l,k}\} \, X_q \Big[(F - e_{(l-1)K+k}) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big] P_{\text{in}}, \; S = \texttt{soft}(X_q^\top W P_{\text{in}}^\top).$$

So final expression for $\nabla_W L_{\text{SFT}}$ can be obtained by taking expectation as follows

$$\nabla_W L_{\text{SFT}}(W)$$
$$= \mathbb{E}_{P_{\text{in}}, X_q, A\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{l=1}^{L}\sum_{k=1}^{K} \mathbb{1}\{A = a_{l,k}\} \, X_q \Big((F - e_{(l-1)K+k}) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P_{\text{in}}\right]$$
$$= \mathbb{E}_{P_{\text{in}}, X_q, A\sim\mathcal{D}_{\text{SFT}}}\left[X_q \Big((F - A) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P_{\text{in}}\right],$$

where

$$S = \texttt{soft}(X_q^\top W P_{\text{in}}^\top), \qquad F = \texttt{soft}(S P_{\text{in}} W_{\text{lm}}).$$

We finish the proof of SFT gradient. □

Suppose the question type is classified as l. We next define the important logits during training. Since only informative tokens from $\mathcal{I}_l$ and context tokens from $\mathcal{C}$ will occur in the prompt of question class l, we only need to consider the following variables.

Definition D.2 (Important Logits and Their Update Rate in SFT)**.** For any $l \in [L]$, $k \in [K]$ and $j \in [J]$, for iteration $t \geq 0$, we denote the weight as $W^{(t)}$ and the loss as $L_{\text{SFT}}^{(t)} = L_{\text{SFT}}(W^{(t)})$. We define the following quantities

$$B_{l,k}^{(t)} = u_l W^{(t)} i_{l,k}^\top \qquad \beta_{l,k}^{(t)} = -u_l \nabla_W L_{\text{SFT}}^{(t)} \, i_{l,k}^\top;$$
$$\widetilde{B}_{l,j}^{(t)} = u_l W^{(t)} c_j^\top \qquad \widetilde{\beta}_{l,j}^{(t)} = -u_l \nabla_W L_{\text{SFT}}^{(t)} \, c_j^\top.$$

By the GD update, we have

$$B_{l,k}^{(t+1)} = B_{l,k}^{(t)} + \eta \beta_{l,k}^{(t)}$$
$$\widetilde{B}_{l,j}^{(t+1)} = \widetilde{B}_{l,j}^{(t)} + \eta \widetilde{\beta}_{l,j}^{(t)}$$

Moreover, by our initialization of $W^{(0)} = \mathbf{0}_{d\times d}$, we have $B_{l,k}^{(0)} = \widetilde{B}_{l,j}^{(0)} = 0$.

In the remainder of the proof, for simplicity, we omit the dependence on the iteration t when it is clear from the context.

Lemma D.3 (Logits Update Rate)**.** *For any $l \in [L]$, $k \in [K]$ and $j \in [J]$, the update rate of logits can be written as*

$$\beta_{l,k} = -\frac{c_{\text{lm}} \log(K)}{L} \mathbb{E}_{P_{\text{in}}}\Big[S_{l,k}\Big((F_{(l-1)K+k} - p_{l,k}) - \sum_{k'=1}^{K} S_{l,k'}(F_{(l-1)K+k'} - p_{l,k'})$$
$$- (2F_{LK+1} + \Phi_{-l})\widetilde{S}\Big)\Big| X_q = u_l\Big],$$
$$\widetilde{\beta}_{l,j} = -\frac{c_{\text{lm}} \log(K)}{L} \mathbb{E}_{P_{\text{in}}}\Big[\widetilde{S}_j\Big((2F_{LK+1} + \Phi_{-l})(1 - \widetilde{S}) - \sum_{k'=1}^{K} S_{l,k'}(F_{(l-1)K+k'} - p_{l,k'})\Big)\Big| X_q = u_l\Big].$$

Proof. We first consider simplifying the gradients as follows.

$$P_{\text{in}}^\top \operatorname{diag}(S) P_{\text{in}} = \sum_{n=1}^{N} S_n X_n X_n^\top,$$

and

$$P_{\text{in}}^\top S^\top S P = \sum_{n=1}^{N}\sum_{m=1}^{N} S_n S_m \, X_n X_m^\top = \left(\sum_{n=1}^{N} S_n X_n\right)\left(\sum_{m=1}^{N} S_m X_m\right)^\top.$$

Hence, we have

$$\begin{aligned} P_{\text{in}}^\top (\operatorname{diag}(S) - S^\top S) P_{\text{in}} &= \sum_{n=1}^{N} S_n X_n X_n^\top - \sum_{n=1}^{N}\sum_{m=1}^{N} S_n S_m \, X_n X_m^\top \\ &= \sum_{n=1}^{N} S_n \left(X_n - \sum_{m=1}^{N} S_m X_m\right) X_n^\top. \end{aligned}$$

Then plugging in $W_{\text{lm}} = c_{\text{lm}} \log(K)\big(\sum_{l=1}^{L}\sum_{k=1}^{K} i_{l,k}^\top (a_{l,k} - A_{-l} - \widetilde{a}) + \sum_{j=1}^{J} c_j^\top \widetilde{a}\big) \in \mathbb{R}^{d\times(LK+1)}$, we have

$$\begin{aligned} & W_{\text{lm}}^\top P_{\text{in}}^\top (\operatorname{diag}(S) - S^\top S) P_{\text{in}} \\ &\quad = c_{\text{lm}} \log(K)\big(\sum_{l=1}^{L}\sum_{k=1}^{K} S_{l,k}(a_{l,k} - A_{-l} - \widetilde{a})^\top i_{l,k} + \tilde{a}^\top \sum_{n=1}^{N} \mathbb{1}\{X_n \in \mathcal{C}\} S_n X_n\big) \\ &\qquad - c_{\text{lm}} \log(K)\big(\sum_{l=1}^{L}\sum_{k=1}^{K} S_{l,k}(a_{l,k} - A_{-l} - \widetilde{a})^\top + \widetilde{S}\tilde{a}^\top\big)\big(\sum_{n=1}^{N} S_n X_n\big). \end{aligned} \tag{1}$$

Then, we have

$$\begin{aligned} \beta_{l,k} &= -u_l \nabla_W L_{\text{SFT}} \, i_{l,k}^\top \\ &= -\mathbb{E}_{P_{\text{in}}, X_q, A\sim\mathcal{D}_{\text{SFT}}}\Big[\, \mathbb{1}\{X_q = u_l\}\Big((F-A) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P i_{l,k}^\top\Big] \\ &= -\frac{1}{L}\mathbb{E}_{P_{\text{in}}, A\sim\mathcal{D}_{\text{SFT}}}\Big[\Big((F-A) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P i_{l,k}^\top \Big| X_q = u_l\Big], \end{aligned}$$

where the last equation follows from that $\mathbb{P}(X_q = u_l) = 1/L$ for all $l \in [L]$. Using Equation (1), we have

$$\begin{aligned} & W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big) P_{\text{in}} \, i_{l,k}^\top \\ &\quad = c_{\text{lm}} \log(K)(S_{l,k}(a_{l,k} - A_{-l} - \widetilde{a})^\top - S_{l,k} \sum_{l'=1}^{L}\sum_{k'=1}^{K} S_{l',k'}(a_{l',k'} - A_{-l'} - \widetilde{a})^\top - S_{l,k}\tilde{S}\,\tilde{a}^\top) \\ &\quad = c_{\text{lm}} \log(K) S_{l,k} \left((a_{l,k} - A_{-l} - \widetilde{a})^\top - \sum_{l'=1}^{L}\sum_{k'=1}^{K} S_{l',k'}(a_{l',k'} - A_{-l'} - \widetilde{a})^\top - \tilde{S}\,\tilde{a}^\top\right). \end{aligned}$$

By definition, when $X_q = u_l$, we have $S_{l',k'} = 0$ for all $l' \neq l$, since question class l contains informative tokens only from $\mathcal{I}_l$. Therefore, we have

$$\begin{aligned} & \beta_{l,k} \\ &= -\frac{c_{\text{lm}} \log(K)}{L}\mathbb{E}_{P_{\text{in}}, A\sim\mathcal{D}_{\text{SFT}}}\Big[(F-A)\, S_{l,k} \\ &\qquad\qquad \left((a_{l,k} - A_{-l} - \widetilde{a})^\top - \sum_{k'=1}^{K} S_{l,k'}(a_{l,k'} - A_{-l} - \widetilde{a})^\top - \tilde{S}\,\tilde{a}^\top\right) \Big| X_q = u_l\Big] \\ &\overset{(i)}{=} -\frac{c_{\text{lm}} \log(K)}{L}\mathbb{E}_{P_{\text{in}}}\Big[S_{l,k} \end{aligned}$$

$$\mathbb{E}_{A\sim\mathcal{D}_{\mathrm{SFT}}}\left[(F-A)\left(a_{l,k}^\top-\sum_{k'=1}^{K}S_{l,k'}a_{l,k'}^\top-\widetilde{S}\,(2\tilde{a}+A_{-l})^\top\right)\Big|P_{\mathrm{in}}\right]\Big|X_q=u_l\Big]$$

$$\overset{(ii)}{=}-\frac{c_{\mathrm{lm}}\log(K)}{L}\mathbb{E}_{P_{\mathrm{in}}}\Big[S_{l,k}$$

$$\Big((F_{(l-1)K+k}-p_{l,k})-\sum_{k'=1}^{K}S_{l,k'}(F_{(l-1)K+k'}-p_{l,k'})-(2F_{LK+1}+\Phi_{-l})\widetilde{S}\Big)\Big|X_q=u_l\Big],$$

where (i) follows from applying the tower property, and (ii) follows from taking expectation on A. Using Equation (1), we have

$$\begin{aligned}
&W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top\big(\mathrm{diag}(S)-S^\top S\big)P_{\mathrm{in}}\,c_j^\top\\
&\quad=c_{\mathrm{lm}}\log(K)(\widetilde{S}_j\tilde{a}^\top-\widetilde{S}_j\sum_{l'=1}^{L}\sum_{k'=1}^{K}S_{l',k'}(a_{l',k'}-A_{-l'}-\widetilde{a})^\top-\widetilde{S}_j\tilde{S}\,\tilde{a}^\top)\\
&\quad=c_{\mathrm{lm}}\log(K)\widetilde{S}_j\big(\tilde{a}^\top-\sum_{l'=1}^{L}\sum_{k'=1}^{K}S_{l',k'}(a_{l',k'}-A_{-l'}-\widetilde{a})^\top-\tilde{S}\,\tilde{a}^\top\big).
\end{aligned}$$

Then similar to $\beta_{l,k}$, we have

$$\begin{aligned}
\widetilde{\beta}_{l,j}&=-u_l\nabla_W L_{\mathrm{SFT}}\,c_j^\top\\
&=-\frac{c_{\mathrm{lm}}\log(K)}{L}\mathbb{E}_{P_{\mathrm{in}}}\Big[\widetilde{S}_j\Big((2F_{LK+1}+\Phi_{-l})(1-\widetilde{S})-\sum_{k'=1}^{K}S_{l,k'}(F_{(l-1)K+k'}-p_{l,k'})\Big)\Big|X_q=u_l\Big].
\end{aligned}$$

We finish the proof of the logit update characterization. □

D.2. Learning Dynamics of SFT

With the gradient expressions derived in Appendix D.1, we analyze the training dynamics of SFT in this section. For each question class $l\in[L]$, we fix l and condition on the event $\{X_q=u_l\}$, and then study the corresponding training dynamics. It is worth noting that, when $X_q=u_l$, we have $S_{l',k'}=0$ for all $l'\neq l$, since question class l contains informative tokens only from $\mathcal{I}_l$. For simplicity, we consider only one step. We take a relatively large learning rate

$$\eta_{\mathrm{SFT}}^{(0)}=\frac{KLc^{(\mathrm{I})}}{p_{l,1}},$$

where $c^{(\mathrm{I})}>0$ is a sufficiently large constant and $p_{l,1}$ denotes the probability of the head answer under the SFT teacher distribution. Since $p_{l,1}=\Omega(1)\gg L/K$, the head-answer token receives the dominant update. Tail answers receive smaller positive updates.

After one step update, the attention score concentrates on the informative token corresponding to the correct answer, namely $i_{l,1},\cdots,i_{l,K}$.

Lemma D.4 (Initialization Variable)**.** *At initialization $t=0$, for all $k\in[K]$ and $j\in[J]$, the key variables satisfy the following estimates.*

- ***Attention Logits.***

$$B_{l,k}^{(0)}=\widetilde{B}_{l,j}^{(0)}=0.$$

- ***Attention Scores.*** *Suppose $P_{\mathrm{in}}\in\mathcal{E}_{\mathrm{P}}^*$ satisfies the high-probability event in Lemma D.7. Then, we have the attention scores of the initial step as follows.*

$$S_{l,k}^{(0)}=\frac{1}{2K}\left(1+O\left(\frac{(K+J)^3}{N}\right)\right),$$

$$\widetilde{S}_j^{(0)} = \frac{1}{2J}\left(1 + O\left(\frac{(K+J)^3}{N}\right)\right),$$
$$\widetilde{S}^{(0)} = \frac{1}{2}\left(1 + O\left(\frac{(K+J)^3}{N}\right)\right). \tag{2}$$

- ***Outputs.*** *Suppose $P_{\text{in}} \in \mathcal{E}_{\text{P}}^*$ satisfies the high-probability event in Lemma D.7. Recall the weight of the language model head is*

$$W_{\text{lm}} = 2\log(K)\left(\sum_{l'=1}^{L}\sum_{k'=1}^{K} i_{l',k'}^\top (a_{l',k'} - A_{-l'} - \widetilde{a}) + \sum_{j=1}^{J} c_j^\top \widetilde{a}\right).$$

Then, for any $l' \neq l$, the output is as follows.

$$F_{(l-1)K+k}^{(0)} = \frac{1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right), \tag{3}$$
$$F_{LK+1}^{(0)} = \frac{1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right), \tag{4}$$
$$F_{(l'-1)K+k}^{(0)} = \frac{1}{K(K+L)}\left(1 + O\left(\frac{\log K}{K}\right)\right). \tag{5}$$

Consequently, we have that

$$\Phi_l^{(0)} = \frac{K}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right), \qquad \Phi_{-l}^{(0)} = \frac{L-1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right).$$

- ***Attention Logits Update.*** *For every $k \in [K]$, the logit update of the informative token is as follows*

$$\beta_{l,k}^{(0)} = \frac{\log K}{LK}\left[p_{l,k} + \frac{L-1}{2(K+L)} - \frac{L}{2K(K+L)} + o\left(\frac{1}{K}\right)\right].$$

Moreover, for the context token, we have

$$\widetilde{\beta}_{l,j}^{(0)} = -\frac{\log K}{LJ}\left[\frac{L+1}{2(K+L)} + \frac{L}{2K(K+L)} + o\left(\frac{1}{K}\right)\right] \leq 0.$$

Proof. The attention-logit result follows directly from the initialization $W^{(0)} = \mathbf{0}_{d\times d}$. The attention-score estimates follow from Lemma D.7.

We next compute the output probabilities. Conditional on $X_q = u_l$, the logits of output before softmax are

$$2\log K\left(\sum_{k=1}^{K} S_{l,k} a_{l,k}^\top + (2\widetilde{S} - 1)\widetilde{a}^\top - (1 - \widetilde{S})A_{-l}\right),$$

where $A_{-l} = \sum_{l' \neq l}\sum_{k'=1}^{K} a_{l',k'}$. From the attention scores result, at initialization,

$$S_{l,k}^{(0)} = \frac{1}{2K}(1 + o(1)), \qquad \widetilde{S}^{(0)} = \frac{1}{2}(1 + o(1)).$$

Therefore,

$$K^{2S_{l,k}^{(0)}} = 1 + O\left(\frac{\log K}{K}\right), \qquad K^{2(2\widetilde{S}^{(0)}-1)} = 1 + o(1),$$

and

$$K^{-2(1-\widetilde{S}^{(0)})} = K^{-1}(1 + o(1)).$$

Thus, the softmax denominator is

$$\sum_{k=1}^{K} K^{2S_{l,k}^{(0)}} + K^{2(2\widetilde{S}^{(0)}-1)} + K^{-2(1-\widetilde{S}^{(0)})}$$

$$= K\big(1 + O\big(\frac{\log K}{K}\big)\big) + 1 + o(1)(L-1)K \cdot K^{-1}(1+o(1))$$
$$= K + L + O(\log K).$$

This gives Equations (3) to (5). The estimates for $\Phi_l^{(0)}$ and $\Phi_{-l}^{(0)}$ follow by summing the corresponding output probabilities.

It remains to compute the logit updates. From Lemma D.3, with $c_{\mathrm{lm}} = 2$, we have

$$\beta_{l,k}^{(0)} = -\frac{2\log K}{L}\mathbb{E}_{P_{\mathrm{in}}}\Bigg[S_{l,k}^{(0)}\Bigg(F_{(l-1)K+k}^{(0)} - p_{l,k} - \sum_{k'=1}^{K} S_{l,k'}^{(0)}\big(F_{(l-1)K+k'}^{(0)} - p_{l,k'}\big) - \big(2F_{LK+1}^{(0)} + \Phi_{-l}^{(0)}\big)\widetilde{S}^{(0)}\Bigg)\Bigg| X_q = u_l\Bigg].$$

Using

$$S_{l,k}^{(0)} = \frac{1}{2K}(1+o(1)), \qquad \widetilde{S}^{(0)} = \frac{1}{2}(1+o(1)),$$

and

$$F_{(l-1)K+k}^{(0)} = F_{LK+1}^{(0)} = \frac{1}{K+L}(1+o(1)),$$

we have

$$\sum_{k'=1}^{K} S_{l,k'}^{(0)}\big(F_{(l-1)K+k'}^{(0)} - p_{l,k'}\big) = \frac{1}{2K}\left(\frac{K}{K+L} - 1\right)(1+o(1)) = -\frac{L}{2K(K+L)}(1+o(1)).$$

Moreover,

$$\big(2F_{LK+1}^{(0)} + \Phi_{-l}^{(0)}\big)\widetilde{S}^{(0)} = \frac{L+1}{2(K+L)}(1+o(1)).$$

Substituting these estimates gives

$$\beta_{l,k}^{(0)} = -\frac{2\log K}{L}\cdot\frac{1}{2K}\left[\frac{1}{K+L} - p_{l,k} + \frac{L}{2K(K+L)} - \frac{L+1}{2(K+L)} + o\left(\frac{1}{K}\right)\right]$$
$$= \frac{\log K}{LK}\left[p_{l,k} + \frac{L-1}{2(K+L)} - \frac{L}{2K(K+L)} + o\left(\frac{1}{K}\right)\right].$$

Similarly, from Lemma D.3, we can get

$$\widetilde{\beta}_{l,j}^{(0)} = -\frac{2\log K}{L}\mathbb{E}_{P_{\mathrm{in}}}\Bigg[\widetilde{S}_j^{(0)}\Bigg(\big(2F_{LK+1}^{(0)} + \Phi_{-l}^{(0)}\big)(1-\widetilde{S}^{(0)}) - \sum_{k'=1}^{K} S_{l,k'}^{(0)}\big(F_{(l-1)K+k'}^{(0)} - p_{l,k'}\big)\Bigg)\Bigg| X_q = u_l\Bigg]$$
$$= -\frac{\log K}{LJ}\left[\frac{L+1}{2(K+L)} + \frac{L}{2K(K+L)} + o\left(\frac{1}{K}\right)\right].$$

This completes the proof of the variable's initialization. □

Then, after one step of gradient descent, we have the following lemma.

Lemma D.5 (Variables after One Step Gradient Descent)**.** *Under the same condition as Lemma D.4, suppose $P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^*$ and choose the first-step learning rate*

$$\eta_{\mathrm{SFT}}^{(0)} = \frac{KLc^{(\mathrm{I})}}{p_{l,1}}.$$

Further assume $c^{(\mathrm{I})}$ is large enough so that

$$c^{(\mathrm{I})}\frac{K}{Jp_{l,1}}\left[\frac{L+1}{2(K+L)} + \frac{L}{2K(K+L)}\right] \geq 2, c^{(\mathrm{I})}\big(1 - \frac{\max_{k\geq 2} p_{l,k}}{p_{l,1}}\big) \gg 1.$$

Then, after one gradient-descent update, the key variables satisfy the following estimates.

- ***Attention Logits.*** *For the logit corresponding to the head answer, we have*

$$B_{l,1}^{(1)} = c^{(\mathrm{I})} \log K(1 + o(1)).$$

For tails $2 \le k \le K$, we have

$$B_{l,k}^{(1)} = \frac{c^{(\mathrm{I})} \log K}{p_{l,1}} \left(p_{l,k} + \frac{L-2}{2(K+L)} + O\left(\frac{1}{K}\right) \right).$$

Moreover, we have that

$$\widetilde{B}_{l,j}^{(1)} = -\frac{c^{(\mathrm{I})} K \log K}{J p_{l,1}} \left[\frac{L+2}{2(K+L)} + O\left(\frac{1}{K}\right) \right].$$

- ***Attention Scores.*** *From the logits, we see that the context tokens' contribution is lower order. Hence, the attention scores are as follows.*

$$S_{l,1}^{(1)} = \frac{K^{c^{(\mathrm{I})}(1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}}) + o(1)}}{K^{c^{(\mathrm{I})}(1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}}) + o(1)} + K},$$

and for every $k \ge 2$, we have that

$$S_{l,k}^{(1)} = \frac{1 + o(1)}{K^{c^{(\mathrm{I})}(1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}}) + o(1)} + K}.$$

In particular, when $c^{(\mathrm{I})} > 1$, we have that

$$S_{l,1}^{(1)} = 1 - O\left(K^{1 - c^{(\mathrm{I})}(1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}})} \right), \qquad S_{l,k}^{(1)} = K^{-c^{(\mathrm{I})}(1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}}) + o(1)}, \quad k \ge 2.$$

The context attention satisfies

$$\widetilde{S}_j^{(1)} = o(S_{l,k}^{(1)}), \qquad \widetilde{S}^{(1)} = o(1).$$

- ***Outputs.*** *Under the modified output layer with $c_{\mathrm{lm}} = 2$, we have that*

$$1 - F_{(l-1)K+1}^{(1)} = \frac{1}{K}(1 + o(1)).$$

For every non-head same-class answer $k \ge 2$, including both covered tails and uncovered answers,

$$F_{(l-1)K+k}^{(1)} = K^{-2+o(1)}.$$

Moreover, we have that

$$F_{LK+1}^{(1)} = K^{-4+o(1)}, \qquad F_{(l'-1)K+k}^{(1)} = K^{-4+o(1)}, \qquad \forall l' \ne l,\ k \in [K].$$

Proof. From Lemma D.4 and the choice $\eta_{\mathrm{SFT}}^{(0)} = KLc^{(\mathrm{I})}/p_{l,1}$, we have

$$B_{l,1}^{(1)} = B_{l,1}^{(0)} + \eta_{\mathrm{SFT}}^{(0)} \beta_{l,1}^{(0)} = c^{(\mathrm{I})} \log K(1 + o(1)).$$

For covered tails $2 \le k \le K$, we obtain

$$B_{l,k}^{(1)} = \frac{c^{(\mathrm{I})} \log K}{p_{l,1}} \left(p_{l,k} + \frac{L-2}{2(K+L)} + O\left(\frac{1}{K}\right) \right).$$

The estimate for $\widetilde{B}_{l,j}^{(1)}$ follows from Lemma D.4.

Because $p_{l,1} \gg L/K$, the non-head logits are lower order compared with $B_{l,1}^{(1)}$. Hence

$$B_{l,1}^{(1)} - \max_{k \ge 2} B_{l,k}^{(1)} = c^{(\mathrm{I})} \log(K) \cdot (1 - \frac{\max_{k \ge 2} p_{l,k}}{p_{l,1}} + o(1)).$$

Moreover, the context tokens contribution to the attention denominator is lower order than the non-head informative tokens contribution. Using the high-probability count bounds in Lemma D.7, we obtain

$$S_{l,1}^{(1)} = \frac{K^{c^{(\mathrm{I})}(1-\frac{\max_{k\geq 2} p_{l,k}}{p_{l,1}})+o(1)}}{K^{c^{(\mathrm{I})}(1-\frac{\max_{k\geq 2} p_{l,k}}{p_{l,1}})+o(1)} + K + o(1)},$$

and for every $k \geq 2$, the attention scores are as follows.

$$S_{l,k}^{(1)} = \frac{1+o(1)}{K^{c^{(\mathrm{I})}(1-\frac{\max_{k\geq 2} p_{l,k}}{p_{l,1}})+o(1)} + K}.$$

In addition, the logits satisfy

$$\widetilde{B}_{l,j}^{(1)} - \max_{k\geq 2} B_{l,k}^{(1)} = \frac{c^{(\mathrm{I})}\log K}{p_{l,1}}\left(-\frac{K(L+2)}{2J(K+L)} - \max_{k\geq 2} p_{l,k} - \frac{L-2}{2(K+L)} + o(\frac{1}{K})\right).$$

As a result, we have

$$\widetilde{S}^{(1)} = \frac{K^{\frac{c^{(\mathrm{I})}}{p_{l,1}}(-\max_{k\geq 2} p_{l,k} - \frac{L+2}{K+L})+o(\frac{1}{K})}}{K^{c^{(\mathrm{I})}(1-\frac{\max_{k\geq 2} p_{l,k}}{p_{l,1}})+o(1)} + K + o(1)} = K^{\frac{c^{(\mathrm{I})}}{p_{l,1}}(-\max_{k\geq 2} p_{l,k} - \frac{L+2}{K+L})+o(\frac{1}{K})} \max_{k\geq 2} S_{l,k}^{(1)},$$

which proves the lemma's claim on attention scores.

The remaining output estimates follow from the same softmax calculation as in the initialization in Lemma D.4. In particular, we have

$$K^{2S_{l,1}^{(1)}} = K^{2-o(1)}, \qquad K^{2S_{l,k}^{(1)}} = 1+o(1), \quad k \geq 2,$$

and

$$K^{2(2\widetilde{S}^{(1)}-1)} = K^{-2+o(1)}, \qquad K^{-2(1-\widetilde{S}^{(1)})} = K^{-2+o(1)}.$$

Thus, for the head answer, we have

$$F_{(l-1)K+1}^{(1)} = 1 - \frac{1}{K}(1+o(1)),$$

and for every $k \geq 2$,

$$F_{(l-1)K+k}^{(1)} = K^{-2+o(1)}.$$

The wrong-answer and other-class estimates follow similarly. Hence we finish the proof. □

Proof of Theorem 4.2. As a direct result of Lemma D.5, we have

$$1 - \Phi_l^{(1)} = 1 - \sum_{k=1}^{K} F_{(l-1)K+k}^{(1)} = K^{-4+o(1)} + O\left(\frac{L}{K^3}\right).$$

We finish the proof of SFT. □

D.3. Useful Concentration Bound

Recall that given a prompt $P_{\text{in}} = (X_1^\top, X_2^\top, \ldots, X_N^\top)^\top$, the distribution of X_n is identical independent. Similar to (Huang et al., 2024; Cheng et al., 2026), it is worth noting that, for any question class $l \in [L]$, the occurrence count of the l, k-th informative and j-th context tokens in P_{in}, denoted as $|\mathcal{V}_{l,k}|$ and $|\widetilde{\mathcal{V}}_j|$, follows a multinomial distribution. Leveraging the concentration property inherent to multinomial distributions, we can identify a high-probability event to which P belongs.

We first introduce the following tail bound for multinomial distributions.

Lemma D.6 (Tail Bound of Multinomial Distribution (Devroye, 1983))**.** *Let $(X_1, \cdots, X_K)$ be a multinomial $(N, p_1, \cdots, p_K)$ random vector. For all $\varepsilon \in (0,1)$ and all K satisfying $K/N \leq \varepsilon^2/20$, we have*

$$P\left(\sum_{i=1}^{K} |X_i - \mathbb{E}(X_i)| > N\varepsilon\right) \leq 3\exp\left(-N\varepsilon^2/25\right).$$

Lemma D.7 (High-probability Event for P_{in})**.** *For any question class $l \in [L]$, suppose that the probability of l, k-th informative token $p_{l,k} = \Theta(1/2K)$ for any $k \in [K]$, and the probability of context token $\widetilde{p}_j = \Theta(1/2J)$ for any $j \in [J]$ and $(K+J)^3 \ll N$. For some constant $\sqrt{40(K+J)^3/N} \geq c \geq \sqrt{20(K+J)^3/N}$, define*

$$\mathcal{E}_{\text{P}}^* = \Big\{ P_{\text{in}} | X_q = u_l : \textit{ for all } l \in [L], |\mathcal{V}_{l,k}| \in \Big[p_{l,k}N - \frac{cN}{K+J}, p_{l,k}N + \frac{cN}{K+J} \Big] \textit{ for } k \in [K] \\ |\widetilde{\mathcal{V}}_j| \in \Big[\widetilde{p}_j N - \frac{cN}{K+J}, \widetilde{p}_j N + \frac{cN}{K+J} \Big] \textit{ for } j \in [J] \Big\}.$$

Then , we have

$$\mathbb{P}(P_{\text{in}} \in \mathcal{E}_{\text{P}}^*) \geq 1 - 3\exp\left(-\frac{c^2 N}{25L(K+J)^2} \right).$$

Let us denote $L_{l,k} = p_{l,k}K - \frac{cK}{K+J}$, $U_{l,k} = p_{l,k}K + \frac{cK}{K+J}$, $\widetilde{L}_j = \widetilde{p}_j J - \frac{cJ}{K+J}$, $\widetilde{U}_j = \widetilde{p}_j J + \frac{cJ}{K+J}$. Note that $L_{l,k}, U_{l,k}$ are at the order of the constant level since $p_{l,k} = \Theta(1/K)$, $\widetilde{L}_j, \widetilde{U}_j$ are at the order of the constant level since $\widetilde{p}_j = \Theta(1/J)$. Then for any P belonging to $\mathcal{E}_{\text{P}}^*$, $|\mathcal{V}_{l,k}| \in [L_{l,k}N/K, U_{l,k}N/K] = \Theta(N/K)$ and $|\widetilde{\mathcal{V}}_j| \in [\widetilde{L}_j N/J, \widetilde{U}_j N/J] = \Theta(N/J)$. Note that we can properly choose c to guarantee $L_{l,k} > 0$ for $k \in [K], l \in [L]$ and $\widetilde{L}_j > 0$ for $j \in [J]$.

E. Proofs of RL

E.1. Gradient Computation for RL

In this section, we analyze the training dynamics of RL. The proof follows a similar structure to the SFT analysis in Appendix D, but the dynamics are different in an important way. In SFT, the teacher distribution is long-tailed, and therefore, the head answer receives a much larger initial positive update. In contrast, the RL objective rewards every *correct* answer symmetrically. As a result, all informative tokens receive nearly the same positive update, while the context tokens, which correspond to the incorrect answer, receive a negative update.

Throughout this section, we fix a question class $l \in [L]$ and condition on $X_q = u_l$. Similar to SFT, we write

$$P_{\text{in}} = (X_1, \ldots, X_N) \in \mathbb{R}^{N \times d}$$

for the prompt tokens, and define

$$S = \texttt{soft}(X_q^\top W P_{\text{in}}^\top) \in \mathbb{R}^{1 \times N}, \qquad F = \texttt{soft}(S P_{\text{in}} W_{\text{lm}}) \in \mathbb{R}^{1 \times (LK+1)}.$$

Similar to SFT, conditional on $X_q = u_l$, the output can be written as

$$F = \texttt{soft}\left(c_{\text{lm}} \log K \left(\sum_{k=1}^{K} S_{l,k} a_{l,k}^\top + \widetilde{S}\widetilde{a}^\top - (1 - \widetilde{S}) \sum_{l' \neq l} \sum_{k'=1}^{K} a_{l',k'}^\top \right) \right).$$

First, we calculate the gradient as follows.

Lemma E.1 (RL Gradient)**.** *The gradient of the RL objective with respect to W is*

$$\nabla_W J_{\text{RL}} = \mathbb{E}_{P_{\text{in}}, X_q} \left[\sum_{l=1}^{L} \mathbb{1}\{X_q = u_l\} X_q \Big(\big(F \odot (A_l - \Phi_l \mathbf{1}_{LK+1}) \big) W_{\text{lm}}^\top P_{\text{in}}^\top \big(\operatorname{diag}(S) - S^\top S \big) \Big) P_{\text{in}} \right],$$

where $\mathbf{1}_{LK+1} \in \mathbb{R}^{1 \times (LK+1)}$ is the all-one vector and $\odot$ denotes entrywise product.

Proof. For a single sample (P_{in}, X_q), define the intermediate variables as

$$s = X_q^\top W P_{\text{in}}^\top \in \mathbb{R}^{1 \times N}, \qquad S = \texttt{soft}(s) \in \mathbb{R}^{1 \times N},$$

$$z = S P_{\text{in}} W_{\text{lm}} \in \mathbb{R}^{1 \times (LK+1)}, \qquad F = \texttt{soft}(z) \in \mathbb{R}^{1 \times (LK+1)}.$$

Here $P_{\text{in}} \in \mathbb{R}^{N\times d}$, $X_q \in \mathbb{R}^{d\times 1}$, $W \in \mathbb{R}^{d\times d}$, and $W_{\text{lm}} \in \mathbb{R}^{d\times(LK+1)}$.

For each question class $l \in [L]$, define the class-specific correct-answer indicator vector

$$A_l = \sum_{k=1}^{K} a_{(l-1)K+k} \in \mathbb{R}^{1\times(LK+1)}.$$

Then the total probability assigned to correct answers of class l is

$$\Phi_l = \sum_{k=1}^{K} F_{(l-1)K+k} = F A_l^\top.$$

Conditional on $X_q = u_l$, the single-sample RL objective is

$$j_{\text{RL}}(W; P_{\text{in}}, X_q) = \Phi_l = \sum_{k=1}^{K} F_{(l-1)K+k}.$$

Therefore,

$$J_{\text{RL}}(W) = \mathbb{E}_{P_{\text{in}}, X_q}\left[\sum_{l=1}^{L} \mathbb{1}\{X_q = u_l\}\Phi_l\right].$$

We compute the gradient for one sample (P_{in}, X_q) conditional on $X_q = u_l$. The unconditional gradient follows by multiplying by $\mathbb{1}\{X_q = u_l\}$ and taking expectation over l.

We first differentiate with respect to the output logits z. Recall that

$$F = \texttt{soft}(z), \qquad \Phi_l = \sum_{k=1}^{K} F_{(l-1)K+k}.$$

For any coordinate $m \in [LK+1]$, the softmax derivative gives

$$\frac{\partial F_r}{\partial z_m} = F_r\big(\mathbb{1}\{r = m\} - F_m\big).$$

Therefore,

$$\begin{aligned}
\frac{\partial \Phi_l}{\partial z_m} &= \sum_{k=1}^{K} \frac{\partial F_{(l-1)K+k}}{\partial z_m} \\
&= \sum_{k=1}^{K} F_{(l-1)K+k}\big(\mathbb{1}\{(l-1)K + k = m\} - F_m\big) \\
&= F_m \mathbb{1}\{m \in \{(l-1)K+1, \ldots, lK\}\} - F_m \sum_{k=1}^{K} F_{(l-1)K+k} \\
&= F_m\left((A_l)_m - \Phi_l\right).
\end{aligned}$$

Thus, in row-vector form,

$$g_z = \nabla_z j_{\text{RL}} = F \odot (A_l - \Phi_l \mathbf{1}_{LK+1}) \in \mathbb{R}^{1\times(LK+1)}.$$

Then, we derive with respect to the attention vector S. Since

$$z = S P_{\text{in}} W_{\text{lm}},$$

we have

$$g_S = \nabla_S j_{\text{RL}} = g_z W_{\text{lm}}^\top P_{\text{in}}^\top \in \mathbb{R}^{1\times N}.$$

Furthermore, we derive through the inner softmax. Since

$$S = \texttt{soft}(s),$$

we have,

$$g_s = \nabla_s j_{\mathrm{RL}} = g_S\big(\operatorname{diag}(S) - S^\top S\big) \in \mathbb{R}^{1\times N}.$$

Substituting the expression for g_S, we obtain

$$g_s = \big(F \odot (A_l - \Phi_l \mathbf{1}_{LK+1})\big) W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big).$$

Finally, we derive with respect to W. Recall that

$$s = X_q^\top W P_{\mathrm{in}}^\top.$$

Taking differential gives

$$ds = X_q^\top \, dW \, P_{\mathrm{in}}^\top.$$

Therefore,

$$\nabla_W j_{\mathrm{RL}} = X_q g_s P_{\mathrm{in}}.$$

Substituting the expression for g_s, we get

$$\nabla_W j_{\mathrm{RL}} = X_q\Big(\big(F \odot (A_l - \Phi_l \mathbf{1}_{LK+1})\big) W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P_{\mathrm{in}}.$$

Finally, taking expectation over (P_{in}, X_q) and summing over the possible question classes gives

$$\begin{aligned}
&\nabla_W J_{\mathrm{RL}}(W) \\
&= \mathbb{E}_{P_{\mathrm{in}}, X_q}\left[\sum_{l=1}^{L} \mathbb{1}\{X_q = u_l\} X_q\Big(\big(F \odot (A_l - \Phi_l \mathbf{1}_{LK+1})\big) W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top \big(\operatorname{diag}(S) - S^\top S\big)\Big) P_{\mathrm{in}}\right].
\end{aligned}$$

This completes the proof of RL gradient. □

Similar to the SFT analysis, we overload the notation and define the important logits and their corresponding update rates for RL.

Definition E.2 (Important Logits and Their Update Rate in RL)**.** For any $l \in [L]$, $k \in [K]$ and $j \in [J]$, for iteration $t \geq 0$, we denote the weight as $W^{(t)}$ and the loss as $J_{\mathrm{RL}}^{(t)} = J_{\mathrm{RL}}(W^{(t)})$. We define the following quantities

$$\begin{aligned}
B_{l,k}^{(t)} &= u_l W^{(t)} i_{l,k}^\top \qquad \beta_{l,k}^{(t)} = u_l \nabla_W J_{\mathrm{RL}}^{(t)} \, i_{l,k}^\top; \\
\widetilde{B}_{l,j}^{(t)} &= u_l W^{(t)} c_j^\top \qquad \widetilde{\beta}_{l,j}^{(t)} = u_l \nabla_W J_{\mathrm{RL}}^{(t)} \, c_j^\top.
\end{aligned}$$

By the GD update, we have

$$\begin{aligned}
B_{l,k}^{(t+1)} &= B_{l,k}^{(t)} + \eta_{\mathrm{RL}} \beta_{l,k}^{(t)} \\
\widetilde{B}_{l,j}^{(t+1)} &= \widetilde{B}_{l,j}^{(t)} + \eta_{\mathrm{RL}} \widetilde{\beta}_{l,j}^{(t)}
\end{aligned}$$

Moreover, by our initialization of $W^{(0)} = \mathbf{0}_{d\times d}$, we have $B_{l,k}^{(0)} = \widetilde{B}_{l,j}^{(0)} = 0$.

Lemma E.3 (Logits Update Rate for RL)**.** *For any $l \in [L]$, $k \in [K]$ and $j \in [J]$, the RL update rates of the attention logits can be written as*

$$\begin{aligned}
\beta_{l,k} = \frac{c_{\mathrm{lm}} \log(K)}{L} \mathbb{E}_{P_{\mathrm{in}}}\Big[S_{l,k}\Big((1-\Phi_l)\Big(F_{(l-1)K+k} - \sum_{k'=1}^{K} S_{l,k'} F_{(l-1)K+k'}\Big) \\
+ \Phi_l (2F_{LK+1} + \Phi_{-l})\widetilde{S}\Big)\Big| X_q = u_l\Big], \\
\widetilde{\beta}_{l,j} = -\frac{c_{\mathrm{lm}} \log(K)}{L} \mathbb{E}_{P_{\mathrm{in}}}\Big[\widetilde{S}_j\Big((1-\Phi_l) \sum_{k'=1}^{K} S_{l,k'} F_{(l-1)K+k'} \\
+ \Phi_l (2F_{LK+1} + \Phi_{-l})(1-\widetilde{S})\Big)\Big| X_q = u_l\Big].
\end{aligned}$$

Proof. The proof follows the same calculation as Lemma D.3, but with the RL output-gradient vector

$$g_z = F \odot (A_l - \Phi_l \mathbf{1}_{LK+1}).$$

Recall from Lemma E.1 that, conditional on $X_q = u_l$,

$$\nabla_W J_{\mathrm{RL}} = \frac{1}{L}\mathbb{E}_{P_{\mathrm{in}}}\Big[u_l^\top \Big(g_z W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top (\mathrm{diag}(S) - S^\top S)\Big) P_{\mathrm{in}} \Big| X_q = u_l\Big].$$

Since

$$W_{\mathrm{lm}} = c_{\mathrm{lm}} \log(K) \left(\sum_{l'=1}^{L} \sum_{k'=1}^{K} i_{l',k'}^\top (a_{l',k'} - A_{-l'} - \widetilde{a}) + \sum_{j=1}^{J} c_j^\top \widetilde{a} \right),$$

we have

$$\begin{aligned} & W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top (\mathrm{diag}(S) - S^\top S) P_{\mathrm{in}} \\ &= c_{\mathrm{lm}} \log(K) \left(\sum_{l'=1}^{L} \sum_{k'=1}^{K} S_{l',k'} (a_{l',k'} - A_{-l'} - \widetilde{a})^\top i_{l',k'} + \widetilde{a}^\top \sum_{n=1}^{N} \mathbb{1}\{X_n \in \mathcal{C}\} S_n X_n \right) \\ &\quad - c_{\mathrm{lm}} \log(K) \left(\sum_{l'=1}^{L} \sum_{k'=1}^{K} S_{l',k'} (a_{l',k'} - A_{-l'} - \widetilde{a})^\top + \widetilde{S}\widetilde{a}^\top \right) \left(\sum_{n=1}^{N} S_n X_n \right). \end{aligned} \tag{6}$$

When $X_q = u_l$, we have $S_{l',k'} = 0$ for all $l' \neq l$. Hence,

$$\begin{aligned} & W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top (\mathrm{diag}(S) - S^\top S) P_{\mathrm{in}}\, i_{l,k}^\top \\ &= c_{\mathrm{lm}} \log(K) S_{l,k} \left((a_{l,k} - A_{-l} - \widetilde{a})^\top - \sum_{k'=1}^{K} S_{l,k'} (a_{l,k'} - A_{-l} - \widetilde{a})^\top - \widetilde{S}\widetilde{a}^\top \right). \end{aligned}$$

Using $\sum_{k'=1}^{K} S_{l,k'} = 1 - \widetilde{S}$, this becomes

$$c_{\mathrm{lm}} \log(K) S_{l,k} \left(a_{l,k}^\top - \sum_{k'=1}^{K} S_{l,k'} a_{l,k'}^\top - \widetilde{S} A_{-l}^\top - 2\widetilde{S}\widetilde{a}^\top \right).$$

Moreover,

$$g_z a_{l,k}^\top = (1 - \Phi_l) F_{(l-1)K+k}, \qquad g_z A_{-l}^\top = -\Phi_l \Phi_{-l}, \qquad g_z \widetilde{a}^\top = -\Phi_l F_{LK+1}.$$

Therefore,

$$\begin{aligned} \beta_{l,k} = \frac{c_{\mathrm{lm}} \log(K)}{L} \mathbb{E}_{P_{\mathrm{in}}} \Big[S_{l,k} \Big((1 - \Phi_l) \Big(F_{(l-1)K+k} - \sum_{k'=1}^{K} S_{l,k'} F_{(l-1)K+k'} \Big) \\ + \Phi_l (2F_{LK+1} + \Phi_{-l}) \widetilde{S} \Big) \Big| X_q = u_l \Big]. \end{aligned}$$

Similarly, for the context token c_j,

$$\begin{aligned} & W_{\mathrm{lm}}^\top P_{\mathrm{in}}^\top (\mathrm{diag}(S) - S^\top S) P_{\mathrm{in}}\, c_j^\top \\ &= c_{\mathrm{lm}} \log(K) \widetilde{S}_j \left(\widetilde{a}^\top - \sum_{k'=1}^{K} S_{l,k'} (a_{l,k'} - A_{-l} - \widetilde{a})^\top - \widetilde{S}\widetilde{a}^\top \right) \\ &= c_{\mathrm{lm}} \log(K) \widetilde{S}_j \left(- \sum_{k'=1}^{K} S_{l,k'} a_{l,k'}^\top + (1 - \widetilde{S}) A_{-l}^\top + 2(1 - \widetilde{S}) \widetilde{a}^\top \right). \end{aligned}$$

Taking inner product with g_z gives

$$\begin{aligned} \widetilde{\beta}_{l,j} = -\frac{c_{\mathrm{lm}} \log(K)}{L} \mathbb{E}_{P_{\mathrm{in}}} \Big[\widetilde{S}_j \Big((1 - \Phi_l) \sum_{k'=1}^{K} S_{l,k'} F_{(l-1)K+k'} \\ + \Phi_l (2F_{LK+1} + \Phi_{-l})(1 - \widetilde{S}) \Big) \Big| X_q = u_l \Big]. \end{aligned}$$

This proves the lemma. □

E.2. Learning Dynamics of RL

With the gradient expressions derived in Appendix E.1, we analyze the training dynamics of RL in this section. For each question class $l \in [L]$, we fix l and condition on $\{X_q = u_l\}$. As before, when $X_q = u_l$, only informative tokens from $\mathcal{I}_l$ and context tokens from $\mathcal{C}$ appear in the prompt, so $S_{l',k'} = 0$ for all $l' \neq l$.

Different from SFT, the RL objective is symmetric over all correct answers in the same question class. Thus, all informative tokens receive the same leading positive update, while context tokens receive a negative update. Consequently, RL does not concentrate attention on the head answer; instead, it learns all correct informative tokens nearly uniformly.

Recall that $c_{\mathrm{lm}} = 2$, under the language model head, we have

$$W_{\mathrm{lm}} = 2\log(K)\left(\sum_{l'=1}^{L}\sum_{k'=1}^{K} i_{l',k'}^{\top}(a_{l',k'} - A_{-l'} - \widetilde{a}) + \sum_{j=1}^{J} c_j^{\top}\widetilde{a}\right),$$

conditional on $X_q = u_l$, the output can be written as

$$F = \texttt{soft}\left(2\log K\left(\sum_{k=1}^{K} S_{l,k}a_{l,k}^{\top} + (2\widetilde{S}-1)\widetilde{a}^{\top} - (1-\widetilde{S})A_{-l}\right)\right),$$

where

$$A_{-l} = \sum_{l'\neq l}\sum_{k'=1}^{K} a_{l',k'}.$$

We first calculate the important estimates at initialization.

Lemma E.4 (Initialization Variable for RL)**.** *At initialization $t = 0$, for all $k \in [K]$ and $j \in [J]$, the key variables satisfy the following estimates.*

- ***Attention Logits.***

$$B_{l,k}^{(0)} = \widetilde{B}_{l,j}^{(0)} = 0.$$

- ***Attention Scores.*** *Suppose $P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^{*}$ satisfies the high-probability event in Lemma D.7. Then*

$$S_{l,k}^{(0)} = \frac{1}{2K}\left(1 + O\left(\frac{(K+J)^3}{N}\right)\right), \qquad \widetilde{S}_j^{(0)} = \frac{1}{2J}\left(1 + O\left(\frac{(K+J)^3}{N}\right)\right),$$

  and

$$\widetilde{S}^{(0)} = \frac{1}{2}\left(1 + O\left(\frac{(K+J)^3}{N}\right)\right).$$

- ***Outputs.*** *Suppose $P_{\mathrm{in}} \in \mathcal{E}_{\mathrm{P}}^{*}$. Then*

$$F_{(l-1)K+k}^{(0)} = \frac{1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right), \qquad F_{LK+1}^{(0)} = \frac{1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right),$$

  and for every $l' \neq l$,

$$F_{(l'-1)K+k}^{(0)} = \frac{1}{K(K+L)}\left(1 + O\left(\frac{\log K}{K}\right)\right).$$

  Consequently,

$$\Phi_l^{(0)} = \frac{K}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right), \quad \Phi_{-l}^{(0)} = \frac{L-1}{K+L}\left(1 + O\left(\frac{\log K}{K}\right)\right).$$

- ***Attention Logits Update.*** *For every $k \in [K]$ and $j \in [J]$,*

$$\beta_{l,k}^{(0)} = \Theta\left(\frac{(L+1)\log K}{LK^2}\right) > 0, \qquad \widetilde{\beta}_{l,j}^{(0)} = -\Theta\left(\frac{(L+1)\log K}{LJK}\right) < 0.$$

Proof. The estimates for attention logits and attention scores, and the calculations of the output probabilities, follow from $W^{(0)} = \mathbf{0}_{d\times d}$ and Lemma D.7, which is similar to the proof of Lemma D.4. It remains to compute the attention logits update.

Applying Lemma E.3, we immediately obtain

$$\beta_{l,k}^{(0)} = \Theta\left(\frac{(L+1)\log K}{LK^2}\right) > 0, \qquad \widetilde{\beta}_{l,j}^{(0)} = -\Theta\left(\frac{(L+1)\log K}{LJK}\right) < 0.$$

This completes the proof. □

Recall we use a one-step RL learning rate

$$\eta_{\mathrm{RL}}^{(0)} = \Theta\left(\frac{LK^2}{L+1}c_\epsilon\right),$$

where $c_\epsilon > 0$ is a sufficiently large constant. More precisely, we choose $\eta_{\mathrm{RL}}^{(0)}$ so that

$$\eta_{\mathrm{RL}}^{(0)}\beta_{l,k}^{(0)} = c_\epsilon \log K(1+o(1)).$$

After the one-step update, the attention weights on context tokens become negligible, while the attention weights on the K informative tokens remain nearly uniform and large enough as in the following lemma.

Lemma E.5 (Variables after One Step Update for RL)**.** *Under the same condition as Lemma E.4, set $b_K = \frac{(L+1)\log K}{LK^2}$, then from Lemma E.4 $\beta_{l,k}^{(0)} = b_K(1+o(1)), \widetilde{\beta}_{l,j}^{(0)} = -\frac{K}{J}b_K(1+o(1))$. For any target accuracy $\epsilon \in (0, 1/4)$, choose constant*

$$c_\epsilon \geq \frac{J}{K}\cdot\frac{\log(C_0 \log K/\epsilon)}{\log K},$$

where $C_0 > 0$ is a sufficiently large universal constant, and set

$$\eta_{\mathrm{RL}}^{(0)} = \frac{2c_\epsilon \log K}{b_K} = \frac{2LK^2}{L+1}c_\epsilon.$$

Then, after one RL update, for every $k \in [K]$ and $j \in [J]$, we have the following estimates.

- ***Attention Logits.*** *We have*

$$B_{l,k}^{(1)} = c_\epsilon \log K(1+o(1)), \qquad \widetilde{B}_{l,j}^{(1)} = -\frac{K}{J}c_\epsilon \log K(1+o(1)).$$

- ***Attention Scores.*** *For the informative tokens, we have the corresponding attention scores as*

$$\max_{k\in[K]}\left|S_{l,k}^{(1)} - \frac{1}{K}\right| \leq \frac{\epsilon}{K},$$

  and for the context tokens, we have

$$\widetilde{S}_j^{(1)} = O\left(\frac{1}{J}K^{-c_\epsilon K/J}\right), \qquad \widetilde{S}^{(1)} \leq \frac{\epsilon}{C_0 \log K}.$$

- ***Outputs.*** *For the output corresponding to the correct answers, we have*

$$\max_{k\in[K]}\left|F_{(l-1)K+k}^{(1)} - \frac{1}{K}\right| \leq \frac{C}{K}\left(\epsilon + \frac{L}{K^2} + \frac{1}{K^3}\right),$$

  for some universal constant $C > 0$. And the total probability of correct answers satisfies that

$$\Phi_l^{(1)} \geq 1 - C\left(\epsilon + \frac{L}{K^2} + \frac{1}{K^3}\right).$$

  Moreover, for the incorrect answers, we have

$$F_{LK+1}^{(1)} = O(K^{-3}), \qquad \Phi_{-l}^{(1)} = O\left(\frac{L}{K^2}\right).$$

Proof. By Lemma E.4 and Lemma E.3, at initialization,

$$\beta_{l,k}^{(0)} = b_K(1+o(1)), \qquad \widetilde{\beta}_{l,j}^{(0)} = -\frac{K}{J} b_K(1+o(1)),$$

where

$$b_K = \frac{(L+1)\log K}{LK^2}.$$

With the learning rate

$$\eta_{\mathrm{RL}}^{(0)} = \frac{2c_\epsilon \log K}{b_K},$$

the informative logits satisfy

$$B_{l,k}^{(1)} = B_{l,k}^{(0)} + \eta_{\mathrm{RL}}^{(0)} \beta_{l,k}^{(0)} = c_\epsilon \log K(1+o(1)), \qquad \forall k \in [K].$$

Similarly, the context logits satisfy

$$\widetilde{B}_{l,j}^{(1)} = \widetilde{B}_{l,j}^{(0)} + \eta_{\mathrm{RL}}^{(0)} \widetilde{\beta}_{l,j}^{(0)} = -\frac{K}{J} c_\epsilon \log K(1+o(1)), \qquad \forall j \in [J].$$

We next control the attention scores. Since all informative logits are equal up to lower-order terms and every context logit is smaller than the informative logits by order

$$\left(1+\frac{K}{J}\right) c_\epsilon \log K,$$

the total context attention mass satisfies

$$\widetilde{S}^{(1)} = O\left(K^{-c_\epsilon K/J}\right).$$

By the choice of c_ϵ, we have

$$\widetilde{S}^{(1)} \le K^{-c_\epsilon K/J} \le \frac{\epsilon}{C_0 \log K}.$$

Since the remaining attention mass is distributed symmetrically over the K informative tokens, we have

$$S_{l,k}^{(1)} = \frac{1-\widetilde{S}^{(1)}}{K}(1+o(1)), \qquad \forall k \in [K].$$

Hence, we have,

$$\max_{k\in[K]} \left| S_{l,k}^{(1)} - \frac{1}{K} \right| \le \frac{\epsilon}{K}.$$

It remains to control the output probabilities. Recall that $c_{\mathrm{lm}} = 2$,

$$F_{(l-1)K+k}^{(1)} = \frac{K^{2S_{l,k}^{(1)}}}{\sum_{k'=1}^{K} K^{2S_{l,k'}^{(1)}} + K^{2(2\widetilde{S}^{(1)}-1)} + \sum_{l'\neq l}\sum_{k'=1}^{K} K^{-2(1-\widetilde{S}^{(1)})}}.$$

From the attention bound,

$$S_{l,k}^{(1)} = \frac{1}{K} + O\left(\frac{\epsilon}{K}\right), \qquad \widetilde{S}^{(1)} \le \frac{\epsilon}{C_0 \log K}.$$

Therefore,

$$K^{2S_{l,k}^{(1)}} = \exp\left(2S_{l,k}^{(1)} \log K\right) = 1 + O\left(\frac{\log K}{K}\right) + O(\epsilon),$$

where the last step uses $\epsilon \in (0, 1/4)$ and absorbs constants into the $O(\cdot)$ term. Also,

$$K^{2(2\widetilde{S}^{(1)}-1)} = K^{-2} \exp\left(4\widetilde{S}^{(1)} \log K\right) = O(K^{-2}),$$

and

$$K^{-2(1-\widetilde{S}^{(1)})} = K^{-2}\exp\left(2\widetilde{S}^{(1)}\log K\right) = O(K^{-2}).$$

Thus, the softmax denominator is

$$\sum_{k'=1}^{K} K^{2S^{(1)}_{l,k'}} + K^{2(2\widetilde{S}^{(1)}-1)} + \sum_{l'\neq l}\sum_{k'=1}^{K} K^{-2(1-\widetilde{S}^{(1)})}$$
$$= K(1+O(\epsilon)) + O(K^{-2}) + O\left(\frac{L}{K}\right).$$

Therefore,

$$F^{(1)}_{(l-1)K+k} = \frac{1}{K}\left(1 + O\left(\epsilon + \frac{\log K}{K}\right)\right).$$

This implies

$$\max_{k\in[K]}\left|F^{(1)}_{(l-1)K+k} - \frac{1}{K}\right| \leq \frac{C}{K}\left(\epsilon + \frac{\log K}{K}\right),$$

for some constant C. Moreover, due to the term $\frac{\log K}{K}$, we can choose $\epsilon = O(\frac{\log K}{K})$ at most.

$$F^{(1)}_{LK+1} = \frac{K^{2(2\widetilde{S}^{(1)}-1)}}{K(1+O(\epsilon)) + O(K^{-2}) + O(L/K)} = O(K^{-3}).$$

For answers from other question classes, each output probability is also $O(K^{-3})$. Since there are $(L-1)K$ such answers,

$$\Phi^{(1)}_{-l} = O\left(\frac{L}{K^2}\right).$$

Finally,

$$\Phi^{(1)}_{l} = 1 - F^{(1)}_{LK+1} - \Phi^{(1)}_{-l} \geq 1 - O(\frac{L}{K^2} + \frac{1}{K^3}) \geq 1-\epsilon,$$

where we enlarge $C > 0$ if necessary. This completes the proof. □

Proof of Theorem 4.2. Following from Lemma E.5, we prove Theorem 4.2 immediately. □

F. Proof of Theorem 4.3

Proof. First, for **SFT**, from the proof of Lemma D.5, for any $k \in [K]$,

$$B^{\mathrm{SFT}}_{l,k} = \frac{c^{(\mathrm{I})}\log K}{p_{l,1}}\left(p_{l,k} + \frac{L-2}{2(K+L)} + o\left(\frac{1}{K}\right)\right),$$
$$\widetilde{B}^{\mathrm{SFT}}_{l,j} = -\frac{c^{(\mathrm{I})}K\log K}{Jp_{l,1}}\left(\frac{L+2}{2(K+L)} + o\left(\frac{1}{K}\right)\right).$$

If $p_{l,k} \geq 1/\log K$, then

$$S^{\mathrm{SFT}}_{l,k} = \frac{(N/2)\exp(B^{\mathrm{SFT}}_{l,k})}{(N/2)\exp(B^{\mathrm{SFT}}_{l,k}) + \sum_{j=1}^{J}|\widetilde{\mathcal{V}}_j|\exp(\widetilde{B}^{\mathrm{SFT}}_{l,j})}$$
$$\geq \frac{\exp\left(c^{(\mathrm{I})}/p_{l,1}\right)}{\exp\left(c^{(\mathrm{I})}/p_{l,1}\right)+1}, \qquad \widetilde{S}^{\mathrm{SFT}} \leq \frac{1}{\exp\left(c^{(\mathrm{I})}/p_{l,1}\right)+1}.$$

Therefore,

$$F^{\mathrm{SFT}}_{l,k} = \frac{K^{2S^{\mathrm{SFT}}_{l,k}}}{K^{2S^{\mathrm{SFT}}_{l,k}} + (L-1)K\cdot K^{-2S^{\mathrm{SFT}}_{l,k}} + K^{2(2\widetilde{S}^{\mathrm{SFT}}-1)}}$$

$$= \frac{K^{4S_{l,k}^{\mathrm{SFT}}-1}}{K^{4S_{l,k}^{\mathrm{SFT}}-1} + (L-1) + K^{1-2S_{l,k}^{\mathrm{SFT}}}}.$$

It follows that

$$1 - F_{l,k}^{\mathrm{SFT}} \leq \frac{L}{K^{4S_{l,k}^{\mathrm{SFT}}-1}} = O\left(K^{1-\frac{4\exp(c^{(\mathrm{I})}/p_{l,1})}{\exp(c^{(\mathrm{I})}/p_{l,1})+1}}\right) = O\left(\frac{1}{K^3}\right),$$

where the last step follows from $c^{(\mathrm{I})}$ can be sufficiently large.

If $p_{l,k} \leq 1/K$, then similarly,

$$S_{l,k}^{\mathrm{SFT}} = \widetilde{S}^{\mathrm{SFT}} = \frac{1}{2}.$$

Hence,

$$\begin{aligned} F_{l,k}^{\mathrm{SFT}} &= \frac{K^{2S_{l,k}^{\mathrm{SFT}}}}{K^{2S_{l,k}^{\mathrm{SFT}}} + (L-1)K \cdot K^{-2S_{l,k}^{\mathrm{SFT}}} + K^{2(2\widetilde{S}^{\mathrm{SFT}}-1)}} \\ &= \frac{K}{K + (L-1) + K} \leq \frac{1}{2}. \end{aligned}$$

Hence, we prove the result for SFT.

On the other hand, for **RL**, from Lemma E.5, for all $k \in [K]$,

$$\begin{aligned} B_{l,k}^{\mathrm{RL}} &= c_\epsilon \log K(1+o(1)), \\ \widetilde{B}_{l,j}^{\mathrm{RL}} &= -\frac{K}{J} c_\epsilon \log K(1+o(1)). \end{aligned}$$

Thus,

$$S_{l,k}^{\mathrm{RL}} = \frac{K^{c_\epsilon}}{K^{-\frac{K}{J}c_\epsilon} + K^{c_\epsilon}}.$$

Therefore,

$$\begin{aligned} F_{l,k}^{\mathrm{RL}} &= \frac{K^{2S_{l,k}^{\mathrm{RL}}}}{K^{2S_{l,k}^{\mathrm{RL}}} + (L-1)K \cdot K^{-2S_{l,k}^{\mathrm{RL}}} + K^{2(2\widetilde{S}^{\mathrm{RL}}-1)}} \\ &= \frac{K^{4S_{l,k}^{\mathrm{RL}}-1}}{K^{4S_{l,k}^{\mathrm{RL}}-1} + (L-1) + K^{1-2S_{l,k}^{\mathrm{RL}}}}. \end{aligned}$$

It follows that

$$1 - F_{l,k}^{\mathrm{RL}} \leq \frac{L}{K^{4S_{l,k}^{\mathrm{RL}}-1}} = O\left(K^{1-\frac{4K^{c_\epsilon}}{K^{-\frac{K}{J}c_\epsilon}+K^{c_\epsilon}}}\right) = O\left(\frac{1}{K^3}\right),$$

where the last step follows from c_ϵ can be sufficiently large. Taking the sum on k, l, we complete the proof. □