

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

Yongcan Yu^{1,2*}, Kuangpu Guo^{1,3*}, Jian Liang^{1,2†},
Lingxiao He⁴, Meng Wang⁴, Ran He^{1,2}

¹ NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences

² School of Artificial Intelligence, University of Chinese Academy of Sciences

³ University of Science and Technology of China ⁴ Meituan Inc.

✉ yuyongcan/Awesome-SFT-for-LRMs {yuyongcan0223, liangjian92}@gmail.com

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; Wang et al., 2026f, 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., 2026b; Lu et al., 2026d), agents (Xu et al.,

2026b; Lu et al., 2025b; Xu et al., 2026a), question-answering (Xue et al., 2024; Yu et al., 2025c), 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 reasoning 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

*These authors contributed equally to this work.

†Corresponding author.

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 attempt 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., 2026a). Another stream of literature addresses reasoning-oriented SFT, yet focuses almost exclusively on early-stage data construction (Chen et al., 2025d; Wang et al., 2026d; Li et al., 2025g). 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 θ 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 π_{θ} 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., 2026d) 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

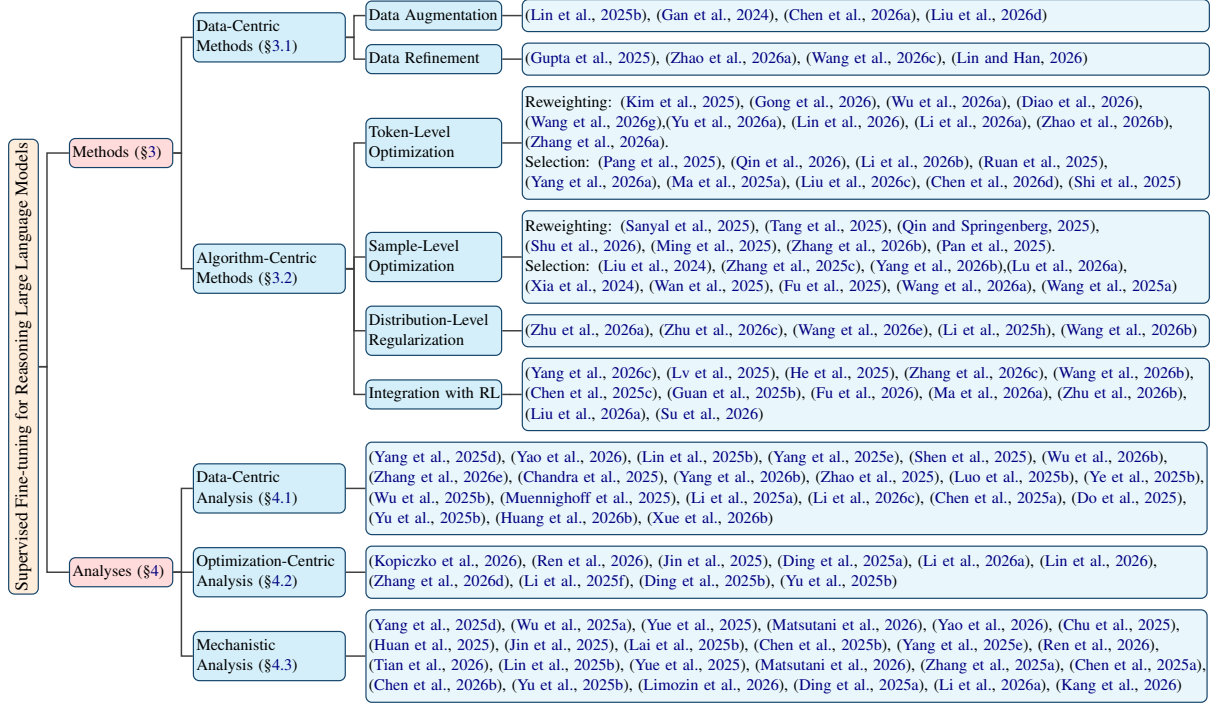


Figure 1: Taxonomy of methodological and analytical research in SFT for reasoning LLMs.

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 reinforcement 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. To ensure a mutually exclusive and clean taxonomy, we adopt *optimization granularity*—progressing from token-level and sample-level modulation to distribution-level regularization and RL integration—as our primary split criterion. While alternative perspectives could categorize these methods based on their underlying data sources or heuristics (e.g., model confidence, sample difficulty, semantic rules, or reference policy divergence), structuring by granularity provides a unified mathematical lens. To bridge these perspectives and clarify potential conceptual overlaps across heuristic motivations, we provide a comprehensive 2D map as shown in Table 1 in the supplementary material, illustrating the intersection of optimization granularity and data heuristics. Accordingly, this subsection system-

atically reviews these algorithmic advancements along the granularity axis.

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

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

where $\mathbf{w}_t^{(i)}$ and α_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 dynamics are modulated by the token weight $\mathbf{w}_t^{(i)}$ 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). However, these methods expose a critical limita-

tion: relying solely on probability or loss struggles to distinguish inherently hard reasoning steps from mere noisy tokens. To address the ambiguity of low-probability tokens, DEFT (Wang et al., 2026g) introduces an entropy-driven trust gate that transitions the optimization focus from exploring uncertain regions to sharpening confident predictions. Building on a similar motivation, RANK-TUNER (Yu et al., 2026a) 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 ℓ_i is the token-level loss and τ 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. These methods recognize a fundamental trade-off: while aggressively upweighting specific tokens maximizes SFT imitation, it risks distributional collapse that destroys the exploration potential crucial for the RL cold start. To mitigate this collapse and preserve exploration potential, 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. This highlights a crucial shift: from sharp, loss-driven imitation for standalone SFT, to soft, distribution-preserving scaling for RL initialization. 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., 2026c) 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 dynamics are governed exclusively by the dynamic sample weight α_i . Within the continuous reweighting framework, one prominent branch calibrates the scalar α_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 auto-

mates 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., 2025c). 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., 2025a) 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) di-

vergence 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., 2025h), 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. To ensure a clear structural boundary, this subsection is strictly confined to approaches utilizing an explicit SFT loss ($\mathcal{L}_{\text{SFT}}$) alongside RL objectives. Frameworks that use demonstrations merely as exploration anchors are discussed in Appendix C.

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., 2025b) dynamically modulates learning weights via entropy scheduling to balance exploration and path refinement. Building upon this uncertainty-aware perspective, SRFT (Fu et al., 2026) 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., 2025e; 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); Liu et al. (2026b) 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) find that smaller learning rates better balance in-distribution and OOD performance.

Model Capability. Reasoning SFT is strongly capability-dependent: Several studies reveal that weak models may struggle to learn the complex patterns encoded in long CoTs (Zhang et al., 2026d; Ren et al., 2026; Li et al., 2025f; Ding et al., 2025b). However, within the domain of VLM reasoning, Yu et al. (2025b) observe that SFT with short CoT disproportionately benefits weaker models, yielding more pronounced improvements. To improve learnability for smaller models, existing work explores easier supervision signals, such as mixing

short-CoT data or trajectories from smaller teachers (Li et al., 2025f), or fine-tuning a medium-sized model as an intermediate teacher to generate data for smaller models (Ding et al., 2025b).

4.3 Mechanistic Analysis

What Does SFT Actually Teach? A central question in reasoning SFT is whether it genuinely teaches new reasoning capabilities or mainly induces models to imitate the surface forms and behavioral patterns of reasoning traces.

Several studies suggest that SFT may primarily reshape model behavior rather than create new reasoning abilities. Yang et al. (2025d) show that even 10 content-corrupted examples can substantially improve reasoning performance, indicating that SFT may mainly induce format imitation. Similarly, models trained on golden shortest paths fail to acquire backtracking capabilities, in contrast to those trained with RLVR (Wei and Kim, 2026; Wen et al., 2026). Wu et al. (2025a) further argue that SFT mainly injects task knowledge into the model and may even compromise its reasoning capability.

In contrast, another line of work argues that SFT can expand the model’s capability boundary rather than merely reshape its output format. From the perspective of Pass@k, several studies suggest that SFT can increase the set of problems that the model is capable of solving (Yue et al., 2025; Matsutani et al., 2026). Moreover, Yao et al. (2026) show that CoT-trained models can compose learned simple skills to solve unseen or more complex problems, providing evidence for genuine reasoning.

Memorization or Generalization? A pivotal inquiry in reasoning SFT is whether the observed improvements stem from the memorization of specific trajectories or the acquisition of transferable reasoning patterns. A critical view is that SFT tends to memorize rather than generalize: Chu et al. (2025) observe a continuous decline in OOD performance on Points24 (Zhai et al., 2024) and navigation tasks (Yang et al., 2024b), leading to the conclusion that “SFT memorizes, RL generalizes”. Related studies report similar limitations, including poor OOD generalization (Jin et al., 2025), degradation on general tasks after math SFT (Huan et al., 2025), and forgetting caused by off-policy supervision (Lai et al., 2025b; Chen et al., 2025b).

However, other studies suggest that SFT can generalize under suitable conditions. Long CoTs, sufficient optimization, high-quality data, and strong

base models have all been identified as important factors for SFT generalization (Yang et al., 2025e; Ren et al., 2026). Moreover, Tian et al. (2026) surprisingly find that training on incorrect CoTs can improve OOD generalization, and Lin et al. (2025b) observes that CoT supervision with diverse prompts can even surpass RL in generalization performance. These findings suggest that poor SFT generalization may reflect unfavorable training conditions rather than an inherent inability.

Interaction between SFT and RL. With the success of DeepSeek-R1 (Guo et al., 2025), combining SFT and RL has become a common post-training paradigm for modern reasoning models (Yu et al., 2025a; Zheng et al., 2025a; Shao et al., 2024; Jaech et al., 2024; Yang et al., 2025a). A mainstream view is that SFT provides basic reasoning behaviors, while RL further refines them through exploration and feedback (Yue et al., 2025; Matsutani et al., 2026; Zhang et al., 2025a). Under this view, Zhang et al. (2025a) argue that post-training should place greater emphasis on SFT, with RL serving mainly as a lightweight refinement stage.

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

Recent studies therefore shift the focus from whether to combine SFT and RL to how SFT should be used for RL initialization. Although various methods have been proposed to better integrate the two paradigms (Fu et al., 2026; Zhang 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.

Limitations

Despite our effort to provide a comprehensive review, this survey has several limitations. First, reasoning SFT is a rapidly evolving field, and new methods and analyses continue to emerge at a fast pace; therefore, some very recent or concurrent studies may not be fully covered. Additionally, this survey primarily focuses on recent studies, especially those published since 2025, in order to capture the latest methodological and analytical developments in reasoning-oriented SFT. As a result, we do not provide a systematic review of earlier

works that mainly concentrated on constructing reasoning or CoT-style training data. Although these efforts laid important foundations for reasoning SFT, a comprehensive historical account of data construction methods is beyond the scope of this survey.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123501), and the National Natural Science Foundation of China under Grants U2441251 and 62276256. We also extend our sincere thanks to the anonymous reviewers for their constructive suggestions.

References

- Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. 2025. Opencodereasoning: Advancing data distillation for competitive coding. *arXiv preprint arXiv:2504.01943*.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proc. EACL*.
- Syeda Nahida Akter, Shrimai Prabhumoye, John Kamalu, Sanjeev Satheesh, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Mind: Math informed synthetic dialogues for pretraining llms. In *Proc. ICLR*.
- American Institute of Mathematics. 2024. Aime 2024 competition mathematical problems.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.
- Hritik Bansal, Devendra Singh Sachan, Kai-Wei Chang, Aditya Grover, Gargi Ghosh, Wen-tau Yih, and Ramakanth Pasunuru. 2026. Honeybee: Data recipes for vision-language reasoners. In *Proc. CVPR*.
- Huanqia Cai, Yijun Yang, and Zhifeng Li. 2024. System-2 mathematical reasoning via enriched instruction tuning. *arXiv preprint arXiv:2412.16964*.
- Abhranil Chandra, Ayush Agrawal, Arian Hosseini, Sebastian Fischmeister, Rishabh Agarwal, Navin Goyal, and Aaron Courville. 2025. Shape of thought: When distribution matters more than correctness in reasoning tasks. *arXiv preprint arXiv:2512.22255*.

- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proc. CVPR*.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. SFT or RL? an early investigation into training rl-like reasoning large vision-language models. *Transactions on Machine Learning Research*.
- Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025b. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*.
- Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. 2026a. Nft: Bridging supervised learning and reinforcement learning in math reasoning. In *Proc. ICLR*.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proc. EMNLP*.
- 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 Findings*.
- Jierun Chen, Tiezheng YU, Haoli Bai, Lewei Yao, Jianan 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 techniques for reasoning VLMs. *Transactions on Machine Learning Research*.
- Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025c. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *arXiv preprint arXiv:2509.06948*.
- 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*.
- 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*.
- 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*.
- Xiaoshu Chen, Sihang Zhou, Ke Liang, Duanyang Yuan, Haoyuan Chen, Xiaoyu Sun, Lingyuan Meng, and Xinwang Liu. 2025e. Putting on the thinking hats: A survey on chain of thought fine-tuning from the perspective of human reasoning mechanism. *arXiv preprint arXiv:2510.13170*.
- Yijie Chen, Yijin Liu, and Fandong Meng. 2026d. Sed-sft: Selectively encouraging diversity in supervised fine-tuning. *arXiv preprint arXiv:2602.07464*.
- 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*.
- Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.
- 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*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Yihe Deng, I Hsu, Jun Yan, Zifeng Wang, Rujun Han, Gufeng Zhang, Yanfei Chen, Wei Wang, Tomas Pfister, Chen-Yu Lee, and 1 others. 2026. Supervised reinforcement learning: From expert trajectories to step-wise reasoning. In *Proc. ICLR*.
- Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang, Zhonghao Yan, Yufei Han, Kongming Liang, Weiran Xu, and Zhanyu Ma. 2026. Entropy-adaptive fine-tuning: Resolving confident conflicts to mitigate forgetting. *arXiv preprint arXiv:2601.02151*.
- Bowen Ding, Yuhang Chen, Jiayang Lv, Jiyao Yuan, Qi Zhu, Shuangshuang Tian, Dantong Zhu, Futing Wang, Heyuan Deng, Fei Mi, and 1 others. 2025a. Rethinking expert trajectory utilization in llm post-training. *arXiv preprint arXiv:2512.11470*.
- Dongyi Ding, Tiannan Wang, Chenghao Zhu, Meiling Tao, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025b. Micota: Bridging the learnability gap with intermediate cot and teacher assistants. *arXiv preprint arXiv:2507.01887*.
- Heejin Do, Jaehui Hwang, Dongyoon Han, Seong Joon Oh, and Sangdoo Yun. 2025. What defines good reasoning in llms? dissecting reasoning steps with multi-aspect evaluation. *arXiv preprint arXiv:2510.20603*.
- Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proc. ACL*.

- Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, and 1 others. 2026. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. In *Proc. NeurIPS*.
- Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. 2024. Sci-knoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*.
- YanJun Fu, Faisal Hamman, and Sanghamitra Dutta. 2025. T-shirt: Token-selective hierarchical data selection for instruction tuning. In *Proc. NeurIPS*.
- Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2026. SRFT: A single-stage method with supervised and reinforcement fine-tuning for reasoning. In *Proc. ICLR*.
- Wangjie Gan, Miao Pan, Linbo Xi, Wenqi Zhang, Jintao Chen, Jianwei Yin, and Xuhong Zhang. 2024. Gft: From imitation to reward fine-tuning with unbiased group advantages and dynamic coefficient rectification. In *Proc. ACL*.
- Xuan Gong, Senmiao Wang, Hanbo Huang, Ruoyu Sun, and Shiyu Liang. 2026. Vcore: Variance-controlled optimization-based reweighting for chain-of-thought supervision. In *Proc. ACL*.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujie Yang, Minlie Huang, Nan Duan, Weizhu Chen, and 1 others. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *Proc. ICLR*.
- Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025a. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.
- Zhong Guan, Likang Wu, Hongke Zhao, Jiahui Wang, and Le Wu. 2025b. Recall-extend dynamics: Enhancing small language models through controlled exploration and refined offline integration. *arXiv preprint arXiv:2508.16677*.
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, and 1 others. 2026. Openthoughts: Data recipes for reasoning models. In *Proc. ICLR*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu, and Sachindra Joshi. 2025. Selective self-to-supervised fine-tuning for generalization in large language models. In *Proc. NAACL*.
- Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhen-ting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, and 1 others. 2024. Folio: Natural language reasoning with first-order logic. In *Proc. EMNLP*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proc. ACL*.
- Lixuan He, Jie Feng, and Yong Li. 2025. Amft: Aligning llm reasoners by meta-learning the optimal imitation-exploration balance. *arXiv preprint arXiv:2508.06944*.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and 1 others. 2021a. Measuring coding challenge competence with apps. In *Proc. NeurIPS*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. In *Proc. NeurIPS*.
- Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. 2025. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Proc. ACL Findings*.
- 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 *Proc. ICML*.
- 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. *arXiv preprint arXiv:2604.14164*.
- 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. *arXiv preprint arXiv:2412.16720*.
- 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 *Proc. ICLR*.

- Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2025. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*.
- 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. *arXiv preprint arXiv:2603.13985*.
- Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2026b. A survey on large language models for code generation. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–72.
- 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 *Proc. ICLR*.
- Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Doina Precup, and Mohammad Hamdaqa. 2025. RL fine-tuning heals the OOD forgetting in SFT. In *Proc. NeurIPS Workshops*.
- Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.
- 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 *Proc. ICLR*.
- Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proc. AAAI*.
- 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. *arXiv preprint arXiv:2506.15021*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proc. NeurIPS*.
- Dawid J Kopiczko, Sagar Vaze, Tijmen Blankevoort, and Yuki M Asano. 2026. Data repetition beats data scaling in long-cot supervised fine-tuning. *arXiv preprint arXiv:2602.11149*.
- Solomon Kullback. 1951. Kullback-leibler divergence. *Encyclopedia of Machine Learning*, pages 581–583.
- Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. 2026. Llm post-training: A deep dive into reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- 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 *Proc. ACL*.
- 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. *arXiv preprint arXiv:2507.05386*.
- 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 *Proc. EMNLP Findings*.
- 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. *Hugging Face repository*.
- 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 *Proc. ACM-MM*.
- Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. 2025c. (mis)fitting scaling laws: A survey of scaling law fitting techniques in deep learning. In *Proc. ICLR*.
- Peiji Li, Kai Lv, Yunfan Shao, Yichuan Ma, Linyang Li, Xiaoqing Zheng, Xipeng Qiu, and Qipeng Guo. 2025d. Fastmcts: A simple sampling strategy for data synthesis. In *Proc. ACL*.
- Xinran Li, Guangda Huzhang, Siqi Shen, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Jun Zhang. 2026a. Getting your LLMs ready for reinforcement learning with lightweight SFT. In *Proc. ICLR*.
- Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and 1 others. 2025e. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. *arXiv preprint arXiv:2507.01921*.
- Yanping Li, Zhening Liu, Zijian Li, Zehong Lin, and Jun Zhang. 2026b. Token-level data selection for safe llm fine-tuning. In *Proc. ICLR*.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025f. Small models struggle to learn from strong reasoners. In *Proc. ACL Findings*.

- Zhaoyi Li, Xiangyu Xi, Zhengyu Chen, Wei Wang, Gangwei Jiang, Ranran Shen, Linqi Song, Ying Wei, and Defu Lian. 2026c. On the role of reasoning patterns in the generalization discrepancy of long chain-of-thought supervised fine-tuning. *arXiv preprint arXiv:2604.01702*.
- Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025g. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.
- Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025h. Preserving diversity in supervised fine-tuning of large language models. In *Proc. ICLR*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *Proc. ICLR*.
- Alexis Limozin, Eduard Durech, Torsten Hoefler, Imanol Schlag, and Valentina Pyatkin. 2026. Sft-then-rl outperforms mixed-policy methods for llm reasoning. *arXiv preprint arXiv:2604.23747*.
- Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025a. Zebralogic: On the scaling limits of LLMs for logical reasoning. In *Proc. ICML*.
- Jiacheng Lin, Zhongruo Wang, Kun Qian, Tian Wang, Arvind Srinivasan, Hansi Zeng, Ruochen Jiao, Xie Zhou, Jiri Gesi, Dakuo Wang, and 1 others. 2026. Sft doesn’t always hurt general capabilities: Revisiting domain-specific fine-tuning in llms. In *Proc. ICLR*.
- Wenye Lin and Kai Han. 2026. Surgical post-training: Cutting errors, keeping knowledge. *arXiv preprint arXiv:2603.01683*.
- Xiaofeng Lin, Hejian Sang, Zhipeng Wang, and Xuezhou Zhang. 2025b. Debunk the myth of sft generalization. *arXiv preprint arXiv:2510.00237*.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Proc. NeurIPS*.
- Jiazhen Liu, Yuchuan Deng, and Long Chen. 2026a. Empowering small vlms to think with dynamic memorization and exploration. In *Proc. ICLR*.
- Jingwen Liu, Ezra Edelman, Surbhi Goel, and Bingbin Liu. 2026b. [Less data, faster training: repeating smaller datasets speeds up learning via sampling biases](#). In *Proc. ICML*.
- Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In *Proc. NeurIPS*.
- Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025a. Uft: Unifying supervised and reinforcement fine-tuning. In *Proc. NeurIPS*.
- Tao Liu, Taiqiang Wu, Runming Yang, Shaoning Sun, Junjie Wang, and Yujiu Yang. 2026c. Profit: Leveraging high-value signals in sft via probability-guided token selection. *arXiv preprint arXiv:2601.09195*.
- Zehua Liu, Shuqi Liu, Tao Zhong, and Mingxuan Yuan. 2026d. Rift: Repurposing negative samples via reward-informed fine-tuning. *arXiv preprint arXiv:2601.09253*.
- Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025b. Acemath: Advancing frontier math reasoning with post-training and reward modeling. In *Proc. ACL Findings*.
- Aojun Lu, Tao Feng, Hangjie Yuan, Wei Li, and Yanan Sun. 2026a. Why does rl generalize better than sft? a data-centric perspective on vlm post-training. *arXiv preprint arXiv:2602.10815*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proc. ICLR*.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proc. ACL*.
- Shuo Lu, Yanyin Chen, Wei Feng, Jiahao Fan, Fengheng Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, and Jian Liang. 2025a. Uni-layout: Integrating human feedback in unified layout generation and evaluation. In *Proc. MM*.
- Shuo Lu, Jianjie Cheng, Yinyu Xu, Yongcan Yu, Lijun Sheng, Peijie Wang, Siru Jiang, Yongguan Hu, Run Ling, Yihua Shao, and 1 others. 2026b. Do mllms really understand space? a mathematical reasoning evaluation. In *Proc. Findings of EMNLP*.
- Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, and 1 others. 2026c. One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation. In *Proc. CIKM*.
- Shuo Lu, Yinyu Xu, Jianjie Cheng, Lingxiao He, Meng Wang, and Jian Liang. 2025b. Deepresearch-slice: Bridging the retrieval-utilization gap via explicit text slicing. *arXiv preprint arXiv:2601.03261*.

- Shuo Lu, Yinyu Xu, Kecheng Yu, Siru Jiang, Yongcan Yu, Yubin Wang, Haitao Yang, Yuxiang Zhang, Bin Wang, Ran He, and 1 others. 2026d. Worldcoder-bench: Benchmarking physically grounded 3d world synthesis. *arXiv preprint arXiv:2606.01869*.
- Zimu Lu, Aojun Zhou, Houxing Ren, Ke Wang, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024b. Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. In *Proc. ACL*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *Proc. ICLR*.
- Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, GengRu Chen, Wenbo Su, and Bo Zheng. 2025b. Deconstructing long chain-of-thought: A structured reasoning optimization framework for long cot distillation. *arXiv preprint arXiv:2503.16385*.
- Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhikai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and 1 others. 2025. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*.
- 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. *arXiv preprint arXiv:2511.16147*.
- 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 *Proc. ICLR*.
- 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 *Proc. ICCV*.
- Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2026b. A survey on vision-language-action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*.
- Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99.
- Mathematical Association of America. 2023. Amc 2023 competition problems.
- 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 *Proc. ICLR*.
- Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. *arXiv preprint arXiv:2509.26313*.
- Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In *Proc. NeurIPS*.
- 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 *Proc. EMNLP*.
- 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 *Proc. ACL*.
- 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 *Proc. ICML*.
- Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. 2024. Openwebmath: An open dataset of high-quality mathematical web text. In *Proc. ICLR*.
- 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. *arXiv preprint arXiv:2501.14249*.
- Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Bäck. 2025. Multi-step reasoning with large language models, a survey. *ACM Computing Surveys*, 58(6):1–35.
- 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? *arXiv preprint arXiv:2407.01284*.
- Chongli Qin and Jost Tobias Springenberg. 2025. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*.
- 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 *Proc. ICLR*.
- 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 *Proc. COLM*.

- 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. *arXiv preprint arXiv:2604.06628*.
- 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. *arXiv preprint arXiv:2308.12950*.
- 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. *arXiv preprint arXiv:2510.10974*.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 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 *Proc. ICML*.
- Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.
- Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *Proc. ICLR*.
- Ranjan Sapkota, Yang Cao, Konstantinos I Rounteliotis, and Manoj Karkee. 2025. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*.
- 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. *arXiv preprint arXiv:2402.03300*.
- Ming Shen. 2024. Rethinking data selection for supervised fine-tuning. *arXiv preprint arXiv:2402.06094*.
- Si Shen, Fei Huang, Zhixiao Zhao, Chang Liu, Tiansheng Zheng, and Danhao Zhu. 2025. Long is more important than difficult for training reasoning models. *arXiv preprint arXiv:2503.18069*.
- Xiaofeng Shi, Qian Kou, Yuduo Li, and Hua Zhou. 2025. Rethinking supervised fine-tuning: Emphasizing key answer tokens for improved llm accuracy. *arXiv preprint arXiv:2512.21017*.
- 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. *arXiv preprint arXiv:2605.02469*.
- 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 *Proc. NAACL*.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. In *Proc. EMNLP*.
- Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. 2026. Trust-region adaptive policy optimization. In *Proc. ICLR*.
- 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. *arXiv preprint arXiv:2505.18536*.
- 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. *ACM Computing Surveys*, 57(11):1–43.
- Emilia Szymańska, Mihai Dusmanu, Jan-Willem Burchage, Mahdi Rad, and Marc Pollefeys. 2024. Space3d-bench: Spatial 3d question answering benchmark. In *Proc. ECCV*.
- Oyvind Taffjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Proc. ACL Findings*.
- 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. *arXiv preprint arXiv:2511.21075*.
- 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. *arXiv preprint arXiv:2601.04992*.
- Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. In *Proc. NeurIPS*.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacani, Alexan Ayrapetyan, and Igor Gitman. 2025. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. In *Proc. ICLR*.
- 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 *Proc. NeurIPS*.

- Weilin Wan, Weizhong Zhang, and Cheng Jin. 2025. Computational budget should be considered in data selection. In *Proc. NeurIPS*.
- 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 *Proc. ACL*.
- 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 *Proc. ACL*.
- 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 *Proc. ACL*.
- 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. *arXiv preprint arXiv:2602.11220*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *Proc. NeurIPS*.
- Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024b. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. In *Proc. ICLR*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proc. ACL*.
- 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. *ACM Computing Surveys*, 58(8):1–35.
- Shaobo Wang, Jiaming Wang, Jiajun Zhang, Cong Wang, Yue Min, Zichen Wen, Xingzhang Ren, Fei Huang, Huiqiang Jiang, Junyang Lin, and 1 others. 2025a. Winning the pruning gamble: A unified approach to joint sample and token pruning for efficient supervised fine-tuning. *arXiv preprint arXiv:2509.23873*.
- Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. 2026e. Stabilizing llm supervised fine-tuning via explicit distributional control. *arXiv preprint arXiv:2605.04468*.
- Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Proc. NeurIPS*.
- Yanbo Wang, Minzheng Wang, Jian Liang, Lu Wang, Yongcan Yu, and Ran He. 2026f. Mitigating the safety-utility trade-off in llm alignment via adaptive safe context learning. In *Proc. ICLR*.
- Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025b. A comprehensive survey on trustworthiness in reasoning with large language models. *arXiv preprint arXiv:2509.03871*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024d. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Proc. NeurIPS*.
- Zecheng Wang, Deyuan Liu, Chunshan Li, Yupeng Zhang, Zhengyun Zhao, Dianhui Chu, Bingning Wang, and Dianbo Sui. 2026g. Gradients must earn their influence: Unifying sft with generalized entropic objectives. *arXiv preprint arXiv:2602.11424*.
- 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 *Proc. NeurIPS*.
- Stanley Wei and Juno Kim. 2026. Provable benefits of RLVR over SFT for reasoning models: Learning to backtrack efficiently. In *Proc. ICLR Workshops*.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*.
- 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 *Proc. ICLR*.
- 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. *arXiv preprint arXiv:2506.02126*.
- 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. *arXiv preprint arXiv:2508.09883*.
- 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 *Proc. ICLR*.
- 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 *Proc. ICLR*.

- 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. *arXiv preprint arXiv:2402.05808*.
- Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. In *Proc. ICML*.
- Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.
- Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*.
- Fengli Xu, Qian Yue 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. *Patterns*, 6(10).
- 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 *Proc. ICCV*.
- Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025c. Redstar: does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.
- 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. 2025d. Muslr: Multimodal symbolic logical reasoning. In *Proc. NeurIPS*.
- Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, and 1 others. 2025e. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*.
- Yinuo Xu, Yuwei Liang, Jianjie Cheng, Meng Wang, Yongcan Yu, Shuo Lu, and Jian Liang. 2026a. Dualstake: Dual-path confidence calibration in deep research agent. In *Proc. EMNLP*.
- Yinuo Xu, Shuo Lu, Jianjie Cheng, Meng Wang, Qianlong Xie, Xingxing Wang, Ran He, and Jian Liang. 2026b. How to train your deep research agent? prompt, reward, and policy optimization in search-r1. *arXiv preprint arXiv:2602.19526*.
- Chao Xue and Ziyuan Gao. 2025. Structcoh: Structured contrastive learning for context-aware text semantic matching. In *Pacific Rim International Conference on Artificial Intelligence*, pages 300–315. Springer.
- Chao Xue, Di Liang, Pengfei Wang, and Jing Zhang. 2024. Question calibration and multi-hop modeling for temporal question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19332–19340.
- Chao Xue, Di Liang, Sirui Wang, Jing Zhang, and Wei Wu. 2023. Dual path modeling for semantic matching by perceiving subtle conflicts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, and 1 others. 2026a. Reason only when needed: Efficient generative reward modeling via model-internal uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 23302–23319.
- Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, and 1 others. 2026b. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30186–30213.
- 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 *Proc. NeurIPS*.
- 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. *arXiv preprint arXiv:2604.16557*.
- 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. *arXiv preprint arXiv:2505.09388*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- 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 *Proc. ICLR*.
- Jihan Yang, Runyu Ding, Ellis Brown, Xiaojuan Qi, and Saining Xie. 2024b. V-irl: Grounding virtual intelligence in real life. In *Proc. ECCV*.

- 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 *Proc. CVPR*.
- 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 *Proc. EMNLP*.
- Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. 2025e. Demystifying long chain-of-thought reasoning. In *Proc. ICML*.
- Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, and 1 others. 2025f. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*.
- 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 *Proc. ICLR*.
- 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 *Proc. ACL*.
- 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.
- 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 *Proc. ICLR*.
- Xinwu Ye, Chengfan Li, Siming Chen, Wei Wei, and Robert Tang. 2025a. Mmsci-bench: Benchmarking language models on chinese multimodal scientific problems. In *Proc. ACL Findings*.
- Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025b. Limo: Less is more for reasoning. In *Proc. COLM*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *Proc. ICLR*.
- 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 *Proc. NeurIPS*.
- Wenhao Yu, Shaohang Wei, Jiahong Liu, Yifan Li, Minda Hu, Aiwei Liu, Hao Zhang, and Irwin King. 2026a. Probability-entropy calibration: An elastic indicator for adaptive fine-tuning. In *Proc. ICML*.
- Yongcan Yu, Lingxiao He, Jian Liang, Kuangpu Guo, Meng Wang, Qianlong Xie, Xingxing Wang, and Ran He. 2026b. Understanding and mitigating spurious signal amplification in test-time reinforcement learning for math reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 37424–37436.
- 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. *arXiv preprint arXiv:2512.12690*.
- Yongcan Yu, Yanbo Wang, Ran He, and Jian Liang. 2025c. Test-time immunization: A universal defense framework against jailbreaks for (multimodal) large language models. *arXiv preprint arXiv:2505.22271*.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruofei Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proc. CVPR*.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2024b. Mammoth: Building math generalist models through hybrid instruction tuning. In *Proc. ICLR*.
- 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 *Proc. NeurIPS*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Proc. NeurIPS*.
- Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Sheng-bang 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 *Proc. NeurIPS*.
- Charlie Zhang, Graham Neubig, and Xiang Yue. 2025a. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*.
- Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhan-feng Mo, Qi Zhang, and 1 others. 2025b. 100 days after deepseek-rl: A survey on replication studies and more directions for reasoning language models. *arXiv preprint arXiv:2505.00551*.

- Dylan Zhang, Qirun Dai, and Hao Peng. 2025c. The best instruction-tuning data are those that fit. In *Proc. NeurIPS*.
- Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. 2026a. Good sft optimizes for sft, better sft prepares for reinforcement learning. *arXiv preprint arXiv:2602.01058*.
- Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and 1 others. 2025d. Matscibench: Benchmarking the reasoning ability of large language models in materials science. *arXiv preprint arXiv:2510.12171*.
- 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 *Proc. ICLR*.
- 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 *Proc. ECCV*. Springer.
- Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025e. Improve vision language model chain-of-thought reasoning. In *Proc. ACL*.
- 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. *arXiv preprint arXiv:2405.04520*.
- 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 *Proc. ICLR*.
- 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 *Proc. NeurIPS*.
- 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 *Proc. ICLR*.
- 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 *Proc. ICASSP*.
- 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 *Proc. EMNLP*.
- 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. *arXiv preprint arXiv:2601.09233*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025a. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. Processbench: Identifying process errors in mathematical reasoning. In *Proc. ACL*.
- 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. *arXiv preprint arXiv:2311.10372*.
- 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. *arXiv preprint arXiv:2504.21277*.
- Kun Zhou, Beichen Zhang, Jiapeng Wang, Zhipeng Chen, Wayne X Zhao, Jing Sha, Zhichao Sheng, Shijin Wang, and Ji-Rong Wen. 2024. Jiuzhang3.0: Efficiently improving mathematical reasoning by training small data synthesis models. In *Proc. NeurIPS*.
- He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2026a. Anchored supervised fine-tuning. In *Proc. ICLR*.
- 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. *arXiv preprint arXiv:2502.16906*.
- 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 *Proc. ACL*.
- Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2026c. Proximal supervised fine-tuning. In *Proc. ICLR*.
- 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 *Proc. ICLR*.

A Data-Centric Methods

Reasoning SFT is shaped by how problems are sourced, reasoning trajectories are constructed and verified, and training examples are selected. We therefore organize data-centric methods according to the pipeline:

Problem Sourcing $\rightarrow$ *CoT Trajectory Synthesis* $\rightarrow$
Trajectory Verification $\rightarrow$
Data Filtering and Curation.

When a work spans multiple stages, we assign it to the stage of its primary contribution.

Problem Sourcing. Human-authored datasets such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) provide reliable mathematical problems, multi-step solutions, and broad difficulty coverage, but are expensive to construct and scale. Synthetic resources, including principled logic corpora and ProofWriter (Tafjord et al., 2021), generate theories, questions, and deduction chains with controlled structures (Morishita et al., 2024). PrOntoQA (Saparov and He, 2022) also uses synthetic logical instances, but primarily serves as a benchmark for analyzing deductive reasoning rather than as a standard SFT corpus. Other works mine or transform public resources: DeepSeekMath (Shao et al., 2024) constructs a large mathematical corpus from filtered web data, NuminaMath (Li et al., 2024) releases a large collection of competition-math problems and solutions, and MIND (Akter et al., 2025) converts OpenWebMath (Paster et al., 2024) into synthetic dialogues for mathematical pretraining. Problem reformulation methods further expand existing distributions through question rewriting (Yu et al., 2024), back-translation (Lu et al., 2024b), or distilled problem synthesis (Zhou et al., 2024).

CoT Trajectory Synthesis. Trajectory synthesis converts problems into explicit reasoning demonstrations. (Cai et al., 2024) adds reasoning plans and intermediate steps to existing answers, while STaR (Zelikman et al., 2022) bootstraps rationales from question-answer pairs with known final answers but without rationale annotations. Teacher-model distillation and self-improvement produce large-scale reasoning data, including hybrid natural-language and program-based solutions in MAMmoTH (Yue et al., 2024b), tool-integrated trajectories in ToRA (Gou et al., 2024)

and MathCoder (Wang et al., 2024b), and mathematical and coding reasoning traces in Qwen2.5-Math (Yang et al., 2024a) and OpenCodeReasoning (Ahmad et al., 2025), respectively. Sampling-based methods explore multiple candidate paths through search or self-evolution, as in FastMCTS (Li et al., 2025d) and rStar-Math (Guan et al., 2025a). DeepSeek-Prover (Xin et al., 2024) extends this approach to Lean theorem-proving trajectories.

Trajectory Verification. Verification can target final outcomes or intermediate reasoning steps. Exact answer matching (Cobbe et al., 2021), code execution (Gou et al., 2024; Wang et al., 2024b), unit tests (Ahmad et al., 2025), and Lean proof checking (Xin et al., 2024) provide deterministic signals for mathematical, executable, and formal solutions. However, execution-based filtering is not always beneficial: OpenCodeReasoning reports that discarding all execution-failing solutions can reduce downstream performance (Ahmad et al., 2025). Process-level verification is studied through human step annotations in PRM800K (Lightman et al., 2024), automatic process supervision in Math-Shepherd (Wang et al., 2024c), and error localization in ProcessBench (Zheng et al., 2025b), the latter being primarily an evaluation benchmark rather than an SFT dataset. AceMath (Liu et al., 2025b) and Qwen2.5-Math (Yang et al., 2024a) additionally use outcome-level reward models to rank complete solutions, which differs from process-level verification.

Data Filtering and Curation. The final stage determines which candidate examples enter the training corpus. DART-Math (Tong et al., 2024) allocates additional generation budget to difficult problems through difficulty-aware rejection tuning. In contrast, RedStar (Xu et al., 2025c) primarily studies the effects of Long-CoT data scale and problem difficulty rather than introducing a dedicated filtering algorithm. s1 (Muennighoff et al., 2025) constructs a compact dataset based explicitly on quality, difficulty, and diversity, while LIMO (Ye et al., 2025b) demonstrates the effectiveness of a small collection of difficult, diverse problems paired with high-quality reasoning traces. HoneyBee (Bansal et al., 2026) investigates analogous data construction and scaling dimensions for multimodal reasoning, while OpenThoughts (Guha et al., 2026) systematically analyzes question sourcing, answer generation, filtering, deduplication, and

Table 1: The 2D Map: Gradient Granularity vs. Data Source/Heuristic.

Granularity	Model Confidence / Entropy	Sample Difficulty	Semantic / Structural Rules	Reference Policy Divergence
Token-Level	DFT, EAFT, DEFT, RANKTUNER, ProFit, SED-SFT	SFT-GO, TS-PEFT, TALR	SFTKey, ssToken, CFT, XTF	PEAR, AESL, GIFT, Token Cleaning, TOSS, OTR
Sample-Level	SelectIT, GRAPE, Q-Tuning	FLOW, BFT, RSR, DC-SFT	T-SHIRT, ASLEC	iw-SFT, BOLT, ICA, LESS, CADS, ScaleBiO
Distribution-Level	CurioSFT	(N/A)	(N/A)	ASFT, PSFT, Anchored Learning, GEM
RL Integration	RED, SRFT, Prefix-RFT	LESE, ReLIFT, DYPO, AMFT, HPT	BRIDGE, CHORD, TRAPO, BREAD, SRL, DyME, UFT, and R^3	LUFFY, S-GRPO

teacher selection. OpenMathInstruct-2 (Toshniwal et al., 2025) investigates large-scale mathematical instruction data, with particular attention to solution quality and question diversity, whereas NaturalThoughts (Li et al., 2025e) studies how to select teacher-generated reasoning trajectories that transfer efficiently to student models. Outcome-level reward models in AceMath (Liu et al., 2025b) and Qwen2.5-Math (Yang et al., 2024a) can additionally score and rank complete candidate solutions.

Overall, problem sourcing determines the input distribution, trajectory synthesis determines the reasoning behaviors available for supervision, verification estimates trajectory reliability, and filtering determines the final training distribution. Difficulty, diversity, model provenance, and reward scores are therefore criteria within the pipeline rather than independent top-level categories.

B A Two-Dimensional Taxonomy of Algorithm-Centric Methods

As detailed in Section 3.2 of the main paper, our primary taxonomy categorizes algorithm-centric methods based on their *optimization granularity*—encompassing Token-Level, Sample-Level, Distribution-Level, and RL Integration. These core dimensions constitute the Y-axis of Table 1. To provide a more comprehensive perspective and clarify how these methods intersect with various heuristic motivations, this 2D map cross-references the optimization granularities with their underlying data sources and algorithmic heuristics (the X-axis).

The definitions for each heuristic category along the X-axis are explicitly stated as follows: **Model**

Confidence / Entropy: Modulating optimization signals via internal probability distributions. **Sample Difficulty:** Allocating training focus by diagnosing whether a sample aligns with the model’s own distribution, rather than its general ambiguity. **Semantic / Structural Rules:** Applying structural constraints or filters via the semantic content of trajectories. **Reference Policy Divergence:** Guiding learning dynamics by measuring deviations against a reference policy.

C 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) (Xue et al., 2026a; Lu et al., 2026c, 2025a; Yu et al., 2026b). 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., 2025a) 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.

D 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 or retrieval tasks (Xue et al., 2023; Xue and Gao, 2025), 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 2. 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.

E 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.

F Bridging Analytical Insights and Methodological Design

To explicitly trace how empirical observations motivate methodological innovations, we outline four key connections between our analytical findings and the corresponding method designs:

Distributional Alignment Drives Data Refinement. Section 4.1 show that LLM-paraphrased CoTs outperform manual ones despite errors, indicating that distributional alignment matters more than surface correctness. These findings directly motivate the data refinement methods in Section 3. Self-alignment frameworks replace expert demonstrations with the model’s own in-distribution generations, explicitly optimizing for distributional alignment rather than importing off-policy trajectories.

Diversity Requirements Motivate Hybrid Trajectory Augmentation. Studies in Section 4.1

Instruction Tuning:	Reasoning SFT:
Make a persuasive argument to promote recycling.	Let $f(x) = 1 + x + x^2 + \cdots + x^{100}$. Find $f'(1)$.
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.	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	Text-only: 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). Multi-modal: MathVista (Lu et al., 2024a), 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	Multi-modal: 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	Text-only: MMLU-Pro (Wang et al., 2024d), GPQA (Rein et al., 2024), SuperGQA (Du et al., 2026), SciQ (Welbl et al., 2017), HLE (Phan et al., 2025), MatSciBench (Zhang et al., 2025d), SciKnowEval (Feng et al., 2024). Multi-modal: ScienceQA (Saikh et al., 2022), MMMU (Yue et al., 2024a), MMSciBench (Ye et al., 2025a).
Logical reasoning	Text-only: LogiQA (Liu et al., 2020), AutoLogi (Zhu et al., 2025), ZebraLogic (Lin et al., 2025a), ProofWriter (Tafford et al., 2021), FOLIO (Han et al., 2024), PrOntoQA (Saparov and He, 2023), CLUTTR (Sinha et al., 2019). Multi-modal: LogicVista (Xiao et al., 2024), VisuLogic (Xu et al., 2025e), MuSLR (Xu et al., 2025d).
Code reasoning	Text-only: 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). Multi-modal: Design2Code (Si et al., 2025), ChartMimic (Yang et al., 2025b), UniSVG (Li et al., 2025b).

Table 2: Representative benchmarks for evaluating reasoning capabilities.

highlight the importance of prompt and response diversity for generalization, with purely imitative training leading to overfitting and limited robustness. This finding drives data augmentation methods such as GFT (Gan et al., 2024), which mitigates single-path dependency by pairing queries with hybrid ensembles of expert, teacher, and self-generated trajectories to actively enrich training diversity.

Overfitting Risks Motivate Selective Token Optimization. Section 4.3 indicate that excessive SFT can compromise reasoning capabilities and lead to poor generalization. These risks motivate token-level interventions that constrain the scope and intensity of SFT updates. Token-masking strategies like SFTKey (Shi et al., 2025) restrict optimization to answer tokens. Similarly, reweighting methods such as DEFT (Wang et al., 2026g) and VCORE (Gong et al., 2026) modulate per-token gradients to prioritize high-entropy or high-loss tokens, selectively limiting the degree of behavioral change imposed by SFT.

SFT-RL Complementarity Drives Integration Methods. Section 4.3 reveal that SFT expands the model’s capability boundary (Pass@k), while RL refines solution quality through exploration. This complementary relationship motivates the

SFT+RL integration methods: SRFT (Fu et al., 2026) preserves exploration potential via entropy-regularized SFT, and CHORD (Zhang et al., 2026c) schedules smooth transitions from SFT to RL.

These four connections demonstrate how empirical findings expose fundamental bottlenecks, which in turn motivate specific methodological innovations. By making these causal links explicit, we explain not only what methods exist, but why they were developed and how they address observed phenomena.

G A Unified Taxonomy of Optimization Objectives

To deeply dissect the structural disparities across various optimization paradigms, Table 3, Table 4, Table 5, Table 6, Table 7, and Table 8 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.

H Comparison with Existing Surveys

To better position our contributions and address related literature, we provide a detailed comparison between our work and existing related surveys in Table 9.

Summary: Among 12 related surveys, **none** provide systematic coverage of SFT methodology for reasoning. Prior surveys either: (a) focus on RL rather than SFT (Chen et al., 2025d; Xu et al., 2025a; Zhou et al., 2025; Sun et al., 2025a), (b) cover SFT only peripherally (Chen et al., 2025e; Zhang et al., 2025b; Kumar et al., 2026; Li et al., 2025g; Sun et al., 2025b; Wang et al., 2026d), (c) focus on inference-time prompting (Plaat et al., 2025), or (d) are domain-specific (Wang et al., 2026d). Our survey fills this gap as the **first dedicated treatment of reasoning SFT methodology**.

I Quantitative Meta-Analysis on a Shared Baseline

We originally refrained from including a quantitative meta-analysis because **unconstrained cross-paper comparisons are scientifically unreliable due to vastly different base models, training datasets, and evaluation setups across the literature**. We were concerned that an uncuration aggregation might present misleading conclusions about relative method performance.

However, we agree with the reviewer that an empirical synthesis adds substantial practical value. To provide a rigorous and fair comparison, **we have now added a quantitative analysis focused strictly on methods sharing the exact same baseline: Qwen2.5-Math-7B fine-tuned on NuminaMath-CoT**. We report their macro-average performance across four shared benchmarks (Minerva Math, OlympiadBench, AIME 2024, and AMC 2023), collecting results from the DFT, RIFT, RASFT, and RANKTUNER papers. Note that RFT* and RIFT* are evaluated as exceptions, as they train on self-generated responses rather than the static NuminaMath-CoT annotations.

Three main observations emerge from this synthesis:

1. **Modified SFT objectives do not guarantee gains.** EAFT (16.5) and ASFT (26.4) underperform vanilla SFT (26.8), highlighting that the success of reweighting or regularization heavily depends on how well its inductive bias matches the task.
2. **Early baselines remain remarkably strong.** Simpler, earlier methods like DFT (29.4) and iw-SFT (32.0) outperform several recently proposed, more complex variants. Chronological novelty does not strictly translate to better performance.
3. **Self-generated CoTs outperform external trajectories.** RFT* and RIFT*, which train on self-generated responses, achieve the highest averages (30.7 and 35.6). This suggests that self-generated trajectories better match the model’s native policy distribution and are therefore easier to internalize than externally provided CoTs.

We note that these observations are derived strictly from this specific model and dataset configuration, serving as an **empirical reference rather than universal conclusions**.

Table 3: 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)	$\mathcal{L}_{\text{VCORE}} = \sum_{(x, y) \in \mathcal{B}} \left[\sum_{t \geq 1} \frac{\exp(\tau s_t^*(x, y, \theta))}{\sum_{j \geq 1} \exp(\tau s_j^*(x, y, \theta))} \nabla_\theta \ell_t(\theta; x, y) \right]$, where $s_t(x, y, \theta) \triangleq \langle \nabla_\theta \mathcal{L}(\theta), \nabla_\theta \ell_t(\theta; x, y) \rangle$
EAFt	(Diao et al., 2026)	$\mathcal{L}_{\text{EAFt}} = -\sum_{t=1}^T \frac{H_t^{\text{top-20}}}{3.0} \cdot \log P_\theta(y_t \mathbf{x}, \mathbf{y}_{<t})$, 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., 2026g)	$\mathcal{L}_{\text{DEFT}}(\theta) = \mathbb{E}_{(c, \tilde{y}) \sim \mathcal{T}} \left[\frac{1 - p_\theta(\tilde{y} c) \alpha_{\text{DEFT}}(c)}{\alpha_{\text{DEFT}}(c)} \right]$, where $\alpha_{\text{DEFT}}(c) \triangleq \sum_{\nu \in \mathcal{V}} P_\theta(\nu c)^2 \in (0, 1]$
RANKTUNER	(Yu et al., 2026a)	$\mathcal{L}(\theta) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[-\sum_{t=1}^T (p_t \cdot \mathcal{S}_t) \log p_t \right]$, where $\mathcal{S}_t \approx (p_t \cdot s(H_t))^{-K(\xi_t)}$, $K(\xi_t) = [\log_2(\xi_t + 1)]^{-2}$, $\xi_t = \max(R_t, s(H_t))$, R_t is the ground-truth token's rank
TALR	(Lin et al., 2026)	$\mathcal{L}_{\text{TALR}}(\theta) = -\mathbb{E}_{(x, y) \sim \mathcal{D}} \left[\sum_{t=1}^{ y } w(c_t, y_t) \log P_\theta(y_t c_t) \right]$, 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 $p(q, s_t^*, \pi_\theta) = 1 - \text{softmax} \left[\frac{y(s_t^* q, s_{<t})}{-t_{\text{scaling}} - \bar{t}_t} \right]$, $\bar{t}_t = \frac{1}{t} \sum_{i=1}^t -\log \pi_\theta(s_i^* q, s_{<i})$
GIFT	(Zhao et al., 2026b)	$\mathcal{L}_{\text{GIFT}}(\phi) = \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[\sum_{t=1}^T q_t \cdot \ell_t(\phi) \right]$, 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 4: Objective comparison across training paradigms for token selection methods.

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

Table 5: 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)	$\mathcal{L}_{BFT}(\theta) = \frac{1}{B} \sum_{b=1}^B s_b \cdot \frac{\sum_{t=1}^T m_{b,t} w_{b,t} l_{b,t}}{\sum_{t=1}^T m_{b,t} + \epsilon}$, 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)	$\mathcal{L}_{\text{iw-SFT}}(\theta) = -\mathbb{E}_{\tau \in \mathcal{D}} [W(\tau) \cdot \log p(\tau; \theta)]$, 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)	$\mathcal{L}_{\text{BOLT}}(\theta) = -\frac{1}{ \mathcal{E} } \sum_{(x,y,w) \in \mathcal{E}} w \cdot \log \pi_\theta(y x)$, where $w(x, y_n) = \frac{\exp(r(x, y_n)/\beta)}{\hat{Z}_N(x)}$, $\hat{Z}_N(x) = \frac{1}{N} \sum_{n=1}^N \exp(r(x, y_n)/\beta)$
ICA	(Zhang et al., 2026b)	$\mathcal{L}_{\text{ICA}}(\theta) = \mathbb{E}_{z \sim \mathcal{D}} [w(z) \cdot \mathcal{L}_{\text{base}}(z; \theta)]$, 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)	$\mathcal{L}_{\text{ScaleBiO}} = L_1(\lambda, w) + L_1(\lambda, u) + \alpha \cdot \left(L_2(\lambda, w) - L_2(\lambda, u)\right)$, 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 6: 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_{ref} is the quality score generated by the model through self-reflection based on the prompt, and τ denotes the quality threshold.
GRAPE	(Zhang et al., 2025c)	$\mathcal{L}_{\text{GRAPE}}(\theta) = -\mathbb{E}_{(x, y^*) \sim \mathcal{D}_{\text{GRAPE}}} \left[\sum_{t=1}^{ y^* } \log P_{\theta}(y_t^* x, y^* \text{Ref}_{<t}) \right]$, 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)	$\text{Score}(z) = \max_j \inf_{\text{Adam}}(z, \mathcal{D}_{\text{val}}^{(j)}), \inf_{\text{Adam}}(z, \mathcal{D}_{\text{val}}^{(j)}) = \sum_{i=1}^N \bar{w}_i \cdot \cos \left(\bar{\nabla} l(\mathcal{D}_{\text{val}}^{(j)}; \theta_i), \tilde{\Gamma}(z, \theta_i) \right), \Gamma(z, \theta_i) = \frac{m_i}{\sqrt{\sigma_i^2 + \epsilon}}$
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 $w_t = \begin{cases} 1 & \text{if } \Delta_t \text{ ranks in the top } k\% \text{ of the entire dataset,} \\ 0 & \text{otherwise} \end{cases}$, $\Delta_t = \log P_{\theta'}(y_t y_{<t}, x) - \log P_{\theta'}(y_t y_{<t})$. $S_{\text{can}} = \arg \max_{\mathcal{D}, \gamma, b} \hat{\mu}(x, y), S = \arg \min_{S_{\text{can}}, b} \hat{\sigma}^2(x, y)$, 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., 2025a)	$Ent(x, y; \theta) = \frac{1}{ T } \sum_{i \in T} H(\pi_{\theta}(y_i y_{<i}, x)), PPL(x, y; \theta) = \exp \left(\frac{1}{ T } \sum_{i \in T} -\log \pi_{\theta}(y_i y_{<i}, x) \right)$, only Keep samples with high entropy.

Table 7: 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 $q^{(t)} = (1 - \alpha)p_{\theta^{(t)}} + \alpha p_{sft}$
GEM	(Li et al., 2025h)	$\mathcal{L}_{GEM}(\theta) = \sum_i \sum_{t=1}^T \sum_{y^{gene}} q_k(y^{gene} x_i, y_{1:t-1}^{real}) \cdot \left[\log f_\theta(y^{gene} x_i, y_{1:t-1}^{real}) - \log f_\theta(y_k^{real} x_i, y_{1:t-1}^{real}) \right]$, where $q = \text{softmax}\left(\frac{1}{\beta} \cdot \log f\right)$
CurioSFT	(Wang et al., 2026b)	$\mathcal{L}_{total}(\theta) = \mathcal{L}_{SFT}(\theta) + \alpha \cdot \mathcal{L}_{SED}(\theta)$, where $\mathcal{L}_{SED}(\theta) = \frac{1}{2} \sum_{t=1}^T \left(\log \frac{\pi_\theta(y_t s_t; \tau=1.0)}{\pi_\theta^{CH}(y_t s_t; \bar{\tau}_t)} \right)^2$

Table 8: Objective comparison across training paradigms for integration with RL.

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

Table 9: Comparison of our survey with existing related surveys on reasoning and post-training.

Survey	Primary Focus	SFT Focus	Algorithm-Centric	Analytical Findings	Time Coverage	Key Distinction from Our Work
Thinking Hats (Chen et al., 2025e)	CoT fine-tuning (SFT+RFT)	△ Partial (SFT+RFT together)	✗ No	△ Limited (cognitive perspective)	Through 2024	Different framework: cognitive psychology vs. optimization dynamics
100 Days After R1 (Zhang et al., 2025b)	R1 replication studies	△ Partial (distillation focus)	✗ No	△ Minimal (field report)	Jan–Apr 2025	Different scope: data-centric scope vs. systematic methodology
Towards Reasoning Era (Chen et al., 2025d)	Long CoT reasoning	✗ Minimal (auxiliary to RL)	✗ No (RL-centric)	△ Limited (phenomena description)	2024–early 2025	Different paradigm: RL-centric vs. SFT-centric
SFT vs RL (Jiang et al., 2026a)	Post-training paradigms comparison	△ Partial	△ Light (high-level comparison)	✗ No	2022–2025	Different specificity: general post-training vs. reasoning-specific
LLM Post-Training (Kumar et al., 2026)	Post-training pipeline (FT, RL, test-time)	△ Partial (one of many)	△ Minimal (pipeline view)	✗ No	2024–2025	Different granularity: pipeline overview vs. SFT deep-dive
From System 1 to System 2 (Li et al., 2025g)	Reasoning LLM development	✗ Minimal	✗ No	△ Limited (benchmark focus)	2024–mid 2025	Different focus: broad reasoning LLM survey vs. SFT methodology
Foundation Models Reasoning (Sun et al., 2025b)	Multimodal reasoning paradigms	✗ Minimal	✗ No	△ Limited (descriptive taxonomy)	Pre-2024	Different scope: multimodal reasoning applications vs. SFT training methods
Mathematical Reasoning (Wang et al., 2026d)	Math reasoning domain	△ Partial (not focus)	✗ No	△ Minimal	Pre-2025	Different domain: math-specific vs. general reasoning; no SFT methodology focus
Multi-Step Reasoning (Plaat et al., 2025)	In-context learning/prompting	✗ Minimal	△ Yes (prompt algorithms)	△ Limited	2021–2024	Different stage: inference-time prompting vs. training-time SFT
Toward Large Reasoning Models (Xu et al., 2025a)	RL for reasoning (“Reinforced Reasoning”)	✗ Minimal (SFT as prerequisite)	✓ Yes (RL algorithms)	△ Limited	Early 2024–Feb 2025	Complementary: RL-centric (they cover RL, we cover SFT)
Reinforced MLLM (Zhou et al., 2025)	RL-based multimodal reasoning	✗ No (RL only)	△ Yes (RL for MLLM)	✗ No	2024–2025	Different modality: multimodal + RL-only vs. LLM SFT focus
RFT for MLLMs (Sun et al., 2025a)	Reinforcement Fine-Tuning for MLLMs	✗ No (RFT only)	△ Yes (RL for MLLM)	✗ No	2024–2025	Different modality: multimodal + RFT vs. LLM SFT
Our Survey	SFT for reasoning	✓✓ Exclusive focus	✓✓ Core contribution	✓✓ Systematic organization	Jul 2024–May 2026	First dedicated reasoning SFT methodology survey
Legend: ✗ No / Minimal △ Limited / Partial ✓ Yes / Significant ✓✓ Heavy / Core focus						

Table 10: Quantitative comparison of methods sharing the Qwen2.5-Math-7B fine-tuned on NuminaMath-CoT.

Method	GSM8K	MATH	Minerva	Olympiad	AIME 2024	AMC 2023	Avg. over 4
No training	54.8	50.3	14.4	16.4	12.1	36.9	20.0
EAFT	–	–	20.2	19.0	2.5	24.2	16.5
ASFT	–	65.8	24.1	29.3	9.0	43.3	26.4
SFT	71.1	60.9	21.8	32.9	9.2	43.4	26.8
ProFiT	–	68.5	28.7	31.4	7.1	44.1	27.8
DFT	87.0	70.6	30.2	33.8	8.6	45.0	29.4
RASFT	–	70.7	30.0	33.0	8.8	45.8	29.4
TALR	–	–	35.7	32.9	6.7	43.4	29.7
RFT*	83.3	69.8	21.3	31.3	11.2	58.8	30.7
iw-SFT	–	70.3	25.7	34.5	16.5	51.1	32.0
RANKTUNER	–	–	38.3	36.3	10.2	46.6	32.9
RIFT*	86.3	74.7	28.9	34.1	17.1	62.2	35.6