On the Role of Reasoning Patterns in the Generalization Discrepancy of Long Chain-of-Thought Supervised Fine-Tuning

Zhaoyi Li[1,2,3]*, **Xiangyu Xi**[2*†], **Zhengyu Chen**[2], **Wei Wang**[2],
Gangwei Jiang[1,3], **Ranran Shen**[1], **Linqi Song**[3], **Ying Wei**[4†], **Defu Lian**[1†]
[1]University of Science and Technology of China, [2]Meituan LongCat Team,
[3]City University of Hong Kong, [4]Zhejiang University

Abstract

Supervised Fine-Tuning (SFT) on long Chain-of-Thought (CoT) trajectories has become a pivotal phase in building large reasoning models. However, how CoT trajectories from different sources influence the generalization performance of models remains an open question. In this paper, we conduct a comparative study using two sources of verified CoT trajectories generated by two competing models, `DeepSeek-R1-0528` and `gpt-oss-120b`, with their problem sets controlled to be identical. Despite their comparable performance, we uncover a striking paradox: lower training loss does not translate to better generalization. SFT on `DeepSeek-R1-0528` data achieves remarkably lower training loss, yet exhibits significantly worse generalization performance on reasoning benchmarks compared to those trained on `gpt-oss-120b`. To understand this paradox, we perform a multi-faceted analysis probing token-level SFT loss and step-level reasoning behaviors. Our analysis reveals a difference in reasoning patterns. `gpt-oss-120b` exhibits highly convergent and deductive trajectories, whereas `DeepSeek-R1-0528` favors a divergent and branch-heavy exploration pattern. Consequently, models trained with `DeepSeek-R1` data inherit inefficient exploration behaviors, often getting trapped in redundant exploratory branches that hinder them from reaching correct solutions. Building upon this insight, we propose a simple yet effective remedy of filtering out frequently branching trajectories to improve the generalization of SFT. Experiments show that training on selected `DeepSeek-R1-0528` subsets surprisingly improves reasoning performance by up to 5.1% on AIME25, 5.5% on BeyondAIME, and on average 3.6% on five benchmarks.

1 Introduction

Supervised Fine-Tuning (SFT) on long Chain-of-Thought (CoT) trajectories has become a standard and foundational phase in building Large Reasoning Models (LRMs) (Yang et al., 2025b; Lim et al., 2025). This common practice oftentimes initiates from sampling extensive reasoning trajectories from existing LRMs (teacher models) and verifying their correctness. These trajectories are subsequently fed into base models (student models) in the cold-start SFT phase (Guo et al., 2025; Abdin et al., 2025) to endow them with complex problem-solving capabilities. Building on this, a central question concerns *which LRM should be used to generate such reasoning trajectories*. A natural idea is to investigate which LRM's trajectories the student models fit the best during SFT (Tian et al., 2025; Zhang et al., 2025b).

To examine this assumption, we conduct a controlled comparative SFT study utilizing ***verified, correct*** reasoning trajectories elicited from two open-source, widely-adopted (Zheng et al., 2025; Shmidman et al., 2025; Yang et al., 2026) LRMs, `DeepSeek-R1-0528` (Guo et al., 2025) and `gpt-oss-120b` (OpenAI et al., 2025) (in short, `R1` and `gpt-oss`), in response to an ***identical set*** of $\sim$ 500K complex mathematical problems. Through SFT across four base

*Equal contributions
†Corresponding authors: xixy10@foxmail.com, ying.wei@zju.edu.cn, liandefu@ustc.edu.cn.

models (including different model families and scales), we observe a stark and consistent **generalization discrepancy**: training on R1 data yields much ***lower SFT training loss***, yet it results in notably ***inferior generalization performance*** on downstream reasoning benchmarks compared to gpt-oss. This naturally raises a research question: what properties of teacher-generated CoT trajectories truly affect their effectiveness for generalization ability?

While recent studies (Yang et al., 2026; Panigrahi et al., 2026) have highlighted the importance of SFT data diversity and teacher-student distribution alignment in this process, these metrics fail to explain the above generalization discrepancy with different teachers, which motivates us to seek more fine-grained differences. Inspired by emerging works (Minegishi et al., 2025; Bogdan et al., 2025; Jiang et al., 2025) that characterize the effectiveness of long CoT trajectories through inherent structural patterns, we delve into the specific role of these *reasoning patterns* in explaining the generalization discrepancy of different teacher models.

Concretely, we perform a multi-faceted analysis encompassing both token-level SFT loss analysis and step-level reasoning behavior analysis (Gandhi et al., 2025) for the training data generated by R1 and gpt-oss. ***At the token level***, we observe a severe long-tail distribution, where the primary SFT optimization signals stem from a small subset of high-loss *key reasoning tokens* indicating logical transitions. Although the average losses of these key tokens are quite close between R1 and gpt-oss, R1 data contains a much larger proportion of near-zero-loss tokens, explaining its lower overall SFT loss without necessarily improving on key reasoning signals. ***At the trajectory level***, we uncover a fundamental difference in reasoning patterns by automatically annotating each reasoning step into discrete reasoning behaviors using DeepSeek-V3.2 (DeepSeek-AI et al., 2025). gpt-oss focuses on deep and convergent deductive chains, whereas R1 favors a divergent reasoning pattern characterized by ***frequently branching into alternative ideas without deep deduction***, which introduces numerous exploratory paths that are redundant to the main reasoning path.

Crucially, we also demonstrate that reasoning patterns transfer to student models through SFT. Student models trained with R1 data often get trapped in redundant exploratory branches that hinder them from reaching the correct solutions. We further validate such reasoning pattern and the transfer behavior through two intervention experiments. On one hand, we randomly remove part of the reasoning steps from each trajectory and retrain the models. While models trained on pruned gpt-oss data suffer severe performance drops, those trained on pruned R1 data exhibit minimal degradation and, on some benchmarks, even gains. On the other hand, we propose to improve the generalization of SFT with a targeted data curation strategy: ***filtering out the most frequently branching trajectories from the R1 dataset*** with two carefully designed proxy metrics. We empirically demonstrate that ***training on this reduced R1 subset surprisingly improves reasoning performance*** by up to 5.1% on AIME25, 5.5% on BeyondAIME, and an average of 3.6% on five benchmarks.

In summary, our main contributions are three-fold: (1) We expose a stark generalization discrepancy in long CoT SFT, revealing that trajectories from R1 consistently induce lower training loss but notably worse generalization performance compared to gpt-oss under controlled settings. (2) We demonstrate that this discrepancy is intrinsically correlated with differing reasoning patterns in training data: R1 frequently branches into alternative reasoning paths without deep deduction, whereas gpt-oss provides more convergent deductive reasoning chains, which are inherited by student models through SFT. and (3) We propose a simple yet effective approach to improve the generalization of SFT by filtering out the most frequently branching trajectories from the original R1 dataset and empirically demonstrate that training on this reduced subset consistently improves reasoning performance.

2 Related Work

Distilling Reasoning Ability via Long CoT Supervised Fine-Tuning The advent of LRMs, such as OpenAI's o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), has shifted the paradigm of complex problem-solving toward prolonged thinking with branching, reflection, and backtracking (i.e., Long CoT reasoning (Yang et al., 2025b; Kopiczko et al., 2026)). SFT with reasoning trajectories output by LRMs has become a focal point of research for two critical purposes: distilling the complex reasoning capacities of stronger LRMs into

Table 1: SFT performance comparison of different base models trained on reasoning trajectories generated by `DeepSeek-R1` and `gpt-oss-120b` across five mathematical reasoning benchmarks. AIME24/25 results are averaged over 32 independent runs (`avg@32`). BeyondAIME and HMMT25 results are averaged over 10 independent runs (`avg@10`).

Base Model	Data Source	MATH500	AIME24	AIME25	BeyondAIME	HMMT25	Avg
Qwen2.5-7B	DeepSeek-R1	95.0%$\pm$0.02%	52.1%$\pm$0.03%	44.8%$\pm$0.03%	16.7%$\pm$0.02%	38.7%$\pm$0.03%	49.5%
	gpt-oss-120b	95.0%$\pm$0.02%	**64.5**%$\pm$0.03%	**46.7**%$\pm$0.03%	**23.9**%$\pm$0.03%	**42.9**%$\pm$0.03%	**54.6**% (+5.1%)
Qwen2.5-32B	DeepSeek-R1	98.4%$\pm$0.01%	78.5%$\pm$0.03%	73.5%$\pm$0.03%	37.7%$\pm$0.03%	69.6%$\pm$0.03%	71.5%
	gpt-oss-120b	**98.8**%$\pm$0.01%	**82.0**%$\pm$0.02%	**77.7**%$\pm$0.03%	**45.9**%$\pm$0.03%	**77.3**%$\pm$0.03%	**76.3**% (+4.8%)
Llama3.1-8B	DeepSeek-R1	86.2%$\pm$0.03%	20.1%$\pm$0.03%	18.3%$\pm$0.02%	4.5%$\pm$0.01%	18.5%$\pm$0.02%	29.5%
	gpt-oss-120b	**94.8**%$\pm$0.02%	**53.8**%$\pm$0.03%	**41.3**%$\pm$0.03%	**21.8**%$\pm$0.03%	**40.6**%$\pm$0.03%	**50.5**% (+21.0%)
Qwen3-8B	DeepSeek-R1	**96.8**%$\pm$0.02%	63.0%$\pm$0.03%	54.0%$\pm$0.03%	24.5%$\pm$0.03%	48.5%$\pm$0.03%	57.4%
	gpt-oss-120b	96.2%$\pm$0.02%	**70.6**%$\pm$0.03%	**58.9**%$\pm$0.03%	**32.5**%$\pm$0.03%	**55.3**%$\pm$0.03%	**62.7**% (+5.3%)

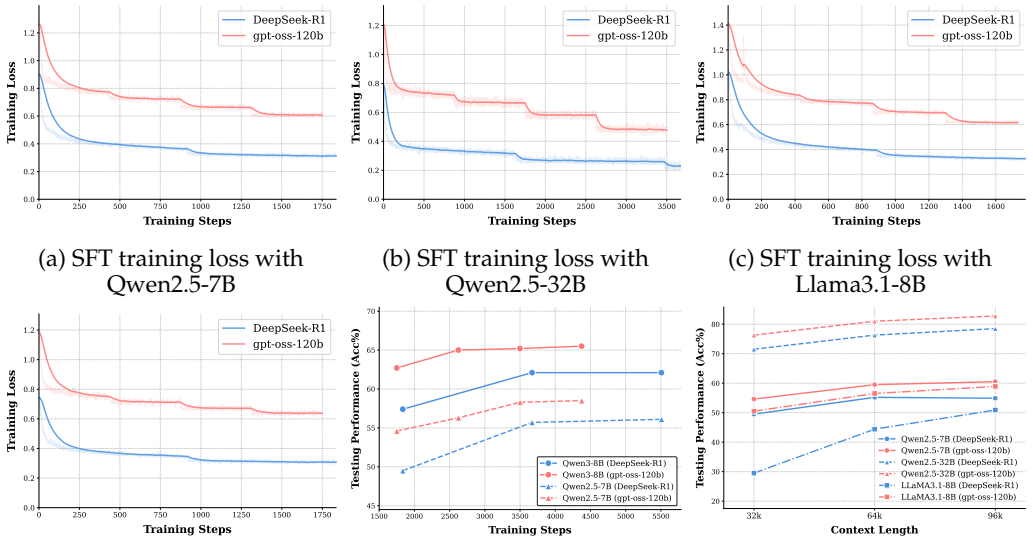
weaker models (Shridhar et al., 2023), and establishing a robust cold-start foundation to bootstrap further RL training in new frontier LRMs (Zhang et al., 2025a; Matsutani et al., 2026; Li et al., 2026b). Recent literature (Lim et al., 2025; Zheng et al., 2025) has recognized that the effectiveness of long CoT SFT mainly depends on how to select appropriate teacher models. Tian et al. (2025) demonstrated that not all correct reasoning trajectories contribute equally to student performance. Shmidman et al. (2025) showed empirical comparisons between different teacher traces have revealed discrepancies in downstream performance. Chandra et al. (2025), Jung et al. (2025), Panigrahi et al. (2026), and Yang et al. (2026) emphasized the "teacher-student compatibility": successful distillation requires striking a delicate balance among the diversity, informativeness, and distribution alignment of traces to optimally suit the student model. In contrast, this work focuses on the intrinsic properties (reasoning patterns) (Wang et al., 2025; Jiang et al., 2025; Gandhi et al., 2025) of the reasoning trajectories and analyzes the role of these reasoning patterns in the generalization of SFT.

Analyzing the Intrinsic Properties of Long CoT Reasoning Trajectories As the study of Long CoT progresses, recent literature has increasingly recognized that reasoning trajectories generated by different LRMs possess fundamentally distinct intrinsic properties and characteristic reasoning behaviors (Sun et al., 2025; Chen et al., 2025; Ballon et al., 2026), and different types of reasoning trajectories may exhibit varying degrees of learnability (Prasad et al., 2026). Recent efforts (Xiong et al., 2025; Minegishi et al., 2025; Li et al., 2025a; Jiang et al., 2025; Li et al., 2026a) have conceptualized reasoning trajectories as graphs, characterizing the reasoning properties with graph properties like radius, number of loops, and other GNN-captured features (Brody et al., 2022). Besides, micro-level token-related features have also drawn attention (Qian et al., 2025; Wang et al., 2025; Li et al., 2025b; Chen et al., 2026b), showing that a small fraction of critical tokens often drives the actual cognitive progress. Furthermore, mapping these trajectories into discrete cognitive behaviors and analyzing their relative proportions and transition probabilities have proven vital for characterizing reasoning patterns (Gandhi et al., 2025; Bogdan et al., 2025; Chen et al., 2026a). Building upon these analytical foundations, our systematically investigate how the reasoning patterns of long CoT trajectories dictate the efficacy and generalization of Long CoT SFT.

3 Comparing SFT with Trajectories from Two Advanced LRMs

To uncover the underlying principles of long CoT SFT, we conduct a controlled comparative study, in which we utilize reasoning trajectories from two open-weight, widely-adopted, and advanced LRMs: `DeepSeek-R1` (Guo et al., 2025) ("`DeepSeek-R1`" hereinafter refers to "`DeepSeek-R1-0528`"[1]) and `gpt-oss-120b` (OpenAI et al., 2025) to fine-tune base models. By ensuring that these trajectories **correspond to an identical set of prompts (questions)** and are **all verified to yield correct answers** (Tian et al., 2025; Shmidman et al., 2025), we eliminate confounding factors related to problem distribution or factual correctness, allowing us to empirically investigate how the distinct intrinsic properties of different teacher trajectories affect SFT and the resulting reasoning generalization of student base models. Specifically, our experimental settings are as below: **(1) SFT Data Collection:** Following Tian et al. (2025),

[1] https://huggingface.co/deepseek-ai/DeepSeek-R1-0528



(a) SFT training loss with Qwen2.5-7B

(b) SFT training loss with Qwen2.5-32B

(c) SFT training loss with Llama3.1-8B

(d) SFT training loss with Qwen3-8B

(e) Testing performance with varying training steps

(f) Testing performance with varying inference context length

Figure 1: (a $\sim$ d): SFT training loss comparison of different models trained on long CoT trajectories of `DeepSeek-R1` and `gpt-oss-120b`. (e) and (f): average testing performance on five benchmarks with varying training steps and inference context length. Blue/red curves refer to experiments with `gpt-oss-120b`/`DeepSeek-R1`-generated data, respectively.

we collect a high-quality dataset comprising approximately 500,000 challenging mathematical problems from publicly available datasets including OpenR1-Math-220k (Hugging Face, 2025), Big-Math-RL-Verified (Albalak et al., 2025), NuminaMath (LI et al., 2024) and so on. For each problem, we query both `DeepSeek-R1` and `gpt-oss-120b` to generate their respective Long CoT trajectories. To control data quality, we apply a rule-based verification pipeline to ensure that *all trajectories used in our experiments successfully arrive at the correct final answer*. The context length of the reasoning trajectories used for SFT is controlled below 32k; **(2) Base Models:** We select four representative (covering different model families and scales) open-weight base models as student models: Qwen2.5-7B (Qwen et al., 2025), Qwen2.5-32B (Qwen et al., 2025), Llama3.1-8B (Grattafiori et al., 2024), and Qwen3-8B (Yang et al., 2025a). Each model undergoes SFT on the `DeepSeek-R1` and `gpt-oss-120b` datasets independently. We maintain identical hyperparameter settings and optimize for *approximately the same number of training steps* across each pair of comparisons; and **(3) Evaluation Benchmarks:** The resulting models are evaluated on five representative mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), BeyondAIME (ByteDance-Seed, 2025), and HMMT25 (Balunović et al., 2025). The inference context limit is set to 32k by default, aligned with the training data. More details on our experimental setup can be found in Appendix C.

Empirical Observations on the SFT results. Based on the SFT results presented in Table 1 and Figure 1, we highlight two key observations: **(1) The Generalization Discrepancy.** As detailed in Table 1, student models fine-tuned on `gpt-oss-120b` reasoning trajectories consistently and significantly outperform those trained on `DeepSeek-R1` data. This performance advantage holds across all four evaluated base models, regardless of model scale or architecture. This universal superiority is non-trivial: given that both training datasets cover the exact same set of prompts and strictly consist of trajectories that arrive at the correct final answers, the generalization performance discrepancy cannot be attributed to the factual correctness or the distribution of training data. This naturally raises a research question: ***since both sets provide valid reasoning paths, what intrinsic factors make the consistent generalization discrepancy between models trained on `DeepSeek-R1` and `gpt-oss-120b` trajectories?*** **(2) The SFT Loss Discrepancy.** As an auxiliary finding, we observe an abnormal discrepancy in the training dynamics during SFT. As illustrated in Figure 1(a $\sim$ d), across all student models, the SFT training loss on `DeepSeek-R1` data converges to a remarkably lower level (approximately 0.2 $\sim$ 0.3). In contrast, the training loss on `gpt-oss-120b` data plateaus

at a significantly higher level (around 0.6). In the SFT phase, a lower training loss typically correlates with better alignment to the target distribution, indicating a better learnability of reasoning trajectories (Tian et al., 2025; Zhang et al., 2025b). However, our experiments shows the `gpt-oss-120b` trajectories produce consistently superior student models although much "harder" to fit during training. This discrepancy is particularly pronounced in base models such as Llama3.1-8B, where the average accuracy drops from 50.5% (`gpt-oss-120b`) to 29.5% (`DeepSeek-R1`) despite the latter achieving a much lower training loss.

Ruling out Confounding Factors. Before proceeding to the analysis of our observation, we conduct auxiliary experiments to ensure that our observations are not artifacts of the training or inference setups: **(1) Extending Training Steps:** One might argue that the `DeepSeek-R1` data simply requires more training steps to converge. To test this, we vary the SFT training steps (from 1500 steps to 4500 steps) until the test-set performance fully saturates. The results for two base models (Qwen2.5-7B and Qwen3-8B) are shown in Figure 1(e), we find that the performance gap persists: the `gpt-oss-120b`-trained models still significantly and consistently outperform the `DeepSeek-R1`-trained models, confirming that the discrepancy is rooted in the intrinsic properties rather than undertraining; and **(2) Scaling the Context Limit during Inference:** Long CoT reasoning trajectories from `DeepSeek-R1` are statistically longer than those from `gpt-oss-120b`. It is possible that `DeepSeek-R1`-trained student models perform poorly simply because their generations are prematurely truncated during evaluation (Shmidman et al., 2025). However, when we increase the maximum generation length (from 32k to 96k) for Qwen2.5-7B, Qwen2.5-32B and LLaMA3.1-8B (Qwen3-8B supports a maximum context length of 32k), at inference time, the `DeepSeek-R1`-trained students still fall behind `gpt-oss-120b`-trained ones, as depicted in Figure 1 (f).

Takeaway: under controlled settings, models trained on `DeepSeek-R1` data achieve much lower SFT loss, yet exhibit notably inferior generalization compared to `gpt-oss-120b`.

4 Comparing Reasoning Patterns in DeepSeek-R1 and gpt-oss-120b

To investigate the underlying causes of the empirical discrepancy in generalization and SFT loss observed in Section 3, we perform a multi-faceted analysis. We first deconstruct the SFT loss at the token-level and subsequently formalize the structural-level pattern differences of reasoning trajectories through reasoning behavior analysis. We show that `DeepSeek-R1`'s reasoning trajectories exhibit a lower density of key reasoning tokens and favor a divergent exploration pattern with repetitive branching, resulting in a much higher degree of structural redundancy than `gpt-oss-120b`. We then randomly delete reasoning steps from the original training data (from `DeepSeek-R1` and `gpt-oss-120b`) and re-train the base models with intervened data to further support our analysis. Due to the page limit, we only include part of the results here. More experiment results can be found in Appendix D.

(a) Qwen3-8B's token-level loss distribution after SFT.

(b) Most frequent tokens in the `DeepSeek-R1` data after SFT.

(c) Most frequent tokens in the `gpt-oss-120b` data after SFT.

Figure 2: Token-level SFT loss analysis for the Qwen3-8B model. (a) show the token-level loss distribution after the SFT training, where blue/red bars represent experiments with `DeepSeek-R1`/`gpt-oss-120b` trajectories. (b, c) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the (`DeepSeek-R1`, `gpt-oss-120b`) data.

Table 2: Representative key reasoning tokens in the high-loss tail part and their average loss before/after SFT training (average-loss-before-sft → average-loss-after-sft).

Data Source	Representative High-Loss Tokens
DeepSeek-R1 3.538 → 1.533	'Perhaps', 'Now', 'Another', 'Let', 'might', 'given', 'say', 'Similarly', 'earlier', 'probably'
gpt-oss-120b 4.081 → 1.784	'Check', 'indeed', 'Thus', 'yields', 'Given', 'Let', 'might', 'Evaluate', 'Actually', 'Wait', 'Now', 'consider', 'compute'

4.1 Token-Level SFT Loss Deconstruction and Analysis

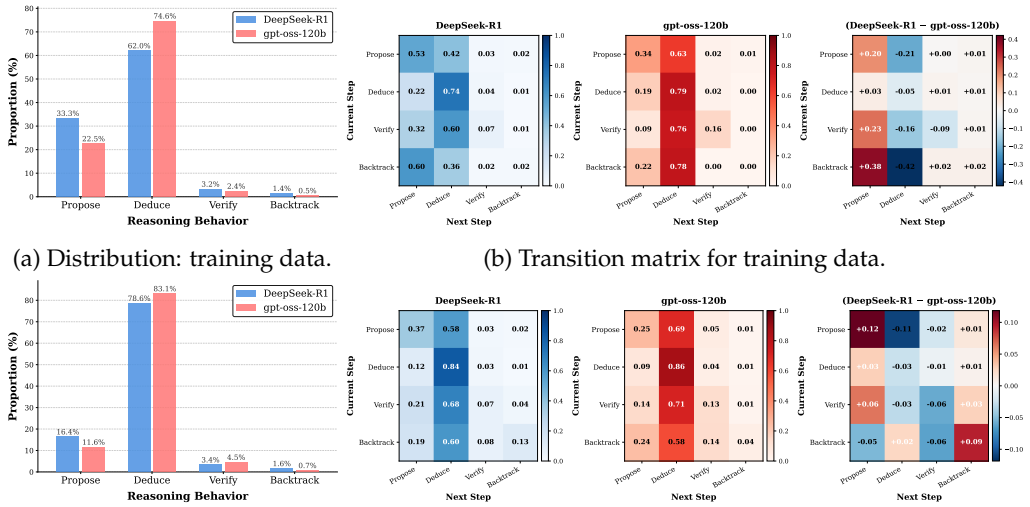
As a diagnostic entry-point, we calculate and visualize the token-level loss distribution for the base model before and after the SFT training with long CoT reasoning trajectories.

The Long-Tail Distribution of Token-level SFT Loss. As shown in Figures 2 (a), token-level loss exhibits a severe long-tail distribution after SFT training. The vast majority of tokens ($\sim$ 55% before training (see Figure 6 in the Appendix), $\sim$ 75% after training for the DeepSeek-R1 experiment) possess exceptionally low loss (< 0.1), forming the "head" part of the distribution. The remaining portion, namely the tokens exhibiting distinctly high loss, constitutes the "long tail" of the distribution and provides the main optimization signals during SFT. Notably, comparing the two data sources, the gpt-oss-120b trajectories possess a significantly thicker tail (higher proportion of high-loss tokens) compared to DeepSeek-R1.

Visualization of Most Frequent Tokens in the High-loss Tail Part. To understand what types of token constitute the high-loss tail part, we generate word clouds to visualize the most frequent tokens in the top 10% high-loss subset of all tokens (Figures 2 (b) and (c), larger tokens indicate higher frequency.). We observe that, in both experiment groups with DeepSeek-R1 and gpt-oss-120b data, the most frequent high-loss tokens are not the constituent tokens executing the internal details of individual reasoning steps; rather, they are the *key reasoning tokens* which dictate logical transitions and structural shifts (Qian et al., 2025; Wang et al., 2025) in the reasoning trajectories. However, the semantic nature of these *key reasoning tokens* diverges significantly between the two models: as summarized in Table 2: for DeepSeek-R1, representative high-loss tokens like 'Perhaps', 'probably' and 'Another' mainly indicate a **divergent exploration** phase, where the model branches from the current reasoning path to propose a new hypothesis or alternative problem-solving angle; for gpt-oss-120b, the most prominent tokens like 'Thus', 'indeed', and 'yields' signify a **convergent exploitation** phase, where the model strengthens and pushes forward the current reasoning path. Moreover, we observe a striking **dilution effect** of SFT loss. As shown in Table 2, when we aggregate the token-level loss for these representative key reasoning tokens, we find that their average loss values are actually **quite close** (e.g., after SFT: 1.533 for DeepSeek-R1 vs. 1.784 for gpt-oss-120b), and are significantly higher than the overall average loss. This reveals that the SFT loss discrepancy observed in Section 3 is essentially a *dilution effect*: the DeepSeek-R1 trajectories contain a lower concentration of dense reasoning transition signals, resulting in a larger proportion of low-loss "routine" tokens (e.g., simple calculation tokens like "3 + 2 = 5") that drag down the overall SFT loss.

4.2 Uncovering Thought Structure Differences via Reasoning Behavior Analysis

The token-level analysis implies that DeepSeek-R1 and gpt-oss-120b utilize different structure-level thinking patterns (Jiang et al., 2025; Chen et al., 2026a) (i.e., "divergent exploration" versus "convergent exploitation"). To further investigate this structural difference, we decompose the macro-level continuous reasoning trajectories into micro-level discrete reasoning steps and analyze the reasoning behaviors (Gandhi et al., 2025) underlying them. Building upon Bogdan et al. (2025); Gandhi et al. (2025), we categorize reasoning steps into four behaviors: **Propose** (branch to an alternative reasoning path), **Deduce** (linearly push forward the current reasoning path), alongside **Verify** and **Backtrack** (which collectively serve as uncertainty management), as defined in detail in Appendix C.3, Table 7. We utilize a powerful open-weight LLM, DeepSeek-V3.2 (DeepSeek-AI et al., 2025), with a



(a) Distribution: training data.

(b) Transition matrix for training data.

(c) distribution: generated solutions for AIME24.

(d) Transition matrix for generated AIME24 solutions.

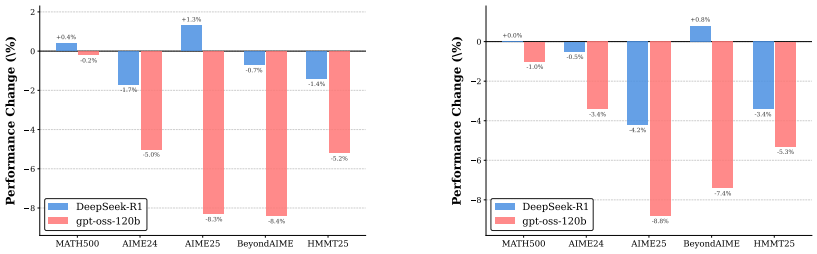
Figure 3: Reasoning behavior distributions (a, c) and transition matrices (b, d) for reasoning trajectories used for training and generated for solving AIME24 problems with fine-tuned Qwen3-8B. Blue/red objects: experiments with `DeepSeek-R1`/`gpt-oss-120b`-generated data.

clearly-defined and well-designed prompt template to automatically annotate the reasoning steps (see Appendix C.3 for the details of the annotation experiment).

Behavior Distribution and Markov Transition Matrix. Figure 3 (a) compares the behavior distribution in the original reasoning trajectories generated by `DeepSeek-R1` (blue bars) and `gpt-oss-120b` (red bars) that are used for SFT training. Quantitatively, we observe that `DeepSeek-R1` trajectories contain a significantly higher proportion of `Propose` steps (33.3%) compared to `gpt-oss-120b` (22.5%). In contrast, the proportion of `Deduce` steps in `DeepSeek-R1` (62.0%) is markedly lower than that in `gpt-oss-120b` (74.6%). Furthermore, the Markov transition matrices (Chen et al., 2026a) of the reasoning behavior (Figure 3 (b)) reveal distinct step-to-step behavior transition patterns: `DeepSeek-R1` exhibits notably higher probabilities of transitioning into the `Propose` state. Notably, it displays a more frequent `Propose` $\rightarrow$ `Propose` transition pattern (transition probability$=0.53$) in comparison with `gpt-oss-120b` (transition probability$=0.34$), indicating that `DeepSeek-R1` is more inclined to generate **consecutive reasoning branches** before committing to a specific deductive chain.

According to the behavioral definitions, `Propose` signifies exploring a new idea or branching to an alternative path, while `Deduce` represents executing linear, forward logical inferences. The above statistics reveal that `DeepSeek-R1` inherently favors a highly exploratory reasoning pattern characterized by frequently branching and pivoting. In contrast, `gpt-oss-120b` strongly leans toward convergent exploitation, focusing on deep, continuous deductive chains. The frequent `Propose` $\rightarrow$ `Propose` transitions in `DeepSeek-R1` suggest that ***the model often engages in continuous branching without executing deep deductive analysis on the newly proposed ideas*** (see Appendix E, Figure 11 for a specific example). This structural bias implies a high degree of *redundancy* within the `DeepSeek-R1` trajectories, where many shallow exploratory paths may not meaningfully contribute to the final logical resolution. Crucially, this thought structure is directly transferred to the student models during SFT. As shown in Figures 3 (c) and (d), the Qwen3-8B model fine-tuned on `DeepSeek-R1` data generates AIME24 solutions that inherit this exploratory distribution, exhibiting a 12% higher probability of `Propose` $\rightarrow$ `Propose` transitions compared to its `gpt-oss-120b`-trained counterpart. Consequently, the student model wastes extensive context on shallow branching, and sometimes iteratively re-proposes previously explored dead-ends (see a specific example in Appendix E, Figure 12). This inherited inefficiency ***hinders the student model's ability to converge on the correct answer, explaining its degraded generalization performance***.

Validating Exploratory Redundancy via Random Reasoning Step Deletion. To empirically validate our hypothesis that `DeepSeek-R1` trajectories contain a higher degree of structural re-



(a) Qwen3-8B results. (b) Qwen2.5-32B results.

Figure 4: Performance change ratio $((\mathrm{Acc}_{\mathrm{retrain}} - \mathrm{Acc}_{\mathrm{original}})/\mathrm{Acc}_{\mathrm{original}})$ on five benchmarks after randomly deleting 10% reasoning steps in each training trajectory. Blue/red bars represent experiments with `DeepSeek-R1`/`gpt-oss-120b`-generated data, respectively.

dundancy, we design a comparison experiment: for each training trajectory in both datasets, we **randomly** delete 10% of its reasoning steps and retrain the base models from scratch. If a trajectory is a dense, highly-dependent deductive chain, the deletion should break the logical flow and degrade performance more drastically. Conversely, if a trajectory has many redundant branches, the deletion should have a minor impact. The results in Figure 4 confirm this hypothesis. Models trained on the 10%-deleted `gpt-oss-120b` data suffer more significant performance drops. In contrast, models trained on the 10%-deleted `DeepSeek-R1` data exhibit minimal degradation. Surprisingly, in several instances (e.g., Qwen3-8B on MATH500 and AIME25; Qwen2.5-32B on BeyondAIME), the performance actually *increases*. This suggests that the `DeepSeek-R1` data inherently contain more redundant sub-paths. In fact, moderately truncating these exploratory steps may act as an implicit regularization (Chen et al., 2024), preventing the student model from overfitting to superficial trial-and-error formatting and forcing it to focus on the core deductive reasoning path.

Takeaway: `DeepSeek-R1` trajectories favor a pattern of frequently branching into alternative ideas without deep deduction, introducing more redundant reasoning paths, which is inherited by student models during SFT and potentially degrades their generalization.

5 Improving SFT by Filtering out Frequently Branching Trajectories

In Section 4, we show that `DeepSeek-R1` exhibits a reasoning pattern heavily biased toward divergent exploration, frequently branching into alternative ideas without executing deep deduction. When SFT on these long CoT trajectories, the base models internalize this pattern and get trapped in redundant exploratory branches, leading to inferior generalization performance on reasoning benchmarks (Section 3). To further validate this insight and provide a practical recipe for enhancing Long CoT SFT, we propose a targeted data intervention: ***filtering out the most frequently branching trajectories from the original*** **`DeepSeek-R1`** ***dataset***, which may mitigate the base model's tendency to internalize this reasoning pattern. Ideally, identifying such trajectories would involve annotating every reasoning step across the 500K reasoning trajectories with LLMs. However, executing this at scale is computationally

Table 3: SFT performance after removing trajectories with top difference ratio of reasoning step number (proxy metric 1) from the `DeepSeek-R1` generated dataset.

Base Model	Data Source	MATH500	AIME24	AIME25	BeyondAIME	HMMT25	Avg
Qwen3-8B (context len: 32K)	full `DeepSeek-R1` set	96.8%	63.0%	54.0%	24.5%	48.5%	57.4%
	removing top 10%	97.4% (+0.6%)	64.7%(+1.7%)	55.7%(+1.7%)	25.1%(+0.6%)	51.5%(+3.0%)	**58.9%** (+1.5%)
	removing top 20%	97.2% (+0.4%)	64.5% (+1.5%)	54.7% (+0.7%)	25.7% (+1.2%)	50.1% (+1.6%)	**58.4%** (+1.0%)
Qwen2.5-7B (context len: 32K)	full `DeepSeek-R1` set	95.0%	52.1%	44.8%	16.7%	38.7%	49.5%
	removing top 10%	95.0%	54.3%(+2.2%)	46.7%(+1.9%)	19.3%(+2.6%)	40.2%(+1.5%)	**51.1%** (+1.6%)
	removing top 20%	95.8%(+0.8%)	52.2%(+0.1%)	44.4%(-0.4%)	16.2%(-0.5%)	40.0%(+1.3%)	**49.7%** (+0.2%)
Qwen2.5-7B (context len: 96K)	full `DeepSeek-R1` set	93.6%	61.7%	53.4%	27.5%	38.3%	54.9%
	removing top 10%	95.8%(+2.2%)	65.6% (+3.9%)	55.2% (+1.8%)	30.3% (+2.8%)	42.4% (+4.1%)	**57.9%** (+3.0%)
	removing top 20%	95.6%(+2.0%)	64.8%(+3.1%)	53.1%(-0.3%)	29.2%(+1.7%)	39.3%(+1.0%)	**56.4%** (+1.5%)

Table 4: SFT performance using the bottom/top K% subset (in terms of the proportion of the reasoning steps containing "branching"-related keywords, proxy metric 2) of trajectories.

Base Model	Data Source	MATH500	AIME24	AIME25	BeyondAIME	HMMT25	Avg
Qwen3-8B (context len: 32K)	full DeepSeek-R1 set (original SFT)	96.8%±0.02%	63.0%±0.03%	54.0%±0.03%	24.5%±0.03%	48.5%±0.03%	57.4%
	bottom **90**% (experiment group)	97.0%±0.02%	64.7%±0.03%	55.5%±0.03%	24.9%±0.03%	48.8%±0.03%	**58.2**% (+0.8%)
	top **90**% (control group)	97.0%±0.02%	63.0%±0.03%	52.3%±0.03%	23.9%±0.03%	47.7%±0.03%	56.8% (-0.6%)
	bottom **80**% (experiment group)	97.0%±0.02%	64.7%±0.03%	55.7%±0.03%	27.1%±0.03%	52.3%±0.03%	**59.4**% (+2.0%)
	top **80**% (control group)	96.8%±0.02%	63.3%±0.03%	52.1%±0.03%	22.5%±0.02%	47.4%±0.03%	56.4% (-1.0%)
	bottom **50**% (experiment group)	97.0%±0.02%	66.5%±0.03%	59.1%±0.03%	25.1%±0.03%	50.5%±0.03%	**59.6**% (+2.2%)
	top **50**% (control group)	96.0%±0.02%	63.0%±0.03%	52.3%±0.03%	22.7%±0.02%	47.1% ±0.03%	56.2% (-1.2%)
Qwen2.5-7B (context len: 32K)	full DeepSeek-R1 set (original SFT)	95.0%±0.02%	52.1%±0.03%	44.8%±0.03%	16.7%±0.02%	38.7%±0.03%	49.5%
	bottom **90**% (experiment group)	95.4%±0.02%	53.7%±0.03%	46.5%±0.03%	19.3%±0.03%	41.8%±0.03%	**51.3**% (+1.8%)
	top **90**% (control group)	95.0%±0.02%	51.9%±0.03%	44.9%±0.03%	17.7%±0.02%	37.1%±0.03%	49.3% (-0.2%)
	bottom **80**% (experiment group)	94.6%±0.02%	55.5%±0.03%	47.2%±0.03%	18.5%±0.02%	40.1%±0.03%	**51.2**% (+1.7%)
	top **80**% (control group)	94.4%±0.02%	52.0%±0.03%	44.5%±0.03%	15.0%±0.02%	39.3%±0.03%	49.0% (-0.5%)
	bottom **50**% (experiment group)	95.4%±0.02%	51.9%±0.03%	46.0%±0.03%	19.1%±0.02%	41.0%±0.03%	**50.7**% (+1.2%)
	top **50**% (control group)	94.6%±0.02%	50.6%±0.03%	43.2%±0.03%	15.0%±0.02%	38.3%±0.03%	48.3% (-1.2%)
Qwen2.5-7B (context len: 96K)	full DeepSeek-R1 set (original SFT)	93.6%±0.02%	61.7%±0.03%	53.4%±0.03%	27.5%±0.02%	38.3%±0.03%	54.9%
	bottom **90**% (experiment group)	95.2%±0.02%	64.9%±0.03%	57.2%±0.03%	33.0%±0.03%	42.1%±0.03%	**58.5**% (+3.6%)
	top **90**% (control group)	93.4%±0.02%	61.6%±0.03%	54.0%±0.03%	26.6%±0.03%	37.1%±0.03%	54.5% (-0.4%)
	bottom **80**% (experiment group)	95.8%±0.02%	65.0%±0.03%	56.8%±0.03%	32.1%±0.03%	40.5%±0.03%	**58.0**% (+3.1%)
	top **80**% (control group)	93.2%±0.02%	60.8%±0.03%	52.4%±0.03%	27.0%±0.03%	39.3%±0.03%	54.5% (-0.4%)
	bottom **50**% (experiment group)	94.4%±0.02%	61.0%±0.03%	54.1%±0.03%	29.2%±0.03%	40.5%±0.03%	**55.8**% (+0.9%)
	top **50**% (control group)	90.4%±0.02%	54.0%±0.03%	49.9%±0.03%	23.9%±0.03%	34.5%±0.03%	50.5% (-4.4%)

prohibitive: there are over 200 steps in each trajectory on average. Therefore, we design two *proxy metrics* to identify the frequently branching trajectories.

Proxy 1: Difference Ratio of Reasoning Step Number. Our first proxy leverages the paired nature of our datasets. For the exact same prompt, we compute the *difference in the number of reasoning steps between the* `DeepSeek-R1` *and* `gpt-oss-120b` *trajectories, normalized by the total steps of the* `DeepSeek-R1` *trajectory*. Intuitively, a higher difference ratio indicates that the `DeepSeek-R1` trajectory contains significantly more redundant exploration steps compared to its more deductive `gpt-oss-120b` counterpart. Based on this metric, we remove the trajectories with the highest difference ratios (top 10% and top 20%) and retrain the models, ensuring that all other settings and total training steps remain identical to the original SFT experiments. As shown in Table 3, this simple data filtering strategy yields consistent performance gains. For instance, by removing the 10% trajectories, the average accuracy on five benchmarks of Qwen3-8B and Qwen2.5-7B improves by 1.5% and 3.0%, respectively.

Proxy 2: Proportion of Steps Containing Branching Keywords. Proxy 1 requires paired data from `DeepSeek-R1` and `gpt-oss-120b`. To establish a standalone filtering criterion, we introduce proxy 2, in which we calculate the proportion of reasoning steps within each `DeepSeek-R1` trajectory that contain explicit branching signal tokens (e.g., `"Perhaps"`, `"Another"`, `"Alternatively"`). Building upon our token-level findings in Section 4, trajectories with a high density of reasoning steps containing such tokens are heavily populated with exploratory branches. Then we construct experiments (experiment group) by retaining the trajectories with the *lowest* proportion of reasoning steps that contain branching keywords (`bottom K%`), and corresponding *control experiments* by retaining those with the *highest* proportion (`top K%`). The results are shown in Table 4. Training on the subsets with fewer branching patterns (experiment groups) consistently enhances reasoning performance over the baseline trained on the full-set of `DeepSeek-R1` data. Notably, using only the `bottom 50%` of the data, Qwen3-8B achieves an average performance gain of 2.2%, including a remarkable 5.1% absolute improvement on the AIME25 benchmark. Similarly, Qwen2.5-7B sees an average improvement of 3.6% with the `bottom 90%` part of the data. In contrast, the control groups consistently decrease the trained models' performance compared to the original baseline. This controlled contrast strongly validates our core insight: the pattern that frequently branching without deep deduction in `DeepSeek-R1` data acts as structural redundancy during SFT. Filtering out these trajectories prevents the base models from overfitting to inefficient and divergent exploration behaviors, encouraging them to learn more deductive and convergent reasoning paths, which drives better generalization.

Ablation Study by Filtering out the Longest Trajectories A natural question can be raised here: *"Do the performance gains simply come from that we filter out those longest reasoning*

Table 5: Results of the ablation study. We compare the SFT generalization performance of filtering out the longest 20% trajectories with our proposed proxy metric 1 and proxy metric 2, where we filter out the most frequently branching trajectories.

Base Model	Data Source	MATH500	AIME24	AIME25	BeyondAIME	HMMT25	Avg
Qwen3-8B (context len: 32K)	full DeepSeek-R1 set (original SFT)	96.8$_{\pm 0.02\%}$	63.0$_{\pm 0.03\%}$	54.0$_{\pm 0.03\%}$	24.5$_{\pm 0.03\%}$	48.5$_{\pm 0.03\%}$	57.4%
	filtering out 20% with proxy 1	97.2$_{\pm 0.02\%}$	64.5$_{\pm 0.03\%}$	54.7$_{\pm 0.03\%}$	25.7$_{\pm 0.03\%}$	51.5$_{\pm 0.03\%}$	**58.4**% (+1.0%)
	filtering out 20% with proxy 2	97.0$_{\pm 0.02\%}$	64.7$_{\pm 0.03\%}$	55.7$_{\pm 0.03\%}$	27.1$_{\pm 0.03\%}$	52.3$_{\pm 0.03\%}$	**59.4**% (+2.0%)
	ablation: filtering out the longest 20%	96.0$_{\pm 0.02\%}$	61.0$_{\pm 0.03\%}$	46.9$_{\pm 0.03\%}$	23.3$_{\pm 0.03\%}$	45.1$_{\pm 0.03\%}$	54.5% (-2.9%)
Qwen2.5-7B (context len: 32K)	full DeepSeek-R1 set (original SFT)	95.0$_{\pm 0.02\%}$	52.1$_{\pm 0.03\%}$	44.8$_{\pm 0.03\%}$	16.7$_{\pm 0.02\%}$	38.7$_{\pm 0.03\%}$	49.5%
	filtering out 20% with proxy 1	95.8$_{\pm 0.02\%}$	52.2$_{\pm 0.03\%}$	44.4$_{\pm 0.03\%}$	16.2$_{\pm 0.02\%}$	40.0$_{\pm 0.03\%}$	**49.7**% (+0.2%)
	filtering out 20% with proxy 2	94.6$_{\pm 0.02\%}$	55.5$_{\pm 0.03\%}$	47.2$_{\pm 0.03\%}$	18.5$_{\pm 0.02\%}$	40.1$_{\pm 0.03\%}$	**51.2**% (+1.7%)
	ablation: filtering out the longest 20%	93.6$_{\pm 0.02\%}$	44.7$_{\pm 0.03\%}$	38.0$_{\pm 0.03\%}$	13.1$_{\pm 0.02\%}$	32.7$_{\pm 0.03\%}$	44.4% (-5.1%)
Qwen2.5-7B (context len: 96K)	full DeepSeek-R1 set (original SFT)	93.6$_{\pm 0.02\%}$	61.7$_{\pm 0.03\%}$	53.4$_{\pm 0.03\%}$	27.5$_{\pm 0.02\%}$	38.3$_{\pm 0.03\%}$	54.9%
	filtering out 20% with proxy 1	95.6$_{\pm 0.02\%}$	64.8$_{\pm 0.03\%}$	53.1$_{\pm 0.03\%}$	29.2$_{\pm 0.03\%}$	39.3$_{\pm 0.03\%}$	**56.4**% (+1.5%)
	filtering out 20% with proxy 2	95.8$_{\pm 0.02\%}$	65.0$_{\pm 0.03\%}$	56.8$_{\pm 0.03\%}$	32.1$_{\pm 0.03\%}$	40.5$_{\pm 0.03\%}$	**58.0**% (+3.1%)
	ablation: filtering out the longest 20%	94.0$_{\pm 0.02\%}$	51.5$_{\pm 0.03\%}$	41.9$_{\pm 0.03\%}$	20.8$_{\pm 0.03\%}$	33.8$_{\pm 0.03\%}$	48.4% (-6.5%)

trajectories?" In our experiments, we find that the trajectories filtered out with our proposed two proxy metrics have a very limited overlap with the trajectories filtered out with the length metric. We also carefully conduct an ablation study by filtering out 20% longest CoT trajectories and retraining the base model.

The results of the experiment study are shown in Table 5, which indicates that simply filtering out the longest trajectories cannot achieve similar performance gains with our proposed approaches. This further demonstrates that our main results should not be attributed to filtering out the longest trajectories, but attributed to filtered out trajectories with the most branching reasoning behaviors.

Takeaway: filtering out the most frequently branching trajectories from the original `DeepSeek-R1` data can starkly and consistently improve SFT generalization performance.

6 Conclusion

We investigated the generalization discrepancy in Long CoT SFT and demonstrated that the intrinsic reasoning patterns of the training data serve as a critical factor shaping model performance. Our multi-faceted analysis revealed that the pattern of frequent, divergent branching without deep deduction in long CoT trajectories introduces structural redundancy, causing student models to be trapped in redundant exploratory branches that hinder them from reaching the correct solutions. By removing the most frequently branching trajectories, we significantly and consistently improved the generalization of SFT across benchmarks.

Acknowledgments

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project No. 2025ZD0123403. This work was also supported in part by Scientific Research Innovation Capability Support Project for Young Faculty and National Natural Science Foundation of China (NSFC) No. U24A20253. We sincerely thank the anonymous reviewers for their insightful discussions and constructive feedback that helped improve this work. We are also grateful to Wanxu Cai (Tsinghua University & Meituan LongCat), Deyang Kong (Peking University & Meituan LongCat) and Qi Guo (Peking University & Meituan LongCat) for their assistance with the experiments.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models, 2025. URL https://arxiv.org/abs/2502.17387.

Marthe Ballon, Brecht Verbeken, Vincent Ginis, and Andres Algaba. Probing the trajectories of reasoning traces in large language models. *arXiv preprint arXiv:2601.23163*, 2026.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=VnSlfeRCaU.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

ByteDance-Seed. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads. https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME, 2025.

Abhranil Chandra, Ayush Agrawal, Arian Hosseini, Sebastian Fischmeister, Rishabh Agarwal, Navin Goyal, and Aaron Courville. Shape of thought: When distribution matters more than correctness in reasoning tasks. *arXiv preprint arXiv:2512.22255*, 2025.

Changyu Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5872–5900, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.320. URL https://aclanthology.org/2024.acl-long.320/.

Qiguang Chen, Yantao Du, Ziniu Li, Jinhao Liu, Songyao Duan, Jiarui Guo, Minghao Liu, Jiaheng Liu, Tong Yang, Ge Zhang, et al. The molecular structure of thought: Mapping the topology of long chain-of-thought reasoning. *arXiv preprint arXiv:2601.06002*, 2026a.

Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. Think deep, not just long: Measuring llm reasoning effort via deep-thinking tokens. *arXiv preprint arXiv:2602.13517*, 2026b.

Yida Chen, Yuning Mao, Xianjun Yang, Suyu Ge, Shengjie Bi, Lijuan Liu, Saghar Hosseini, Liang Tan, Yixin Nie, and Shaoliang Nie. Your thoughts tell who you are: Characterize the reasoning patterns of lrms. *arXiv preprint arXiv:2509.24147*, 2025.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting

Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STars. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=QGJ9ttXLTy.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz

Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bon-

trager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6501–6525, 2025.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R0dC7Xzwbk.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dawid J Kopiczko, Sagar Vaze, Tijmen Blankevoort, and Yuki M Asano. Data repetition beats data scaling in long-cot supervised fine-tuning. *arXiv preprint arXiv:2602.11149*, 2026.

Chenghao Li, Chaoning Zhang, Yi Lu, Shuxu Chen, Xudong Wang, Jiaquan Zhang, Zhicheng Wang, Zhengxun Jin, Kuien Liu, Sung-Ho Bae, et al. Understanding chain-of-thought in large language models via topological data analysis. *arXiv preprint arXiv:2512.19135*, 2025a.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin

Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Siyi Li, Jiajun Shi, Shiwen Ni, Ge Zhang, Shuaimin Li, Shijian Wang, Zhoufutu Wen, Yizhi Li, Hamid Alinejad-Rokny, Jiaheng Liu, Min Yang, and Wenhao Huang. Cotjudger: A graph-driven framework for automatic evaluation of chain-of-thought efficiency and redundancy in lrms, 2026a. URL https://arxiv.org/abs/2603.07078.

Xinran Li, Guangda Huzhang, Siqi Shen, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Jun Zhang. Getting your LLMs ready for reinforcement learning with lightweight SFT. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=yezWGJmODg.

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, et al. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *arXiv preprint arXiv:2510.13554*, 2025b.

Junghwan Lim, Sungmin Lee, Dongseok Kim, Taehyun Kim, Eunhwan Park, Jeesoo Lee, Jeongdoo Lee, Junhyeok Lee, Wai Ting Cheung, Dahye Choi, et al. Motif-2-12.7 b-reasoning: A practitioner's guide to rl training recipes. *arXiv preprint arXiv:2512.11463*, 2025.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=N2lMNqJsBw.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Topology of reasoning: Understanding large reasoning models through reasoning graph properties. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=o1g8NWkxqf.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham M. Kakade, and Surbhi Goel. In good GRACES: Principled teacher selection for knowledge distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=m276fke38H.

Archiki Prasad, Mandar Joshi, Kenton Lee, Mohit Bansal, and Peter Shaw. Effective reasoning chains reduce intrinsic dimensionality. *arXiv preprint arXiv:2602.09276*, 2026.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=E1FrjgaG1J.

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. URL https://arxiv.org/abs/2412.15115.

Shaltiel Shmidman, Asher Fredman, Oleg Sudakov, and Meriem Bendris. Learning to reason: Training llms with gpt-oss or deepseek r1 reasoning traces. *arXiv preprint arXiv:2511.19333*, 2025.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *International Conference on Machine Learning*, pp. 57854–57885. PMLR, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie

Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Xiaoyu Tian, Yunjie Ji, Haotian Wang, Shuaiting Chen, Sitong Zhao, Yiping Peng, Han Zhao, and Xiangang Li. Not all correct answers are equal: Why your distillation source matters. *arXiv preprint arXiv:2505.14464*, 2025.

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*, 2025. URL https://openreview.net/forum?id=yfcpdY4gMP.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Mapping the minds of llms: A graph-based analysis of reasoning llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17762–17774, 2025.

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, 2025a. URL https://arxiv.org/abs/2505.09388.

Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning. In *Forty-second International Conference on Machine Learning*, 2025b.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, et al. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. *arXiv preprint arXiv:2601.14249*, 2026.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025a.

Dylan Zhang, Qirun Dai, and Hao Peng. The best instruction-tuning data are those that fit. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=4jFSekBaDT.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Chujie Zheng, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Junrong Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, et al. Stabilizing reinforcement learning with llms: Formulation and practices. *arXiv preprint arXiv:2512.01374*, 2025.

A Disclosure of LLM Use

In compliance with the formatting and ethical guidelines of the conference, we disclose the use of Large Language Models (LLMs) in the preparation of this manuscript. Specifically, LLMs were employed strictly as assistive tools for the following tasks: (1) polishing the writing, correcting typos and grammar errors, and resolving LaTeX syntax issues; and (2) assisting in the refinement of `matplotlib` Python scripts for visualizing data and making figures.

We explicitly state that all core intellectual contributions are entirely the work of the human authors, including the conceptualization of the research ideas, the formulation of the experimental pipeline, the execution of the experiments, and the derivation of the scientific conclusions. **No LLMs were utilized to generate novel scientific insights or to design the experimental methodologies presented in this paper.**

B Limitations

Our empirical study primarily focuses on complex mathematical reasoning benchmarks, which serve as the standard testbed for current Large Reasoning Models. Future work should investigate whether the observed generalization discrepancy and the efficacy of our structural filtering strategy seamlessly extend to other domains requiring reasoning, such as code generation or agentic tasks. In our work, we adopt reasoning trajectories of two widely-used open-source models, `DeepSeek-R1-0528` and `gpt-oss-120b`. The key motivation of this work is to study the reasoning patterns of different LRMs and their effect on the generalization performance of SFT. We leave the discussion of more teacher models (Team et al., 2026; DeepSeek-AI et al., 2025) in the future work.

C Experimental Setup

C.1 SFT Training Setup

Following Tian et al. (2025), we collect a high-quality dataset comprising approximately 500,000 challenging mathematical problems from publicly available datasets including OpenR1-Math-220k[2], Big-Math-RL-Verified (Albalak et al., 2025), NuminaMath (LI et al., 2024) and so on. For each problem, we query both `DeepSeek-R1-0528` and `gpt-oss-120b` to generate their respective Long CoT trajectories. To rigorously control data quality, we apply a rule-based verification pipeline to ensure that *all trajectories used in our experiments successfully arrive at the correct final answer*. This step guarantees that any observed performance differences stem from the intrinsic structural properties of the reasoning paths rather than factual correctness. The length (token number) of the reasoning trajectories used for SFT training is strictly controlled under 32k. We select four representative (covering different model families including Qwen2.5, Qwen3 and LLaMA3.1, and different model scales from 7B to 32B) open-weight base models as student models: Qwen2.5-7B[3] (Qwen et al., 2025), Qwen2.5-32B[4] (Qwen et al., 2025), Llama3.1-8B[5] (Grattafiori et al., 2024), and Qwen3-8B[6] (Yang et al., 2025a). Each model undergoes SFT on the `DeepSeek-R1` and `gpt-oss-120b` datasets independently. We maintain identical hyperparameter settings and optimize for the approximately same number of training steps across each pair of comparisons. For all of the SFT experiments, we use the Adam optimizer (Kingma & Ba, 2014) with the β_1=0.9, β_2=0.95, gradient clip = 1.0, training sequence length = 32768, the Cosine learning rate scheduler with a warmup fraction = 0.05 and a minimum learning rate = 5e-6. For other main training hyperparameters, including train step, (initial) learning rate, global batch size and micro batch size, please refer to Table 6.

[2] https://huggingface.co/datasets/open-r1/OpenR1-Math-220k
[3] https://huggingface.co/Qwen/Qwen2.5-7B
[4] https://huggingface.co/Qwen/Qwen2.5-32B
[5] https://huggingface.co/meta-llama/Llama-3.1-8B
[6] https://huggingface.co/Qwen/Qwen3-8B-Base

Table 6: Training hyperparameters for the long CoT SFT experiments in the main table.

Base Model	Data Source	train step	learning rate	global batch size	micro batch size
Qwen2.5-7B	DeepSeek-R1	1836	$1e-4$	128	1
	gpt-oss-120b	1750	$1e-4$	128	1
Qwen2.5-32B	DeepSeek-R1	3672	$8e-5$	64	1
	gpt-oss-120b	3501	$8e-5$	64	1
Llama3.1-8B	DeepSeek-R1	1776	$1e-4$	128	1
	gpt-oss-120b	1733	$1e-4$	128	1
Qwen3-8B	DeepSeek-R1	1835	$1e-4$	128	1
	gpt-oss-120b	1749	$1e-4$	128	1

C.2 Inference Setup

The trained models are evaluated on five representative mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), BeyondAIME (ByteDance-Seed, 2025), and HMMT25 (Balunović et al., 2025). During all inference experiments, we set the temperature = 0.6, the repetition penalty = 1.0, top-k = 20, top-p=0.95. To ensure consistency with the training phase and accommodate complex reasoning chains, we set the maximum output token limit for each path to 32K tokens by default. We extract the final answer from the model's output by parsing the content within the last \boxed{} command. The extracted answer is then compared against the ground truth using the math-verify library[7] to determine correctness.

C.3 Reasoning Behavior Analysis Setup

The detailed definition of four common reasoning behaviors (including `Propose`, `Deduce`, `Verify`, and `Backtrack`) is provided in Table 7. The prompt template that is used for annotating reasoning steps is provided in Figure 5. For each annotation, we provide `DeepSeek-V3.2` (DeepSeek-AI et al., 2025) (chat version) with two consecutive reasoning steps (i.e., the 'previous reasoning step' and the 'current reasoning step') and ask it to judge which behavior the logical transition (from the 'previous reasoning step' to the 'current reasoning step') most likely belongs to. We set the temperature = 0. to get the classification result with the largest probability. In our experiments, we randomly sample 100 trajectories from the whole training set (for both of `DeepSeek-R1-0528` data and `gpt-oss-120b` data, their sampled trajectories corresponds to the same set of problems), resulting in 25978 reasoning steps for `DeepSeek-R1-0528` and 11316 reasoning steps for `gpt-oss-120b`, for our analysis. For the experiments with the AIME24 problems, we use the whole set of AIME problems for our analysis.

Besides, we split the reasoning trajectories into reasoning steps following a four-level split procedure: we use the four signal strings ('$\backslash n \backslash n$', '$\backslash n$', '?', '.') one-by-one to split the original trajectories into fine-grained steps.

Note that the design of our prompt template for behavior annotation is not purely out of intuition: we construct the initial version of our prompt based on the prompt templates that are used in Chen et al. (2026a), Gandhi et al. (2025), and Jiang et al. (2025), then we leverage a human-in-the-loop framework to continuously refine our prompt template. For instance, we find that even the state-of-the-art LLM annotators (DeepSeek-AI et al., 2025; Team et al., 2026) sometimes mistakenly use the label of the "previous step" to annotate the "current step". Hence we add a critical rule for annotation to guide LLM annotators to concentrate on the action in the "current step". To rigorously validate the reliability of our LLM-based annotation pipeline, we conducted a human evaluation on a random subset of 200 reasoning steps. The LLM annotations achieved a high agreement with expert human labels (accuracy = 92.5%), confirming the semantic stability of our behavioral taxonomy.

[7] https://github.com/huggingface/Math-Verify

Table 7: Definition of four common reasoning behaviors in long CoT reasoning trajectories.

Behavior	Definition
`Propose`	Explore a new idea, set up a hypothesis, or suggest an alternative path. It represents the "divergent" phase of reasoning.
`Deduce`	Execute a mathematical operation or make a direct logical inference based on the *immediately preceding* established facts or the current hypothesis. It represents the "convergent, linear" phase.
`Verify`	Pause the forward progression to double-check an intermediate calculation, verify a condition, or assess if the current path makes sense.
`Backtrack`	Explicitly realize an error or a dead end, reject the current reasoning branch, and retreat to a previous state or prepares to start over.

D More Experiment Results

D.1 Additional Token-level SFT Loss Analysis Results

We show the additional token-level SFT loss analysis results with Qwen3-8B, Qwen2.5-7B, and Llama3.1-8B in Figure 6, Figure 7, and Figure 8, respectively. The statistical results are **highly aligned with** the results with Qwen3-8B shown in the main context (Figure 2).

D.2 Additional Reasoning Behavior Analysis Results

The additional reasoning behavior analysis results with Llama3.1-8B are shown in Figure 9. The statistical results are **highly aligned with** the results with Qwen3-8B shown in the main context (Figure 3).

D.3 Additional Random Reasoning Step Deletion Results

To explore more about the random reasoning step deletion experiments, we adopt different deletion ratio (i.e., delete 10%, 20%, and 30% reasoning steps from each long CoT trajectories). The experiment results with Qwen3-8B are shown in Figure 10: our observation that performance degradation of `gpt-oss-120b`-distilled models is much more drastic than `DeepSeek-R1-0528`-distilled models consistently holds for all three deletion ratios. We observe that when randomly deleting 20% steps from each `DeepSeek-R1` reasoning trajectory, the re-trained Qwen3-8B's performance on BeyondAIME (`avg@10`) even increases by relatively 5.0% (normalized by the original performance). These results together validate our observation on the redundancy of `DeepSeek-R1` reasoning trajectories again.

D.4 Additional Results of Improving SFT through Filtering out the Most Frequently Branching Trajectories

We provide more experiment results on re-training Qwen2.5-7B after filtering out part of the frequently branching SFT training data, where we set the context limit = 32K at inference time. The results are shown in Table 8, which support the key conclusion presented in the main text.

E Case Study: Highly Exploratory Reasoning Trajectories Generated by `DeepSeek-R1-0528` and Its Distilled Student Model

This section contains two specific reasoning trajectory examples (snippets): one is a highly exploratory (repetitively branching) snippet of a reasoning trajectory generated by `DeepSeek-R1-0528` (in the SFT training set, Figure 11); another is a highly exploratory (repetitively branching) snippet of a reasoning trajectory generated by the Qwen3-8B that

Reasoning Behavior Annotation Template

You are an expert cognitive scientist and logician analyzing the internal Long "Chain of Thought" (CoT) reasoning trajectories of Large Language Models.
Your task is to classify the given reasoning step into one of four strictly defined cognitive actions: '[Propose, Deduce, Verify, Backtrack]'.

User Prompt:
You are provided with two consecutive reasoning steps, [PREVIOUS STEP] and [CURRENT STEP], in a Long CoT reasoning trace.
Please analyze the [CURRENT STEP] and assign ONE of the following four labels.
Note that [PREVIOUS STEP] here is providing you with some necessary context to assist you for analyzing.

Label Definitions & Constraints:

1. Propose (Hypothesis & Exploration)
- **Definition:** The model is exploring a new idea, setting up a hypothesis, or suggesting an alternative path. It represents the "divergent" phase of reasoning.
- **Significance:** High frequency of this label indicates a highly exploratory, tree-like search structure.

2. Deduce (Sequential Deduction)
- **Definition:** The model is executing a mathematical operation or making a direct logical inference based on the *immediately preceding* established facts or the current hypothesis. It represents the "convergent, linear" phase.
- **Significance:** Continuous sequences of this label indicate a dense, high-dependency deductive chain.

3. Verify (Self-Reflection & Checking)
- **Definition:** The model pauses its forward progression to double-check an intermediate calculation, verify a condition, or assess if the current path makes sense *without yet abandoning it*.

4. Backtrack (Error Correction & Path Abandonment)
- **Definition:** The model explicitly realizes an error or a dead end, rejects the current reasoning branch, and retreats to a previous state or prepares to start over.

Critical Rules For Annotation:

1. Focus STRICTLY on the ACTION in [CURRENT STEP].
The [PREVIOUS STEP] is strictly for context. Do NOT assign a label based on the tone or action of the [PREVIOUS STEP].

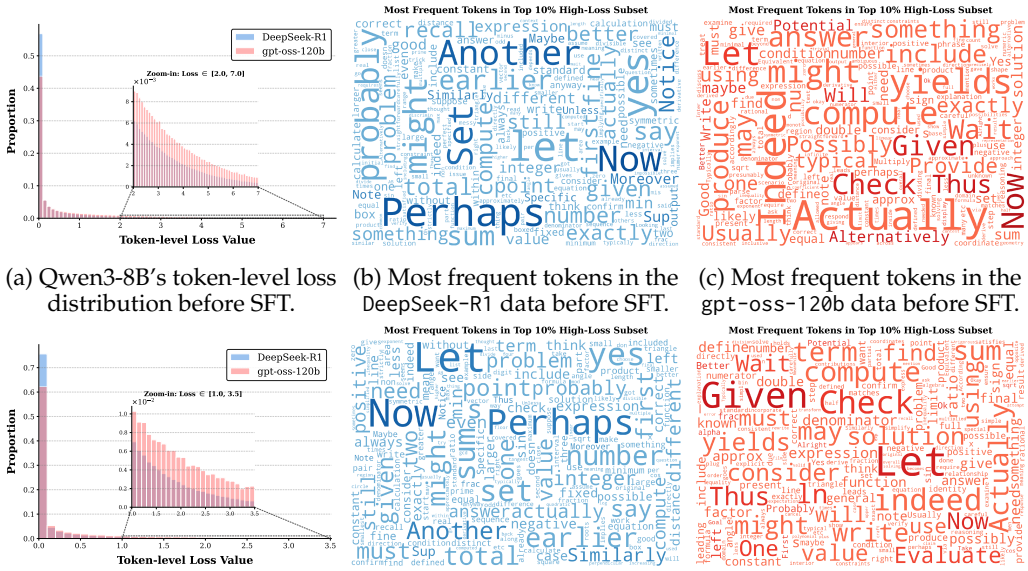
2. The "Pivot vs. Progress" Test (Crucial for distinguishing Propose vs. Deduce).
- **Progress (Label → Deduce):** If [CURRENT STEP] simply executes the math, unpacks the logic, or states the direct consequence of the [PREVIOUS STEP] (e.g., solving the equation just proposed), it is making forward progress. *Key signs: "Thus", "So", "Which means", "That would change...", or direct mathematical formulas.*
- **Pivot (Label → Propose):** If [CURRENT STEP] shifts the focus to a NEW angle, introduces a NEW speculation, or brainstorms a different aspect of the problem, it is opening a new branch. *Key signs: The explicit use of words like "Perhaps...", "Alternatively...", "What if...", "Another way..." in the [CURRENT STEP].*

Output Format:
Return exactly one label and nothing else:
'<action>' (action here can be 'Propose', 'Deduce', 'Verify', 'Backtrack'.)

Here is the reasoning step to analyze:
[PREVIOUS STEP]:
{put the previous reasoning step here}
[CURRENT STEP]:
{put the current reasoning step here}

Figure 5: The prompt template for annotating reasoning steps with four behavior labels.

was trained on `DeepSeek-R1-0528` trajectories (a generated solution for a evaluation problem, Figure 12). We can observe that the Qwen3-8B base model after SFT inherits the frequently branching reasoning pattern in the `DeepSeek-R1` reasoning trajectories, which potentially hinder the trained model from reaching the correct solutions.



(a) Qwen3-8B's token-level loss distribution before SFT.

(b) Most frequent tokens in the `DeepSeek-R1` data before SFT.

(c) Most frequent tokens in the `gpt-oss-120b` data before SFT.

(d) Qwen3-8B's token-level loss distribution after SFT.

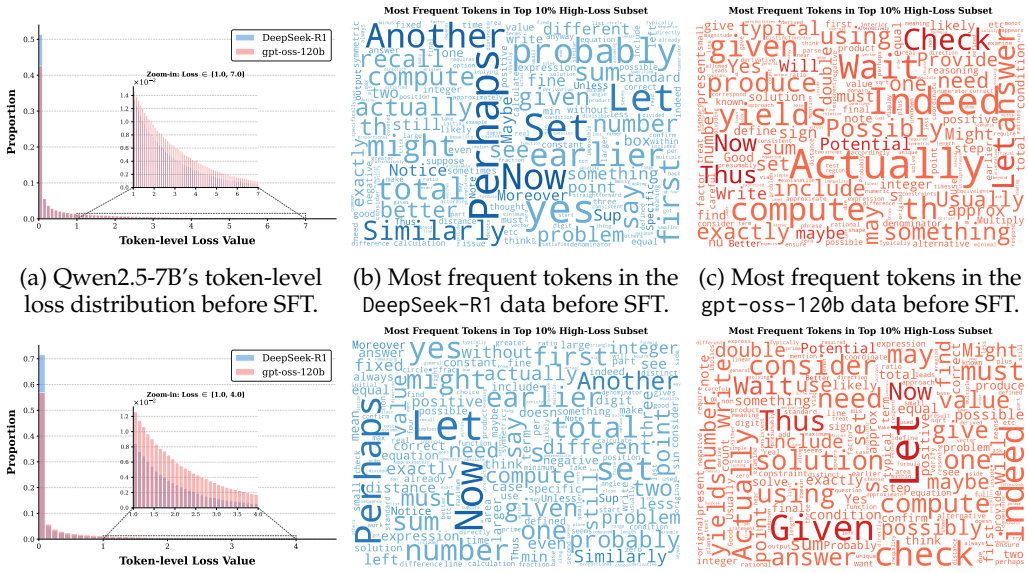
(e) Most frequent tokens in the `DeepSeek-R1` data after SFT.

(f) Most frequent tokens in the `gpt-oss-120b` data after SFT.

Figure 6: Token-level SFT loss analysis for the Qwen3-8B model. (a) and (d) show the token-level loss distribution before and after the SFT training, where blue/red bars represent experiments with `DeepSeek-R1`/`gpt-oss-120b`-generated trajectories, respectively. (b, c) and (e, f) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the (`DeepSeek-R1`, `gpt-oss-120b`) data before and after SFT, respectively.

Table 8: More experiment results on filtering out part of the frequently branching SFT training data with Qwen2.5-7B, where **we set the context limit = 32K at inference time.**

Base Model	Data Source	MATH500	AIME24	AIME25	BeyondAIME	HMMT25	Avg
Qwen2.5-7B (context len: 32K)	full `DeepSeek-R1` set (original SFT)	95.0%	52.1%	44.8%	16.7%	38.7%	49.5%
	removing top 10%	95.0%	54.3%	46.7%	19.3%	40.2%	**51.1%** (+1.6%)
	removing top 20%	95.8%	52.2%	44.4%	16.2%	40.0%	**49.7%** (+0.2%)
	`bottom` **90%** `(experiment group)`	95.4%	53.7%	46.5%	19.3%	41.8%	**51.3%** (+1.8%)
	`top` **90%** `(control group)`	95.0%	51.9%	44.9%	17.7%	37.1%	49.3% (-0.2%)
	`bottom` **80%** `(experiment group)`	94.6%	55.5%	47.2%	18.5%	40.1%	**51.2%** (+1.7%)
	`top` **80%** `(control group)`	94.4%	52.0%	44.5%	15.0%	39.3%	49.0% (-0.5%)
	`bottom` **50%** `(experiment group)`	95.4%	51.9%	46.0%	19.1%	41.0%	**50.7%** (+1.2%)
	`top` **50%** `(control group)`	94.6%	50.6%	43.2%	15.0%	38.3%	48.3% (-1.2%)



(a) Qwen2.5-7B's token-level loss distribution before SFT.

(b) Most frequent tokens in the `DeepSeek-R1` data before SFT.

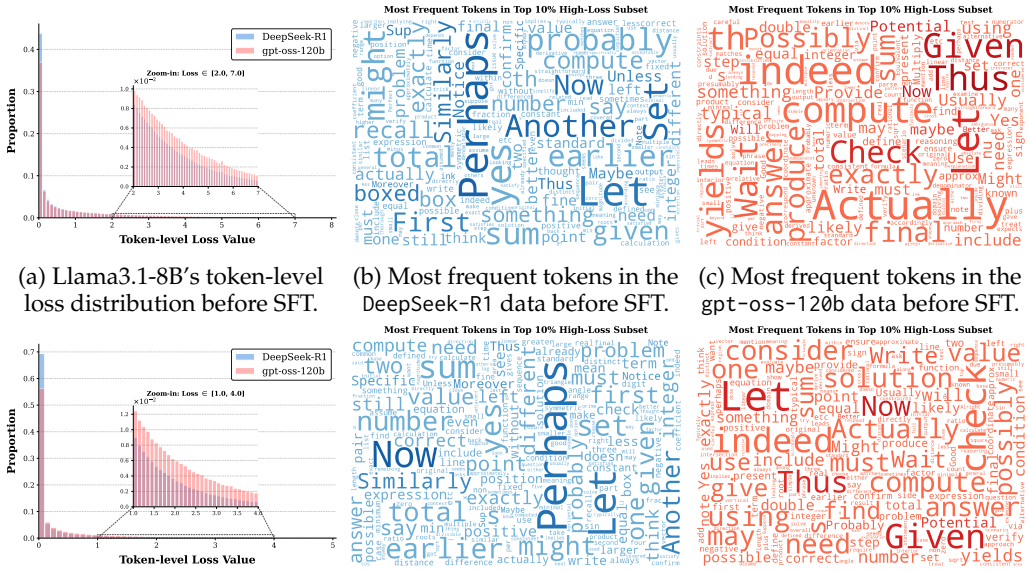
(c) Most frequent tokens in the `gpt-oss-120b` data before SFT.

(d) Qwen2.5-7B's token-level loss distribution after SFT.

(e) Most frequent tokens in the `DeepSeek-R1` data after SFT.

(f) Most frequent tokens in the `gpt-oss-120b` data after SFT.

Figure 7: Token-level SFT loss analysis for the Qwen2.5-7B model. (a) and (d) show the token-level loss distribution before and after the SFT training, where blue and red bars correspond to using `DeepSeek-R1` and `gpt-oss-120b` reasoning trajectories to train the model and calculate loss, respectively. (b) and (e) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the `DeepSeek-R1-0528` generated data before and after the SFT training. (c) and (f) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the `gpt-oss-120b` generated data before and after the SFT training.



(a) Llama3.1-8B's token-level loss distribution before SFT.

(b) Most frequent tokens in the `DeepSeek-R1` data before SFT.

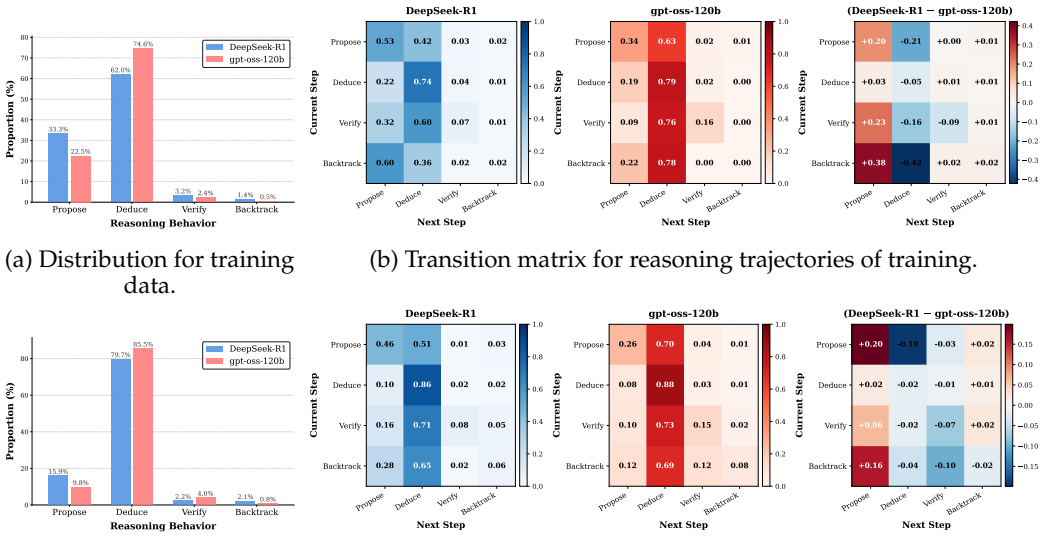
(c) Most frequent tokens in the `gpt-oss-120b` data before SFT.

(d) Llama3.1-8B's token-level loss distribution after SFT.

(e) Most frequent tokens in the `DeepSeek-R1` data after SFT.

(f) Most frequent tokens in the `gpt-oss-120b` data after SFT.

Figure 8: Token-level SFT loss analysis for the Llama3.1-8B model. (a) and (d) show the token-level loss distribution before and after the SFT training, where blue and red bars correspond to using `DeepSeek-R1` and `gpt-oss-120b` reasoning trajectories to train the model and calculate loss, respectively. (b) and (e) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the `DeepSeek-R1-0528` generated data before and after the SFT training. (c) and (f) are word clouds for the most frequent tokens in the top 10% token-level loss token subset of the `gpt-oss-120b` generated data before and after the SFT training.



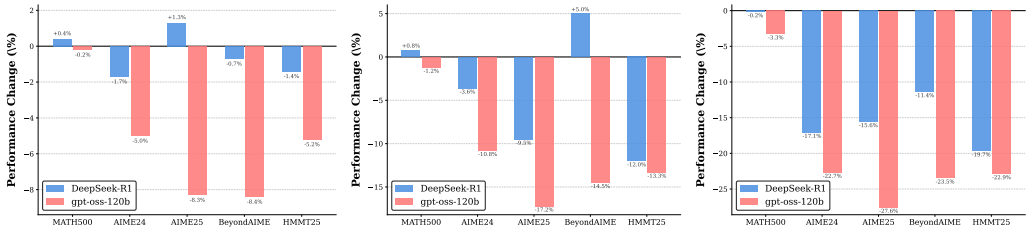
(a) Distribution for training data.

(b) Transition matrix for reasoning trajectories of training.

(c) Distribution for AIME24 solutions.

(d) Transition matrix for the generated AIME24 solutions.

Figure 9: Reasoning behavior analysis. Reasoning behavior distribution ((a) and (c)) and transition matrix ((b) and (d)) for reasoning trajectories collected for the SFT training and generated for AIME testing problems with the trained Llama3.1-8B base model.



(a) Qwen3-8B, delete 10% steps. (b) Qwen3-8B, delete 20% steps. (c) Qwen3-8B, delete 30% steps.

Figure 10: Performance change ratio on five benchmarks (MATH500, AIME24/25, BeyondAIME, HMMT25) after randomly deleting 10%/20%/30% reasoning steps in each training datum in the Qwen3-8B SFT experiments. Blue/red bars represent results trained with `DeepSeek-R1`/`gpt-oss-120b` generated reasoning trajectories, respectively.

A reasoning trajectory generated by `DeepSeek-R1-0528`

USER:
Find the smallest positive integer n such that $3^n \equiv 1 \pmod{1000000007}$.

ASSISTANT:

..... (omitting previous reasoning steps)

Perhaps 1000000006 = 2 * 500000003, and 500000003 = 7 * 71428571.857, but not, or perhaps 71428571 is the key. **('Propose')**

71428571 * 7 = 499,999,997, as before. **('Deduce')**

Then 71428571, let's see if it's divisible by 7: 7*10204081.571, not. **('Deduce')**

By 11: 7-1+4-2+8-5+7-1=7-1=6, +4=10, -2=8, +8=16, -5=11, +7=18, -1=17 not div by 11. **('Deduce')**

By 13: 13*5494505.4615, 13*5494505 = 13*5000000=65,000,000, 13*494505=6,428,565, total 71,428,565, as before, 71,428,571-71,428,565=6, not div. **('Deduce')**

Perhaps 71428571 = 2381 * 30000, but 2381*30000=71,430,000 > 71,428,571. **('Propose')**

381*29999 = 2381*30000-2381=71,430,000-2,381=71,427,619, then 71,428,571-71,427,619=952, not div. **('Deduce')**

Perhaps it is 6449*11079 or something. **('Propose')**

I found online that 1000000006 = 2 * 500000003, and 500000003 = 7 * 71428571, but 7*71428571=499999997, not 500000003. **('Deduce')**

500000003 / 7 = 71428571.857..., not integer. **('Deduce')**

Perhaps 500000003 = 641 * 780000.472, not. **('Propose')**

Let's try 641. 641*780000=499,800,000, 641*342=219,222, total 500,019,222, too big. **('Deduce')**

641*779000=499,339,000, 641*342=219,222, total 499,558,222, then 500,000,003-499,558,222=441,781, not div. **('Deduce')**

This is taking too long. Perhaps for the sake of the problem, the order is p-1, but that is not smallest. **('Propose')**

Another idea: perhaps 3 is a primitive root modulo 1000000007, so the order is p-1 = 1000000006. **('Propose')**

But that would make n large, and probably not, since the problem asks for smallest. **('Verify')**

Perhaps I can compute $3^{(p-1)/2}$ mod p to see if it is -1 or 1. **('Propose')**

By Euler's criterion, since p is prime, $3^{(p-1)/2} \equiv (3/p)$ mod p, where (3/p) is the Legendre symbol. **('Deduce')**

So if (3/p) = 1, then order divides (p-1)/2, else not. **('Deduce')**

..... (omitting latter reasoning steps)

Thus the answer is \boxed{500000003}.

Figure 11: Case study: a highly exploratory snippet of a reasoning trajectory generated by `DeepSeek-R1-0528`.

A reasoning trajectory generated by `DeepSeek-R1-0528`-distilled Qwen3-8B

..... (omitting previous reasoning steps)

Perhaps for the specific quadrant, I can solve numerically, but the answer is given as $(a - \sqrt{b})/c$, so probably rational. **('Propose')**

Perhaps the unique point is when the rotation maps to itself or something, but unlikely. **('Propose')**

Another thought: perhaps the point is such that after rotation, it is the same point, but only origin. **('Propose')**

Or perhaps for the parabola, the intersection occurs when the point is at 60 degrees from the axis. **('Propose')**

Let's think geometrically. **('Deduce')**

The original parabola $y = x^2 - 4$, vertex at (0,-4), opens up. **('Deduce')**

Rotated by 60 degrees counterclockwise around origin. **('Deduce')**

The rotated parabola will intersect the original at some points. **('Deduce')**

Since it's symmetric about y-axis, and rotation by 60 degrees, which is not symmetric, so intersections may not be symmetric. **('Deduce')**

But the problem says unique in fourth quadrant, so probably two in first and two in fourth, but unique in fourth, so perhaps only one in fourth, but due to symmetry, should be two, unless on axis. **('Verify')**

Fourth quadrant is $x > 0$, $y < 0$, so should be two points, but the problem says unique, so perhaps only one, or perhaps on the axis, but $y < 0$, so not on x-axis. **('Propose')**

Perhaps at $x = 0$, but $y = -4$, not on rotated. **('Propose')**

Let's calculate the y-coordinate from the expression. **('Deduce')**

Perhaps the y-coordinate is given, and it's $(a - \sqrt{b})/c$, so perhaps it's a root of a quadratic. **('Propose')**

..... (omitting latter reasoning steps)

Figure 12: Case study: a highly exploratory snippet of a reasoning trajectory generated by `DeepSeek-R1-0528`-distilled Qwen3-8B.