

# Supervised Fine-Tuning for Reasoning Large Language Models: A Survey

Anonymous ACL submission

## Abstract

Large reasoning models have recently demonstrated remarkable capabilities in solving complex tasks, where supervised fine-tuning (SFT) on long chain-of-thought data serves as a crucial foundation for eliciting and enhancing their reasoning abilities. Despite rapid progress in both improving and analyzing reasoning-oriented SFT, the field still lacks a systematic survey that consolidates its fast-growing literature. To fill this gap, we present a comprehensive review of recent advancements in reasoning SFT, examining the literature through the dual lenses of methodological design and analytical investigation. First, we review methodological improvements across the SFT pipeline and categorize them into data-centric approaches and algorithm-centric innovations. Second, we reorganize analytical studies along three dimensions: data characteristics, optimization dynamics, and mechanistic insights. Finally, by synthesizing current research focuses and remaining bottlenecks, we outline promising future directions for reasoning SFT. We hope this survey deepens the understanding of reasoning SFT and paves the way for advanced reasoning models.

## 1 Introduction

Reasoning has become a defining capability of modern large language models (LLMs), enabling them to solve complex tasks that require multi-step problem decomposition, intermediate verification, and logical abstraction (Kahneman, 2011; Huang and Chang, 2023; Sun et al., 2025b). Such capabilities are central to a wide range of domains, including mathematical problem solving (Ahn et al., 2024; Wang et al., 2026d), code generation (Zheng et al., 2023; Jiang et al., 2026c), and robotics (Ma et al., 2026b; Sapkota et al., 2025). Early efforts primarily relied on Chain-of-Thought (CoT) prompting (Wei et al., 2022; Wang and Zhou, 2024; Chia et al., 2023; Kojima et al., 2022) to elicit reason-

ing behaviors from LLMs. Although CoT prompting provides an effective interface for reasoning, it operates only at inference time and suffers from instability (Wang et al., 2023; Turpin et al., 2023).

To overcome these limitations, the paradigm has naturally transitioned toward Supervised Fine-Tuning (SFT) (Lai et al., 2025a; Shen, 2024) on CoT trajectories to internalize reasoning capabilities within the model’s parameters. Early efforts in reasoning SFT predominantly focused on the dimension of data construction, exploring how to curate high-quality step-by-step rationales, typically through knowledge distillation from frontier models or automated synthesis pipelines (Roziere et al., 2023; Li et al., 2024; Xu et al., 2025b; Luo et al., 2025a). While such efforts provide an essential foundation, standard SFT based on vanilla behavior cloning also exhibits intrinsic limitations, including exposure bias (Gan et al., 2024), severe overfitting (Huan et al., 2025), and limited generalization (Wu et al., 2026a).

To address these fundamental bottlenecks, the research community has recently pivoted toward algorithmic innovations that refine the optimization dynamics of SFT across diverse granularities, encompassing token- and sample-level signal modulation (Kim et al., 2025; Sanyal et al., 2025), distribution constraints (Zhu et al., 2026a), and structural integration with reinforcement learning (Jiang et al., 2026a). In parallel, a growing body of work has begun to analyze the behavior and mechanisms of reasoning SFT with long CoT data. These studies investigate diverse factors such as data properties (Ye et al., 2025b), optimization dynamics (Ren et al., 2026), and mechanistic interpretability (Wu et al., 2025a), offering insights into when and why SFT improves reasoning.

Despite the recent surge of literature focusing on reasoning SFT, a systematic survey consolidating this fast-evolving landscape is still conspicuously absent. While several existing reviews at-

tempt to summarize post-training paradigms for reasoning models, they predominantly emphasize RL-based frameworks (Zhou et al., 2025; Sun et al., 2025a; Xu et al., 2025a) or restrict their examination of SFT to its alignment or synergy with RL pipelines (Jiang et al., 2026b). Another stream of literature touches upon reasoning-oriented SFT, yet focuses almost exclusively on early-stage data construction (Chen et al., 2025d; Wang et al., 2026d; Li et al., 2025e). Consequently, a holistic consolidation of recent non-trivial algorithmic innovations and analytical insights remains lacking.

To fill this gap, this survey presents a systematic and timely overview of contemporary research, encompassing both methodological designs and analytical insights. On the methodological front, we establish a structured taxonomy spanning data-centric improvements and algorithm-centric innovations. On the analytical front, we synthesize and reorganize highly fragmented empirical findings into three core dimensions: data characteristics, optimization dynamics, and mechanistic analysis. Furthermore, we identify open challenges and chart promising future directions to guide subsequent explorations in this domain. Ultimately, we hope this work serves as a foundational roadmap to navigate and inspire future advancements in reasoning SFT.

## 2 Preliminary

SFT serves as the foundational paradigm for aligning LLMs with human intent using expert demonstrations. Formally, given a language model parameterized by  $\theta$  and a training dataset  $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ , standard SFT minimizes the Negative Log-Likelihood (NLL) objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \sum_{t=1}^{|y|} \log \pi_{\theta}(y_t | y_{<t}, x) \right],$$

where  $x$  denotes the input prompt,  $y$  represents the target response sequence, and  $\pi_{\theta}$  specifies the model policy. Within the scope of this survey, we center our focus exclusively on *reasoning SFT*. Distinct from standard instruction-following, the target response  $y$  is structurally decomposed into an extensive CoT rationale  $r$  followed by the final answer  $a$ , i.e.,  $y = [r; a]$ . Optimizing  $\mathcal{L}_{\text{SFT}}$  over protracted reasoning paths compels the model to internalize intrinsic causal structures, thereby capturing genuine problem-solving mechanics rather than exploiting surface-level correlations.

## 3 Reasoning SFT: Methods

### 3.1 Data-Centric Methods

**Data Augmentation.** To actively ensure robust internalization and prevent the model from overfitting to specific reasoning formats, data augmentation (Maharana et al., 2022) has emerged as a critical data-centric strategy for enriching the diversity of training distributions. At the input level, (Lin et al., 2025b) leverage systematic prompt diversification to break the model’s anchoring to surface-level linguistics. For target responses, GFT (Gan et al., 2024) mitigates single-path dependency by pairing queries with a hybrid ensemble of expert, teacher, and self-generated trajectories. Furthermore, recent methodologies have shifted from purely positive imitation toward extracting dense contrastive signals from negative samples. For instance, NFT (Chen et al., 2026a) constructs an implicit negative policy by jointly modeling both successful and failed reasoning responses, utilizing the contrast to actively push the policy away from common hallucination patterns and logical fallacies. Taking a reward-calibrated approach, RIFT (Liu et al., 2026c) assigns a constant positive reward to correct reasoning traces, while explicitly allocating a constant negative reward to incorrect trajectories.

**Data Refinement.** Despite the benefits of augmented diversity, directly training on off-policy expert trajectories often introduces a distributional mismatch that can degrade a model’s pre-existing capabilities. To address this, data refinement aims to seamlessly align the training corpus with the LLM’s intrinsic distribution (Gupta et al., 2025; Zhao et al., 2026a; Wang et al., 2026c; Lin and Han, 2026). Recent self-alignment frameworks (Gupta et al., 2025; Zhao et al., 2026a) substitute standard expert demonstrations with a combination of the model’s inherently correct generations and its successfully rewritten correct responses. Specifically, the rewriting process leverages reference answers to guide the model in reconstructing valid, in-distribution reasoning trajectories. Alternatively, SPOT (Lin and Han, 2026) employs a stronger teacher model for minimal-edit step corrections and utilizes a decoupled objective to reinforce rectified reasoning steps and penalize logical errors. To overcome the diversity collapse inherent in heuristic templates, another work (Wang et al., 2026c) formulates data refinement as a policy learning task, optimizing semantic diversity and alignment via re-

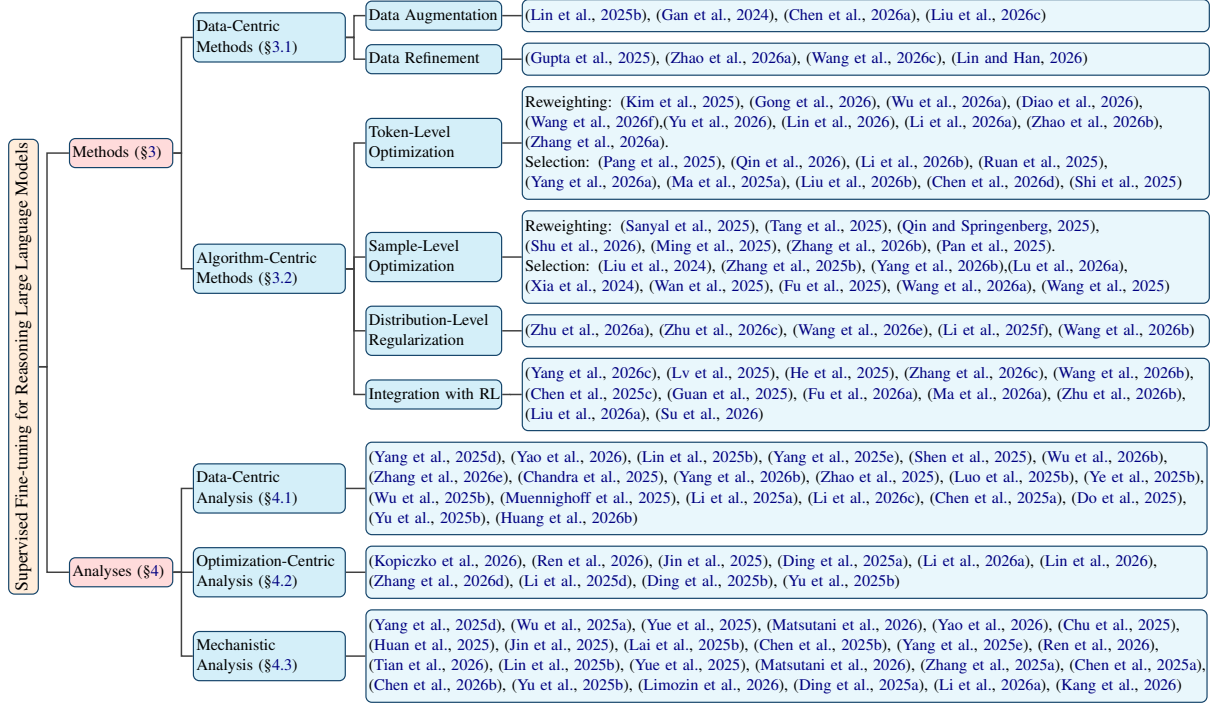


Figure 1: Taxonomy of methodological and analytical research in SFT for reasoning LLMs.

inforcement learning under a task-consistency gate.

### 3.2 Algorithm-Centric Methods

Alongside data-centric strategies, algorithm-centric methodologies offer a complementary axis for enhancing standard SFT by directly modifying the underlying optimization dynamics. Accordingly, this subsection systematically reviews these algorithmic advancements, progressing from fine-grained signal modulation (token-level and sample-level optimization) to global paradigm shifts (distribution-level regularization and integration with RL).

Formally, fine-grained signal modulation can be unified under a generalized optimization objective:

$$\mathcal{L}_{\text{SFT}}^g(\theta) = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \alpha_i \cdot \left[ \sum_{t=1}^{|y^{(i)}|} w_t^{(i)} \log \pi_{\theta}(y_t^{(i)} | y_{<t}^{(i)}, x^{(i)}) \right],$$

where  $w_t^{(i)}$  and  $\alpha_i$  denote the dynamic token-level and sample-level weights, respectively. By actively calibrating these coefficients alongside global regularization and RL integration, recent frameworks systematically mitigate the empirical limitations of standard supervised alignment.

**Token-Level Optimization.** We first review token-level optimization, where the learning dy-

namics are modulated by the token weight  $w_t^{(i)}$  to prioritize suitable reasoning tokens and mitigate overfitting. Within the reweighting paradigm, recent works dynamically scale token updates based on gradient consistency between the individual token and the global batch (Gong et al., 2026), the predicted probabilities (Wu et al., 2026a), or the output distribution entropy (Diao et al., 2026). To address the ambiguity of low-probability tokens, DEFT (Wang et al., 2026f) introduces an entropy-driven trust gate that transitions the optimization focus from exploring uncertain regions to sharpening confident predictions. Conversely, RANK-TUNER (Yu et al., 2026) aggregates probability and entropy into a unified ranking metric, systematically upweighting critical low-probability, low-entropy targets. From an optimization perspective, TALR (Lin et al., 2026) applies a dynamic, normalized weight to each token. Specifically, the token weight is formulated as  $w_i = \frac{\exp(-\ell_i/\tau)}{\sum_j \exp(-\ell_j/\tau)}$ , where  $\ell_i$  is the token-level loss and  $\tau$  represents the median batch loss during training.

A parallel line of research tailors the SFT objective via token reweighting to establish an optimal initialization policy for subsequent RL alignment. To mitigate distributional collapse and preserve exploration potential during the RL cold start, recent frameworks maintain the pre-trained probability landscape by amplifying gradients on uncertain

tokens via sequence-relative weighting (Li et al., 2026a). Alternatively, another approach (Zhao et al., 2026b) upweights correct tokens via a finite-temperature Gibbs distribution, which explicitly preserves the relative probability mass of alternative tokens to mitigate distributional collapse. To further bridge the gap between SFT and RL, PEAR (Zhang et al., 2026a) employs multi-level importance sampling, discounted suffix weighting, and negative sample suppression to construct a superior initialization for online RL.

Within the discrete token selection paradigm, recent methods extract optimal learning tokens through diverse filtering criteria. Some works explicitly retain tokens exhibiting the highest training loss (Kim et al., 2025) or the largest parameter update norms (Ma et al., 2025a). A parallel line of work identifies high-quality tokens via cross-model loss discrepancies, comparing superior and inferior models to preserve critical tokens (Pang et al., 2025), or contrasting safety-degraded and utility-oriented policies to eliminate vulnerabilities (Li et al., 2026b). Alternative frameworks restrict gradient updates exclusively to critical tokens identified through semantic-aware response-to-prompt attention (Qin et al., 2026), counterfactual perturbations (Ruan et al., 2025), or multi-dimensional explainability metrics (Yang et al., 2026a).

Beyond gradient and semantic signals, further approaches define selective optimization strategies based on the model’s predictive confidence, policy-driven sampling, or predefined trajectory structures. ProFit (Liu et al., 2026b) consolidates established knowledge by computing the loss solely on high-probability tokens, whereas SED-SFT (Chen et al., 2026d) encourages generative diversity by selectively optimizing low-confidence tokens associated with multiple plausible branches. To explicitly incorporate on-policy exploration into the SFT phase, OTR (Ming et al., 2025) formulates token generation as a single-step reinforcement learning trajectory, dynamically rewarding ground-truth tokens while penalizing incorrect samples drawn from the current policy distribution. Finally, from a structural standpoint, SFTKey (Shi et al., 2025) completely masks the intermediate CoT rationale to prevent verbose reasoning trajectories from overshadowing core conclusions, thereby explicitly restricting optimization solely to the answer tokens.

**Sample-Level Optimization.** Transitioning to sample-level optimization, the learning dynam-

ics are governed exclusively by the dynamic sample weight  $\alpha_i$ . Within the continuous reweighting framework, one prominent branch calibrates the scalar  $\alpha_i$  based on the intrinsic difficulty of the sample. For example, FLOW (Sanyal et al., 2025) upweights “easy” samples with low pre-training losses to mitigate catastrophic forgetting, whereas BFT (Tang et al., 2025) prioritizes “hard” sequences featuring continuous low-probability spans to force the absorption of novel knowledge.

Another direction establishes a theoretical equivalence between offline SFT and online RL by reweighting static samples to simulate the dynamic distribution shifts inherent to RL. This theoretical equivalence to RL is realized by deriving adaptive importance weights from policy trajectory ratios (Qin and Springenberg, 2025) or scaling prompt-normalized exponential rewards (Shu et al., 2026). Moving beyond predefined mathematical formulations, a final line of research automates sample utility evaluation and data composition. These frameworks dynamically optimize instance weights based on how much their loss drops when shown high-quality reference examples as context (Zhang et al., 2026b), or by framing sample weighting as a minimax bilevel optimization problem to automatically learn optimal multi-source data proportions (Pan et al., 2025).

Complementary to continuous modulation, discrete sample selection employs hard pruning to isolate optimal training instances. To align the training corpus with the model’s inherent distribution, recent frameworks evaluate internal uncertainty to identify high-quality instructions (Liu et al., 2024), or select teacher-generated responses that maximize the student model’s assigned probability (Zhang et al., 2025b). Furthermore, to align with the model’s current learning capacity, other methods (Yang et al., 2026b; Lu et al., 2026a) explicitly filter out excessively challenging instances to focus optimization on moderate-difficulty zones.

Beyond heuristic difficulty assessments, a parallel paradigm formulates sample selection through rigorous mathematical optimization and fine-grained structural dynamics. To curate datasets under strict computational budgets, optimization-driven frameworks prioritize samples whose gradients closely align with predefined target examples (Xia et al., 2024) or by framing efficient data selection as a surrogate-penalized bilevel optimization problem (Wan et al., 2025). To mitigate structural biases, further studies refine sample scoring

by prioritizing informative tokens that are robust to embedding noise (Fu et al., 2025), or by causally disentangling step-length confounders to calibrate branching first-token probabilities (Wang et al., 2026a). Finally, moving beyond isolated sample- or token-level filters, Q-Tuning (Wang et al., 2025) unifies both granularities via an error-uncertainty diagnostic framework, dynamically discarding entirely noisy samples while applying token pruning exclusively to partially misunderstood instances.

**Distribution-Level Regularization.** Stepping beyond localized token and sample-level interventions, distribution-level regularization constrains the global policy distribution to mitigate catastrophic forgetting and preserve pre-trained capabilities. To restrict deviations from the reference policy, recent methodologies explicitly bound the optimization trajectory by incorporating a standard Kullback-Leibler (KL) (Kullback, 1951) divergence penalty (Zhu et al., 2026a), applying hard gradient clipping beyond predefined divergence thresholds (Zhu et al., 2026c), or enforcing a strict upper bound on iterative KL divergence through dynamic moving anchors (Wang et al., 2026e). To prevent diversity degradation and maintain exploration potential, a parallel line of research leverages entropy-aware modulation. These frameworks sustain probability mass across diverse reasoning trajectories by either simulating reverse KL minimization via a sparse meta-controller (Li et al., 2025f), or applying asymmetric regularization that explicitly targets high-entropy reasoning steps while strictly imitating low-entropy factual knowledge (Wang et al., 2026b).

**Integration with RL.** Unifying SFT and RL avoids the imitation bottleneck of pure SFT and the optimization instability of isolated RL, enabling models to expand their capability boundaries while refining reasoning trajectories. One prominent paradigm dynamically harmonizes these objectives via continuous loss weighting. For instance, some works adaptively calibrate SFT and RL coefficients based on real-time metrics such as response correctness (Yang et al., 2026c; Lv et al., 2025) or diversity (Yang et al., 2026c). AMFT (He et al., 2025) employs a meta-gradient weight controller to dynamically learn the optimal balance between supervised imitation and reinforcement exploration. Incorporating predictive uncertainty as a guiding metric, RED (Guan et al., 2025) dynamically modulates learning weights via entropy

scheduling to balance exploration and path refinement. Building upon this uncertainty-aware perspective, SRFT (Fu et al., 2026a) integrates an entropy-regularized SFT objective with off-policy RL, stabilizing knowledge distillation without constraining the policy’s exploratory potential. Alternatively, CHORD (Zhang et al., 2026c) employs deterministic schedules to smoothly transition from SFT to RL, and BRIDGE (Chen et al., 2025c) enforces structural decoupling by restricting SFT updates to lightweight adapters.

Alternatively, discrete sample-wise routing allocates instances based on difficulty or correctness. To systematically overcome performance plateaus, some methods route excessively hard (Ma et al., 2026a) or completely failed prompts (Zhu et al., 2026b) exclusively to SFT or multi-teacher distillation. Conversely, successfully resolved prompts trigger RL to consolidate reasoning pathways, occasionally incorporating visual refiners in DyME (Liu et al., 2026a) or contrastive objectives in DYPO to stabilize exploration. Advancing this integration to a finer granularity, TRAPO (Su et al., 2026) applies supervised learning to adaptively selected expert prefixes, while employing standard RL to optimize the subsequent model-generated rollouts.

## 4 Reasoning SFT: Analyses

### 4.1 Data-Centric Analysis

**The Impact of CoT Length.** Characterized by its intuitive and quantifiable nature, CoT length serves as a natural entry point for investigating the empirical behavior of reasoning SFT. Several studies have shown that CoT supervision is crucial for reasoning: models fine-tuned without CoT suffer substantial performance degradation or poor generalization (Yang et al., 2025d; Yao et al., 2026; Zhang et al., 2025d; Lin et al., 2025b). A common view in the literature is that longer CoTs are more beneficial than shorter ones, yielding stronger performance and better OOD generalization (Yang et al., 2025e). Shen et al. (2025) further shows that CoT length matters more than problem difficulty in determining model performance. However, Wu et al. (2026b) argue that CoT length follows an inverted-U relationship with model accuracy: both overly short and overly long CoTs are suboptimal, with the optimal length increasing for harder tasks but decreasing as model capability improves.

**Distribution, Difficulty, and Diversity.** Data distribution plays a critical role in reasoning

SFT (Zhang et al., 2026e; Chandra et al., 2025; Yang et al., 2025d, 2026b; Zhao et al., 2025; Luo et al., 2025b). Chandra et al. (2025) show that LLM-paraphrased CoTs, even when containing errors, can outperform manually constructed data, suggesting that distributional alignment may matter more than surface-level correctness. This view is further supported by studies emphasizing teacher-student alignment in long-CoT distillation (Yang et al., 2025d; Luo et al., 2025b; Wu et al., 2025b). To mitigate distribution mismatch, recent works further explore stronger or more adaptive data sources, including advanced reasoning models (Zhao et al., 2025), RL-generated trajectories (Zhang et al., 2026e), and teacher-student collaborative data construction (Huang et al., 2026b).

Beyond distributional alignment, problem difficulty and data diversity also shape SFT effectiveness. Ye et al. (2025b); Muennighoff et al. (2025) favor harder and more diverse samples when constructing reasoning datasets, while Wu et al. (2025b) advocate filtering out overly simple problems and maintaining response diversity through Levenshtein-distance filtering. In addition, a recent study highlights the critical role of prompt diversity in improving generalization (Lin et al., 2025b).

**Reasoning Pattern.** Beyond surface-level data properties, another line of work examines how the internal reasoning patterns of CoTs affect SFT, including their structural organization, exploration behaviors, and error-correction processes. An early study shows that disrupting the structure of CoTs leads to larger performance degradation than corrupting their surface content, highlighting the central role of reasoning structure in CoT-based training (Li et al., 2025a). However, more complex or exploratory reasoning patterns are not always beneficial. Li et al. (2026c) find that stronger teacher models do not necessarily yield better distillation, and that teachers producing convergent and deductive trajectories can be more effective than those generating divergent, branch-heavy explorations. Similarly, SFT on CoTs with “Aha moments” may degrade performance, suggesting that explicitly imitating complex patterns is not always helpful (Chen et al., 2025a).

Beyond global reasoning patterns, the quality of individual reasoning steps also matters. Do et al. (2025) show that final-answer correctness alone is insufficient for data filtering, and that selecting examples based on intermediate-step relevance and

step-to-step coherence can further improve performance. Moreover, removing similar or repetitive steps can reduce token costs and improve performance, whereas retaining erroneous steps may help models learn verification, backtracking, and error correction (Luo et al., 2025b).

**Data Quantity.** Although large-scale model training is often guided by the assumption that more data improves performance (Li et al., 2025c), reasoning SFT appears to be less dependent on data quantity. Several studies show that small but carefully curated datasets can substantially improve or elicit reasoning capabilities, with fewer than 1K examples often being sufficient (Ye et al., 2025b; Muennighoff et al., 2025) and even as few as 10 examples working in extreme settings (Yang et al., 2025d). In VLM post-training, SFT performance has also been found to be relatively insensitive to data quantity (Yu et al., 2025b). Overall, these findings suggest that the structure and quality of reasoning data matter more than its scale (Zhao et al., 2025), and increasing SFT data may even hurt performance in some cases (Chen et al., 2025a).

## 4.2 Optimization-Centric Analysis

**Training Dynamics.** Beyond data construction, reasoning SFT is highly sensitive to training dynamics, including optimization duration, checkpoint selection, and learning-rate control. These factors are crucial for balancing reasoning improvement, generalization, and capability preservation.

A central question is how extensively SFT should be performed. For standalone SFT, several studies suggest that sufficient optimization is necessary to internalize reasoning patterns. Kopiczko et al. (2026) show that multi-epoch training is important for reasoning SFT, as insufficient optimization can substantially degrade performance. Ren et al. (2026) further reveal a non-monotonic phenomenon, where OOD performance first decreases and then recovers as training proceeds.

However, when SFT serves as initialization for subsequent RL, excessive optimization may be undesirable. In contrast to the recovery pattern observed by Ren et al. (2026), Jin et al. (2025) find that OOD performance first improves and then declines, possibly due to differences in the evaluated tasks. Consistent with this concern, Ding et al. (2025a) and Li et al. (2026a) show that early-stage SFT checkpoints can provide better initialization for downstream RL. In addition, Lin et al. (2026)

537 find that smaller learning rates better balance in- 587  
538 distribution and OOD performance. 588

539 **Model Capability.** Reasoning SFT is strongly 589  
540 capability-dependent: Several studies reveal that 590  
541 weak models may struggle to learn the complex pat- 591  
542 terns encoded in long CoTs (Zhang et al., 2026d; 592  
543 Ren et al., 2026; Li et al., 2025d; Ding et al., 593  
544 2025b). However, within the domain of VLM rea- 594  
545 soning, Yu et al. (2025b) observe that SFT with 595  
546 short CoT disproportionately benefits weaker mod- 596  
547 els, yielding more pronounced improvements. To 597  
548 improve learnability for smaller models, existing 598  
549 work explores easier supervision signals, such as 599  
550 mixing short-CoT data or trajectories from smaller 600  
551 teachers (Li et al., 2025d), or fine-tuning a medium- 601  
552 sized model as an intermediate teacher to generate 602  
553 data for smaller models (Ding et al., 2025b). 603

### 554 4.3 Mechanistic Analysis 604

555 **What Does SFT Actually Teach?** A central 605  
556 question in reasoning SFT is whether it genuinely 606  
557 teaches new reasoning capabilities or mainly in- 607  
558 duces models to imitate the surface forms and be- 608  
559 havioral patterns of reasoning traces. 609

560 Several studies suggest that SFT may primar- 610  
561 ily reshape model behavior rather than create new 611  
562 reasoning abilities. Yang et al. (2025d) show that 612  
563 even 10 content-corrupted examples can substan- 613  
564 tially improve reasoning performance, indicating 614  
565 that SFT may mainly induce format imitation. Sim- 615  
566 ilarly, models trained on golden shortest paths fail 616  
567 to acquire backtracking capabilities, in contrast to 617  
568 those trained with RLVR (Wei and Kim, 2026; Wen 618  
569 et al., 2026). Wu et al. (2025a) further argue that 619  
570 SFT mainly injects task knowledge into the model 620  
571 and may even compromise its reasoning capability. 621

572 In contrast, another line of work argues that SFT 622  
573 can expand the model’s capability boundary rather 623  
574 than merely reshape its output format. From the 624  
575 perspective of Pass@k, several studies suggest that 625  
576 SFT can increase the set of problems that the model 626  
577 is capable of solving (Yue et al., 2025; Matsutani 627  
578 et al., 2026). Moreover, Yao et al. (2026) show that 628  
579 CoT-trained models can compose learned simple 629  
580 skills to solve unseen or more complex problems, 630  
581 providing evidence for genuine reasoning. 631

582 **Memorization or Generalization?** A pivotal in- 632  
583 quiry in reasoning SFT is whether the observed 633  
584 improvements stem from the memorization of spe- 634  
585 cific trajectories or the acquisition of transferable 635  
586 reasoning patterns. A critical view is that SFT 636

587 tends to memorize rather than generalize: Chu et al. 587  
588 (2025) observe a continuous decline in OOD per- 588  
589 formance on Points24 (Zhai et al., 2024) and nav- 589  
590 igation tasks (Yang et al., 2024), leading to the 590  
591 conclusion that “SFT memorizes, RL generalizes”. 591  
592 Related studies report similar limitations, including 592  
593 poor OOD generalization (Jin et al., 2025), degra- 593  
594 dation on general tasks after math SFT (Huan et al., 594  
595 2025), and forgetting caused by off-policy supervi- 595  
596 sion (Lai et al., 2025b; Chen et al., 2025b). 596

597 However, other studies suggest that SFT can gen- 597  
598 eralize under suitable conditions. Long CoTs, suf- 598  
599 ficient optimization, high-quality data, and strong 599  
600 base models have all been identified as important 600  
601 factors for SFT generalization (Yang et al., 2025e; 601  
602 Ren et al., 2026). Moreover, Tian et al. (2026) 602  
603 surprisingly find that training on incorrect CoTs 603  
604 can improve OOD generalization, and Lin et al. 604  
605 (2025b) observes that CoT supervision with diverse 605  
606 prompts can even surpass RL in generalization per- 606  
607 formance. These findings suggest that poor SFT 607  
608 generalization may reflect unfavorable training con- 608  
609 ditions rather than an inherent inability. 609

610 **Interaction between SFT and RL.** With the suc- 610  
611 cess of DeepSeek-R1 (Guo et al., 2025), combining 611  
612 SFT and RL has become a common post-training 612  
613 paradigm for modern reasoning models (Yu et al., 613  
614 2025a; Zheng et al., 2025; Shao et al., 2024; Jaech 614  
615 et al., 2024; Yang et al., 2025a). A mainstream 615  
616 view is that SFT provides basic reasoning behav- 616  
617 iors, while RL further refines them through explo- 617  
618 ration and feedback (Yue et al., 2025; Matsutani 618  
619 et al., 2026; Zhang et al., 2025a). Under this view, 619  
620 Zhang et al. (2025a) argue that post-training should 620  
621 place greater emphasis on SFT, with RL serving 621  
622 mainly as a lightweight refinement stage. 622

623 However, the benefit of SFT before RL appears 623  
624 to be conditional rather than universal. In VLM 624  
625 reasoning, Chen et al. (2025a) find that SFT can im- 625  
626 pair subsequent RL, and Chen et al. (2026b) show 626  
627 that several hybrid strategies, including SFT-then- 627  
628 RL, interleaved training, and model merging, may 628  
629 underperform RL-only training. Similarly, Yu et al. 629  
630 (2025b) show that SFT benefits RL mainly when 630  
631 the base model is weak, but can become harmful 631  
632 when the base model is already strong. 632

633 Recent studies therefore shift the focus from 633  
634 whether to combine SFT and RL to how SFT 634  
635 should be used for RL initialization. Although 635  
636 various methods have been proposed to better in- 636  
637 tegrate the two paradigms (Fu et al., 2026b; Zhang 637

et al., 2026c), Limozin et al. (2026) show that some reported gains over SFT-then-RL were due to bugs in the training framework; after correction, a simple SFT-then-RL pipeline achieves the best performance. Consistently, Ding et al. (2025a) argue that SFT-then-RL remains a strong post-training paradigm and find that early-stage SFT checkpoints are better starting points for subsequent RL, in line with Li et al. (2026a). This also suggests that stronger standalone SFT performance does not necessarily imply better RL initialization (Zhang et al., 2026a). To select suitable SFT checkpoints, Kang et al. (2026) further propose using validation loss or Pass@k with a large- $k$  as more reliable indicators.

## 5 Challenges and Future Directions

**Harmonizing Reasoning Enhancement with Knowledge Learning.** While confidence-based SFT strategies excel at boosting reasoning, they inadvertently reinforce pre-existing beliefs, suppressing learning signals for novel facts and under-represented domains (Wu et al., 2026a). To overcome this vulnerability in knowledge-intensive tasks, future research should prioritize capability-aware optimization frameworks that seamlessly balance reasoning refinement with factual acquisition. Achieving this synergy is essential for training versatile foundation models that act as both rigorous reasoners and reliable knowledge repositories.

**Inference-Efficient Reasoning Trajectories.** While current reasoning-oriented SFT frameworks have achieved remarkable strides in generalization and task accuracy, this progress often inadvertently sacrifices inference efficiency. By strictly cloning verbose expert demonstrations, models are conditioned to generate excessively long CoT trajectories, incurring severe computational overhead and elevated latency during deployment (Kang et al., 2025). Consequently, explicitly optimizing the length-accuracy trade-off has emerged as a crucial future direction. Future frameworks must transition from passively replicating extensive reasoning chains to actively compressing them, exploring novel mechanisms to distill prolonged logical steps into dense, condensed trajectories without compromising the underlying mathematical rigor or deductive soundness.

**Toward RL-Compatible Supervised Fine-Tuning.** In modern post-training pipelines, SFT is rarely used in isolation, but typically serves

as a preparatory stage for RL (Guo et al., 2025). However, as prior work has observed (Zhang et al., 2026a), better standalone SFT performance does not necessarily translate into a better RL initialization. This suggests that existing SFT methods may have been evaluated under an incomplete criterion: most data construction and algorithmic designs focus on post-SFT performance, while overlooking their compatibility with RL. Therefore, an important future direction is to re-evaluate existing methods under the full SFT-then-RL pipeline, examining which types of data, objectives, and checkpoints provide better entry for RL. Beyond evaluation, future work should also develop RL-compatible SFT objectives that preserve exploration potential, policy diversity, and reward-improvable behavior.

**Toward Causal Analysis of Reasoning SFT.** Existing analyses have shown that factors such as CoT length, data difficulty, diversity, reasoning patterns, and data quantity can affect reasoning SFT (Shen et al., 2025; Muennighoff et al., 2025). However, many conclusions remain correlational and may be confounded by other variables. For example, longer CoTs may improve generalization not only because of their length, but also because they change the reasoning structure, token budget, problem difficulty, or teacher-model distribution (Yang et al., 2025e; Chandra et al., 2025). Future work should therefore conduct more controlled studies that isolate the causal effects of individual data properties, model capability, and training dynamics. Such analyses would lead to a clearer understanding of why and when reasoning SFT improves generalization.

## 6 Conclusion

In this survey, we present a comprehensive review of recent advancements in reasoning SFT, examining the literature through the dual lenses of methodological design and analytical investigation. We established a structured taxonomy categorizing methods into data-centric and algorithm-centric approaches, while synthesizing fragmented findings across data, optimization dynamics, and mechanistic analysis. By outlining key bottlenecks, we further summarize several promising future directions. Ultimately, we hope this survey serves as a foundational roadmap to guide researchers and inspire subsequent breakthroughs in developing advanced reasoning foundation models.

## 737 Limitations

738 Despite our effort to provide a comprehensive re-  
739 view, this survey has several limitations. First, rea-  
740 soning SFT is a rapidly evolving field, and new  
741 methods and analyses continue to emerge at a fast  
742 pace; therefore, some very recent or concurrent  
743 studies may not be fully covered. Additionally,  
744 this survey primarily focuses on recent studies, es-  
745 pecially those published since 2025, in order to  
746 capture the latest methodological and analytical de-  
747 velopments in reasoning-oriented SFT. As a result,  
748 we do not provide a systematic review of earlier  
749 works that mainly concentrated on constructing rea-  
750 soning or CoT-style training data. Although these  
751 efforts laid important foundations for reasoning  
752 SFT, a comprehensive historical account of data  
753 construction methods is beyond the scope of this  
754 survey.

## 755 References

756 Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui  
757 Zhang, and Wenpeng Yin. 2024. Large language  
758 models for mathematical reasoning: Progresses and  
759 challenges. In *Proc. EACL*.

760 American Institute of Mathematics. 2024. Aime 2024  
761 competition mathematical problems.

762 Jacob Austin, Augustus Odena, Maxwell Nye, Maarten  
763 Bosma, Henryk Michalewski, David Dohan, Ellen  
764 Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1  
765 others. 2021. Program synthesis with large language  
766 models. *arXiv preprint arXiv:2108.07732*.

767 Abhranil Chandra, Ayush Agrawal, Arian Hosseini, Se-  
768 bastian Fischmeister, Rishabh Agarwal, Navin Goyal,  
769 and Aaron Courville. 2025. Shape of thought: When  
770 distribution matters more than correctness in reason-  
771 ing tasks. *arXiv preprint arXiv:2512.22255*.

772 Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter,  
773 Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024.  
774 Spatialvlm: Endowing vision-language models with  
775 spatial reasoning capabilities. In *Proc. CVPR*.

776 Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xian-  
777 feng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie.  
778 2025a. SFT or RL? an early investigation into train-  
779 ing rl-like reasoning large vision-language models.  
780 *Transactions on Machine Learning Research*.

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

785 Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu  
786 Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-  
787 Yu Liu, Jun Zhu, and Haoxiang Wang. 2026a.

Nft: Bridging supervised learning and reinforcement  
learning in math reasoning. In *Proc. ICLR*. 788  
789

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin,  
Chongyu Chen, and Xiaodan Liang. 2022. Unigeo:  
Unifying geometry logical reasoning via reformulat-  
ing mathematical expression. In *Proc. EMNLP*. 790  
791  
792  
793

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang,  
Lingbo Liu, Eric Xing, and Liang Lin. 2021a. Geoqa:  
A geometric question answering benchmark towards  
multimodal numerical reasoning. In *Proc. ACL Find-  
ings*. 794  
795  
796  
797  
798

Jierun Chen, Tiezheng YU, Haoli Bai, Lewei Yao, Jian-  
nan Wu, Kaican Li, Fei Mi, Chaofan Tao, Lei Zhu,  
Manyi Zhang, Xiao-Hui Li, Lu Hou, Lifeng Shang,  
and Qun Liu. 2026b. The synergy dilemma of long-  
cot SFT and RL: Investigating post-training tech-  
niques for reasoning VLMs. *Transactions on Ma-  
chine Learning Research*. 799  
800  
801  
802  
803  
804  
805

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-  
Fai Wong. 2025c. Beyond two-stage training: Coop-  
erative sft and rl for llm reasoning. *arXiv preprint  
arXiv:2509.06948*. 806  
807  
808  
809

Longfei Chen, Ji Zhao, Lanxiao Cui, Tong Su,  
Xingbo Pan, Ziyang Li, Yongxing Wu, Qijiang  
Cao, Qiyao Cai, Jing Zhang, and 1 others. 2026c.  
Seccodebench-v2 technical report. *arXiv preprint  
arXiv:2602.15485*. 810  
811  
812  
813  
814

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan,  
Henrique Ponde De Oliveira Pinto, Jared Kaplan,  
Harri Edwards, Yuri Burda, Nicholas Joseph, Greg  
Brockman, and 1 others. 2021b. Evaluating large  
language models trained on code. *arXiv preprint  
arXiv:2107.03374*. 815  
816  
817  
818  
819  
820

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng,  
Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang  
Zhou, Te Gao, and Wanxiang Che. 2025d. Towards  
reasoning era: A survey of long chain-of-thought  
for reasoning large language models. *arXiv preprint  
arXiv:2503.09567*. 821  
822  
823  
824  
825  
826

Yijie Chen, Yijin Liu, and Fandong Meng. 2026d. Sed-  
sft: Selectively encouraging diversity in supervised  
fine-tuning. *arXiv preprint arXiv:2602.07464*. 827  
828  
829

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo,  
Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei  
Liu. 2024. Spatialrgpt: Grounded spatial reasoning  
in vision-language models. In *Proc. NeurIPS*. 830  
831  
832  
833

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan,  
Soujanya Poria, and Lidong Bing. 2023. Con-  
trastive chain-of-thought prompting. *arXiv preprint  
arXiv:2311.09277*. 834  
835  
836  
837

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang  
Tong, Saining Xie, Dale Schuurmans, Quoc V Le,  
Sergey Levine, and Yi Ma. 2025. SFT memorizes,  
RL generalizes: A comparative study of foundation  
model post-training. In *Proc. ICML*. 838  
839  
840  
841  
842

843	Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,	single-stage method with supervised and reinforcement	898
844	Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias	fine-tuning for reasoning. In <i>Proc. ICLR</i> .	899
845	Plappert, Jerry Tworek, Jacob Hilton, Reiichiro		
846	Nakano, and 1 others. 2021. Training verifiers	Wangjie Gan, Miao Pan, Linbo Xi, Wenqi Zhang, Jintao	900
847	to solve math word problems. <i>arXiv preprint</i>	Chen, Jianwei Yin, and Xuhong Zhang. 2024. Gft:	901
848	<i>arXiv:2110.14168</i> .	From imitation to reward fine-tuning with unbiased	902
		group advantages and dynamic coefficient rectifica-	903
		tion. In <i>Proc. ACL</i> .	904
849	Yihe Deng, I Hsu, Jun Yan, Zifeng Wang, Rujun Han,	Xuan Gong, Senmiao Wang, Hanbo Huang, Ruoyu Sun,	905
850	Gufeng Zhang, Yanfei Chen, Wei Wang, Tomas Pfister,	and Shiyu Liang. 2026. Vcore: Variance-controlled	906
851	Chen-Yu Lee, and 1 others. 2026. Supervised	optimization-based reweighting for chain-of-thought	907
852	reinforcement learning: From expert trajectories to	supervision. In <i>Proc. ACL</i> .	908
853	step-wise reasoning. In <i>Proc. ICLR</i> .		
		Zhong Guan, Likang Wu, Hongke Zhao, Jiahui Wang,	909
854	Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang,	and Le Wu. 2025. Recall-extend dynamics: Enhanc-	910
855	Zhonghao Yan, Yufei Han, Kongming Liang, Weiran	ing small language models through controlled explo-	911
856	Xu, and Zhanyu Ma. 2026. Entropy-adaptive fine-	ration and refined offline integration. <i>arXiv preprint</i>	912
857	tuning: Resolving confident conflicts to mitigate for-	<i>arXiv:2508.16677</i> .	913
858	getting. <i>arXiv preprint arXiv:2601.02151</i> .		
		Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song,	914
859	Bowen Ding, Yuhan Chen, Jiayang Lv, Jiyao Yuan,	Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang,	915
860	Qi Zhu, Shuangshuang Tian, Dantong Zhu, Futing	Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-	916
861	Wang, Heyuan Deng, Fei Mi, and 1 others. 2025a.	r1 incentivizes reasoning in llms through reinforce-	917
862	Rethinking expert trajectory utilization in llm post-	ment learning. <i>Nature</i> , 645(8081):633–638.	918
863	training. <i>arXiv preprint arXiv:2512.11470</i> .		
		Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh	919
864	Dongyi Ding, Tiannan Wang, Chenghao Zhu, Meiling	Khandelwal, Dinesh Raghu, and Sachindra Joshi.	920
865	Tao, Yuchen Eleanor Jiang, and Wangchunshu Zhou.	2025. Selective self-to-supervised fine-tuning for	921
866	2025b. Micota: Bridging the learnability gap with in-	generalization in large language models. In <i>Proc.</i>	922
867	intermediate cot and teacher assistants. <i>arXiv preprint</i>	<i>NAACL</i> .	923
868	<i>arXiv:2507.01887</i> .		
		Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhent-	924
869	Heejin Do, Jaehui Hwang, Dongyoon Han, Seong Joon	ing Qi, Martin Riddell, Wenfei Zhou, James Coady,	925
870	Oh, and Sangdoon Yun. 2025. What defines good	David Peng, Yujie Qiao, Luke Benson, and 1 oth-	926
871	reasoning in llms? dissecting reasoning steps with multi-	ers. 2024. Folio: Natural language reasoning with	927
872	aspect evaluation. <i>arXiv preprint arXiv:2510.20603</i> .	first-order logic. In <i>Proc. EMNLP</i> .	928
		Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding	929
873	Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang,	Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han,	930
874	and Zhongyu Wei. 2024. Embspatial-bench: Bench-	Yujie Huang, Yuxiang Zhang, and 1 others. 2024.	931
875	marking spatial understanding for embodied tasks	Olympiadbench: A challenging benchmark for pro-	932
876	with large vision-language models. In <i>Proc. ACL</i> .	moting agi with olympiad-level bilingual multimodal	933
		scientific problems. In <i>Proc. ACL</i> .	934
877	Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu	Lixuan He, Jie Feng, and Yong Li. 2025. Amft:	935
878	Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin,	Aligning llm reasoners by meta-learning the opti-	936
879	Zhenlin Wei, Chujie Zheng, and 1 others. 2026. Su-	mation-imitation-exploration balance. <i>arXiv preprint</i>	937
880	pergppqa: Scaling llm evaluation across 285 graduate	<i>arXiv:2508.06944</i> .	938
881	disciplines. In <i>Proc. NeurIPS</i> .		
		Dan Hendrycks, Steven Basart, Saurav Kadavath, Man-	939
882	Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang,	tas Mazeika, Akul Arora, Ethan Guo, Collin Burns,	940
883	Yuqi Tang, Qiang Zhang, and Keyan Ding. 2024. Sci-	Samir Puranik, Horace He, Dawn Song, and 1 oth-	941
884	knoweval: Evaluating multi-level scientific knowl-	ers. 2021a. Measuring coding challenge competence	942
885	edge of large language models. <i>arXiv preprint</i>	with apps. In <i>Proc. NeurIPS</i> .	943
886	<i>arXiv:2406.09098</i> .		
		Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul	944
887	Yanjun Fu, Faisal Hamman, and Sanghamitra Dutta.	Arora, Steven Basart, Eric Tang, Dawn Song, and	945
888	2025. T-shirt: Token-selective hierarchical data se-	Jacob Steinhardt. 2021b. Measuring mathematical	946
889	lection for instruction tuning. In <i>Proc. NeurIPS</i> .	problem solving with the math dataset. In <i>Proc.</i>	947
		<i>NeurIPS</i> .	948
890	Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang,	Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu,	949
891	Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang,	Seungone Kim, Minxin Du, Radha Poovendran, Gra-	950
892	Yuanheng Zhu, and Dongbin Zhao. 2026a. Srft: A	ham Neubig, and Xiang Yue. 2025. Does math rea-	951
893	single-stage method with supervised and reinforce-	soning improve general llm capabilities? understand-	952
894	ment fine-tuning for reasoning. In <i>Proc. ICLR</i> .	ing transferability of llm reasoning. <i>arXiv preprint</i>	953
		<i>arXiv:2507.00432</i> .	954
895	Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang,		
896	Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang,		
897	Yuanheng Zhu, and Dongbin Zhao. 2026b. SRFT: A		

955	Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In <i>Proc. ACL Findings</i> .	1008
956		1009
957		1010
958	Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M Ponti, and Ivan Titov. 2026a. Blending supervised and reinforcement fine-tuning with prefix sampling. In <i>Proc. ICML</i> .	1011
959		1012
962	Zixian Huang, Kaichen Yang, Xu Huang, Feiyang Hao, Qiming Ge, Bowen Li, He Du, Kai Chen, and Qipeng Guo. 2026b. How to fine-tune a reasoning model? a teacher-student cooperation framework to synthesize student-consistent sft data. <i>arXiv preprint arXiv:2604.14164</i> .	
963		
964		
965		
966		
967		
968	Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. <i>arXiv preprint arXiv:2412.16720</i> .	
969		
970		
971		
972		
973	Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Live-codebench: Holistic and contamination free evaluation of large language models for code. In <i>Proc. ICLR</i> .	
974		
975		
976		
977		
978		
979	Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispacial: Towards comprehensive spatial reasoning benchmark for vision language models. <i>arXiv preprint arXiv:2506.03135</i> .	
980		
981		
982		
983		
984	Haitao Jiang, Wenbo Zhang, Jiarui Yao, Hengrui Cai, Sheng Wang, and Rui Song. 2026a. Supervised fine-tuning versus reinforcement learning: A study of post-training methods for large language models. <i>arXiv preprint arXiv:2603.13985</i> .	
985		
986		
987		
988		
989	Haitao Jiang, Wenbo Zhang, Jiarui Yao, Hengrui Cai, Sheng Wang, and Rui Song. 2026b. Supervised fine-tuning versus reinforcement learning: A study of post-training methods for large language models. <i>arXiv preprint arXiv:2603.13985</i> .	
990		
991		
992		
993		
994	Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026c. A survey on large language models for code generation. <i>ACM Transactions on Software Engineering and Methodology</i> , 35(2):1–72.	
995		
996		
997		
998	Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In <i>Proc. ICLR</i> .	
999		
1000		
1001		
1002	Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Doina Precup, and Mohammad Hamdaqa. 2025. RL fine-tuning heals the OOD forgetting in SFT. In <i>Proc. NeurIPS Workshops</i> .	
1003		
1004		
1005		
1006	Daniel Kahneman. 2011. Thinking, fast and slow. <i>Farrar, Straus and Giroux</i> .	
1007		
	Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. 2026. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. In <i>Proc. ICLR</i> .	1012
		1013
		1014
		1015
	Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In <i>Proc. AAAI</i> .	1016
		1017
		1018
		1019
	Gyuhak Kim, Sumiran Singh Thakur, Su Min Park, Wei Wei, and Yujia Bao. 2025. Sft-go: Supervised fine-tuning with group optimization for large language models. <i>arXiv preprint arXiv:2506.15021</i> .	1020
		1021
		1022
		1023
	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In <i>Proc. NeurIPS</i> .	1024
		1025
		1026
		1027
	Dawid J Kopiczko, Sagar Vaze, Tijmen Blankevoort, and Yuki M Asano. 2026. Data repetition beats data scaling in long-cot supervised fine-tuning. <i>arXiv preprint arXiv:2602.11149</i> .	1028
		1029
	Solomon Kullback. 1951. Kullback-leibler divergence. <i>Encyclopedia of Machine Learning</i> , pages 581–583.	1030
		1031
		1032
		1033
		1034
	Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, and 1 others. 2025a. A survey of post-training scaling in large language models. In <i>Proc. ACL</i> .	1035
		1036
		1037
		1038
		1039
		1040
	Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, and 1 others. 2025b. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. <i>arXiv preprint arXiv:2507.05386</i> .	1041
		1042
		1043
		1044
		1045
		1046
	Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhmaneshi, Shishir G Patil, Matei Zaharia, and 1 others. 2025a. Language models can easily learn to reason from demonstrations. In <i>Proc. EMNLP Findings</i> .	1047
		1048
		1049
		1050
		1051
		1052
	Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. <i>Hugging Face repository</i> .	1053
		1054
		1055
		1056
		1057
	Jinke Li, Jiarui Yu, Chenxing Wei, Hande Dong, Qiang Lin, Liangjing Yang, Zhicai Wang, and Yanbin Hao. 2025b. Unisvg: A unified dataset for vector graphic understanding and generation with multimodal large language models. In <i>Proc. ACM-MM</i> .	1058
		1059
		1060
		1061
	Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. 2025c. (mis)fitting scaling laws: A survey of scaling law fitting techniques in deep learning. In <i>Proc. ICLR</i> .	



1168	Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and 1 others. 2025. Towards a unified view of large language model post-training. <i>arXiv preprint arXiv:2509.04419</i> .	1224
1169		1225
1170		1226
1171		1227
1172		1228
1173	Dabiao Ma, Ziming Dai, Zhimin Xin, Shu Wang, Jian Yang, and Haojun Fei. 2025a. Ts-peft: Unveiling token-level redundancy in parameter-efficient fine-tuning. <i>arXiv preprint arXiv:2511.16147</i> .	1229
1174		1230
1175		1231
1176		1232
1177		1233
1178	Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, and 1 others. 2026a. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. In <i>Proc. ICLR</i> .	1234
1179		1235
1180		1236
1181		1237
1182		1238
1183	Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 2025b. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In <i>Proc. ICCV</i> .	1239
1184		1240
1185		1241
1186		1242
1187	Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2026b. A survey on vision-language-action models for embodied ai. <i>IEEE Transactions on Neural Networks and Learning Systems</i> .	1243
1188		1244
1189		1245
1190		1246
1191		1247
1192	Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A review: Data pre-processing and data augmentation techniques. <i>Global Transitions Proceedings</i> , 3(1):91–99.	1248
1193		1249
1194		1250
1195		1251
1196	Mathematical Association of America. 2023. Amc 2023 competition problems.	1252
1197		1253
1198		1254
1199	Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In <i>Proc. ICLR</i> .	1255
1200		1256
1201		1257
1202		1258
1203	Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. <i>arXiv preprint arXiv:2509.26313</i> .	1259
1204		1260
1205		1261
1206		1262
1207	Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In <i>Proc. EMNLP</i> .	1263
1208		1264
1209		1265
1210		1266
1211		1267
1212	Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. 2025. Scalebio: Scalable bilevel optimization for llm data reweighting. In <i>Proc. ACL</i> .	1268
1213		1269
1214		1270
1215	Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. 2025. Token cleaning: Fine-grained data selection for llm supervised fine-tuning. In <i>Proc. ICML</i> .	1271
1216		1272
1217		1273
1218		1274
1219	Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. <i>arXiv preprint arXiv:2501.14249</i> .	1275
1220		1276
1221		1277
1222		1278
1223		1279
	Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? <i>arXiv preprint arXiv:2407.01284</i> .	
	Chongli Qin and Jost Tobias Springenberg. 2025. Supervised fine tuning on curated data is reinforcement learning (and can be improved). <i>arXiv preprint arXiv:2507.12856</i> .	
	Xiaohan Qin, Xiaoxing Wang, Ning Liao, Cancheng Zhang, Xiangdong Zhang, Mingquan Feng, Jingzhi Wang, and Junchi Yan. 2026. sstoken: Self-modulated and semantic-aware token selection for llm fine-tuning. In <i>Proc. ICLR</i> .	
	David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In <i>Proc. COLM</i> .	
	Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and 1 others. 2026. Rethinking generalization in reasoning sft: A conditional analysis on optimization, data, and model capability. <i>arXiv preprint arXiv:2604.06628</i> .	
	Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, and 1 others. 2023. Code llama: Open foundation models for code. <i>arXiv preprint arXiv:2308.12950</i> .	
	Zhiwen Ruan, Yixia Li, He Zhu, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. 2025. Enhancing large language model reasoning via selective critical token fine-tuning. <i>arXiv preprint arXiv:2510.10974</i> .	
	Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. <i>International Journal on Digital Libraries</i> , 23(3):289–301.	
	Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. 2025. Upweighting easy samples in fine-tuning mitigates forgetting. In <i>Proc. ICML</i> .	
	Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In <i>Proc. ICLR</i> .	
	Ranjan Sapkota, Yang Cao, Konstantinos I Roulmliotis, and Manoj Karkee. 2025. Vision-language-action models: Concepts, progress, applications and challenges. <i>arXiv preprint arXiv:2505.04769</i> .	
	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. <i>arXiv preprint arXiv:2402.03300</i> .	

1280	Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. <i>arXiv preprint arXiv:2402.06094</i> .	Negative reasoning samples enhance out-of-domain generalization. <i>arXiv preprint arXiv:2601.04992</i> .	1334
1281			1335
1282	Si Shen, Fei Huang, Zhixiao Zhao, Chang Liu, Tiansheng Zheng, and Danhao Zhu. 2025. Long is more important than difficult for training reasoning models. <i>arXiv preprint arXiv:2503.18069</i> .	Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In <i>Proc. NeurIPS</i> .	1336
1283			1337
1284			1338
1285			1339
1286	Xiaofeng Shi, Qian Kou, Yuduo Li, and Hua Zhou. 2025. Rethinking supervised fine-tuning: Emphasizing key answer tokens for improved llm accuracy. <i>arXiv preprint arXiv:2512.21017</i> .	Weilin Wan, Weizhong Zhang, and Cheng Jin. 2025. Computational budget should be considered in data selection. In <i>Proc. NeurIPS</i> .	1340
1287			1341
1288			1342
1289			
1290	Yao Shu, Chenxing Wei, Hongbin Lin, Shuang Qiu, and Hui Xiong. 2026. Reference-sampled boltzmann projection for kl-regularized rlvr: Target-matched weighted sft, finite one-shot gaps, and policy mirror descent. <i>arXiv preprint arXiv:2605.02469</i> .	Bing Wang, Rui Miao, Chen Shen, Shaotian Yan, Kaiyuan Liu, Ximing Li, Xiaosong Yuan, Sinan Fan, Jun Zhang, and Jieping Ye. 2026a. On the step length confounding in llm reasoning data selection. In <i>Proc. ACL</i> .	1343
1291			1344
1292			1345
1293			1346
1294			1347
1295	Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. 2025. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In <i>Proc. NAACL</i> .	Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In <i>Proc. ACL</i> .	1348
1296			1349
1297			1350
1298			1351
1299	Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. In <i>Proc. EMNLP</i> .	Hao Wang, Hao Gu, Hongming Piao, Kaixiong Gong, Yuxiao Ye, Xiangyu Yue, Sirui Han, Yike Guo, and Dapeng Wu. 2026b. Learning while staying curious: Entropy-preserving supervised fine-tuning via adaptive self-distillation for large reasoning models. In <i>Proc. ACL</i> .	1352
1300			1353
1301			1354
1302			1355
1303	Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. 2026. Trust-region adaptive policy optimization. In <i>Proc. ICLR</i> .	Jiacheng Wang, Ping Jian, Zhen Yang, Zirong Chen, Keren Liao, and Zhongbin Guo. 2026c. Patch the distribution mismatch: RL rewriting agent for stable off-policy sft. <i>arXiv preprint arXiv:2602.11220</i> .	1356
1304			1357
1305			
1306	Haoyuan Sun, Jiaqi Wu, Bo Xia, Yifu Luo, Yifei Zhao, Kai Qin, Xufei Lv, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. 2025a. Reinforcement fine-tuning powers reasoning capability of multimodal large language models. <i>arXiv preprint arXiv:2505.18536</i> .	Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In <i>Proc. NeurIPS</i> .	1358
1307			1359
1308			1360
1309			1361
1310			1362
1311			1363
1312	Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, and 1 others. 2025b. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. <i>ACM Computing Surveys</i> , 57(11):1–43.	Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and 1 others. 2026d. A survey on large language models for mathematical reasoning. <i>ACM Computing Surveys</i> , 58(8):1–35.	1364
1313			1365
1314			1366
1315			1367
1316			1368
1317			1369
1318	Emilia Szymańska, Mihai Dusmanu, Jan-Willem Burchard, Mahdi Rad, and Marc Pollefeys. 2024. Space3d-bench: Spatial 3d question answering benchmark. In <i>Proc. ECCV</i> .	Shaobo Wang, Jiaming Wang, Jiajun Zhang, Cong Wang, Yue Min, Zichen Wen, Xingzhang Ren, Fei Huang, Huiqiang Jiang, Junyang Lin, and 1 others. 2025. Winning the pruning gamble: A unified approach to joint sample and token pruning for efficient supervised fine-tuning. <i>arXiv preprint arXiv:2509.23873</i> .	1370
1319			1371
1320			1372
1321			1373
1322	Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In <i>Proc. ACL Findings</i> .	Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. 2026e. Stabilizing llm supervised fine-tuning via explicit distributional control. <i>arXiv preprint arXiv:2605.04468</i> .	1374
1323			1375
1324			1376
1325			1377
1326	Zhenchao Tang, Fang Wang, Haohuai He, Jiale Zhou, Tianxu Lv, Jun Zhu, Shouzhi Chen, Minghao Yang, Yu Wang, Jiayang Wu, and 1 others. 2025. Aligning llms with biomedical knowledge using balanced fine-tuning. <i>arXiv preprint arXiv:2511.21075</i> .	Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In <i>Proc. NeurIPS</i> .	1378
1327			1379
1328			1380
1329			1381
1330			1382
1331	Xueyun Tian, Minghua Ma, Bingbing Xu, Nuoyan Lyu, Wei Li, Heng Dong, Zheng Chu, Yuanzhuo Wang, and Huawei Shen. 2026. Learning from mistakes: Negative reasoning samples enhance out-of-domain generalization. <i>arXiv preprint arXiv:2601.04992</i> .	Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others.	1383
1332			1384
1333			1385

1388	2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In <i>Proc. NeurIPS</i> .	Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. <i>arXiv preprint arXiv:2407.04973</i> .	1443
1389			1444
1390			1445
1391	Zecheng Wang, Deyuan Liu, Chunshan Li, Yupeng Zhang, Zhengyun Zhao, Dianhui Chu, Bingning Wang, and Dianbo Sui. 2026f. Gradients must earn their influence: Unifying sft with generalized entropic objectives. <i>arXiv preprint arXiv:2602.11424</i> .	Fengli Xu, Qianye Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, and 1 others. 2025a. Toward large reasoning models: A survey of reinforced reasoning with large language models. <i>Patterns</i> , 6(10).	1447
1392			1448
1393			1449
1394			1450
1395			1451
1396	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In <i>Proc. NeurIPS</i> .	Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025b. Llava-cot: Let vision language models reason step-by-step. In <i>Proc. ICCV</i> .	1452
1397			1453
1398			1454
1399			1455
1400			1456
1401	Stanley Wei and Juno Kim. 2026. Provable benefits of RLVR over SFT for reasoning models: Learning to backtrack efficiently. In <i>Proc. ICLR Workshops</i> .	Jundong Xu, Hao Fei, Yuhui Zhang, Liangming Pan, Qijun Huang, Qian Liu, Preslav Nakov, Min-Yen Kan, William Yang Wang, Mong-Li Lee, and 1 others. 2025c. Muslr: Multimodal symbolic logical reasoning. In <i>Proc. NeurIPS</i> .	1457
1402			1458
1403			1459
1404	Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In <i>Proceedings of the 3rd Workshop on Noisy User-generated Text</i> .	Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025d. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. <i>arXiv preprint arXiv:2504.15279</i> .	1460
1405			1461
1406			1462
1407			1463
1408	Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2026. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In <i>Proc. ICLR</i> .	Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. In <i>Proc. NeurIPS</i> .	1464
1409			1465
1410			1466
1411			1467
1412			1468
1413			1469
1414	Juncheng Wu, Sheng Liu, Haoqin Tu, Hang Yu, Xiaoke Huang, James Zou, Cihang Xie, and Yuyin Zhou. 2025a. Knowledge or reasoning? a close look at how llms think across domains. <i>arXiv preprint arXiv:2506.02126</i> .	Yuming Yan, Kai Tang, Sihong Chen, Ke Xu, Dan Hu, Qun Yu, and Pengfei Hu. 2026. S-grpo: Unified post-training for large vision-language models. <i>arXiv preprint arXiv:2604.16557</i> .	1470
1415			1471
1416			1472
1417			1473
1418			1474
1419	Xiaojun Wu, Xiaoguang Jiang, Huiyang Li, Jucai Zhai, Dengfeng Liu, Qiaobo Hao, Huang Liu, Zhiguo Yang, Ji Xie, Ninglun Gu, and 1 others. 2025b. Beyond scaling law: A data-efficient distillation framework for reasoning. <i>arXiv preprint arXiv:2508.09883</i> .	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. <i>arXiv preprint arXiv:2505.09388</i> .	1475
1420			1476
1421			1477
1422			1478
1423			1479
1424	Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2026a. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In <i>Proc. ICLR</i> .	Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, and 1 others. 2025b. Chartmimic: Evaluating lmm’s cross-modal reasoning capability via chart-to-code generation. In <i>Proc. ICLR</i> .	1480
1425			1481
1426			1482
1427			1483
1428			1484
1429	Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2026b. When more is less: Understanding chain-of-thought length in LLMs. In <i>Proc. ICLR</i> .	Jihan Yang, Runyu Ding, Ellis Brown, Xiaojuan Qi, and Saining Xie. 2024. V-irl: Grounding virtual intelligence in real life. In <i>Proc. ECCV</i> .	1485
1430			1486
1431			1487
1432			1488
1433	Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, and 1 others. 2024. Training large language models for reasoning through reverse curriculum reinforcement learning. <i>arXiv preprint arXiv:2402.05808</i> .	Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025c. Thinking in space: How multimodal large language models see, remember, and recall spaces. In <i>Proc. CVPR</i> .	1489
1434			1490
1435			1491
1436			1492
1437			1493
1438			1494
1439	Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. In <i>Proc. ICML</i> .	Linyao Yang, Jian-Tao Huang, Yafei Lu, Zhenhui Jessie Li, and Guirong Xue. 2025d. The emperor’s new reasoning: Format imitation overshadows genuine mathematical understanding in sft. In <i>Proc. EMNLP</i> .	1495
1440			1496
1441			
1442			

1497	Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. 2025e. Demystifying long chain-of-thought reasoning. In <i>Proc. ICML</i> .		
1498		Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In <i>Proc. CVPR</i> .	1549 1550 1551 1552 1553 1554
1500	Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025f. Mmsibench: A benchmark for multi-image spatial intelligence. <i>arXiv preprint arXiv:2505.23764</i> .		
1501		Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In <i>Proc. NeurIPS</i> .	1555 1556 1557 1558 1559
1502		Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and 1 others. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In <i>Proc. NeurIPS</i> .	1560 1561 1562 1563 1564
1503	Yuchen Yang, Wenze Lin, Enhao Huang, Zhixuan Chu, Hongbin Zhou, Lan Tao, Yiming Li, Zhan Qin, and Kui Ren. 2026a. Explainable token-level noise filtering for llm fine-tuning datasets. In <i>Proc. ICLR</i> .		
1504		Charlie Zhang, Graham Neubig, and Xiang Yue. 2025a. On the interplay of pre-training, mid-training, and rl on reasoning language models. <i>arXiv preprint arXiv:2512.07783</i> .	1565 1566 1567 1568
1505	Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, and 1 others. 2026b. Which reasoning trajectories teach students to reason better? a simple metric of informative alignment. In <i>Proc. ACL</i> .		
1506		Dylan Zhang, Qirun Dai, and Hao Peng. 2025b. The best instruction-tuning data are those that fit. In <i>Proc. NeurIPS</i> .	1569 1570 1571
1507	Zheng Yang, Xiaojie Wang, Caixia Yuan, Zhuoran Li, Wenhao Gao, Bingwei Li, and Yilin Liu. 2026c. Learning from examples and self-exploration: A new paradigm for dynamic fusion.		
1508		Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. 2026a. Good sft optimizes for sft, better sft prepares for reinforcement learning. <i>arXiv preprint arXiv:2602.01058</i> .	1572 1573 1574 1575
1509	Xinhao Yao, Ruifeng Ren, Yun Liao, Lizhong Ding, and Yong Liu. 2026. Compositional generalization from learned skills via cot training: A theoretical and structural analysis for reasoning. In <i>Proc. ICLR</i> .		
1510		Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and 1 others. 2025c. Matscibench: Benchmarking the reasoning ability of large language models in materials science. <i>arXiv preprint arXiv:2510.12171</i> .	1576 1577 1578 1579 1580 1581
1511	Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Robert Tang. 2025a. Mmscibench: Benchmarking language models on chinese multimodal scientific problems. In <i>Proc. ACL Findings</i> .		
1512		Ling Zhang, Xianliang Yang, Juwon Yu, Park Cheonyoung, Miran Lee, Lei Song, and Jiang Bian. 2026b. Holdout-loss-based data selection for llm finetuning via in-context learning. In <i>Proc. ICLR</i> .	1582 1583 1584 1585
1513	Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025b. Limo: Less is more for reasoning. In <i>Proc. COLM</i> .		
1514		Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In <i>Proc. ECCV</i> . Springer.	1586 1587 1588 1589 1590 1591
1515	Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, and 1 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. In <i>Proc. NeurIPS</i> .		
1516		Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025d. Improve vision language model chain-of-thought reasoning. In <i>Proc. ACL</i> .	1592 1593 1594 1595 1596
1517	Wenhao Yu, Shaohang Wei, Jiahong Liu, Yifan Li, Minda Hu, Aiwei Liu, Hao Zhang, and Irwin King. 2026. Probability-entropy calibration: An elastic indicator for adaptive fine-tuning. In <i>Proc. ICML</i> .		
1518		Shudan Zhang, Hanlin Zhao, Xiao Liu, Qinkai Zheng, Zehan Qi, Xiaotao Gu, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. 2024b. Naturalcodebench: Examining coding performance mismatch on humaneval and natural user prompts. <i>arXiv preprint arXiv:2405.04520</i> .	1597 1598 1599 1600 1601
1519	Yongcan Yu, Lingxiao He, Shuo Lu, Lijun Sheng, Yinuo Xu, Yanbo Wang, Kuangpu Guo, Jianjie Cheng, Meng Wang, Qianlong Xie, and 1 others. 2025b. Reassessing the role of supervised fine-tuning: An empirical study in vlm reasoning. <i>arXiv preprint arXiv:2512.12690</i> .		
1520			
1521	Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? <i>arXiv preprint arXiv:2304.02015</i> .		
1522			

1602	Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2026c. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In <i>Proc. ICLR</i> .		
1603			
1604			
1605			
1606			
1607			
1608	Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2026d. BREAD: Branched rollouts from expert anchors bridge SFT & RL for reasoning. In <i>Proc. NeurIPS</i> .		
1609			
1610			
1611			
1612	Zhihao Zhang, Qiaole Dong, Qi Zhang, Enyu Zhou, Jun Zhao, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Mingqi Wu, Yanwei Fu, Tao Ji, Tao Gui, Xuanjing Huang, and Kai Chen. 2026e. Why reinforcement fine-tuning enables MLLMs preserve prior knowledge better: A data perspective. In <i>Proc. ICLR</i> .		
1613			
1614			
1615			
1616			
1617			
1618	Shiwan Zhao, Xuyang Zhao, Jiaming Zhou, Aobo Kong, Qicheng Li, and Yong Qin. 2026a. Mind the gap: Data rewriting for stable off-policy supervised fine-tuning. In <i>Proc. ICASSP</i> .		
1619			
1620			
1621			
1622	Yike Zhao, Simin Guo, Ziqing Yang, Shifan Han, Dahua Lin, and Fei Tan. 2025. More data or better data? a critical analysis of data selection and synthesis for mathematical reasoning. In <i>Proc. EMNLP</i> .		
1623			
1624			
1625			
1626	Zhengyang Zhao, Lu Ma, Yizhen Jiang, Xiaochen Ma, Zimo Meng, Chengyu Shen, Lexiang Tang, Haoze Sun, Peng Pei, and Wentao Zhang. 2026b. Gift: Unlocking global optimality in post-training via finite-temperature gibbs initialization. <i>arXiv preprint arXiv:2601.09233</i> .		
1627			
1628			
1629			
1630			
1631			
1632	Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. <i>arXiv preprint arXiv:2507.18071</i> .		
1633			
1634			
1635			
1636			
1637	Zibin Zheng, Kaiwen Ning, Yanlin Wang, Jingwen Zhang, Dewu Zheng, Mingxi Ye, and Jiachi Chen. 2023. A survey of large language models for code: Evolution, benchmarking, and future trends. <i>arXiv preprint arXiv:2311.10372</i> .		
1638			
1639			
1640			
1641			
1642	Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. 2025. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. <i>arXiv preprint arXiv:2504.21277</i> .		
1643			
1644			
1645			
1646			
1647	He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2026a. Anchored supervised fine-tuning. In <i>Proc. ICLR</i> .		
1648			
1649			
1650	Qin Zhu, Fei Huang, Runyu Peng, Keming Lu, Bowen Yu, Qinyuan Cheng, Xipeng Qiu, Xuanjing Huang, and Junyang Lin. 2025. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. <i>arXiv preprint arXiv:2502.16906</i> .		
1651			
1652			
1653			
1654			
1655			
	Taojie Zhu, Dongyang Xu, Ding Zou, Sen Zhao, Qiaobo Hao, Zhiguo Yang, and Yonghong He. 2026b. Bridging sft and rl: Dynamic policy optimization for robust reasoning. In <i>Proc. ACL</i> .	1656	
		1657	
		1658	
		1659	
	Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2026c. Proximal supervised fine-tuning. In <i>Proc. ICLR</i> .	1660	
		1661	
		1662	
	Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widayarsi, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, and 1 others. 2025. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In <i>Proc. ICLR</i> .	1663	
		1664	
		1665	
		1666	
		1667	
		1668	

1669  
1670  
  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719

## A Integration with RL via Off-Policy Expert Demonstrations

Beyond joint optimization of SFT and RL objectives, a distinct research paradigm leverages offline expert demonstrations as auxiliary exploration anchors to systematically resolve the reward sparsity and imitation bottlenecks inherent in online Reinforcement Learning (RL). By integrating off-policy trajectories directly into the online policy-gradient loop rather than forcing direct cross-entropy minimization, these methods bridge the gap between static imitation and autonomous exploration.

Among these approaches, LUFFY (Yan et al., 2025) and S-GRPO (Yan et al., 2026) explicitly integrate full expert trajectories into the candidate pool of standard relative group optimization. LUFFY (Yan et al., 2025) introduces Mixed-Policy GRPO, which combines off-policy demonstrations with on-policy rollouts for advantage computation, employing entropic policy shaping to reinforce low-probability yet critical actions. Similarly, S-GRPO (Yan et al., 2026) addresses optimization collapse in visual tasks by injecting a verified ground-truth anchor into the candidate pool whenever exploratory failure is detected, ensuring positive advantage signals to break training deadlocks.

To grant the policy higher exploratory autonomy, another branch of research truncates expert demonstrations into prefixes or hints, prompting the model to explore and complete the remaining paths. UFT (Liu et al., 2025) and  $R^3$  (Xi et al., 2024) establish reverse curricula to reduce exploration difficulty. Specifically,  $R^3$  progressively shifts the exploration starting state backwards from the demonstration’s end, while UFT smoothly decays hint lengths to zero, blending log-likelihood objectives with reward optimization. To stabilize this hybrid learning, Prefix-RFT (Huang et al., 2026a) reinforces offline prefixes using dynamic advantages and entropy-based clipping, preventing off-policy gradients from dominating optimization. Advancing this, BREAD (Zhang et al., 2026d) employs Episode Anchor Search (EAS) to adaptively select branching points, providing self-paced curriculum guidance where standard SFT-then-RL pipelines stall. Finally, to extract dense signals without rigid token-level imitation, SRL (Deng et al., 2026) reformulates reasoning as a sequential decision-making process of discrete logical "actions." By generating internal monologues before committing to each

step, SRL computes sequence similarity rewards strictly between generated and expert actions.

## B Reasoning Benchmarks

The rapid development of large reasoning models has been accompanied by an increasingly diverse set of benchmarks for evaluating reasoning capabilities. Unlike general instruction-following evaluations, reasoning benchmarks emphasize whether models can decompose complex problems, perform multi-step inference, maintain logical consistency, and arrive at verifiable final answers. These benchmarks therefore serve as an essential empirical foundation for studying reasoning-oriented supervised fine-tuning: they not only quantify the effectiveness of different SFT data and objectives, but also expose limitations such as memorization, poor out-of-distribution generalization, and insufficient transfer across reasoning domains.

In this section, we summarize representative benchmarks commonly used to evaluate reasoning models in Table 1. We organize them according to the type of reasoning skill they primarily assess, including mathematical reasoning, code reasoning, general and logical reasoning, scientific reasoning, and spatial reasoning. This overview is intended to provide a practical reference for interpreting experimental results in the literature and for selecting appropriate evaluation suites when developing or analyzing reasoning SFT methods.

## C Comparison between Reasoning SFT and Non-Reasoning SFT

In the study of LLMs, instruct tuning and Reasoning SFT represent two distinct training paradigms. Instruct Tuning focuses on enabling models to comprehend and execute a wide variety of human instructions directly; its outputs are typically final, fluent textual responses tailored to specific tasks such as creative writing, information extraction, or general Q&A. Conversely, Reasoning SFT is dedicated to cultivating the model’s deep logical thinking and problem-solving capabilities. Rather than just delivering the final answer, it trains the model to generate a structured "Chain of Thought" that explicitly details the step-by-step analysis, mathematical derivations, and self-correction processes required to tackle complex mathematical, scientific, or logical challenges. We provide a demonstration of the two in Figure 2.

1720  
1721  
  
1722  
  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
  
1749  
1750  
  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767

**Instruction Tuning:**

Make a persuasive argument to promote recycling.

Recycling is an incredibly important action that we must all take to help preserve our planet for future generations. By recycling, we can reduce the amount of waste that goes to landfills, conserve natural resources and energy, and reduce pollution from production of new materials.

When we don't recycle, our garbage ends up in landfills where it can take hundreds of years to decompose. Landfills can release harmful chemicals and greenhouse gases, polluting the soil, air and water. However, by recycling and reusing materials, we minimize the amount of garbage that ends up in landfills, and therefore reduce the potential harm it can cause to the environment.

Recycling also helps conserve our natural resources by reducing the demand for new materials. For example, recycling paper means fewer trees need to be cut down, and reusing metal and plastic means less mining and oil extraction. This, in turn, conserves energy, minimizes deforestation and reduces the environmental impact associated with producing new materials.

Additionally, recycling also has a positive impact on the economy. It creates jobs in the recycling and manufacturing industries, and reduces the costs of waste disposal for companies and taxpayers. Recycling is good for the environment and the economy.

In conclusion, recycling is an easy and effective way to take action and make a difference for the future of our planet. By simply separating our recyclable materials and disposing of them properly, we can help reduce waste, conserve resources, and protect the environment. We all have a part to play and every little action counts. So, let's make the smart choice and recycle today for a better tomorrow.

**Reasoning SFT:**

Let  $f(x) = 1 + x + x^2 + \dots + x^{100}$ . Find  $f'(1)$ .

**Step 1: Recognize the form of the function**

This is a geometric series:

$$f(x) = \sum_{k=0}^{100} x^k$$

This is the sum of a geometric series with first term 1, common ratio  $x$ , and 101 terms.

The sum of a finite geometric series is:

$$f(x) = \frac{1 - x^{101}}{1 - x} \quad \text{for } x \neq 1$$

This is a standard formula for the sum of a finite geometric series.

**Step 2: Differentiate the function**

We differentiate the expression for  $f(x)$ :

$$f(x) = \frac{1 - x^{101}}{1 - x}$$

We will use the **quotient rule**:

$$f'(x) = \frac{(1 - x)^2 \cdot \frac{d}{dx}(1 - x^{101}) - (1 - x^{101}) \cdot \frac{d}{dx}(1 - x)}{(1 - x)^2}$$

Let's compute the derivatives:

$$-\frac{d}{dx}(1 - x^{101}) = -101x^{100} - \frac{d}{dx}(1 - x) = -1$$

Now plug these into the quotient rule:

$$f'(x) = \frac{(1 - x)^2(-101x^{100}) - (1 - x^{101})(-1)}{(1 - x)^2}$$

Simplify the numerator:

$$(1 - x)^2(-101x^{100}) + (1 - x^{101})$$

Now, evaluate this at  $x = 1$ :

Note that the original expression for  $f(x)$  uses a division by  $1 - x$ , which is zero at  $x = 1$ , so the derivative formula is not directly valid at  $x = 1$ . Instead, we should **differentiate the original series directly**, which is easier.

**Step 3: Differentiate the series directly**

We have:

$$f(x) = \sum_{k=0}^{100} x^k$$

Differentiate term by term:

$$f'(x) = \sum_{k=1}^{100} kx^{k-1}$$

So:

$$f'(1) = \sum_{k=1}^{100} k \cdot 1^{k-1} = \sum_{k=1}^{100} k = \frac{100 \cdot 101}{2} = 5050$$

**Final Answer:**

5050

Figure 2: Comparison of Instruction Tuning and Reasoning SFT.

Task Type	Benchmarks
Mathematical reasoning	<b>Text-only:</b> GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), MATH-401 (Yuan et al., 2023), AIME24/25 (American Institute of Mathematics, 2024), AMC (Mathematical Association of America, 2023), Olympiad Bench (He et al., 2024). <b>Multi-modal:</b> MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a), Geometry3K (Lu et al., 2021), GEOQA (Chen et al., 2021a), UniGEO (Chen et al., 2022), MathVerse (Zhang et al., 2024a), We-MATH (Qiao et al., 2024).
Spatial reasoning	<b>Multi-modal:</b> EmbSpatial-Bench (Du et al., 2024), Space3D-Bench (Szymańska et al., 2024), SpatialRGPT-Bench (Cheng et al., 2024), SpatialVLM (Chen et al., 2024), VSI-Bench (Yang et al., 2025c), OmniSpatial (Jia et al., 2025), 3DSRBench (Ma et al., 2025b), MathSpatial (Lu et al., 2026b), Mmsi-bench (Yang et al., 2025f).
Scientific reasoning	<b>Text-only:</b> MMLU-Pro (Wang et al., 2024b), GPQA (Rein et al., 2024), SuperGPQA (Du et al., 2026), SciQ (Welbl et al., 2017), HLE (Phan et al., 2025), MatSciBench (Zhang et al., 2025c), SciKnowEval (Feng et al., 2024). <b>Multi-modal:</b> ScienceQA (Saikh et al., 2022), MMMU (Yue et al., 2024), MMSciBench (Ye et al., 2025a).
Logical reasoning	<b>Text-only:</b> LogiQA (Liu et al., 2020), AutoLogi (Zhu et al., 2025), ZebraLogic (Lin et al., 2025a), ProofWriter (Taffjord et al., 2021), FOLIO (Han et al., 2024), PrOntoQA (Saparov and He, 2023), CLUTTR (Sinha et al., 2019). <b>Multi-modal:</b> LogicVista (Xiao et al., 2024), VisuLogic (Xu et al., 2025d), MuSLR (Xu et al., 2025c).
Code reasoning	<b>Text-only:</b> SWE-bench (Jimenez et al., 2024), LiveCodeBench (Jain et al., 2025), Sec-CodeBench (Chen et al., 2026c), APPS (Hendrycks et al., 2021a), MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021b), HumanEval+ (Liu et al., 2023), BigCodeBench (Zhuo et al., 2025), NaturalCodeBench (Zhang et al., 2024b). <b>Multi-modal:</b> Design2Code (Si et al., 2025), ChartMimic (Yang et al., 2025b), UniSVG (Li et al., 2025b).

Table 1: Representative benchmarks for evaluating reasoning capabilities.

## D A Unified Taxonomy of Optimization Objectives

To deeply dissect the structural disparities across various optimization paradigms, Table 2, Table 3, Table 4, Table 5, Table 6, and Table 7 systematically formalize the loss functions and gradient expressions of contemporary reasoning-oriented SFT methodologies. By unifying these disparate formulations into a standardized mathematical framework, we crystallize the evolutionary trajectory that spans from fine-grained gradient adjustments (*Token/Sample-Level*) to macroscopic policy constraints (*Distribution-Level*). This taxonomy not only provides an intuitive quantitative baseline for comparing diverse optimization objectives, but also enables readers to clearly demarcate different paradigms based on their data utilization strategies (*offline vs. online*) and the mechanics of their SFT-RL integration.

Table 2: Objective comparison across training paradigms for token reweighting methods.

Method	Paper	Objective Expression
DFT	(Wu et al., 2026a)	$\mathcal{L}_{\text{DFT}} = \mathbb{E}_{(x, y^*) \sim \mathcal{D}} \left[ -\sum_{t=1}^{ y^* } \text{sg}(\pi_\theta(y_t^*   y_{<t}^*, x)) \log \pi_\theta(y_t^*   y_{<t}^*, x) \right]$
VCORE	(Gong et al., 2026)	$\mathcal{L}_{\text{VCORE}} = \sum_{(x, y) \in \mathcal{B}} \left[ \sum_{t \geq 1} \frac{\exp(\tau s_t^i(x, y, \theta))}{\sum_{j \geq 1} \exp(\tau s_j^i(x, y, \theta))} \nabla_\theta \ell_t(\theta; x, y) \right]$ , where $s_t(x, y, \theta) \triangleq \langle \nabla_\theta \mathcal{L}(\theta), \nabla_\theta \ell_t(\theta; x, y) \rangle$
EAFIT	(Diao et al., 2026)	$\mathcal{L}_{\text{EAFIT}} = -\sum_{t=1}^T \frac{H_t^{\text{top-}K}}{3.0} \cdot \log P_\theta(y_t   \mathbf{x}, \mathbf{y}_{<t})$ , where $H_t = -\sum_{v \in \mathcal{V}} P_\theta(v   \mathbf{x}, \mathbf{y}_{<t}) \log P_\theta(v   \mathbf{x}, \mathbf{y}_{<t})$ , $H_t^{\text{top-}K}$ is the entropy of the top- $K$ probability distribution.
DEFT	(Wang et al., 2026f)	$\mathcal{L}_{\text{DEFT}}(\theta) = \mathbb{E}_{(c, \bar{y}) \sim \mathcal{T}} \left[ \frac{1 - p_\theta(\bar{y} c)^{\alpha_{\text{DEFT}}(c)}}{\alpha_{\text{DEFT}}(c)} \right]$ , where $\alpha_{\text{DEFT}}(c) \triangleq \sum_{\nu \in \mathcal{V}} P_\theta(\nu c)^2 \in (0, 1]$
RANKTUNER	(Yu et al., 2026)	$\mathcal{L}(\theta) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ -\sum_{t=1}^T (p_t \cdot \mathbf{S}_t \cdot \log p_t) \right]$ , where $\mathbf{S}_t \approx (p_t \cdot s(H_t))^{-K} (\xi_t)$ , $K(\xi_t) = \lceil \log_2(\xi_t + 1) \rceil^{-2}$ , $\xi_t = \max(R_t, s(H_t))$ , $R_t$ is the ground-truth token's rank
TARL	(Lin et al., 2026)	$\mathcal{L}_{\text{TARL}}(\theta) = -\mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \sum_{t=1}^{ y } w(c_t, y_t) \log P_\theta(y_t   c_t) \right]$ , where $w(c_t, y_t) = \ \nabla_\theta \log P_\theta(y_t   c_t)\  \cdot \exp(-\lambda \cdot \text{KL}(P_\theta(\cdot   c_t) \  Q_{\text{base}}(\cdot   c_t)))$
AESL	(Li et al., 2026a)	$\mathcal{L}_{\text{AESL}}(\theta) = -\mathbb{E}_{q, s^* \sim \mathcal{D}_{\text{SFT}}} \left[ \sum_{t=1}^{ s^* } p(q, s_t^*, \pi_\theta) \cdot \log \pi_\theta(s_t^*   q, s_{<t}) \right]$ , where $p(q, s_t^*, \pi_\theta) = 1 - \text{softmax} \left[ \frac{y(s_t^*   q, s_{<t})}{-t_{\text{caching}} - t_t} \right]$ , $\bar{t}_t = \frac{1}{t} \sum_{i=1}^t -\log \pi_\theta(s_i^*   q, s_{<i})$
GIFT	(Zhao et al., 2026b)	$\mathcal{L}_{\text{GIFT}}(\phi) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \sum_{t=1}^T q_t \cdot \ell_t(\phi) \right]$ [cite: 8600] $\mathcal{L}_{\text{GIFT}}(\phi) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \sum_{t=1}^T q_t \cdot \ell_t(\phi) \right]$ , where $\ell_t(\phi) = -\log p_{\theta_{B, \phi}}(y_t   x, y_{<t})$ , $q_t = p_{\theta_I}(y_t   x, y_{<t})$ , $t = 1, \dots, T$
PEAR	(Zhang et al., 2026a)	$\mathcal{L}_{\text{PEAR}}(\theta) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \sum_{t=1}^T \text{sg}[\hat{G}_t] \cdot l_\theta(x, y_{<t}, y_t) \right]$ , where $G_t = w_{1:T} = \prod_{j=1}^T \Delta_j$ ( $\forall t = 1, \dots, T$ )

Table 3: Objective comparison across training paradigms for token selection methods.

Method	Paper	Objective Expression
SFT-GO	(Kim et al., 2025)	$\mathcal{L}_{\text{SFT-GO}}(\theta) = (1 - \lambda) \ell_{\text{SFT}}(\theta) + \lambda L_{\text{worst}}(\theta)$ , where $\ell_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{t=1}^N \log \pi_{\theta}(y_t   x, y_{<t})$ . Grouping tokens into $g \in \mathcal{G}$ by importance.
TS-PEFT	(Ma et al., 2025a)	Output = $Y_{\text{ backbone}} + Y_{\text{ adapter}}$ , $Y_{\text{ adapter}} = (M \odot X) \cdot A \cdot B$ , $I_t = \text{Selector}(X_t)$ , the selector evaluates token importance scores based on metrics such as hidden state norms or cumulative attention scores.
Token Cleaning	(Pang et al., 2025)	$\mathcal{L}_{\text{Cleaning}}(\theta) = -\frac{1}{\sum_{t=1}^N M_t} \sum_{t=1}^N M_t \cdot \log \pi_{\theta}(y_t   x, y_{<t})$ , $M_t = \begin{cases} 1, & \text{if } I_t \geq \tau \\ 0, & \text{if } I_t < \tau \end{cases}$ , where $I_t = \log \mathbb{P}(x x; \theta') - \log \mathbb{P}(x x; \theta)$ , where $\theta$ represents the base model while $\theta'$ serves as a task-aware reference model.
TOSS	(Li et al., 2026b)	$\mathcal{L}_{\text{TOSS}}(\theta) = -\frac{1}{\sum_{t=1}^N M_t} \sum_{t=1}^N M_t \cdot \log \pi_{\theta}(y_t   x, y_{<t})$ , where $S_t = \text{Evaluator}(y_t   x, y_{<t})$ serves as a safety probe based on a pre-trained model.
SSTOKEN	(Qin et al., 2026)	$\mathcal{L}_{\text{SSTOKEN}}(\theta) = -\frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^T w_t \cdot \log P_{\theta}(y_t   x, y_{<t})$ , $w_t = \sigma\left(\frac{S_t - \tau}{\tau}\right) = \frac{1}{1 + \exp\left(-\frac{S_t - \tau}{\tau}\right)}$ , where $S_t = \text{MLP}\left(\text{LayerNorm}\left(\mathbf{h}_t^{(L)}\right)\right)$ , where $\mathbf{h}_t^{(L)}$ is the raw hidden state vector at the final layer at token position $t$ .
CFT	(Ruan et al., 2025)	$\mathcal{L}_{\text{CFT}}(\theta) = -\frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^T w_t \cdot \log P_{\theta}(y_t   x, y_{<t})$ , $M_t = \begin{cases} 1, & \text{if } S_t < \tau \\ 0, & \text{if } S_t \geq \tau \end{cases}$ , where $S_t = D_{\text{KL}}(P_{\theta}(\cdot   x, y_{<t}) \  P_{\theta, \text{permuted}}(\cdot   x, y_{<t}))$
XTF	(Yang et al., 2026a)	$\mathcal{L}_{\text{XTF}} = -\sum_{O_k \notin D_{\text{noise}}} \log P(O_k   I + [O_0, O_1, \dots, O_{k-1}])$ , $D_{\text{noise}} = (D_{\text{RT}}) \cup (D_{\text{KN}\downarrow}) \cup (D_{\text{TR}\downarrow})$ , $D_{\text{RT}}(O_k) = \mathcal{A}(\theta, I + O)[I + k]$ , $D_{\text{KN}}(O_k) = 1 - P(O_k   I + [O_0, O_1, \dots, O_{k-1}])$ , $D_{\text{TR}}(O_k) = 1 - \text{Normalize}\left(D(\mathcal{E}(O_k), \sum_{\tau, w} \mathcal{E}(\theta, \mathcal{E}(P, w))\right)$ .
ProFit	(Liu et al., 2026b)	$\mathcal{L}_{\text{ProFit}}(\theta) = \mathbb{E}_{(x, y^*) \sim \mathcal{D}} \left[ -\sum_{t=1}^{ y^* } M_t \cdot \log \pi_{\theta}(y_t^*   x, y_{<t}^*) \right]$ , $M_t = \mathbb{I}[sg(\pi_{\theta}(y_t^*   x, y_{<t}^*)) > \tau]$
SED-SFT	(Chen et al., 2026d)	$\mathcal{L}_{\text{SED-SFT}}(\theta) = \sum_{t=1}^{ y^* } [-\log \pi_{\theta}(y_t^*   x, y_{<t}^*) + \lambda \cdot M_t \cdot \mathcal{L}_{\text{DE}}(\pi_{\theta}(y_t^*   x, y_{<t}^*))]$ , $\mathcal{L}_{\text{DE}}(p) = (p - 0.5)^2$ , $M_t = \mathbb{I}[P_{\text{top-k}}(t) < \tau]$
OTR	(Ming et al., 2025)	$\mathcal{L}_{\text{OTR}}(\theta) = \frac{1}{T} \sum_{t=1}^T \mathcal{L}_{\text{OTR}}^t(\theta)$ , $\mathcal{L}_{\text{OTR}}^t(\theta) = -\frac{N^{\text{gt}}}{K} \log \pi_{\theta}(x_t   s_t) - \frac{\beta}{K} \sum_{j \text{ s.t. } a'_{t,j} \neq x_t} a'_{t,j} \log \pi_{\theta}(a'_{t,j}   s_t)$ , $R(a'_{t,j}, x_t) = \begin{cases} 1 & \text{if } a'_{t,j} = x_t \\ \beta & \text{if } a'_{t,j} \neq x_t \end{cases}$
SFTKey	(Shi et al., 2025)	First stage: $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^N \sum_{t=1}^{L_i} \log P(y_{i,t}   x_i, y_{i,<t}; \theta)$ , $y_{i,t}$ includes both the CoT steps and the final answer. Second stage: $\mathcal{L}_{\text{Answer}}(\theta) = -\sum_{i=1}^N \sum_{t=T_i+1}^{L_i} \log P(y_{i,t}   x_i, y_{i,<t}; \theta)$ , $y_{i,t}$ excludes the CoT steps and contains only the final answer.

Table 4: Objective comparison across training paradigms for sample reweighting methods.

Method	Paper	Objective Expression
FLOW	(Sanyal et al., 2025)	$\mathcal{L}_{FLOW}(\theta) = \sum_{i=1}^n w_i \cdot f_i(\theta)$ , $w_i = \exp\left(-\frac{f_i(\theta^*)}{\tau}\right)$ , where $f_i(\theta^*)$ denotes the cross-entropy loss of the $i$ -th sample computed on the pre-trained model.
BFT	(Tang et al., 2025)	$\mathcal{L}_{BFT}(\theta) = \frac{1}{B} \sum_{b=1}^B s_b \cdot \frac{\sum_t m_{b,t} w_{b,t} l_{b,t}}{\sum_t m_{b,t} + \epsilon}$ , where $l_{b,t} = -\log \pi_\theta(y_{b,t}   y_{b,<t}, x_b)$ , $p_b^{\text{conf}} = \min_{G_i \in \mathcal{G}_b} C_{G_i}$ , and $s_b = sg(1 - p_b^{\text{conf}})$ , $w_{b,t} = sg(\pi_\theta(y_{b,t}   y_{b,<t}, x_b)) = sg(c_{b,t})$
iW-SFT	(Qin and Springenberg, 2025)	$\mathcal{L}_{iW-SFT}(\theta) = -\mathbb{E}_x \log p(\tau; \theta)$ , where $W(\tau) = \exp\left(\sum_i \left(\log \pi_q(a_i   s_i) - \log \pi_{\text{ref}}(a_i   s_i)\right)\right)$
BOLT	(Shu et al., 2026)	$\mathcal{L}_{BOLT}(\theta) = -\frac{1}{ \mathcal{E} } \sum_{(x,y,w) \in \mathcal{E}} w \cdot \log \pi_\theta(y   x)$ , where $w(x, y_n) = \frac{\exp(\tau(x, y_n)/\beta)}{Z_N(x)}$ , $\hat{Z}_N(x) = \frac{1}{N} \sum_{n=1}^N \exp(\tau(x, y_n)/\beta)$
ICA	(Zhang et al., 2026b)	$\mathcal{L}_{ICA}(\theta) = \mathbb{E}_{z \sim \mathcal{D}} [w(z) \cdot \mathcal{L}_{\text{base}}(z; \theta)]$ , where $w(z) = \sigma\left(\frac{\text{ScoreICA}(z) - \mu}{\tau}\right)$ , $\text{ScoreICA}(z) = \sum_{z_h \in \mathcal{H}} (\log P_\theta(z_h) - \log P_\theta(z_h   z))$
ScaleBIO	(Pan et al., 2025)	$\mathcal{L}_{\text{ScaleBIO}} = L_1(\lambda, w) + L_1(\lambda, u) + \alpha \cdot (L_2(\lambda, w) - L_2(\lambda, u))$ , where $L_1(\lambda, w) = -\frac{1}{ \mathcal{D}_{\text{train}} } \sum_{i \in \mathcal{D}_{\text{train}}} \lambda_i \cdot \log P_w(y_i   x_i)$ and $L_2(\lambda, w^*(\lambda)) = -\frac{1}{ \mathcal{D}_{\text{val}} } \sum_{j \in \mathcal{D}_{\text{val}}} \log P_{w^*(\lambda)}(y_j   x_j)$

Table 5: Objective comparison across training paradigms for sample selection methods.

Method	Paper	Objective Expression
SelectIT	(Liu et al., 2024)	$Selectedsample : M(X, Y) = \mathbb{I}[\alpha_{\min} \leq \text{Unc}(Y   X) \leq \alpha_{\max}] \wedge \mathbb{I}[S_{\text{ref}}(X, Y) \geq \tau]$ , where $\text{Unc}(Y   X) = -\frac{1}{T} \sum_{t=1}^T \log P_{\theta}(y_t   X, y_{<t})$ . $S_{\text{ref}}$ is the quality score generated by the model through self-reflection based on the prompt, and $\tau$ denotes the quality threshold.
GRAPE	(Zhang et al., 2025b)	$\mathcal{L}_{\text{GRAPE}}(\theta) = -\mathbb{E}_{(x, y^*) \sim \mathcal{D}_{\text{GRAPE}}} \left[ \sum_{t=1}^{ y^* } \log P_{\theta}(y_t^*   x, y^* \text{Ref}_{<t}) \right]$ , where $y^* = \arg \max_{y \in \mathcal{Y}_x} S(y   x)$ , $\mathcal{D}_{\text{GRAPE}} = \{(x, y^*)\}$ , $S(y   x) = \exp \left( \frac{1}{ Y } \sum_{t=1}^{ Y } \log P_{\theta_0}(y_t   x, y_{<t}) \right)$
RSR	(Yang et al., 2026b)	Selected sample: $M(X, Y) = \mathbb{I}[\text{RSR}(Y   X) \geq \tau]$ , $\text{RSR}(Y   X) = \frac{1}{T} \sum_{t=1}^T \text{RSR}_t$ , where $\text{RSR}_t = \frac{-\log P_{\theta}(y_t   X, y_{<t})}{\log \text{Rank}_{\theta}(y_t   X, y_{<t}) + \epsilon}$
DC-SFT	(Lu et al., 2026a)	$Selectedsample : \mathcal{D}_{\text{curated}} = \mathcal{D}_{\text{Easy}} \cup \mathcal{D}_{\text{Medium}} = \{(x, y) \in \mathcal{D} \mid \text{Correct}(x) > 0\}$
LESS	(Xia et al., 2024)	$\text{Score}(z) = \max_j \text{Inf}_{\text{Adam}}(z, \mathcal{D}_{\text{val}}^{(j)}), \text{Inf}_{\text{Adam}}(z, \mathcal{D}_{\text{val}}^{(j)}) = \sum_{i=1}^N \bar{w}_i \cdot \cos \left( \nabla l(\mathcal{D}_{\text{val}}^{(j)}; \theta_i), \tilde{\Gamma}(z, \theta_i) \right), \Gamma(z, \theta_i) = \frac{m_i}{\sqrt{m_i + \epsilon}}$
CADS	(Wan et al., 2025)	$\mathcal{L}_{\text{CADS}}^{\alpha}(\theta, s) = \mathbb{E}_{m \sim p(m s)} \left[ \mathcal{L}_{\text{val}}(\theta) + \alpha \left( \mathcal{L}_{\text{trn}}(\theta, m) - l( m ) \right)^2 \right]$ , $\nabla_s \mathcal{L}_{\text{CADS}}^{\alpha}(\theta, s) = \mathbb{E}_{m \sim p(m s)} \left[ \left( \mathcal{L}_{\text{val}}(\theta) + \alpha \left( \mathcal{L}_{\text{trn}}(\theta, m) - l( m ) \right)^2 \right) \cdot \nabla_s \log p(m s) \right]$
T-SHIRT	(Fu et al., 2025)	$S\text{-IFD}_k(x, y) = \exp \left\{ -\frac{1}{\sum_{t=1}^T w_t} \sum_{t=1}^T w_t \cdot \Delta_t \right\}$ , where $w_t = \begin{cases} 1 & \text{if } \Delta_t \text{ ranks in the top } k\% \text{ of the entire dataset,} \\ 0 & \text{otherwise} \end{cases}$ , $\Delta_t = \log P_{\theta'}(y_t   y_{<t}, x) - \log P_{\theta'}(y_t   y_{<t})$ . $S_{\text{can}} = \arg \max_{\mathcal{D}, \gamma, b} \hat{\mu}(x, y)$ , $S = \arg \min_{\text{scan}, b} \hat{\sigma}^2(x, y)$ , where $\hat{\mu}(x, y) = \frac{1}{M} \sum_{i=1}^M S\text{-IFD}_k(x^{(i)}, y^{(i)})$ , $\hat{\sigma}^2(x, y) = \frac{1}{M} \sum_{i=1}^M \left( S\text{-IFD}_k(x^{(i)}, y^{(i)}) - \hat{\mu}(x, y) \right)^2$ .
Q-Tuning	(Wang et al., 2025)	$\text{Ent}(x, y; \theta) = \frac{1}{ T } \sum_{i \in T} H(\pi_{\theta}(y_i   y_{<i}, x))$ , $PPL(x, y; \theta) = \exp \left( \frac{1}{ T } \sum_{i \in T} -\log \pi_{\theta}(y_i   y_{<i}, x) \right)$ , only Keep samples with high entropy.

Table 6: Objective comparison across training paradigms for distribution-level regularization.

Method	Paper	Objective Expression
ASFT	(Zhu et al., 2026a)	$\mathcal{L}_{ASFT}(\theta) = \mathcal{L}_{DFT}(\theta) + \lambda \cdot D_{KL}(\pi_\theta \parallel \pi_{anchor})$
PSFT	(Zhu et al., 2026c)	$\mathcal{L}_{PSFT}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} [\min(r_t(\theta) \cdot A(x, y), \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)) \cdot A(x, y)]$ , where $r_t(\theta) = \frac{\pi_\theta(y x)}{\pi_{\theta_{old}}(y x)}$
Anchored Learning	(Wang et al., 2026e)	$\theta^{(t+1)} = \arg \min_{\theta} \mathcal{L}_{distill}(\theta; q^{(t)}) \approx \arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [KL(q^{(t)}(\cdot x) \parallel p_\theta(\cdot x))]$ , where $q^{(t)} = (1 - \alpha)p_{\theta^{(t)}} + \alpha p_{sft_t}$
GEM	(Li et al., 2025f)	$\mathcal{L}_{GEM}(\theta) = \sum_i \sum_{t=1}^T \sum_{y^{gene}} q_k(y^{gene}   x_i, y_{1:t-1}^{real}) \cdot [\log f_\theta(y^{gene}   x_i, y_{1:t-1}^{real}) - \log f_\theta(y_i^{real}   x_i, y_{1:t-1}^{real})]$ , where $q = \text{softmax}(\frac{1}{\beta} \cdot \log f)$
CurioSFT	(Wang et al., 2026b)	$\mathcal{L}_{total}(\theta) = \mathcal{L}_{SFT}(\theta) + \alpha \cdot \mathcal{L}_{SED}(\theta)$ , where $\mathcal{L}_{SED}(\theta) = \frac{1}{2} \sum_{t=1}^T \left( \log \frac{\pi_\theta(y_t   s_t; \tau=1.0)}{\pi_{\theta'}(y_t   s_t; \tau_t)} \right)^2$

Table 7: Objective comparison across training paradigms for integration with RL.

Method	Paper	Objective Expression
LESE	(Yang et al., 2026c)	$L = (1 - \alpha) \cdot L_{\text{SFT}} + \alpha \cdot L_{\text{GRPO}}, \alpha = \begin{cases} \min(1, l \cdot (e^M - 1) \cdot \frac{D}{D_{\max}}), & \text{if } M \geq c \\ 0, & \text{if } M < c \end{cases}$ <p>is the theoretical upper bound of the standard deviation of in-group rewards.</p>
HPT	(Lv et al., 2025)	$\mathcal{L} = \alpha \mathcal{L}_{\text{RL}} + \beta \mathcal{L}_{\text{SFT}}, (\alpha, \beta) = \begin{cases} (0, 1), & \text{if } P \leq \gamma \\ (1, 0), & \text{if } P > \gamma \end{cases}$ <p>where <math>P</math> is the mean pass rate of <math>n</math> parallel rollouts for the current question.</p>
AMFT	(He et al., 2025)	$L = (1 - \mu) \cdot L_{\text{RL}}(\theta) + \mu \cdot L_{\text{SFT}}(\theta), \text{ where } \mu_{t+1} = \text{clip}(\mu_t + \eta_{\mu} \nabla_{\mu} U(\theta_t) + \eta_H (H^* - H(\pi_{\theta_t})), \mu_{\min}, \mu_{\max}), \nabla_{\mu} U(\theta_t) \approx -\alpha_{\phi} \nabla_{\theta} U(\theta_t)^{\top} (\nabla_{\theta} L_{\text{SFT}}(\theta_{t-1}) - \nabla_{\theta} L_{\text{RL}}(\theta_{t-1})), H(\pi_{\theta_t}) = \mathbb{E}_{x \sim \mathcal{D}_{\text{SFT}}} [H(\pi_{\theta_W}(\cdot   x))], H^* = \mathbb{E}_{x \sim \mathcal{D}_{\text{SFT}}} [H(\pi_{\theta_W}(\cdot   x))].$
CHORD	(Zhang et al., 2026c)	$\mathcal{L} = (1 - \mu) \mathcal{L}_{\text{GRPO}}(\theta) + \mu \mathcal{L}_{\text{SFT-}\phi}(\theta), \mathcal{L}_{\text{SFT-}\phi}(\theta) = -\mathbb{E}_{(x, y^*) \sim \mathcal{D}_{\text{SFT}}} \left[ \sum_{t=1}^{ y^* } \phi(y_t^*; \pi_{\theta}) \cdot \log \pi_{\theta}(y_t^*   x, y_{<t}^*) \right], \phi(y_t^*; \pi_{\theta}) = p_t(1 - p_t).$
BRIDGE	(Chen et al., 2025c)	$\bar{\theta} = (\theta, w), \text{ where } \theta \text{ is the base model parameters and } w \text{ denotes LoRA parameters. } \mathcal{L}_{\text{loft}}(\theta; w) = -(1 - \lambda) J_{\text{SFT}}(\theta; w) - \lambda J_{\text{RL}}(\theta; w), \mathcal{L}_{\text{clim}}(w) = -(1 - \lambda) J_{\text{SFT}}(\theta, w) - \lambda [J_{\text{RL}}(\theta, w) - J_{\text{RL}}(\hat{\theta}, w)].$ <p><math>J_{\text{SFT}}</math> denotes the supervised fine-tuning objective and <math>J_{\text{RL}}</math> represents the outcome-based reinforcement learning objective.</p>
SRFT	(Fu et al., 2026a)	$\mathcal{L}_{\text{SRFT}}(\theta) = \mathcal{L}_{\text{SFT}}^{\text{demo}}(\theta) + \mathcal{L}_{\text{RL}}^{\text{demo}}(\theta) + \mathcal{L}_{\text{RL}}^{\text{self-rollout}}(\theta), \text{ where } \mathcal{L}_{\text{SFT}}^{\text{demo}}(\theta) = w_{\text{SFT}} \cdot \mathbb{E}_{(x, y) \sim \mathcal{D}_{\text{demo}}} [-\log \pi_{\theta}(y   x)], \mathcal{L}_{\text{RL}}^{\text{demo}}(\theta) = -\mathbb{E}_{(x, y) \sim \mathcal{D}_{\text{demo}}} \left[ \sum_{t=1}^{ y } r_{k,t}(\theta) \cdot \hat{A}_k \right], w_{\text{SFT}} = 0.5 \cdot \text{stop\_grad} \left( e^{-\mathcal{H}(\pi_{\theta})} \right), \mathcal{L}_{\text{RL}}^{\text{self-rollout}}(\theta) = w_{\text{RL}} \cdot \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}} [-\log \pi_{\theta}(y^+   x)] + \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}} [\log \pi_{\theta}(y^-   x)]; w_{\text{RL}} = 0.1 \cdot \text{stop\_grad} \left( e^{\mathcal{H}(\pi_{\theta})} \right).$
ReLIPT	(Ma et al., 2026a)	$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G} \sum_{i=1}^G \frac{1}{ \hat{o}_i } \sum_{t=1}^{ \hat{o}_i } \min [r_{i,t}(\theta) \hat{A}_i, \text{clip}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i], \mathcal{D}_{\text{hard}} = \{q \mid \max_{x_i \in [1, G]} R(o_i(q)) = 0\}, \mathcal{L}_{\text{SFT}}^{\text{hard}}(\theta) = -\frac{1}{ \tau^* } \sum_{t=1}^{ \tau^* } \log \pi_{\theta}(\tau_t^*   q, \tau_{<t}^*), \text{ where } q \in \mathcal{D}_{\text{hard}}$
DYPO	(Zhu et al., 2026b)	$\mathcal{L} = \sum_{q \in \mathcal{B}} [\mathbb{I}(\text{All-Wrong}) \cdot \mathcal{L}_{\text{SFT}}^{\text{Multi-Teacher}}(\theta; q) + \mathbb{I}(\text{Any-Correct}) \cdot \mathcal{L}_{\text{RL}}^{\text{IS}}(\theta; q)], \mathcal{L}_{\text{SFT}}^{\text{Multi-Teacher}}(\theta; q) = -\sum_{m=1}^M w_m \cdot \left[ \frac{1}{ \tau_m^* } \sum_{t=1}^{ \tau_m^* } \log \pi_{\theta}(\tau_{m,t}^*   q, \tau_{m,<t}^*) \right], \mathcal{L}_{\text{RL}}^{\text{IS}}(\theta; q) = -\frac{1}{G} \sum_{i=1}^G \frac{1}{ \hat{o}_i } \sum_{t=1}^{ \hat{o}_i } \min [\bar{r}_{i,t}(\theta) \hat{A}_i, \text{clip}(\bar{r}_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i]$