A Survey of Joint Online-Offline Fine-tuning for Large Language Models

Taihang Zhen[1], Guang Yang[2], Chenzhang Li[3], Nuo Yan[2], Shilong Zhou[3], Guangyu Liu[2], Xiaotong Tang[2], Jing Huo[2*], Boyan Wang[3], Junlan Feng[4*] and **Yuyao Zhang[4]**
[1]School of Electronic Science and Engineering, Nanjing University
[2]State Key Laboratory for Novel Software Technology, Nanjing University
[3]School of Intelligence Science and Technology, Nanjing University
[4]JIUTIAN Research, China Mobile
huojing@nju.edu.cn, fengjunlan@cmjt.chinamobile.com

Abstract

Post-training for Large Language Models (LLMs) can be mainly categorized into offline Supervised Fine-Tuning (SFT) for knowledge acquisition and online Reinforcement Fine-Tuning (RFT) for adaptive refinement. Current state-of-the-art approaches typically employ a sequential cold-start pipeline (SFT-Then-RFT). However, we argue that this disjoint transition imposes an "alignment tax", leading to catastrophic forgetting and reward hacking during the unregularized exploration phase. In this work, we advocate for Joint Online-Offline Fine-Tuning as a superior paradigm that breaks the convention of restricting offline data to SFT and online data to RFT. By integrating full offline response generation with online rollouts—particularly within the realm of Reinforcement Learning with Verifiable Rewards (RLVR)—this approach mitigates the limitations of isolated training phases. We provide the first comprehensive survey focusing specifically on the synchronization of data provenance. We introduce a novel taxonomy for related works, analyze their theoretical advantages in balancing stability with plasticity, and outline a roadmap for next-generation post-training frameworks.

1 Introduction

Supervised Fine-Tuning (SFT) establishes the cornerstone of Large Language Models (LLMs) post-training. By minimizing the negative log-likelihood on curated offline datasets, SFT aligns models with human intent and instills foundational instruction-following capabilities. This paradigm leverages high-quality demonstrations to "cold-start" the policy with expert priors. However, standard SFT is fundamentally constrained by its static nature: it suffers from distribution shifts between training and inference and lacks the exploratory mechanism required to surpass the capabilities of the data annotators, effectively placing a performance ceiling defined by the offline supervision.

To transcend these limitations, Reinforcement Fine-Tuning (RFT) introduces online interaction, refining model behavior through rollouts guided by reward signals. particularly in domains with ground-truth verification like math or code—often framed as Reinforcement Learning with Verifiable Rewards (RLVR)—RFT fosters adaptive learning and self-correction. By actively exploring the generation space, the model can discover novel solution paths absent in the offline corpus. Yet, pure RFT faces significant challenges, including high sample variance, training instability, and the risk of reward hacking, where models exploit verifiers rather than developing robust reasoning chains.

The prevailing consensus seeks to combine these paradigms, typically manifesting as an asynchronous, multi-stage pipeline: SFT-Then-RFT. In this sequential regime, exemplified by large reasoning models, SFT provides the necessary behavioral initialization (e.g., injecting Chain-of-Thought patterns), which is subsequently refined through RL-driven exploration. Despite its popularity, this rigid sequentiality imposes a significant "alignment tax" [Ouyang *et al.*, 2022]. The disjoint transition from supervised imitation to reinforcement learning often triggers catastrophic forgetting [Luo *et al.*, 2025], degrading general linguistic proficiency. Furthermore, without the continuous regularization of offline expert priors during the RFT phase, exploration becomes brittle and susceptible to mode collapse.

In response to these inefficiencies, a paradigm shift is emerging towards **joint online-offline fine-tuning**. Unlike cold-start strategies that segregate data usage by stage, this paradigm integrates offline demonstrations and online rollouts into a *unified* optimization objective. By jointly leveraging the stability of offline expert priors and the adaptability of online exploration, these methods mitigate catastrophic forgetting while accelerating convergence. This way effectively bridges the gap between *offline learning* (exploiting historical repositories) and *online learning* (interacting with environments), offering a robust pathway for holistic post-training.

Despite the transformative potential of this integration, joint online-offline fine-tuning remains underexplored in existing literature. Current surveys primarily focus on SFT or RFT in isolation [Liu *et al.*, 2025b; Kumar *et al.*, 2025], or treat their combination strictly as a sequential hand-off [Zhang *et al.*, 2025a], lacking a systematic formulation of the unified perspective proposed herein. To bridge this gap, this

*Corresponding author

work advocates for joint online-offline fine-tuning as a pivotal direction for next-generation LLM post-training. We aim to answer: *How can offline stability and online adaptability be mathematically and empirically unified?*

Our contributions are summarized as follows: (1) We provide a comprehensive review of existing paradigms, systematically dissecting the mechanisms and limitations of isolated SFT, isolated RFT, and disjoint SFT-Then-RFT pipelines. (2) We introduce a structured taxonomy for **joint online-offline** methodologies, analyzing how recent works decouple data usage from training stages to achieve unified optimization. (3) We discuss critical applications and future research directions, identifying the technical barriers that must be overcome to fully realize the potential of this integrated paradigm.

2 Related Work

The rapid evolution of LLMs has catalyzed a comprehensive body of literature summarizing post-training methodologies. To demarcate the distinct contributions of this work, we categorize existing surveys into three primary streams and delineate the critical research gap this survey addresses. A comparative summary is provided in Table 1.

2.1 Existing Surveys

Surveys on General Post-Training. Broad overviews [Kumar *et al.*, 2025; Tie *et al.*, 2025] systematically review the LLM lifecycle from pre-training to Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). While these works offer extensive breadth—covering reasoning capabilities [Kumar *et al.*, 2025] and scaling laws [Lai *et al.*, 2025]—they predominantly adopt a *modular perspective*. They characterize SFT and RFT as sequential, independent stages (e.g., the standard "SFT-then-RLHF" pipeline), thereby overlooking the algorithmic synergies between offline data utilization and online exploration. Consequently, these paradigms are often presented as mutually exclusive alternatives rather than complementary components of a unified system.

Surveys on RLHF and Alignment. A second stream focuses on the mechanics of alignment and reward modeling [Kaufmann *et al.*, 2023; Wang *et al.*, 2024]. These works elucidate the mathematical underpinnings of Reward Modeling (RM) and policy optimization, contrasting algorithms (e.g., PPO vs. DPO) or feedback modalities (e.g., RLAIF vs. RLHF). Recent studies [Ji *et al.*, 2026] further categorize progress through the lens of reward design. However, these surveys largely operate under a static data regime: DPO is typically discussed strictly as an offline method, while PPO is framed purely as an online method. They rarely address the emerging class of *joint* approaches that blur these boundaries, such as online iterative DPO or synchronous replay strategies.

Surveys on Instruction Tuning. Literature centered on instruction tuning [Zhang *et al.*, 2026a; Han *et al.*, 2025] emphasizes data engineering, curriculum learning, and instruction synthesis. While these surveys underscore the critical role of offline data quality, they typically stop short of discussing how such static priors can be dynamically integrated into RL loops to mitigate catastrophic forgetting during exploration.

2.2 Our Scope: The Joint Online-Offline Paradigm

Distinct from prior works, this survey pivots the analytical lens from algorithmic selection to structural integration. Specifically, we distinguish ourselves by: (1) **Taxonomizing by Data Interaction**, categorizing methods into *Disjoint* vs. *Joint* strategies rather than by loss functions; (2) **Focusing on Synergy**, analyzing the reciprocal stabilization between offline priors and online exploration; and (3) **Bridging the Gap**, reviewing unified objectives that chart the path for next-generation post-training.

3 Preliminary

We formalize the post-training of LLMs through the lens of data distribution dynamics. Let $\mathcal{X}$ denote the prompt space and $\mathcal{Y}$ the response space. A policy $\pi_\theta(y|x)$ maps prompts to a probability distribution over responses. The fundamental bifurcation among existing paradigms lies in the dependency of the training data distribution on the evolving policy π_θ.

3.1 Offline Fine-tuning

Offline SFT. Standard instruction tuning is formally equivalent to behavioral cloning. The objective minimizes the negative log-likelihood of expert demonstrations:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{offline}}}\left[\log \pi_\theta(y|x)\right] \tag{1}$$

While sample-efficient, offline SFT suffers from *distribution shift*. During inference, the model may drift into states undefined by the training support, leading to error accumulation [Ouyang *et al.*, 2022].

Offline RFT. Methods such as Direct Preference Optimization (DPO) align models using a static set of preference pairs $\mathcal{D}_{\text{pref}} = \{(x, y_w, y_l)\}$. The objective typically optimizes a classification loss based on the Bradley-Terry model [Rafailov *et al.*, 2023]:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{2}$$

Since π_θ does not interact with the environment, strictly offline methods are prone to over-optimization, often failing to generalize to out-of-distribution prompts [Xu *et al.*, 2024].

3.2 Online Fine-tuning

Online learning introduces dynamic distribution shifts. At training step t, data is generated by the current policy π_{θ_t} interacting with the environment (or a reward function R), allowing the model to actively explore the output space $\mathcal{Y}$.

Online SFT (Iterative SFT). Often referred to as *Rejection Sampling Fine-Tuning*, this method approximates an Expectation-Maximization (EM) procedure. For a prompt x, the policy generates candidates, and a verifier selects the optimal response y^* based on reward $R(x, y)$. The policy is to maximize the likelihood of these high-reward samples:

$$\mathcal{L}_{\text{Iter-SFT}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{\text{prompt}}}\mathbb{E}_{y^*\sim p(\cdot|x,\pi_{\theta_{\text{old}}},R)}\left[\log \pi_\theta(y^*|x)\right] \tag{3}$$

where $p(\cdot|x, \pi_{\theta_{\text{old}}}, R)$ denotes the distribution of filtered responses selected by the reward function [Yuan *et al.*, 2024].

Survey Category	Taxonomy Basis	View on Integration	Data Strategy	Core Limitation
General Post-Training *(e.g., [Kumar* et al.*, 2025])*	By *Training Stage* (Pretrain → SFT → RL)	Sequential Pipeline (SFT initializes RL)	Isolated Phases (Static → Dynamic)	Ignores algorithmic synergy (Treats SFT/RL as separate tasks)
RLHF & Alignment *(e.g., [Kaufmann* et al.*, 2023])*	By *Algorithm/Objective* (PPO vs. DPO vs. RLAIF)	Reward Optimization (Max Reward)	Online-Dominant (Focus on Rollouts)	Lacks offline stabilization (Neglects catastrophic forgetting)
Instruction Tuning *(e.g., [Zhang* et al.*, 2026a])*	By *Data Engineering* (Data Selection, Synthesis)	Behavior Cloning (Min Loss)	Offline-Only (Static Datasets)	No exploration capability (Upper-bound by data quality)
Ours (This Work)	**By *Coupling Mechanism* (Disjoint vs. Joint)**	**Joint Synergy (SFT ↔ RL Cycle)**	**Synchronous Co-flow (Offline + Online)**	**– (Unified Framework)**

Table 1: Ecological position of this survey. Unlike prior works that categorize methods by training stages or specific algorithms, we introduce a taxonomy based on the **structural coupling** of data streams, highlighting the shift from sequential pipelines to joint synergy.

Online RFT. Methods like PPO and GRPO directly optimize the expected reward over the policy's trajectory:

$$\mathcal{J}_{\text{RFT}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{prompt}}, y\sim\pi_\theta(\cdot|x)}\left[R(x,y) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right] \quad (4)$$

Online RFT facilitates exploration, effectively addressing the generalization limits of offline learning, but is notoriously sample-inefficient and sensitive to hyperparameter tuning.

3.3 Disjoint Online-Offline Fine-tuning

To bridge the gap between static instruction following and dynamic environment interaction, hybrid strategies have emerged. The core motivation is to combine the stability of offline priors with the exploration of online learning. However, prevalent approaches typically treat these data sources as mutually exclusive optimization phases. We categorize these methods as **disjoint strategies**, where the gradient update at any timestep t is derived exclusively from either offline or online objectives, strictly preventing the fusion of supervision signals.

The Disjoint Formulation. Formally, disjoint fine-tuning decomposes the learning process into non-overlapping stages. Let $\mathcal{T}$ be the total training steps. The training schedule partitions $\mathcal{T}$ into subsets $\mathcal{T}_{\text{off}}$ and $\mathcal{T}_{\text{on}}$, such that $\mathcal{T}_{\text{off}} \cap \mathcal{T}_{\text{on}} = \emptyset$. The update rule toggles between distinct loss landscapes:

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta \mathcal{L}(\theta_t), \quad \text{where } \mathcal{L} \in \{\mathcal{L}_{\text{SFT}}, \mathcal{L}_{\text{RFT}}\} \quad (5)$$

This formulation encompasses both coarse-grained (Sequential) and fine-grained (Interleaved) separation paradigms.

Sequential Integration (SFT-Then-RFT). The most standard practice strictly decouples the process into two phases.

Phase 1 (Offline Initialization): The model minimizes $\mathcal{L}_{\text{SFT}}$ on $\mathcal{D}_{\text{offline}}$ to obtain a reference policy π_{ref}.

Phase 2 (Online Alignment): The model switches permanently to minimizing $\mathcal{L}_{\text{RFT}}$ using online trajectories.

This "Cold Start" strategy assumes that Phase 1 provides sufficient structural pre-conditioning. However, the rigid separation often leads to the "alignment tax" where the aggressive distribution shifts in Phase 2 degrade the general capabilities acquired in Phase 1.

Interleaved Integration (SFT-Alternating-RFT). Recent works seek to mitigate the forgetting problem by alternating between phases at a finer granularity [Yan *et al.*, 2025; Yuan *et al.*, 2025a]. These methods maintain a dynamic replay buffer $\mathcal{D}_{\text{buffer}}$. The training iterates through cycles:

Online Collection: Run Online RFT and populate $\mathcal{D}_{\text{buffer}}$ with hard or high-value samples.

Offline Replay: Pause exploration and run Offline SFT on $\mathcal{D}_{\text{buffer}}$ to consolidate knowledge.

Although this approach improves data efficiency, the optimization remains asynchronous: the model still oscillates between conflicting objectives, creating a "seesaw effect" where stability and exploration trade off against each other rather than synergizing in real-time.

Deficiencies of Disjoint Optimization. The fundamental limitation of disjoint fine-tuning lies in the temporal isolation of supervision signals. (1) During offline updates, the model lacks feedback on its current policy's hallucinations. (2) During online updates, the absence of direct supervision risks catastrophic forgetting and mode collapse [Jin *et al.*, 2025]. These challenges necessitate a joint learning framework for LLMs where offline constraints and online exploration are integrated synchronously.

4 Taxonomy

Building upon the above limitations, this motivates a shift from when to how offline and online data are integrated. Joint online-offline fine-tuning directly couples offline supervision and online exploration within a single optimization process. Such coupling enables real-time interaction between expert priors and policy-induced data.

As shown in Figure 1, we show the taxonomy of joint online-offline fine-tuning based on the workflow which includes SFT, RFT, hybrid fine-tuning (HFT) and alternating fine-tuning (AFT). Offline and online data is integrated into a unified framework. Joint online-offline learning is then directly implemented via a single-stage optimization process. We also summarize the datasets, objectives and involved models of these works in Table 2.

4.1 Joint Online-Offline SFT

Learning solely from static offline data in standard SFT can lead distribution shift which undermines LLMs self-correction. In joint online-offline SFT, recent studies have re-examined SFT framework. Rather than modifying the standard likelihood-based training objective, these methods focus on reshaping the construction and interpretation of supervision signals with online and offline data. Because of different negative log-likelihood shape drivers, existing works can be broadly categorized into two directions in Figure 2:

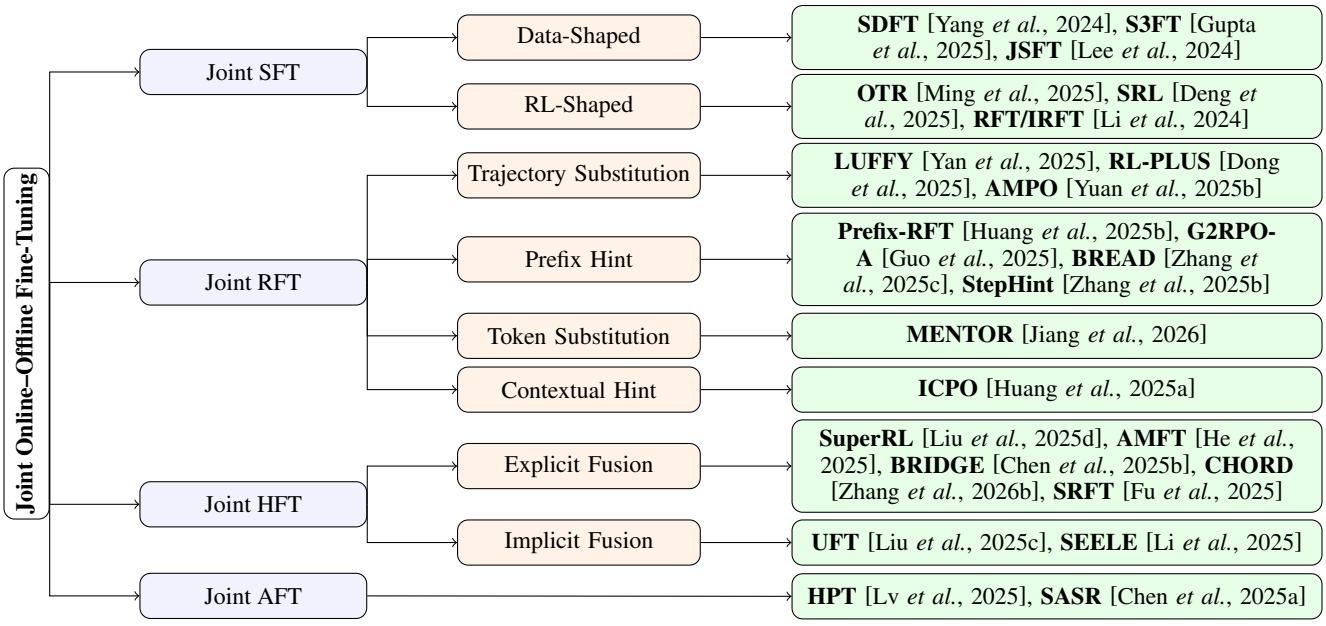

Figure 1: Overview of joint online-offline fine-tuning methods.

Data-Shaped SFT. Arising from data engineering, the minimal design is to mix distinct samples. A series of works unifies offline and online SFT which depend on data source to maintain the standard SFT loss. They align and amplify the model's inherent knowledge structure under the constraint of external expert knowledge. This facilitates a learning process driven by LLMs self-consistency, moving away from forced imitation of offline data distributions toward a more natural alignment with the model's pre-existing capabilities:

$$\mathcal{L}_{\text{Data-Shaped SFT}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{\text{offline}},y\sim(\pi_{\theta_{\text{old}}},\mathcal{D}_{\text{offline}})}\left[\log \pi_\theta(y|x)\right] \quad (6)$$

SDFT [Yang *et al.*, 2024] rewrites offline expert traces to generate distilled responses. These distilled responses are evaluated with simple heuristic rules to decide whether they can replace the original expert responses for standard SFT. S3FT [Gupta *et al.*, 2025] follow SDFT and use a judge to evaluate whether the self-generation is equivalent to the expert response. JSFT [Lee *et al.*, 2024] merge standard offline SFT data with preference judgment examples. LLM acts as the judge by itself. LLM performs pairwise preference comparisons between offline collections and online rollouts to choose better responses for SFT.

RL-Shaped SFT. Simple data-driven pattern struggle to capture differences in solution quality between offline and online data. These methods reconceptualizes standard SFT as an optimization process constrained by RFT principle. This perspective effectively understand SFT under a RL framework to rectify optimization biases. Consequently, while the supervision signal remains formally grounded in likelihood learning, its semantics are elevated to a strategy optimization process that is more stable and generalizable:

$$\mathcal{L}_{\text{RL-Shaped SFT}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{\text{offline}},\, y\sim(\pi_{\theta_{\text{old}}},\mathcal{D}_{\text{offline}})}\left[\sum_{i=1}^{t} w_t \log \pi_\theta(y_t \mid x, y_{<t})\right] \quad (7)$$

w_t is used to characterize the components (reward, advantage, or importance sampling) in RL. When $w_t = 1$, this objective naturally degenerates into standard SFT. OTR [Ming *et al.*, 2025] treats each token generation step in the autoregressive generation process of LLM as an independent, single-step RL trajectory. OTR samples multiple candidate tokens from the distribution of the current policy, and uses the offline tokens to provide a reward signal for the candidate tokens. SRL [Deng *et al.*, 2025] trains model to generate internal inferential monologues before performing each action, and based on the similarity between model actions and expert actions, provides smoother rewards in a stepwise manner. This supervisory signal can provide a richer learning signal even when all rollouts are incorrect, while encouraging flexible inference guided by expert demonstrations. RFT and IRFT [Li *et al.*, 2024] utilize the inverse reinforcement learning, taking offline data as the positive alignment target and the model's self-generation as the contrastive negative example. This method enables the model to learn rewards from the demonstration dataset and improve SFT for alignment.

Takeaway #1. These methods are particularly well-suited for scenarios where LLMs have acquired robust foundational capabilities. Under such circumstances, they can further tap into the LLMs latent capabilities while preserving high sample efficiency. However, if LLMs are insufficient, they will struggle to systematically correct the fundamental biases. Worse still, they may amplify existing errors and preferences during the self-generation process. **They are more suitable as means for refining capabilities and enhancing alignment in the middle and late stages of LLMs development.**

4.2 Joint Online-Offline RFT

For challenging problems, frequent reasoning failures cause most sampled paths to collapse to zero-reward. In joint online-offline RFT, these methods interleave online rollouts with selectively injected offline collections in each group, thereby preserving the exploratory of online RFT while sta-

Method	Applicability	Works	Datasets	Objectives	Models
Joint SFT	Sample quality control	SDFT	GSM8K/OpenFunction/Magicoder/LIMA	SFT	LLaMA2
		S3FT	GSM8K/MBPP/NQ	SFT	Mistral-Instruct-v2
		JSFT	Anthropic-HH/UltraFeedback	SFT+DPO	LLaMA2
		OTR	OpenR1-Math-220k		Qwen2.5/Qwen3
		SRL	s1K-1.1	/	Qwen2.5
		RFT	Anthropic-HH		pythia
		IRFT	Ultrachat200k		pythia/zephyr
Joint RFT	Sparse reward	LUFFY	OpenR1-Math-220k	GRPO	Qwen2.5/LLaMA3.1
		RL-PLUS	OpenR1-Math-220k		Qwen2.5/LLaMA3.1/DeepSeek-Math
		AMPO	OpenR1-Math-46k-8192		Qwen2.5/LLaMA3.2
		Prefix-RFT	OpenR1-Math-220k		Qwen2.5/LLaMA3.1
		G2RPO-A	OpenR1-Math-220k/Verifiable-Coding-Problems-Python		Qwen3/DeepSeek-Math/DeepSeek-Code
		StepHint	DAPO-Math-17K/DeepMath		Qwen2.5
		BREAD	NuminaMath-CoT		Qwen2.5
		MENTOR	MATH/OpenR1-Math-220k		Qwen2.5/LLaMA3.1
		ICPO	OpenR1-Math-220k/Skywork-OR1-RL-Data		Qwen3
Joint HFT	Discontinuous signal	SuperRL	GSM8K/MetaMath/PRM12K/LIMO/OpenR1/MetaMath	SFT+GRPO	Qwen2.5/LLaMA3.2
		AMFT	OpenR1-Math-46k-8192		Qwen2.5-Math
		BRIDGE	LIMR/MATH		Qwen2/Qwen2.5/LLaMA- 3.2
		CHORD	OpenR1-Math-220k		Qwen2.5
		SRFT	OpenR1-Math-46k-8192		Qwen2.5-Math
		UFT	Countdown/MATH/Logic		Qwen2.5/LLaMA3.2
		SEELE	DeepMath-103K		Qwen2.5
Joint AFT	Real-time scheduling	HPT	OpenR1-Math-46k-8192	SFT+GRPO	Qwen2.5-Math/LLaMA3.1
		SASR	GSM8K/MATH/KK		Qwen2.5/DeepSeek-R1-Distill-Qwen

Table 2: Summary of the datasets, objectives, and involved models of joint online-offline fine-tuning works.

bilizing optimization through anchored supervision:

$$\begin{aligned}\mathcal{J}_{\text{Joint RFT}}(\theta) = \lambda_{\text{online}} \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{online}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \underbrace{\mathbb{M}_{t,on}^{(i)}}_{\text{Mask}} \cdot \mathcal{F}_{on}\left(\rho_{t,on}^{(i)}, \hat{A}_{t,on}^{(i)}\right) \right] \\ + \lambda_{\text{offline}} \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{offline}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \underbrace{\mathbb{M}_{t,off}^{(i)}}_{\text{Mask}} \cdot \mathcal{F}_{off}\left(\rho_{t,off}^{(i)}, \hat{A}_{t,off}^{(i)}\right) \right]\end{aligned} \tag{8}$$

where λ is a weighting coefficient controlling the trade-off between modes, $\mathbb{M}$ serves as a token-level routing mask to explicitly identify whether the data source is online or offline, while $\mathcal{F}$, ρ, and $\hat{A}$ represent the mode-specific objective function, importance sampling ratio, and group-relative advantage, respectively. Owing to the granularity of integrating offline data into online rollouts, existing works can be broadly categorized into four directions in Figure 2:

Trajectory Substitution. A common strategy among these approaches is to directly substitute part of each rollout group with offline trajectories. Specifically, LUFFY [Yan *et al.*, 2025] substitutes one trajectory in each online rollout group with the offline response generated by an expert model, extending the RFT framework to avoid overly rapid convergence and entropy collapse. RL-PLUS [Dong *et al.*, 2025] refines the objective of LUFFY by introducing Multiple Importance Sampling and an Exploration-Based Advantage Function. Instead of always injecting one type of offline expert trajectories, AMPO [Yuan *et al.*, 2025b] leverages several types of offline data from multiple expert models, and dynamically decides when and how strongly to apply expert supervision based on online reward signals.

Prefix Hint. Different from directly bringing complete expert data, these methods guide policy learning by conditioning rollouts on offline reasoning prefixes and letting LLMs complete the remaining steps autonomously. Inspired by LUFFY, Prefix-RFT [Huang *et al.*, 2025b] replaces offline full-trajectory hints with partial prefixes, allowing the model to complete rollouts conditioned on these prefixes. It employs cosine decay to gradually reduce the hint length from a high initial value to near zero over the course of training. BREAD [Zhang *et al.*, 2025c] inserts prefix hint when self-generated traces fail until ensuring at least one successful path. G2PRO-A [Guo *et al.*, 2025] extends this idea by dynamically adjusting the guidance length based on the average reward from previous training steps and further investigates how the number of trajectories receiving prefix hints within each rollout group affects training performance. StepHint [Zhang *et al.*, 2025b] goes even further by providing multiple levels of guidance for a single sample. It segments the CoT trajectory into steps based on token entropy and offers step-level prefix hints, where varying the number of hinted steps yields fine-grained control over guidance strength.

Token substitution. When the distributional gap between offline reference and online response becomes large, direct trajectory-level guidance may destabilize RFT. This way restricts expert intervention to token-level corrections. MENTOR [Jiang *et al.*, 2026] observes that overly strong supervision can suppress policy exploration, leading to issues such as entropy collapse. To address this, MENTOR proposes a dynamic correction mechanism: during rollout, an expert model computes logits in parallel for each token generated by the training model. If the divergence between the two sets of logits exceeds a dynamic threshold and indicates that the training model has deviated from the desired distribution, the expert's logits are used to correct the problematic tokens.

Contextual Hint. Instead of intervening in rollouts, offline in-context learning (ICL) can provide soft guidance. ICPO [Huang *et al.*, 2025a] induces rollout by including questions and corresponding offline expert responses as few-shot examples in the prompt.

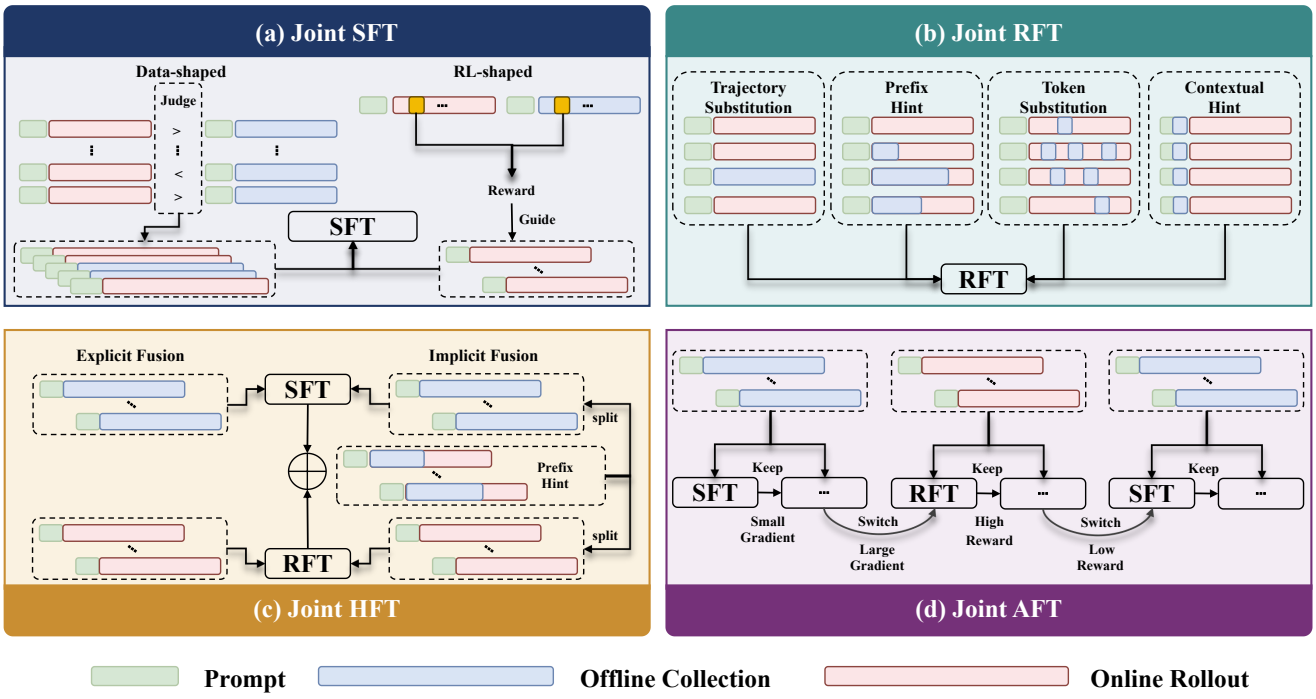

Figure 2: Workflows of joint online-offline fine-tuning.

Takeaway #2. By injecting dense offline supervision into rollout at different granularities, these methods trade some exploration and long-term adaptability for improved stability, efficiency, and convergence reliability. Actually, they can be regarded as controlled trade-off. Notably, this pipeline may be sensitivity to data selection criteria, and intervention frequency. **Therefore, they are particularly well suited to sparse-reward tasks, and settings where high-quality demonstrations are available but cannot fully specify the desired behavior space.**

4.3 Joint Online-Offline HFT

The training signal between SFT and RFT is inherently discontinuous. In joint online-offline HFT, offline SFT and online RFT are co-activated within a single training loop. Both workflows are coupled through the structured, training-time coordination mechanism with their influence regulated:

$$\mathcal{L}_{\text{Joint HFT}}(\theta) = \lambda_{\text{offline}}\mathcal{L}_{\text{SFT}}(\theta) - \lambda_{\text{online}}\mathcal{J}_{\text{RFT}}(\theta) \quad (9)$$

where λ_{online} and λ_{offline} respectively is the fusion weight of online and offline learning objectives. Existing works can be broadly categorized into two directions in Figure 2:

Explicit Fusion. These works integrates offline and online fine-tuning through explicit objective-level strategy, treating offline SFT and online RFT as jointly optimized components. One of the key points lies in the weight control. SuperRL [Liu *et al.*, 2025d] adopts reward-gated instance-level objective fusion, prioritising RL updates when reward signals exist and adaptively strengthening SFT supervision only when exploration fails. AMFT [He *et al.*, 2025] formulates the SFT–RL balance as a bi-level optimisation problem, using a meta-gradient controller to proactively learn optimal loss weights and schedule a smooth transition from imitation to exploration. BRIDGE [Chen *et al.*, 2025b] formulates SFT–RL integration as a bilevel optimization problem, where supervised learning acts as an upper-level controller that meta-learns how to guide lower-level RL updates by explicitly maximizing cooperative gain, enabling adaptive and principled objective fusion throughout training. CHORD [Zhang *et al.*, 2026b] reframes SFT as a dynamically weighted auxiliary objective within RL, combining global and token-wise control to harmonise online and offline learning while mitigating overfitting under distribution shift. Unlike the above methods that maintain a strict separation in HFT—offline data for SFT and online data for RFT, SRFT [Fu *et al.*, 2025] leverages offline data for both SFT and RFT with the entropy-aware weighting.

Implicit Fusion. By operating entirely at the input and behavior level, implicit fusion avoids complex loss coupling. UFT [Liu *et al.*, 2025c] unifies SFT and RL by appending partial expert reasoning as hint prefixes and gradually annealing their length, while incorporating prefix likelihood into the RL objective to reduce training–inference distribution shift. SEELE [Li *et al.*, 2025] introduces a capability-adaptive hint scaffolding strategy that dynamically adjusts solution prefix length via multi-round sampling to maintain training difficulty near an optimal learning regime.

Takeaway #3. These methods continuously adjust the relative influence of offline and online signals and better connect SFT and RFT. **They are better for situations where the**

challenge is not the absence of supervision or rewards, but how strongly each should influence learning. The shared limitation also stems from this design choice. Their performance is highly sensitive to this balance. Miscalibration can lead to over-reliance on demonstrations, suppressed exploration, or conversely, premature exposure to noisy rewards and unstable updates.

4.4 Joint Online-Offline AFT

To better independently design SFT and RFT, in joint online-offline AFT, signal-driven scheduling rather than isolated-stage interleaved pipelines interchangeably employ offline SFT and online RFT in Figure 2 in a single stage. SFT and RFT are complementary gradient sources within a single training process. The key point lies in using online training signals to decide when offline supervision should intervene:

$$\mathcal{L}_{\mathrm{JointAFT}}(\theta) = \alpha_t \cdot \mathcal{L}_{\mathrm{SFT}}(\theta) - (1-\alpha_t) \cdot \mathcal{J}_{\mathrm{RFT}}(\theta) \quad (10)$$

$\alpha_t \in \{0,1\}$ at training iteration t to decide SFT or RFT. HPT [Lv *et al.*, 2025] theoretically derives the gradient estimators of various post-training algorithms and further dynamically switches offline SFT and online RFT with online reward feedback. Apart from reward-based selection, SASR [Chen *et al.*, 2025a] schedules SFT and RFT in a gradient-level. It demonstrates a high correlation between divergence from the expert policy and gradient norm, implying that monitoring the LLMs training status can adjust the ratio of SFT steps to GRPO steps.

Takeaway #4. In essence, adaptive scheduling provides a powerful mechanism to coordinate and isolate imitation and exploration on real-time training indicators. This enables LLMs to benefit from offline guidance without constraining exploration unnecessarily. But their effectiveness depends critically on the design of the switching or adaptation criteria. **They are suited to datasets with widely varying difficulty and knowledge distribution.**

5 Applications

5.1 Multi-modal Reasoning

In the multi-modal reasoning domain, as a joint online-offline AFT pipeline, DyME [Liu *et al.*, 2025a] takes RFT as the primary objective and employs a selective supervision strategy in which SFT updates are triggered only when rollout trajectories fail to obtain positive rewards. This approach successfully mitigates advantage collapse of small vision-language models on vision reasoning tasks.

5.2 Vertical Domain

In the specialized vertical domain, MentraSuite [Xiao *et al.*, 2025] advances mental health reasoning within a hybrid fine-tuning, enforcing clinical consistency across diagnosis and intervention tasks. TaoSR-AGRL [Yang *et al.*, 2026] follows the joint online-offline RFT and addresses e-commerce relevance challenges by utilizing an adaptive guided replay mechanism that dynamically steers online learning using offline trajectories when autonomous exploration stagnates.

6 Future Directions

Competence-Aware Data Routing. Poor sample efficiency often stems from a mismatch between data freshness and policy competence. While offline data provides stable expert priors, online data reflects the model's evolving inductive biases. We argue that joint fine-tuning implicitly performs data-level curriculum learning: offline samples anchor the policy in stable regions, while online samples densify coverage near decision boundaries where learning progress is maximal. Future work should formalize this via data routing, where the contribution of offline versus online gradients is adaptively weighted based on real-time metrics (e.g., policy uncertainty or value estimation error). This allows the model to dynamically shift focus from imitating priors to exploring novel trajectories as its competence matures.

Latent Reward Priors and Shaping. In online RFT (particularly RLVR), valid reasoning paths are scarce, leading to zero-reward trajectories. Here, offline data should function less as rigid imitation targets and more as dense reward shaping mechanisms. Joint fine-tuning mitigates sparsity by injecting dense supervision signals, derived from offline likelihoods, directly into the reinforcement process. A promising direction is to treat offline data as a latent reward prior that regularizes policy updates under partial feedback. This framework effectively reshapes the optimization landscape, constraining exploration to regions supported by expert structures. Crucially, future research should explore annealing strategies, where offline priors are gradually decayed as online reward density increases, enabling a smooth transition from guided learning to autonomous discovery.

Imitation and Generalization Balance. The imitation and generalization trade-off is fundamentally a state-dependent property rather than a global hyperparameter. Early in training or under high uncertainty, imitation provides low-variance gradients to stabilize learning; as competence increases, excessive imitation constrains expressivity, making exploration more valuable. Joint fine-tuning enables the policy to navigate this continuum. We envision future research framing this as a meta-decision problem, where the model (or a meta-controller) learns not only the task solution but also when to rely on demonstrations versus self-discovery. Moving toward such self-regulating post-training paradigms, potentially via meta-learning objectives, remains a critical frontier for achieving human-transcending intelligence.

7 Conclusion

In this survey, moving beyond the conventional dichotomy between offline SFT and online RFT, we identify joint online-offline fine-tuning as a principled paradigm that fundamentally rethinks how expert priors and on-policy exploration should interact during learning. Importantly, our analysis highlights that joint online-offline fine-tuning is not merely a heuristic combination of objectives. The effectiveness therefore hinges on when, where, and to what extent each signal intervenes. Such systems blur the line between supervision and reinforcement, paving the way for post-training pipelines that are more adaptive, interpretable, and aligned with the long-horizon objectives of general-purpose reasoning models.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (62192783, 62276128, 62406111), Jiangsu Natural Science Foundation (BK20243051), Jiangsu Science and Technology Major Project (BG2024031), the Fundamental Research Funds for the Central Universities(14380128, KG202514), the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University-China Mobile Communications Group Co.,Ltd. Joint Institute and the "111 Center" (No.B26023).

Contribution Statement

Taihang Zhen and Guang Yang contribute equally.

References

[Chen *et al.*, 2025a] Jack Chen, Fazhong Liu, Naruto Liu, Yuhan Luo, Erqu Qin, Harry Zheng, Tian Dong, Haojin Zhu, Yan Meng, and Xiao Wang. Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. *arXiv preprint arXiv:2505.13026*, 2025.

[Chen *et al.*, 2025b] Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *arXiv preprint arXiv:2509.06948*, 2025.

[Deng *et al.*, 2025] Yihe Deng, I Hsu, Jun Yan, Zifeng Wang, Rujun Han, Gufeng Zhang, Yanfei Chen, Wei Wang, Tomas Pfister, Chen-Yu Lee, et al. Supervised reinforcement learning: From expert trajectories to step-wise reasoning. *arXiv preprint arXiv:2510.25992*, 2025.

[Dong *et al.*, 2025] Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, et al. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*, 2025.

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

[Guo *et al.*, 2025] Yongxin Guo, Wenbo Deng, Zhenglin Cheng, and Xiaoying Tang. G^2rpo-a: Guided group relative policy optimization with adaptive guidance. *arXiv preprint arXiv:2508.13023*, 2025.

[Gupta *et al.*, 2025] Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Dinesh Khandelwal, Dinesh Raghu, and Sachindra Joshi. Selective self-to-supervised fine-tuning for generalization in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6240–6249, 2025.

[Han *et al.*, 2025] Xudong Han, Junjie Yang, Tianyang Wang, Ziqian Bi, Xinyuan Song, Junfeng Hao, and Junhao Song. Towards alignment-centric paradigm: A survey of instruction tuning in large language models. *arXiv preprint arXiv:2508.17184*, 2025.

[He *et al.*, 2025] Lixuan He, Jie Feng, and Yong Li. Amft: Aligning llm reasoners by meta-learning the optimal imitation-exploration balance. *arXiv preprint arXiv:2508.06944*, 2025.

[Huang *et al.*, 2025a] Hsiu-Yuan Huang, Chenming Tang, Weijie Liu, Clive Bai, Saiyong Yang, and Yunfang Wu. Think outside the policy: In-context steered policy optimization. *arXiv preprint arXiv:2510.26519*, 2025.

[Huang *et al.*, 2025b] Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling. *arXiv preprint arXiv:2507.01679*, 2025.

[Ji *et al.*, 2026] Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. A survey of progress in llm alignment from the perspective of reward design. *IEEE Transactions on Artificial Intelligence*, 2026.

[Jiang *et al.*, 2026] Zishang Jiang, Jinyi Han, tingyun li, Xinyi Wang, Sihang Jiang, Zhaoqian Dai, Ma Shuguang, Fei Yu, Jiaqing Liang, and Yanghua Xiao. Selective expert guidance for effective and diverse exploration in reinforcement learning of LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Jin *et al.*, 2025] Hangzhan Jin, Sicheng Lv, Sifan Wu, and Mohammad Hamdaqa. Rl is neither a panacea nor a mirage: Understanding supervised vs. reinforcement learning fine-tuning for llms. *arXiv preprint arXiv:2508.16546*, 2025.

[Kaufmann *et al.*, 2023] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

[Kumar *et al.*, 2025] 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. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[Lai *et al.*, 2025] Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, et al. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, 2025.

[Lee *et al.*, 2024] Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. Aligning large language models by on-policy self-judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11442–11459, 2024.

[Li *et al.*, 2024] Jiaxiang Li, Siliang Zeng, Hoi-To Wai, Chenliang Li, Alfredo Garcia, and Mingyi Hong. Getting more juice out of the sft data: Reward learning from human demonstration improves sft for llm alignment. *Advances in Neural Information Processing Systems*, 37:124292–124318, 2024.

[Li *et al.*, 2025] Ziheng Li, Zexu Sun, Jinman Zhao, Erxue Min, Yongcheng Zeng, Hui Wu, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Xu Chen, et al. Staying in the sweet spot: Responsive reasoning evolution via capability-adaptive hint scaffolding. *arXiv preprint arXiv:2509.06923*, 2025.

[Liu *et al.*, 2025a] Jiazhen Liu, Yuchuan Deng, and Long Chen. Empowering small vlms to think with dynamic memorization and exploration. *arXiv preprint arXiv:2506.23061*, 2025.

[Liu *et al.*, 2025b] Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*, 2025.

[Liu *et al.*, 2025c] Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*, 2025.

[Liu *et al.*, 2025d] Yihao Liu, Shuocheng Li, Lang Cao, Yuhang Xie, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. Superrl: Reinforcement learning with supervision to boost language model reasoning. *arXiv preprint arXiv:2506.01096*, 2025.

[Luo *et al.*, 2025] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[Lv *et al.*, 2025] Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, et al. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*, 2025.

[Ming *et al.*, 2025] Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. *arXiv preprint arXiv:2509.26313*, 2025.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[Tie *et al.*, 2025] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

[Wang *et al.*, 2024] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024.

[Xiao *et al.*, 2025] Mengxi Xiao, Kailai Yang, Pengde Zhao, Enze Zhang, Ziyan Kuang, Zhiwei Liu, Weiguang Han, Shu Liao, Lianting Huang, Jinpeng Hu, et al. Mentrasuite: Post-training large language models for mental health reasoning and assessment. *arXiv preprint arXiv:2512.09636*, 2025.

[Xu *et al.*, 2024] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

[Yan *et al.*, 2025] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

[Yang *et al.*, 2024] Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, 2024.

[Yang *et al.*, 2026] Jianhui Yang, Yiming Jin, Pengkun Jiao, Chenhe Dong, Zerui Huang, Shaowei Yao, Xiaojiang Zhou, Dan Ou, and Haihong Tang. Taosr-agrl: Adaptive guided reinforcement learning framework for e-commerce search relevance. In *Proceedings of the ACM Web Conference 2026*, pages 7955–7966, 2026.

[Yuan *et al.*, 2024] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[Yuan *et al.*, 2025a] Xiangchi Yuan, Xiang Chen, Tong Yu, Dachuan Shi, Can Jin, Wenke Lee, and Saayan Mitra. Mitigating forgetting between supervised and reinforcement learning yields stronger reasoners. *arXiv preprint arXiv:2510.04454*, 2025.

[Yuan *et al.*, 2025b] Xiaoyang Yuan, Yujuan Ding, Yi Bin, Wenqi Shao, Jinyu Cai, Jingkuan Song, Yang Yang, and Heng Tao Shen. More than one teacher: Adaptive multi-guidance policy optimization for diverse exploration. *arXiv preprint arXiv:2510.02227*, 2025.

[Zhang *et al.*, 2025a] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

[Zhang *et al.*, 2025b] Kaiyi Zhang, Ang Lv, Jinpeng Li, Yongbo Wang, Feng Wang, Haoyuan Hu, and Rui Yan. Stephint: Multi-level stepwise hints enhance reinforcement learning to reason. *arXiv preprint arXiv:2507.02841*, 2025.

[Zhang *et al.*, 2025c] Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. Bread: Branched rollouts from expert anchors bridge sft & rl for reasoning. *arXiv preprint arXiv:2506.17211*, 2025.

[Zhang *et al.*, 2026a] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

[Zhang *et al.*, 2026b] Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *The Fourteenth International Conference on Learning Representations*, 2026.