# Reinforcement Learning Meets Large Language Models: A Survey of Advancements and Applications Across the LLM Lifecycle

KELIANG LIU\*, Fudan University, China

DINGKANG YANG\*†, Fudan University, China and ByteDance SAIL Team, China

ZIYUN QIAN, Fudan University, China

WEIJIE YIN, ByteDance SAIL Team, China

YUCHI WANG and HONGSHENG LI, The Chinese University of Hong Kong, MMLab, China

JUN LIU, Lancaster University, UK

PENG ZHAI<sup>‡</sup>, Fudan University, China

YANG LIU<sup>‡</sup>, Tongji University, China and The University of Toronto, Canada

LIHUA ZHANG<sup>‡</sup>, Fudan University, China

In recent years, training methods centered on Reinforcement Learning (RL) have markedly enhanced the reasoning and alignment performance of Large Language Models (LLMs), particularly in understanding human intents, following user instructions, and bolstering inferential strength. Although existing surveys offer overviews of RL augmented LLMs, their scope is often limited, failing to provide a comprehensive summary of how RL operates across the full lifecycle of LLMs. We systematically review the theoretical and practical advancements whereby RL empowers LLMs, especially Reinforcement Learning with Verifiable Rewards (RLVR). First, we briefly introduce the basic theory of RL. Second, we thoroughly detail application strategies for RL across various phases of the LLM lifecycle, including pre-training, alignment fine-tuning, and reinforced reasoning. In particular, we emphasize that RL methods in the "reinforced reasoning" phase serve as a pivotal driving force for advancing model reasoning to its limits. Next, we collate existing datasets and evaluation benchmarks currently used for RL fine-tuning, spanning human-annotated datasets, AI-assisted preference data, and program-verification-style corpora. Subsequently, we review the mainstream open-source tools and training frameworks available, providing clear practical references for subsequent research. Finally, we analyse the future challenges and trends in the field of RL-enhanced LLMs. This survey aims to present researchers and practitioners with the latest developments and frontier trends at the intersection of RL and LLMs, with the goal of fostering the evolution of LLMs that are more intelligent, generalizable, and secure.

CCS Concepts: • General and reference → Surveys and overviews; • Computing methodologies → Natural language processing; Reinforcement learning.

Authors' Contact Information: Keliang Liu, klliu25@m.fudan.edu.cn, Fudan University, Shanghai, China; Dingkang Yang, dkyang20@fudan.edu.cn, Fudan University, Shanghai, China and ByteDance SAIL Team, Shanghai, China; Ziyun Qian, zyqian22@m.fudan.edu.cn, Fudan University, Shanghai, China; Weijie Yin, yinwj2021@163.com, ByteDance SAIL Team, Shanghai, China; Yuchi Wang, wangyuchi@link.cuhk.edu.hk; Hongsheng Li, hsli@ee.cuhk.edu.hk, The Chinese University of Hong Kong, MMLab, Hongkong, China; Jun Liu, j.liu81@lancaster.ac.uk, Lancaster University, Lancaster, UK; Peng Zhai, pzhai@fudan.edu.cn, Fudan University, Shanghai, China; Yang Liu, yangliu@cs.toronto.edu, Tongji University, Shanghai, China and The University of Toronto, Toronto, Canada; Lihua Zhang, lihuazhang@fudan.edu.cn, Fudan University, Shanghai, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

<sup>\*</sup>Both authors contributed equally to this research.

<sup>&</sup>lt;sup>†</sup>Project leader.

<sup>‡</sup>Corresponding authors.

Additional Key Words and Phrases: Reinforcement Learning, Large Language Models, Reasoning, Alignment, Reinforcement Learning from Human Feedback.

#### **ACM Reference Format:**

#### 1 Introduction

Large Language Models such as ChatGPT [126] have risen rapidly demonstrating remarkable performance across various tasks, including general dialogue [9], code generation [105], and mathematical reasoning [40], and have gradually become essential cornerstones for interactive artificial intelligence systems [20, 21, 89, 98, 205, 206]. Despite their broad generalization capabilities, current LLMs still struggle with crucial shortcomings: they often fail to reliably capture nuanced human intentions and can produce misleading or unsafe outputs [11, 14, 43, 81, 158, 185]. Moreover, several recent studies [65, 123, 151] have indicated that the reasoning capabilities of LLMs still exhibit substantial shortcomings. Therefore, effectively aligning the generative capabilities of LLMs with human preferences, values, and specific task requirements, as well as enhancing their reasoning abilities for addressing complex problems, has emerged as one of the significant challenges in current LLM research. In response, RL has been introduced as a powerful framework to address these challenges by directly optimizing model behavior through interactive feedback and reward signals. Table 1 shows the performance improvement of typical models after being trained with RL compared to their baselines.

Table 1. This table compares representative models trained with RL against their baseline counterparts, showing that RL substantially enhances the performance of foundation models and underscoring the critical importance of reinforcement learning. Among them, Magistral Small-SC\* and Magistral Small-RL# refer to Magistral Small-24B-Starting Checkpoint and the result of this model trained only through reinforcement learning, respectively.

Model / Benchmark	AIME2024	GPQA-Diamond	LiveCodeBench	MATH-500	MMLU	SWE-benchVerified
DeepSeek-V3 [102]	39.2	59.1	36.2	90.2	88.5	42.0
DeepSeek-R1-Zero [48]	71.0 <b>(+31.8)</b>	73.3 <b>(+14.2)</b>	50.0 <b>(+13.8)</b>	95.9 <b>(+5.7)</b>	_	-
DeepSeek-R1 [48]	79.8 <b>(+40.6)</b>	71.5 <b>(+12.4)</b>	65.9 <b>(+29.7)</b>	97.3 <b>(+7.1)</b>	90.8 (+2.3)	49.2 <b>(+7.2)</b>
Magistral Small-SC* [139]	32.2	63.4 (GPQA, +SFT)	22.7 (v5)	93.2 (+SFT)	_	-
Magistral Small-RL <sup>#</sup> [139]	65.8 <b>(+33.6)</b>	68.8 (GPQA, <b>+5.4</b> )	46.4 (v5, <b>+23.7</b> )	95.4 <b>(+2.2)</b>	_	-
GPT-40-0513 [69]	9.3	49.9	32.9	74.6	87.2	38.8
OpenAI-o1-1217 [70]	79.2 <b>(+70.2)</b>	75.7 <b>(+25.8)</b>	63.4 <b>(+30.5)</b>	96.4 <b>(+21.8)</b>	91.8 <b>(+4.6)</b>	48.9 <b>(+10.1)</b>

Since the seminal introduction of Reinforcement Learning from Human Feedback (RLHF) by Ouyang *et al.* [129], RL-based fine-tuning has become a cornerstone method for improving LLM alignment with human instructions and preferences. By leveraging human evaluative feedback or learned reward models, RLHF enables models to iteratively adjust their outputs toward more preferred and helpful responses, going beyond what supervised training alone can achieve. Building on the success of RLHF for alignment, researchers have more recently begun to apply RL paradigms to bolster reasoning capabilities. Notably, starting around 2024, a series of advanced LLMs demonstrated substantial improvements on complex reasoning tasks (*e.g.*, in mathematics and programming) by employing test-time or post-training RL techniques. High-profile examples include OpenAI's o1 system [70], Anthropic's Claude 3.7/4 [3], Manuscript submitted to ACM

DeepSeek R1 [48], the Kimi K1.5 [160], and Owen 3 [204] etc., all of which integrate reinforcement-driven reasoning strategies during inference. These successes suggest that reinforcement learning, when applied at the inference or post-training stage, can unlock new problem-solving abilities in LLMs beyond their pre-trained knowledge. A key innovation underlying these recent advances is the paradigm of Reinforcement Learning with Verifiable Rewards (RLVR) [48, 87, 204], which augments the standard RL loop with objective, automatically verifiable reward signals, such as programmatic checks or proofs of correctness on the model's output. By rewarding an LLM for producing outputs that pass rigorous correctness tests (e.g., unit tests for code or theorem verifications for math), RLVR directly incentivises the model to generate reliably correct and logically sound solutions. This approach has been a driving force behind the aforementioned reasoning improvements, effectively pushing models to reason through multi-step problems until a verifiable correct result is found. Nevertheless, the integration of RL into LLM training and usage raises several open questions and limitations. First, it remains under debate to what extent RLVR truly expands the LLM's reasoning capabilities beyond what was learned during pre-training [190, 218, 235]. Second, there is no clear consensus on how different RL techniques should be best applied at various stages of the LLM lifecycle, ranging from pre-training and instruction alignment to post-training inference optimization. Third, practical issues of data curation and optimization strategy in RL remain challenging; e.g., constructing high-quality reward datasets via human preference labels, AI-assistant preferences, or programmatic rewards and choosing appropriate RL algorithms such as policy gradients versus reward model optimization are non-trivial design decisions. Finally, the question of how to implement RL fine-tuning efficiently at scale without destabilizing the model's performance is still not fully resolved.

In light of these gaps, this survey aims to provide a systematic and comprehensive review of recent progress in RL-enhanced LLMs, with particular focus on developments in the highly influential RLVR paradigm, especially with rapid developments since 2025. We aim to clarify the role of RL methods in the entire LLM training pipeline and their contributions to advancing the frontiers of model alignment and reasoning. Specifically, we offer in-depth analysis and discussion along multiple dimensions: (1) the theoretical foundations of applying RL to LLMs; (2) application strategies detailing how RL is integrated at different training stages, including initial pre-training, alignment fine-tuning, and post-training inference-time reasoning; (3) the datasets and benchmarks used to train and evaluate RL-fine-tuned LLMs; and (4) the emerging tools and frameworks that support large-scale RL training for LLMs. By organizing the survey along these axes, we aim to provide researchers and practitioners with a clear roadmap of the field's current state, insights into the efficacy and limitations of various RL techniques especially RLVR, and well-supported guidance for future work in leveraging RL to make LLMs more aligned, powerful, and reliable.

## 1.1 Related Surveys

In recent years, numerous surveys [8, 12, 15, 16, 51, 72, 75, 78, 80, 85, 134, 154, 169, 178, 197, 224, 239, 240, 244] have reviewed reinforcement learning research related to large language models and proposed various classification schemes. Existing surveys have proposed a variety of classification schemes, but often with a limited scope. For example, some studies [78, 178, 239] narrowly focus only on RL-based alignment techniques, organizing their taxonomies primarily around the use of reward models while overlooking important emerging approaches. Although several works in 2025 have attempted to summarize research on RL at inference time [8, 12, 80, 197, 240], these reviews are often partial and fail to provide a holistic examination of reinforcement-at-inference across its multiple dimensions. Pternea *et al.* [134] discuss the synergy between RL and LLMs, but their analysis is largely limited to the perspective of bidirectional RL–LLM collaboration. Zhu *et al.* [244] focuses exclusively on the narrow domain of Concise and Adaptive Thinking. While these survey frameworks offer value, they remain constrained to specific viewpoints and lack a unified, end-to-end lifecycle

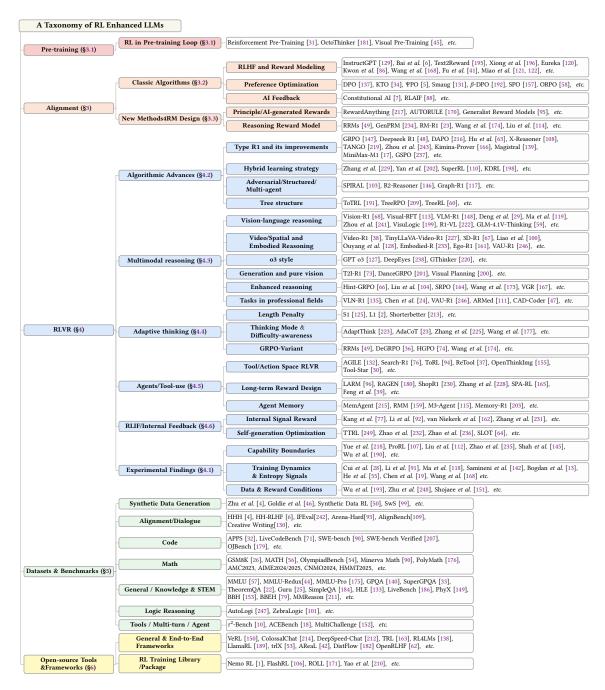


Fig. 1. A taxonomy of RL enhanced LLMs. This figure presents a taxonomy of the key stages and resources involved in creating RL-enhanced LLMs, organized into five branches: pre-training, alignment, RLVR, datasets & benchmarks, and open-source frameworks. The taxonomy clarifies the interconnections between stages, serving as a roadmap for understanding methodological advancements and resources discussed in the survey.

Table 2. Comparative Analysis Table of Representative Surveys: The comparison is conducted across five dimensions—lifecycle coverage, dataset and benchmark summarization, tool/framework collection and practicality, breadth and timeliness of citations, and future outlook and challenges.

Survey $\downarrow$ / Dimension $\rightarrow$	Lifecycle Coverage	Datasets&Benchmarks	Tools / Frameworks	Citation Breadth & Timeliness	Future Directions & Challenges
Wang et al. [178]	× (Alignment only)	× (Limited mention)	× (Not covered)	× (Insufficiently updated)	✓ (Future directions mentioned)
Srivastava et al. [154]	× (Alignment + Reasoning)	× (For demonstrating performance only)	× (Not covered)	$\checkmark$ (Covers up to 2025)	$\checkmark$ (Dedicated section)
Wang et al. [169]	× (Mainly Alignment)	× (Limited mention)	× (Not covered)	√ (Covers early 2025)	√ (Simpler discussion)
Cao et al. [15]	× (Not all covered)	× (Not covered)	× (Not covered)	× (Not included)	$\checkmark$ (Dedicated section)
Chaudhari et al. [16]	× (Alignment only)	× (Limited mention)	× (Not covered)	× (Focus on RLHF only, outdated)	$\checkmark$ (In-depth analysis)
Kaufmann et al. [78]	× (Alignment only)	$\checkmark$ (Dedicated section)	✓ (Library support mentioned)	× (Relatively early)	✓ (Brief analysis)
Our Survey	$\checkmark$ (Full coverage)	$\checkmark$ (Dedicated section)	$\checkmark$ (Well organized)	$\checkmark$ (Covers latest work)	$\checkmark$ (In-depth analysis)

perspective on RL–LLM interactions. In contrast, our survey systematically investigates the role of RL throughout the entire LLM training pipeline (ranging from pre-training and alignment fine-tuning reasoning) and proposes an organizational framework that, to the best of our knowledge, has not been comprehensively addressed in prior research. Table 2 summarizes the advantages and disadvantages of our survey compared with other representative surveys.

### 1.2 Contribution Summary

This survey provides a structured review of RL techniques for LLMs, with three distinctive contributions:

- Lifecycle Organization: We systematically cover the full lifecycle of RL for LLMs, detailing each stage of the process, from pre-training, alignment, to reinforcement for reasoning. In doing so, we clarify the objectives, methodologies, and challenges encountered at each phase. This organization helps in understanding how RL techniques are applied and refined throughout the LLM development lifecycle.
- Advanced RLVR Technology Focus: This paper highlights state-of-the-art approaches in RLVR. We provide
  an in-depth analysis of the experimental phenomena and cutting-edge applications of RLVR, exploring the
  methodologies used to ensure that rewards are objective and verifiable. Additionally, we discuss how verifiable
  rewards contribute to improved model performance and alignment, showcasing the strengths and limitations of
  RLVR in real-world applications.
- Consolidated Resources: We summarize the datasets, benchmarks, and open-source frameworks that are critical for RL-based experimentation, evaluation, and practical implementation in LLMs. By aggregating this information, we provide a valuable resource for future researchers looking to experiment with RL techniques in the context of LLMs. The inclusion of these resources enhances the reproducibility and transparency of RL-driven LLM research.

To provide an organizational roadmap, Figure 1 presents a comprehensive taxonomy, which divides existing approaches into five branches: pre-training, alignment, RLVR, datasets & benchmarks, and open-source frameworks. As outlined in Figure 2, our review is organized around the full RL lifecycle for LLMs, with a particular emphasis on RL with verifiable

rewards. In summary, this survey delivers a lifecycle-based synthesis of methods, with particular emphasis on RLVR, complemented by practical resources for research and application.

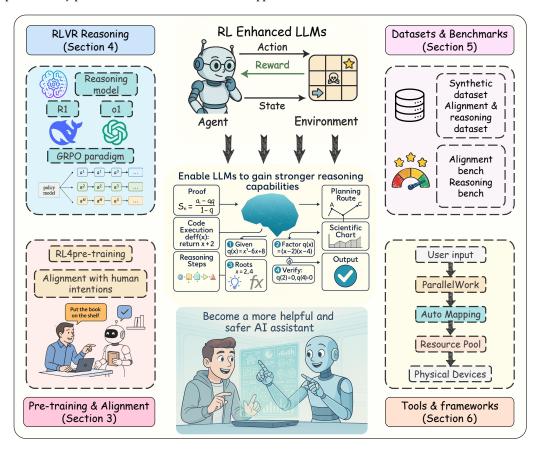


Fig. 2. Key components in RL-enhanced LLMs. This figure illustrates the key components and their interactions within the lifecycle of RL-enhanced LLMs. Driven by RL frameworks and toolkits, RL algorithms participate in the pre-training, alignment, and reasoning enhancement training of LLMs, and are validated through test benchmarks.

# 2 Preliminaries of Reinforcement Learning

Reinforcement learning enables agents to learn optimal policies through interaction with the environment, aiming to maximize cumulative rewards. A typical RL problem can be modelled as a Markov Decision Process (MDP), which consists of a state space, an action space, a state transition probability distribution, and a reward function. At each timestep, the agent selects an action a based on the current state s, receives an immediate reward r, and transitions to a new state s' according to the environment dynamics. The objective of the agent is to learn an optimal policy  $\pi^*$  that maximizes the expected long-term cumulative reward over the course of interactions. To achieve this objective, RL algorithms have evolved along two primary paradigms: policy-based and value-based learning. The former directly focuses on optimizing the policy, often through policy gradient methods; the latter emphasizes estimating the value of states or actions, from which the policy is derived indirectly. This section introduces representative algorithms and theoretical foundations of these two RL paradigms, and further discusses their applications in LLM training. Manuscript submitted to ACM

#### 2.1 Policy Learning

Policy learning methods directly optimize the policy  $\pi(a|s;\theta)$ , typically without explicitly learning an environment model or value function. A common approach is the policy gradient method, which adjusts policy parameters  $\theta$  in the parameter space through gradient ascent to maximize expected returns. REINFORCE [187] is the most fundamental Monte Carlo policy gradient method. It directly estimates the gradient of the expected return with respect to the policy parameters, where the objective is defined as  $J(\theta) = \mathbb{E}[R]$ , with R denoting the cumulative return. By applying the log-derivative trick for stochastic policies, an unbiased estimator of the policy gradient can be derived as:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[ \sum_{t=0}^{T} \nabla_{\theta} \log \pi_{\theta}(a_{t}|s_{t}) R_{t} \right]. \tag{1}$$

Here,  $\tau = (s_0, a_0, r_0, \dots, s_T)$  denotes a trajectory, and  $R_t = \sum_{k=t}^T \gamma^{k-t} r_k$  is the discounted return starting from timestep t, where  $\gamma$  is the discount factor. Intuitively, this formulation implies that increasing the probability of taking action  $a_t$  in state  $s_t$  is positively correlated with the cumulative reward obtained thereafter [156]. After sampling a complete trajectory, REINFORCE updates the policy parameters using the estimated gradient as  $\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$ , where  $\alpha$  is the learning rate. To reduce the variance of the gradient estimator, a common technique is to introduce a baseline function b(s), which depends only on the state. Subtracting this baseline from the return does not bias the gradient estimate but can significantly reduce its variance. The policy gradient estimator with a baseline is thus written as:

$$\nabla_{\theta} J(\theta) = \mathbb{E}\left[\sum_{t} \nabla_{\theta} \log \pi_{\theta}(a_{t}|s_{t}) \left(R_{t} - b(s_{t})\right)\right]. \tag{2}$$

 $A_t = R_t - b(s_t)$  is the advantage function, which measures how much the actual return exceeds the baseline, reducing the variance of the gradient estimator without altering its expectation.

The Actor-Critic (AC) method [82] combines policy gradient techniques with value function approximation by integrating two components within a unified framework: the actor, which selects actions according to a parameterized policy  $\pi_{\theta}(a|s)$ , and the critic, which evaluates the policy using a parameterized value function  $V_{\phi}(s)$  or action-value function  $Q_{\phi}(s,a)$ . At each timestep, the AC algorithm alternates between: (1) the critic estimating the advantage A(s,a) = Q(s,a) - V(s) or the Temporal-Difference (TD) error  $\delta = r + \gamma V(s') - V(s)$ ; and (2) the actor updating the policy parameters along the direction of the policy gradient, weighted by the low-variance estimate provided by the critic, *i.e.*, using  $\nabla_{\theta} \log \pi_{\theta}(a|s) \cdot A(s,a)$  as the gradient.

Trust Region Policy Optimization (TRPO) [143] aims to address the instability that can arise from large policy updates. TRPO formulates policy optimization as a constrained optimization problem: it maximizes the expected advantage under the old policy  $\pi_{\text{old}}$ , while constraining the KL divergence between the new and old policies to remain below a threshold  $\delta$ . The formal objective is given by:

$$\max_{\theta} L(\theta) = \mathbb{E}_{s \sim \pi_{\text{old}}} \left[ \sum_{a} \frac{\pi_{\theta}(a|s)}{\pi_{\text{old}}(a|s)} A^{\pi_{\text{old}}}(s, a) \right], \tag{3}$$

s.t. 
$$\mathbb{E}_{s \sim \pi_{\text{old}}} \left[ D_{\text{KL}}(\pi_{\text{old}}(\cdot|s) \parallel \pi_{\theta}(\cdot|s)) \right] \leq \delta$$
.

Proximal Policy Optimization (PPO) [144] is a landmark innovation of traditional policy gradient algorithms in the era of deep reinforcement learning. The core contribution is a clipped surrogate objective that allows multi-step gradient

updates without policy collapse. Specifically, PPO uses the objective function:

$$L^{\text{PPO}}(\theta) = \mathbb{E}_t \left[ \min(r_t(\theta) \hat{A}_t, \operatorname{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right]. \tag{4}$$

Here,  $r_t(\theta)$  represents the probability ratio of the old and new policies on the action at time t,  $\epsilon$  is the threshold, and  $\hat{A}_t$  is the advantage estimate. This objective enables optimization according to the standard policy gradient when the magnitude of the policy change is within the threshold; once it exceeds the threshold, the gradient is weakened, thereby ensuring that a single update will not cause the policy to deviate too far from the original policy.

In LLM fine-tuning, PPO optimizes parameters via reward scores with a value network baseline for efficient advantage estimation, but despite its stability and success in RLHF alignment, reasoning tasks face added memory/computation costs from the extra value network and instability in long sequences due to inaccurate value estimates. To address this issue, as well as the low learning efficiency in traditional RLHF settings where only one response is scored at a time, the DeepSeek team proposed Group Relative Policy Optimization (GRPO) in DeepSeekMath [147]. The core idea of GRPO is to sample a set of outputs for each prompt and use the relative differences in intra-group feedback to guide policy updates. Specifically, for each question-answer pair, the behavioral policy of GRPO generates G different answers at once (forming a group). Then, each answer is assigned a reward value  $R_i$  through a reward model or predefined rules. Instead of training a separate value function to estimate a global baseline, GRPO adopted the intra-group average reward as the benchmark: it calculates the average or a certain statistic of the rewards of all answers in the group, and defines the advantage of each answer as  $A_i = R_i - \bar{R}_{group}$ . In this way, answers with rewards higher than the group average gain positive advantages, while those lower than the average gain negative advantages. Subsequently, GRPO constructs a clipped policy objective function similar to PPO, and through gradient ascent, it increases the probability of answers with positive advantages and decreases the probability of answers with negative advantages. Since the group average serves as a dynamic baseline, advantages can be calculated without additional training of a value network, thereby simplifying the algorithm structure. For a specific question-answer pair (q, a), the behavioral policy  $\pi_{\theta_{\text{old}}}$  samples and generates G independent responses  $\{o_i\}_{i=1}^G$ . Subsequently, the advantage value of the i-th response is calculated by normalizing the group-level rewards:

$$\hat{A}_{i,t} = \frac{r_i - \max(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}.$$
 (5)

Similar to PPO, GRPO employs a clipped objective function and directly introduces a KL penalty term. The objective function of GRPO is as follows:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \\
\left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left( \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{ clip } \left( r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right) - \beta D_{\text{KL}}(\pi_{\theta} | | \pi_{\text{ref}}) \right).$$
(6)

#### 2.2 Value Learning

Value-based methods aim to indirectly derive the optimal policy by estimating value functions. A value function quantifies the expected long-term utility of a state or state-action pair under a given policy. Typical examples include the state-value function  $V^{\pi}(s) = \mathbb{E}_{\pi}[R \mid s]$  and the action-value function  $Q^{\pi}(s, a) = \mathbb{E}_{\pi}[R \mid s, a]$ . Value-based algorithms focus on approximating the optimal value function V(s) or Q(s, a), and then deriving the optimal policy by following the value maximization principle—e.g., selecting the action with the highest estimated value at each state.

Q-learning [183] adopted a model-free, off-policy learning approach to approximate the optimal action-value function  $Q^*(s, a)$ . The core idea is to iteratively update value estimates for state-action pairs, guided by the Bellman optimality equation. The basic Q-learning update rule is given by:

$$Q_{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[ r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right], \tag{7}$$

where  $\alpha$  denotes the learning rate, and  $\gamma$  is the discount factor. The update rule corrects the current estimate  $Q(s_t, a_t)$  by incorporating the temporal-difference (TD) error:  $\delta = r + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$ , which reflects the discrepancy between the newly estimated return and the previous estimate of  $Q(s_t, a_t)$ . In each step, Q-learning collects experience tuples  $(s_t, a_t, r_t, s_{t+1})$  using exploration strategies such as  $\epsilon$ -greedy, and then updates the corresponding Q value. Under suitable conditions, the Q(s, a) values converge to the true optimal action-value function  $Q^*(s, a)$  in the long run. Since each update relies on the estimated maximum reward at the next state, *i.e.*,  $\max_{a'} Q(s_{t+1}, a')$ , rather than the action actually taken, Q-learning is categorized as an off-policy method. This allows it to learn from historical data or samples generated by different policies. However, this also introduces an overestimation bias: the maximization step can yield upward-biased value estimates. In practice, several improvements have been proposed, such as Double Q-learning [52], which mitigates overestimation by maintaining two separate value estimators.

SARSA [141] is another temporal-difference-based value learning method. In contrast to Q-learning, SARSA is an on-policy algorithm: it evaluates the action values according to the currently executed policy and updates the estimates using samples collected from the same policy. The SARSA update rule is given by:

$$Q_{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[ r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]. \tag{8}$$

Unlike Q-learning, which uses the maximal action at the next state  $s_{t+1}$ , SARSA relies on the action  $a_{t+1}$  actually taken by the agent according to the current policy (e.g.,  $\epsilon$ -greedy). This implies that SARSA updates the estimate of  $Q^{\pi}(s,a)$  under the current policy  $\pi$ . As the policy gradually improves toward a greedy strategy, the SARSA estimates of Q progressively approach the optimal  $Q^*$ .

Deep Q-Network (DQN) [124] represents a breakthrough in value-based methods by introducing neural function approximation into Q-learning. The core idea of DQN is to use a deep neural network  $Q(s, a; \theta)$ , parameterized by  $\theta$ , to approximate the action-value function. By adjusting the network parameters,  $Q(s, a; \theta)$  outputs value estimates for all possible actions given a state input. DQN adopted the Q-learning target to train this network, aiming to make the predicted Q-values satisfy the Bellman optimality equation. Specifically, given a transition (s, a, r, s') sampled from the experience replay buffer, DQN minimizes the difference between the predicted Q-value and the TD target using the mean squared error loss:

$$L(\theta) = \left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta)\right)^2. \tag{9}$$

Here,  $\theta^-$  denotes the parameters of the target network, which is periodically synchronized from the training network parameters  $\theta$ , but kept fixed between updates. This dual-network mechanism enhances training stability by preventing divergence caused by simultaneous changes in both the target and prediction values.

In the domain of LLMs, value-based methods are not the primary components of training frameworks such as RLHF. This is largely due to the immense and complex action space in LLMs, making it infeasible to explicitly construct Q-tables or networks that evaluate the value of every possible output, as in standard reinforcement learning environments. Nonetheless, the conceptual foundations of value learning still manifest in LLM reinforcement learning. For instance, Wang *et al.* [172] adopted a Q-learning-based framework to dynamically select in-context exemplars. By computing a

diversity score over label distributions among selected demonstrations, the framework jointly maximizes both diversity and task relevance, effectively guiding LLMs to generate more informative references for text classification.

### 3 Reinforcement Learning Methods in the Pre-training Phase and Alignment Phase

### 3.1 Reinforcement Learning Methods in the Pre-training Phase

Most current reinforcement learning enabled tasks for LLMs are mainly focused on the alignment and fine-tuning phases of the models. However, Dong et al. [31] reconstructed the next-token prediction task in pre-training into an RL-based reasoning task, allowing the model to obtain verifiable rewards when correctly predicting the next token in a given context, thereby introducing reinforcement learning into pre-training tasks. Nevertheless, this method consumes excessive resources and requires an excellent model with existing reasoning capabilities as the base model for training. Ghosh et al. [45] introduced RL into visual pre-training with Annotation Bootstrapping, framing unlabeled image pre-training as an RL problem and arguing that common self-supervised methods like crop-consistency resemble value learning. OctoThinker [181] proposed to significantly improve the compatibility of basic language models with reinforcement learning through a two-stage mid-training strategy, enabling the Llama series, which was originally unsuitable for RL, to reach the same level as Qwen in mathematical reasoning tasks. It reveals the key role of mid-training data quality, style, and scheduling strategies in RL scaling. Mid-training refers to self-supervised training conducted in the same way as pre-training, i.e., through next-word prediction, but with a different goal. The goal of mid-training is to transform the pre-trained model to make it suitable for RL training, and the training data is converted from massive amounts of text to high-quality, task-related data.

### 3.2 Classic Algorithms in the Alignment Phase

Christiano et al. [129] established a foundational paradigm for modern LLM alignment, demonstrating that incorporating human preferences into fine-tuning significantly improves the helpfulness and safety of instruction following behaviour. Bai et al. [6] found that RLHF-based alignment training enhances performance across nearly all NLP evaluation tasks and is fully compatible with domain-specific skills such as Python programming and summarisation. Xiong et al. [196] reinterpreted RLHF from an information theoretic perspective, proposing an iterative optimization framework that controls alignment bias via KL regularization and drives data acquisition based on uncertainty estimates. SPO [157] modeled human preference as a minimax winner in a zero-sum game and directly optimized the policy through self-play, providing theoretical guarantees of robust convergence under non-transitive, non-Markovian, and stochastic preferences. Wang, Fu, and Miao et al. [41, 121, 122, 168] explored methods for mitigating the issue of reward hacking. Bai et al. [7] proposed Constitutional AI, a framework in which artificial intelligence systems supervise other AI agents to train a harmless assistant through self-improvement without relying on human-labelled data identifying harmful outputs. The only human oversight is provided through a predefined set of rules or principles. RLAIF [88]

Traditional RLHF requires first training a reward model and then optimizing the policy through RL. This two-stage process is complex and may be unstable. In 2023, Rafailov et al. [137] introduced the Direct Preference Optimization (DPO) method, a training paradigm that eliminates the need for explicit reinforcement learning. It has been proven that under certain assumptions, it can bypass explicit reward modeling and RL optimization, and directly fine-tune the Manuscript submitted to ACM

(Reinforcement Learning with AI Feedback) leveraged existing large language models to generate preference labels without the involvement of human annotators, achieving performance improvements comparable to those of RLHF.

pre-trained language model to the optimal policy indicated by preference data through a simple loss function. Azar et al. proposed a unified theoretical framework  $\Psi$ PO [5], which systematically characterizes the fundamental connections and limitations between RLHF and DPO. Smaug [131] provided a theoretical analysis showing that DPO can fail to preserve the likelihood of preferred sequences when the preference data involves small edit distances. To address this, they introduced a regularized variant, DPOP, which augments the loss with a lower bound constraint on the preferred sequence likelihood.  $\beta$ -DPO [192] improved the robustness and alignment performance of DPO under diverse data conditions by dynamically adjusting the KL regularization coefficient  $\beta$  based on intra-batch preference quality and employing  $\beta$ -guided data filtering.

Kwon et al. [86] proposed a method that leverages LLMs as proxy reward functions by generating reinforcement learning signals from natural language prompts, thereby enabling efficient training of agents aligned with user intentions. Eureka [120] achieved automated reward function design without task-specific prompting by incorporating environment source code as contextual input, combined with reflective reward mechanisms and evolutionary search. Text2Reward [195] utilized LLMs to automatically generate interpretable and dense reward functions from natural language instructions. KTO [34] directly optimized for human utility over generated content rather than the log-likelihood of preferences, achieving performance comparable to or better than existing methods using only binary feedback. ORPO [58] integrated Supervised Fine-Tuning (SFT) with dynamic preference optimization, significantly enhancing instruction following capabilities and generation quality of language models without requiring a reference model. Since comprehensive surveys on classical RL-based alignment methods have already been published, this subsection provides only a brief overview of representative approaches during the alignment phase. The latest advances in reward model design will be discussed in detail in the next subsection.

# 3.3 Emerging Methods for Reward Model Design

The reward model plays a key role in guiding large language models to generate outputs that align with human expectations. Recently, a series of studies [23, 49, 114, 174, 234] have utilized testing-phase computational resources to enhance the performance of reward models. Guo et al. present the Reward Reasoning Model (RRM) [49], which is trained using a reinforcement learning framework, Through the Chain-of-Thought (CoT) reasoning mechanism, RRM can adaptively call additional testing phase computational resources for complex queries where reward judgments are unclear. Zhao et al. [234] introduced a generative process reward model, which performs explicit CoT reasoning and code verification before making judgments for each reasoning step. Chen et al. [23] framed reward modeling as a reasoning task. By distilling high-quality reasoning chains and training in two phases—reinforcement learning based on verifiable rewards—the model achieves state-of-the-art performance on three major reward model benchmarks. Wang et al. [174] presented the first unified reward model, UNIFIEDREWARD-THINK, based on multimodal CoT reasoning. This model uses exploration-driven reinforcement fine-tuning to unlock the model's latent complex reasoning capabilities. Liu et al. [114] proposed a rule-based online RL trained Pointwise Generative Reward Modeling (GRM) method, which can evaluate single, paired, and multiple responses, thereby overcoming the limitations of traditional reward models in input flexibility. At the same time, Yu et al. [217] explored how to improve the generalization of reward models, advocating that reward models should understand and follow dynamically provided natural language reward principles, similar to instruction following in large language models. A novel reward model is introduced that is designed and trained by explicitly following natural language principles. Li et al. [95] found that any LLM trained using standard next-token prediction inherently contains a powerful general-purpose reward model. AUTORULE [170] builds rule-based rewards

by extracting rules from preference feedback through interpretation, candidate selection, and merging, then using a verifier to measure rule satisfaction as an auxiliary reward alongside a learned reward model.

# 4 Reinforcement Learning Methods in the Reasoning Phase

With the release of GPT-o1 [70] and DeepSeek R1 [48], the focus of research on reinforcement learning for large language models has gradually shifted towards RLVR technology in 2025. This chapter will introduce the latest advancements in the algorithms of RLVR technology and will discuss its applications in multimodal reasoning, adaptive thinking, agents, as well as the integration of RL with other techniques in the fine-tuning phase. Figure 3 illustrates the overall framework diagram of RLVR technology and the key technical points that can be improved.

### 4.1 Experimental Findings of RLVR in Improving the Reasoning Ability of LLMs

Reinforcement Learning with Verifiable Rewards has recently demonstrated notable success in enhancing the reasoning performance of LLMs, particularly in mathematics and programming tasks. However, there is controversy in academic circles over whether RL truly expands the model's reasoning ability, or merely amplifies the high-reward outputs already present in the base model's distribution, as well as whether continuously increasing computational power for reinforcement learning can reliably improve reasoning performance. Liu et al. [107] found that when sufficient training time is provided and applied to new reasoning tasks, reinforcement learning can indeed discover entirely new solution paths that the base model lacks entirely. Although RLVR improves the sampling efficiency of correct reasoning paths, Yue et al. [218], using pass@k (the probability of generating at least one correct solution among k samples) as an evaluation metric, revealed that the current training paradigm has not stimulated truly novel reasoning patterns. Data shows that RLVR models outperform the base model when k is small (e.g., k=1), but the base model performs better when k increases. The boundary of reasoning ability in LLMs often narrows as RLVR training progresses. Analysis of coverage and perplexity indicates that all reasoning paths generated by RLVR exist in the sampling distribution of the base model, suggesting that its capabilities are inherently derived from and limited by the base model. Prior to this, several studies [112, 145, 235] have pointed out that the reflective behavior in RLVR models stems from the base model itself, rather than being acquired through reinforcement learning. To address the above controversies, Wu et al. [190] pointed out that RLVR is mainly an efficient sampler. Although it can occasionally explore capabilities beyond the base model, it often fails to solve new problems due to diversity collapse, and it also forgets the problems that the base model already knows how to solve. Regarding the phenomenon of policy entropy collapse that continuously occurs in massive reinforcement learning experiments without entropy intervention, Cui et al. [28] established a conversion equation between entropy value H and downstream performance R:

$$R = -a\exp(\mathcal{H}) + b. \tag{10}$$

This equation indicates that policy performance is achieved at the cost of entropy consumption, and thus is limited by the depletion of entropy, with a fully predictable theoretical upper limit (R = -a + b when H = 0). This finding suggests that continuous exploration must be achieved through entropy management to break through the computational power expansion bottleneck of reinforcement learning.

There is disagreement regarding whether to retain entropy regularization [63, 129, 147]. Through experiments, Cui et al. [28] showed that although entropy values need to be controlled, it is possible to design objective functions that are superior to entropy loss. The study found that the covariance exhibited by a small number of tokens is extremely high, far exceeding the average level. This means that these abnormal tokens play a dominant role in triggering entropy Manuscript submitted to ACM

collapse. Separating the gradients of high-covariance tokens and preventing their updates from reducing entropy can suppress entropy collapse. Coincidentally, Wang *et al.* [168] found that only a few high-entropy tokens guide the model's reasoning paths, and training on just 80% of low-entropy tokens significantly reduces performance. High-entropy tokens (*e.g.*, "however," "thus," "because," "suppose") handle logical connections, while low-entropy tokens (*e.g.*, affixes, code snippets, mathematical expression components) are used for sentence construction, with other tokens blending both roles. Limiting policy gradient updates to branching tokens can improve RLVR.

Li et al. [91] systematically investigated the necessity of explicit thinking processes in rule-based reinforcement fine-tuning for multimodal large language models. Research has shown that for some tasks and models, removing or adjusting the thinking process can actually improve performance and efficiency. Ma et al. [118] questioned the necessity of explicit thinking. This study reveals that using simple prompts to bypass the thinking process yields better results, and it is found that as the k-value increases, the competitiveness of the non-thinking method in the pass@k metric continues to strengthen. Samineni et al. [142] critically analyzed the structural assumptions of reinforcement learning in the post-training of large language models, pointing out that these assumptions reduce RL to filtered iterative supervised fine-tuning. Through theoretical and empirical evidence, it reveals that the increase in response length is a side effect of training settings rather than an improvement in reasoning ability. Zhu et al. [248] decomposed learning signals into two categories: reinforcing correct answers and punishing incorrect ones. Subsequently, it is found that training using only negative samples without reinforcing correct answers may be more effective. Bogdan et al. [13] proposed three methods-black box resampling, attention aggregation for receiver heads, and causal suppression-to identify "thought anchors" pivotal sentences guiding LLM CoT reasoning, validated by the case study and visualization tool. Wu et al. [193] systematically analyzed that Owen2.5 [136] conducted RLVR training on the model by constructing a clean mathematical dataset under different reward settings. The authors found that Owen can truly improve its mathematical reasoning ability only when trained on clean data with accurate rewards, emphasizing the importance of data cleanliness and the quality of reward design in evaluating RL methods. Chen et al. [19] analyzed the limitations of the traditional "supervised fine-tuning + reinforcement learning" strategy and found that SFT is prone to generating "pseudo reasoning paths". This not only fails to effectively enhance complex reasoning abilities but also significantly impairs the performance of the subsequent RL stage. He et al. [55] presented the Trajectory Policy Gradient Theorem (TPGT), showing response-level rewards can estimate token-level rewards for LLM reinforcement learning, introducing the efficient TRePO algorithm, and comparing it to methods like PPO and DPO. Shojaee et al. [151] investigated Large Reasoning Models through controlled puzzle environments, demonstrating their effectiveness in moderately complex tasks but revealing significant limitations, including accuracy collapse at high complexity and inefficient reasoning processes, questioning their generalizable reasoning capabilities. Liu et al. [104] found that stronger reasoning often increases hallucinations, as longer chains shift attention from image content to language priors, with attention analysis showing weakened visual focus exacerbates this effect.

### 4.2 Recent Advances in RL Algorithms for LLMs

This section chronologically introduces the main algorithms of RLVR, focusing on Group Relative Policy Optimization (GRPO) [147] and its improved algorithms. As RL training for large models on long-chain reasoning tasks such as mathematics competitions and code generation has matured, new challenges and corresponding algorithms have emerged. ByteDance and collaborators released Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) [216], an open-source framework for large-scale long-sequence RL training of LLMs. DAPO, built on GRPO, enhances long-CoT performance with four techniques: Clip-Higher, Dynamic Sampling, Token-Level Policy Gradient Loss, and Overlong

Reward Shaping. Unlike PPO/GRPO's symmetric clipping that restricts low probability actions and causes entropy collapse, DAPO relaxes the upper bound to maintain exploration. It resamples extreme reward prompts to avoid wasted samples and speed up convergence. By weighting loss by sequence length, it discourages verbose, low-quality outputs while retaining high-quality, long ones. Finally, it penalizes or truncates overly long outputs, curbing uncontrolled growth and stabilizing training.

Open-Reasoner-Zero [63] adopted a minimalist training strategy to achieve efficient and scalable improvement in reasoning ability on foundation models without pre-training fine-tuning, significantly outperforming DeepSeek-R1-Zero with only one-tenth of the training steps. Zhang et al. [229] generated pseudo-rewards in a self-supervised manner by leveraging the intrinsic structure of responses from teacher and student models, enabling reward learning without explicit external evaluation. Kimina-Prover [166] is an RL-trained LLM that enhances reasoning capabilities in Lean 4 theorem proving by constructing structured formal reasoning patterns, achieving performance improvements that scale with model size without relying on external search algorithms. SRPO [226] enhanced the reasoning capabilities of LLMs in mathematics and programming tasks through a two-stage reinforcement learning-centric training strategy combined with a historical resampling mechanism, providing a viable pathway for improving cross-task reasoning capabilities. Yan et al. [202] introduced off-policy reasoning trajectories from external models and, through a combination of mixed-policy optimization and regularized importance sampling, enabled the model to learn from both its own generated data and external demonstrations during training. X-Reasoner [108] through general text post-training (i.e., SFT + RL only), has been experimentally proven to have reasoning abilities that can generalize across modalities and domains. TANGO [219] jointly trained the generator and generative process-level verifier of large language models through reinforcement learning, aiming to enable the two to promote each other and evolve synergistically without step-by-step annotations. Zhou et al. [243] extended R1-Zero-style training to tasks without rule-verifiable answers by generating reasoning trajectories, concatenating them with reference answers, and evaluating the likelihood of the reference answer. This likelihood serves both as a reward for trajectory optimization and as a weight for supervised training. REINFORCE++ [61] achieves PPO-like training stability and efficiency without relying on a value network, by incorporating clipped policy updates, KL divergence penalties, and advantage normalization.

SuperRL [110] detected reward sparsity via an adaptive switching mechanism and, when sparsity is identified, activates a hybrid executor that combines policy gradients with offline supervision to stabilize learning. KDRL [198] explored the integration of teacher supervision and RL by constructing a unified objective function that incorporates GRPO and KD. Graph-R1 [117] proposed a proxy-based GraphRAG framework that enhances the accuracy, efficiency, and generation quality of LLMs in knowledge-intensive tasks through lightweight knowledge hypergraph construction, multi-turn interactive retrieval, and end-to-end reinforcement learning optimization. R2-Reasoner [146] employed a reinforced router to decompose queries and allocate subtasks across heterogeneous LLMs, enabling collaborative reasoning that balances accuracy, efficiency, and cost. ToTRL [191] improved the reasoning efficiency of language models in multi-path logical problems by guiding them to transition from linear chain reasoning to tree-structured thinking. TreeRPO [209] constructs tree-structured reasoning paths and optimizes relative node rewards via tree-based sampling, providing LLMs with step-level dense feedback without a reward model. TreeRL [60] introduced the entropy-based tree search strategy EPTree into the reinforcement learning training process of large language models to improve the diversity of reasoning paths and the quality of process supervision signals.

Magistral [139] proposed a method for enhancing the reasoning capabilities of large language models through reinforcement learning alone, without distilling reasoning trajectories. MiniMax-M1 [17] introduced a large language

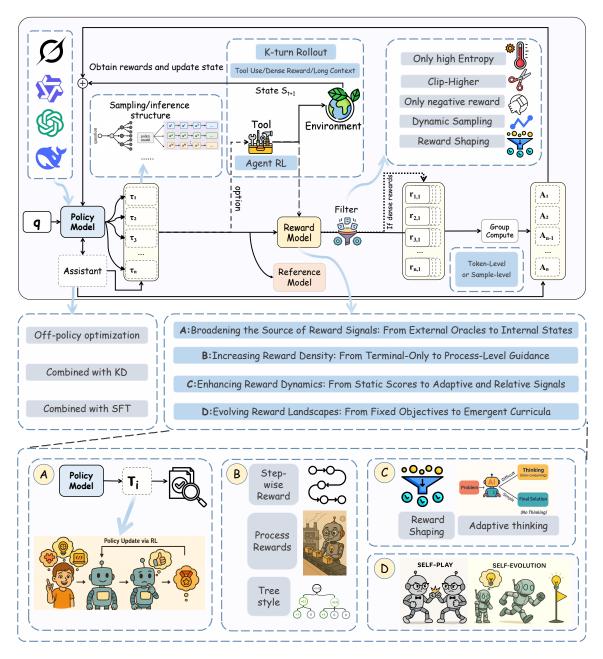


Fig. 3. Technical architecture of the RLVR methods. It depicts the overall workflow of the RLVR and expands on the design methods for the reward model, off-policy assistance, reward filtering, sampling and reasoning strategies, Agent RL, and reward update hierarchy.

model with a mixture-of-experts architecture and a Lightning Attention mechanism, incorporating the CISPO reinforcement learning algorithm to improve training stability and efficiency in RL, while supporting million-token contexts and enabling efficient inference. SPIRAL [103] introduced a multi-agent self-play RL framework with zero-sum

Manuscript submitted to ACM

language games that improves LLM reasoning without supervision or domain data by generating evolving tasks through adversarial opponents and stabilizing training via role-conditioned advantage estimation. The GSPO algorithm [237] improved RL training stability and efficiency by replacing token-level importance weights with sequence-level ratios for reward computation and policy updates, alleviating granularity mismatch.

### 4.3 Application of RLVR in Multimodal Reasoning

DeepSeek-R1-Zero has successfully demonstrated that reasoning capabilities can emerge in large language models solely through reinforcement learning. A large body of work has begun to explore how to leverage reinforcement learning to foster multimodal reasoning abilities. Vision-R1 [68] took the lead in applying reinforcement learning to enhance the reasoning capabilities of Multimodal Large Language Models (MLLMs), while systematically analyzing the differences between direct reinforcement learning training and the combined approach of "cold-start initialization + reinforcement learning training". Visual-RFT [113] proposed a reinforcement fine-tuning framework for Vision-Language Models (VLMs), which is applied in fields such as detection, localization, and classification. VLM-R1 [148] observed that many vision understanding tasks have explicitly annotated ground truths, making them suitable for rule-based reward mechanisms, and thereby extended R1-style reinforcement learning to vision-language models. R1-onevision [208] designed a cross-modal reasoning process, which converts images into standardized text representations, thereby enabling precise language-based reasoning. Deng et al. [29] specifically designed a curriculum reinforcement fine-tuning algorithm for small-scale Vision-Language Models. Ma et al. [119] explored integrating reasoning capabilities into visual perception. Zhou et al. [241] successfully reproduced the emergent properties of multimodal reasoning for the first time on an unsupervised fine-tuned model with only 2 billion parameters. Through experiments, VisuLogic [199] was found that RL is an effective way to enhance the visual reasoning ability of multimodal large language models. R1-VL [222] proposed the StepGRPO algorithm, which provides rewards for paths containing necessary intermediate reasoning steps through soft key step matching technology, and incentivizes reasoning processes that are structured and logically consistent through reasoning completeness and logicality evaluation strategies, thereby achieving dense rewards for visual-language reasoning tasks. SophiaVL-R1 [35] incorporated reasoning-process reward signals by computing the difference in reasoning rewards between correct and incorrect answers, and dynamically assigning confidence weights to these reasoning rewards, thereby mitigating the impact of unreliable reasoning rewards.

Video-R1 [38] and TinyLLaVA-Video-R1 [227] successively applied reinforcement learning to the video domain. Video-R1 established a dataset for video reasoning and designed a mechanism where the model receives positive rewards only when its current reasoning strategy for specific questions demonstrates dependence on temporal information, thereby strengthening the large model's temporal modeling ability. TinyLLaVA-Video-R1 [227], on the other hand, explored how to use reinforcement learning to enhance the video reasoning capabilities of small-scale models. Liao *et al.* [100] discussed methods to enhance the visual-spatial reasoning abilities of MLLMs through R1-Zero-style training. SpaceR [128] extended GRPO through a map imagination mechanism, prompting the model to infer spatial layouts during the thinking process, thereby enhancing the video spatial reasoning ability of MLLMs. The essence of humans' and robots' cognition of the environment lies in perceiving and understanding spatial relationships through first-person perspective video streams. Therefore, a crucial aspect of embodied intelligence tasks is that the model needs to possess the ability to perceive and understand spatial relationships from first-person perspective video streams. Zhao *et al.* [233] proposed the Embodied-R framework, which realizes collaborative work by combining the perceptual capabilities of large-scale VLMs with the reasoning capabilities of small-scale Language Models. Ego-R1 [161] explored a new framework for reasoning over ultra-long (measured in days/weeks) first-person videos. By decomposing complex Manuscript submitted to ACM

reasoning into modular steps, the RL agent iteratively collaborates by calling specific tools at each step, sequentially solving sub-problems such as temporal retrieval and multimodal understanding, thereby significantly extending the time coverage from several hours to a week. VAU-R1 [246] extended RLVR to the field of Video Anomaly Understanding (VAU), enhances anomaly reasoning capabilities through Reinforcement Fine-Tuning (RFT), and introduces VAUBench, the first CoT benchmark specifically designed for video anomaly reasoning. VLN-R1 [135] extended RLVR to the field of Vision-Language Navigation (VLN). CAD-Coder [47] introduced RL into the field of artificial intelligence-assisted CAD. Chen *et al.* [24] combined SFT with GRPO and introduce a rule-based reward tailored to scene graph structures, improving the structural validity, relationship recall, and long-tail category performance of MLLMs in end-to-end scene graph generation. ARMed [111] explored the application of RLVR techniques in the domain of medical imaging. 3D-R1 [67] enhanced the reasoning and generalization capabilities of 3D VLM by combining cold-start initialization with high-quality CoT datasets, RL-based training, and a dynamic view selection strategy.

Huang et al. [66] proposed Hint-GRPO to address GRPO's low data utilization—where models struggle to gain positive updates from hard samples—and text bias—where models ignore images and rely only on text—by supplying partial correct reasoning steps as hints and leveraging prediction differences between inputs with and without image conditions to enforce visual grounding. SRPO [164] explored the problem that MLLMs still lag significantly behind unimodal text models in complex problems that require explicit self-reflection and correction. Wang et al. [173] extended the successful pattern of RL in mathematical reasoning and code generation to the field of visual perception by injecting subtle synthetic visual hallucinations into manually written image description paragraphs and training VLMs to locate these errors.

T2I-R1 [73] took the lead in introducing RLVR strategies into the field of visual generation, proposing a new text-to-image generation model based on a two-level CoT reasoning framework and reinforcement learning, which even surpassed FLUX.1 on benchmarks. DanceGRPO [201] is the first unified framework that adapts GRPO to visual generation paradigms. It enables the universal deployment of a single reinforcement learning algorithm across diffusion models and rectified flows generative paradigms; text-to-image, text-to-video, and image-to-video tasks; foundation models such as Stable Diffusion, HunyuanVideo, FLUX, and SkyReels-I2V; and reward models including image/video aesthetics, text-image alignment, video motion quality, and binary reward. Xu *et al.* [200] explored planning through visual representations completely independent of text, confirming that pure visual planning can serve as a feasible alternative to language-based reasoning.

Although RL has already been applied to multimodal reasoning, its reasoning process is still mainly limited to text forms [208]. There is still substantial room for exploration in approaches that deeply integrate multimodal information and external tools. GPT o3 [127] has demonstrated a strong ability to reason by relying on visual cues, setting off a new upsurge in visual reasoning research, with researchers scrambling to explore methods to achieve GPT o3-style visual reasoning. DeepEyes [238] proposed an interleaved multimodal reasoning paradigm, where the model decides at each step whether to continue reasoning with text or call tools to crop regions of the image as historical context, and proceeds with reasoning in this way to form an interleaved reasoning sequence. Several methods have been proposed to address the limitation that models often struggle to effectively anchor their reasoning on visual cues. For instance, GThinker [220] introduced CueRethinking, which anchors the inference process to visual cues and resolves inconsistencies through iterative cue reinterpretation. GLM-4.1V-Thinking [59] proposed a unified multimodal reasoning framework that integrates multi-source pre-training, supervised fine-tuning, and cross-task reinforcement learning with RLCS and a refined multi-domain reward system to enhance reasoning across diverse tasks.

### 4.4 Adaptive Reasoning

RLVR technology has achieved improved performance during testing by consuming more computing resources to generate longer chain-of-thought sequences. However, the length of its CoT reasoning is uncontrollable, making it impossible to achieve the expected performance level by allocating computing resources during testing. It often occurs that excessive testing time is allocated to simple problems or insufficient testing time is allocated to difficult problems [2, 116, 213]. This has inspired researchers to study adaptive length reasoning methods for large language models. S1 [125] extended computation through "budget forcing" technique. L1 [2] proposed Length Controlled Policy Optimization, an RL-based method. After training, the model can generate outputs that meet the required length constraints given in the prompts, and its performance is better than that of S1. Yi et al. [213] defined Sample Optimal Length as the length of the shortest correct response among multiple generated results, and uses it as a dynamic reward signal to guide the model to achieve efficient reasoning. AdaCoT [116], based on the PPO method, dynamically controlled the CoT trigger decision boundary by adjusting the penalty coefficient, enabling the model to judge the necessity of CoT according to the implicit query complexity. Thinkless [36] is trained under the reinforcement learning paradigm and uses two control symbols: <short> (concise) and <think> (deliberate). It decomposes the hybrid reasoning learning objective into a control symbol loss function (managing the selection of reasoning modes) and an answer loss function, thereby enhancing the hybrid-length reasoning ability. AdaptThink [223] showed that skipping reasoning benefits simple tasks in performance and efficiency, and by constraining the objective with importance sampling, it balances "thinking" and "non-thinking" samples during training to enable adaptive mode selection. Jiang et al. [74] designed a two-stage training framework of first cold start and then reinforcement learning to realize a model that can adaptively judge whether to start the thinking process according to the context information of user queries. Zhang et al. [225] systematically quantified the performance upper bounds of Large Reasoning Models (LRMs) in "long thinking" and "non-thinking" modes. By introducing a precision-aware length reward adjustment mechanism, it adaptively allocates reasoning resources according to problem difficulty to achieve efficient reasoning. Wang et al. [177] first performed supervised fine-tuning on the base model to simultaneously acquire long and short chain reasoning abilities. Then, it adopted a long-short adaptive grouping reward strategy to evaluate the prompt complexity and give corresponding incentives, and implements a logic-based reasoning mode switching loss function to optimize the initial token selection of the model, thereby guiding the reasoning type decision. Zhang et al. [221] used the parameter  $\alpha$  to characterize the scaled thinking phase. After the  $\alpha$  moment ends,  $\alpha$ 1 deterministically terminates slow thinking through a thinking termination marker, thereby promoting rapid reasoning and efficient answer generation. This study models the insertion of reasoning transition markers as a Bernoulli random process to dynamically schedule slow-thinking transitions.

### 4.5 RLVR for Agent

The interplay between long-horizon decision-making and stochastic environmental feedback introduces unique challenges in training large language models as interactive agents. Unlike typical single-turn interaction tasks common in standard LLM applications, interactive agents necessitate extended, multi-turn interactions with their environment. Despite the advancements of reinforcement learning in static tasks, multi-turn agent training in reinforcement learning remains understudied, particularly due to issues such as delayed rewards [39, 96, 165]. AGILE [132] integrated large language models with memory systems, tool utilization, and expert interactions. Here, the LLM serves as the policy model and is fine-tuned using annotated action data through the Proximal Policy Optimization algorithm. LARM [96] employed lightweight LLMs (fewer than 5 billion parameters) to directly output executable actions rather than textual descriptions. Utilizing the PPO algorithm for training and introducing a referee mechanism based on large-scale Manuscript submitted to ACM

LLMs for providing intermediate rewards, LARM successfully addresses the challenge of long-term reward vanishing, demonstrated by obtaining enchanted diamond equipment in Minecraft. Search-R1 [76] optimized the reasoning trajectory of LLMs through multi-round search interactions and adopts retrieval token masking technology to ensure the stability of reinforcement learning training. ToRL [94] extended reinforcement learning directly from foundational models (i.e., without additional post-training), enabling large language models to autonomously leverage computational tools. ReTool [37] enhanced long-chain reasoning by generating synthetic cold-start data to build code-enhanced reasoning trajectories for fine-tuning, and then applying iterative reinforcement learning with task success rewards to autonomously optimize tool-use strategies without manual priors. RAGEN [180] investigated four environments and identified the "Echo Trap" of reward fluctuations and gradient spikes, highlighting that robust reasoning in multi-turn RL requires diversified initial states, moderate interaction granularity, and higher sampling frequency, while lacking fine-grained reasoning-aware rewards leads agents to superficial or hallucinatory strategies. OpenThinkImg [155] optimized task success rates directly through interactive tool feedback, enabling large vision-language models (LVLMs) to autonomously identify optimal tool-use strategies. Feng et al. [39] addressed the sparse and delayed reward issues emerging from extended multi-step interactions by computing macro-level relative advantages at the episodic level based on full trajectory groups, and step-level groupings through an anchored state grouping mechanism. Zhang et al. [228] adopted a heuristic approach to group games based on features such as rules and difficulty, subsequently training specialized models for each group. It then merges the parameters from these group-specific models into a unified model, which is further trained across multiple groups until effective generalization across diverse game scenarios is achieved. Tool-Star [30] proposed a two-stage training framework integrating six tool categories, employing a hierarchical reward design through a multi-tool self-assessment RL algorithm. SPA-RL [165] decomposed final rewards into stepwise contributions aligned with task progress, used a progress evaluator to match cumulative values to completion, and combined them with baseline signals to yield fine-grained intermediate rewards that mitigate delay and enhance training. Shop-R1 [230] improved LLMs' shopping behavior simulation by rewarding action correctness proportional to difficulty, while noting that limited context windows constrain long-term dynamic reasoning. Memory-R1 [203] learned to manage external memory and utilize it for long-term reasoning through two dedicated agents: a memory manager and a response agent. MemAgent [215] employed segmented processing and selective memory mechanisms to manage extended contextual information. In contrast, RMM [159] and M3-Agent [115] explored strategies for managing multi-turn long-term memory.

#### 4.6 Reinforcement Learning from Internal Feedback

While methods such as RLHF and RLVR have achieved remarkable results, they require extensive external supervision. However, a series of studies [190, 218] have found that RLVR does not inspire truly novel reasoning patterns; it merely improves the sampling efficiency of correct reasoning paths. Since RLVR does not bring new external information to LLMs but only stimulates the knowledge learned during pre-training, can we find a way to enable LLMs to activate such pre-trained knowledge without external supervision? Kang *et al.* [77] hypothesized that higher distributed self-certainty across samples correlates with response accuracy and thus proposed self-certainty as a metric for evaluating response quality without external rewards. TTRL [249] utilized prior knowledge from pre-trained models during reinforcement learning training and employed a majority voting method to address the lack of explicitly labeled data, achieving performance improvements. Zhao *et al.* [232] proposed a self-evolving model that autonomously generates and solves tasks to maximize learning progress without external data, using a code executor to verify both tasks and solutions and provide unified rewards for open-ended learning. SLOT [64] is a lightweight test-time method that, without altering

the main model, optimizes an additive vector  $\delta$  for each prompt using self-constructed supervision signals to minimize language modeling loss and thereby improve accuracy and reasoning under complex instructions. Zhao *et al.* [236] used the model's own confidence as the sole reward signal and replaced the external reward in GRPO with a self-certainty score, achieving fully unsupervised learning. Kang *et al.* [77] hypothesized that higher aggregated self-certainty across samples correlates with response accuracy and proposed it as a metric for evaluating quality without external rewards. Li *et al.* [92] used the model's own confidence as the reward signal, eliminating the need for manual annotation, preference models, or reward function design. Zhang *et al.* [231] showed that RLIF, using unsupervised reward proxies (entropy and self-certainty), initially boosts base LLM reasoning to rival RLVR but later degrades below pre-finetuning levels, yields limited gains for instruction-tuned models, and exposes intrinsic causes of these behaviors. RLSF [162] builds preference data from models' self-evaluated answer confidence and fine-tunes them with RL to improve calibration and reasoning in math and multiple-choice tasks.

#### 5 Datasets and Benchmarks

# 5.1 Synthetic Data Generation

Zhu et al. [245] proposed a synthetic data framework for abstract visual reasoning that generates structured QA pairs and reasoning chains via A-SIG for regular patterns and crawls, templates, and manual annotations for irregular ones, providing diverse, well-defined samples to enhance both perception and reasoning. Goldie et al. [46] designed a synthetic data pipeline for multi-step reasoning and tool use by leveraging LLM-tool interactions to build stepwise trajectories with context, actions, and feedback, and applying filtering based on process plausibility and final correctness to yield high-quality offline data for reinforcement learning. Guo et al. [50] proposed a task-definition-based synthetic data RL method that generates QA pairs from task definitions, adapts difficulty to model performance, and reinforces with high-potential samples solvable but not yet mastered, enabling task adaptation without human labels. SwS [99] targeted reasoning tasks in reinforcement learning with large language models. It identifies problems that the model consistently fails to solve during training, extracts the underlying concepts involved, and synthesizes targeted follow-up questions to enhance training.

### 5.2 Datasets and Benchmarks

This section will introduce the common datasets and test benchmarks in reinforcement learning for large language models. Table 3 summarizes commonly used datasets and benchmarks spanning alignment/dialogue, code, math, and general knowledge. HHH [4] is a dialogue-based benchmark designed to evaluate large language models along three critical dimensions: Helpful, Honest, and Harmless (HHH). It is intended to assess the degree of alignment between the model and human expectations during interaction. HH-RLHF [6] focused on training preference (or reward) models as a precursor to RLHF. The test benchmarks for alignment tasks also include IFEval [242], Arena-Hard [93], AlignBench [109], Creative Writing [130], and so on. APPS [56] evaluated a model's ability to generate satisfactory Python code given arbitrary natural language specifications. APPS+ [32] built upon the original APPS dataset with manual verification and refinement, containing 7,456 examples, each including a programming task description, reference solution, function signature, unit tests (input/output), and starter code, specifically designed for code generation. LiveCodeBench [71] extended evaluation beyond code generation to broader code-related skills, including self-repair, code execution, and output prediction. GSM8K [26] contained 8,500 high-quality, linguistically Manuscript submitted to ACM

Manuscript submitted to ACM

Table 3. Datasets and Benchmarks Overview. This table categorizes and lists common datasets and benchmarks for RL-LLM research, covering fields such as alignment, code, mathematics, knowledge, logical reasoning, and agentic tasks.

Category	Dataset & Benchmark	Summary	
Alignment / Dialogue	HHH [4], HH-RLHF [6], IFEval [242], Arena-Hard [93], AlignBench [109], Creative Writing [130]	Evaluates alignment in dialogue, focusing on helpfulness, honesty, and harmlessness.	
Code	APPS [32], LiveCodeBench [71], SWE-bench [90], SWE-bench Verified [207], OJBench [179]	Programming tasks involve code generation and debugging, with automated or real-time evaluation.	
Math	GSM8K [26], MATH [56], OlympiadBench [54], Minerva Math [90], OlympiadBench [54], PolyMath [176], AMC2023, AIME2024/2025, CNMO2024, HMMT2025	Benchmarks for solving mathematical problems, from elementary to advanced levels, including competition and Olympiad-level tasks.	
General Exams / Knowledge & STEM	MMLU [57], MMLU-Pro [175], GPQA [140], SuperGPQA [33], TheoremQA [22], Guru [25], SimpleQA [184], HLE [133], LiveBench [186], PhyX [149], BBH [153], BBEH [79], MMReason [211]	General knowledge benchmarks covering various fields, including STEM and human-level exam comparisons.	
Logic Reasoning	AutoLogi [247], ZebraLogic [101]	Logic Reasoning Evaluation.	
Tools / Multi-turn / Agent	$ au^2$ -Bench [10], ACEBench [18], MultiChallenge [152]	Benchmarks testing multi-turn interaction with tools, agent-based reasoning.	

diverse elementary school math word problems. It also demonstrates that verification strategies significantly improve model performance on GSM8K, especially when scaling data. MATH [56] introduced 12,500 challenging competitionstyle math problems, each accompanied by detailed step-by-step solutions for training models to produce answer derivations and explanations. MMLU [57] measured multitask accuracy of language models across 57 subjects, including elementary math, U.S. history, computer science, and law. Achieving high accuracy requires extensive world knowledge and problem-solving abilities. MMLU-Redux [44] adopted a novel error annotation protocol to identify errors in the dataset, thereby improving MMLU [57]. MMLU-Pro [175] extended MMLU by incorporating more challenging, reasoning-focused questions and increasing the number of answer choices from four to ten. It also removed ambiguous or noisy items present in the original MMLU. GPQA [140] consisted of 448 multiple-choice questions written by domain experts across biology, physics, and chemistry. Expert accuracy is around 65%, while non-experts (with web access) achieve 34%, and GPT-4 scores 39%. Designed to be "Google-proof," the benchmark evaluated deep scientific reasoning. SuperGPQA [33] assessed postgraduate-level reasoning and domain knowledge across 285 disciplines. It employed a novel human-LLM co-filtering strategy to iteratively refine questions using model outputs and expert feedback, eliminating vague or ill-formed items. TheoremQA [22] is the first theorem-driven QA dataset, targeting a model's ability to apply formal theorems to solve complex scientific problems. It includes 800 expert-curated highquality questions covering 350 theorems from mathematics, physics, electrical engineering, computer science, and finance. Guru [25] is a reinforcement learning corpus for reasoning, containing 92,000 verifiable examples across six domains: mathematics, code, science, logic, simulations, and tabular data. BBH [153] included 204 tasks spanning linguistics, child development, mathematics, commonsense reasoning, biology, physics, social bias, and software engineering. BBEH [79] replaced every BBH task with a new one testing similar reasoning skills but at significantly higher difficulty. OlympiadBench [54] is a bilingual, multimodal science benchmark containing 8,476 Olympiad-level math and physics problems, each accompanied by expert-authored step-by-step reasoning. Minerva Math [90] consisted of 272 undergraduate-level STEM problems designed to test multi-step scientific reasoning in language models. SWEbench [90] included 2,294 software engineering tasks derived from real GitHub issues and corresponding pull requests across 12 popular Python repositories. SWE-bench Verified [207] contained 50,000 instances collected from 128 GitHub

Table 4. The Performance of Some Well-Known Reasoning Large Language Models on Test Benchmarks, respectively. The role of these test benchmarks is to compare the performance of mainstream reasoning models in terms of general tasks, alignment, mathematics/programming, and logical reasoning benchmarks. The \* denotes the 14-language version.

		OpenAI-o1 [70]	DeepSeek-R1 [48]	Grok-3-Beta (Think) [194]	Gemini2.5-Pro [27]	Qwen3-235B-A22B [204]
	Architecture	-	MoE	-	-	MoE
	# Activated Params	-	37B	-	-	22B
	# Total Params	-	671B	-	-	235B
General Tasks	MMLU-Redux [44]	92.8	92.9	-	93.7	92.7
	MMMLU* [57]	88.4	86.4	_	86.9	84.3
	GPQA-Diamond [140]	78.0	71.5	80.2	84.0	71.1
	LiveBench [186]	75.7	71.6	-	82.4	<u>77.1</u>
Alignment Tasks	IFEval [242]	92.6	83.3	-	89.5	83.4
	Arena-Hard [93]	92.1	92.3		96.4	<u>95.6</u>
	AlignBench v1.1 [109]	8.86	8.76		9.03	8.94
	Creative Writing v3 [130]	81.7	<u>85.5</u>	-	86.0	84.6
Math&Coding Reasoning	MATH-500 [56]	96.4	97.3		98.8	98.0
	AIME'24	74.3	79.8	83.9	92.0	<u>85.7</u>
	AIME'25	79.2	70.0	77.3	86.7	<u>81.5</u>
	PolyMath [176]	38.9	47.1	-	52.2	54.7
	LiveCodeBench v5 [71]	63.9	64.3	<u>70.6</u>	70.4	70.7
Logic Reasoning	ZebraLogic [101]	81.0	78.7	-	87.4	80.3
	AutoLogi [247]	79.8	86.1	-	85.4	89.0

repositories. SimpleQA [184] evaluated model performance on answering concise factual questions. LiveBench [186] addressed concerns of test set contamination and human/model evaluation biases, covering diverse and challenging tasks across math, programming, reasoning, language, instruction-following, and data analysis. OJBench [179] contained 232 competitive programming problems drawn from national and international contests (e.g., NOI and ICPC), providing a more rigorous test of reasoning under competitive conditions. AutoLogi [247] and ZebraLogic [101] are dedicated to evaluating logical reasoning.  $\tau^2$ -Bench [10] and ACEBench [18] are two complementary benchmarks designed to assess multi-turn tool usage. MultiChallenge [152] evaluated the ability of LLMs to engage in multi-turn dialogue with human users. MMReason [211] encompassed complex problems spanning multiple domains and difficulty levels—ranging from pre-university to higher education, and from foundational to competition-level tasks—all of which require multi-step reasoning to solve. PolyMath [176] is a multilingual benchmark for mathematical reasoning across 18 languages and four difficulty levels. AMC 2023, AIME 2024/2025, CNMO 2024, and HMMT 2025 are high-difficulty mathematical competition benchmarks commonly used in the era of advanced reasoning models. MATH500 is a 500-question subset sampled from the full MATH [56] benchmark. Humanity's Last Exam (HLE) [133] is a multimodal benchmark situated at the frontier of human knowledge, aiming to be the final comprehensive academic evaluation of its kind. HLE contains 2,500 questions spanning dozens of disciplines, including mathematics, the humanities, and natural sciences. Developed in collaboration with global experts, it includes both multiple-choice and short-answer formats suitable for automatic evaluation. Each question has a clear and verifiable answer that is not easily searchable online. Cutting-edge language models currently exhibit relatively low accuracy and calibration on this benchmark. PhyX [149] introduced a large-scale multimodal benchmark covering six physical domains and reasoning types, with tasks built from real visual scenes requiring image understanding, physical modeling, and symbolic reasoning, revealing comprehension bias and weak visual grounding in current MLLMs. Table 4 contrasts several well-known reasoning LLMs across tasks and capacities. Manuscript submitted to ACM

#### 6 Open-source Tools and Frameworks

VeRL [150] is a system framework for efficient RLHF training and scheduling that integrates single- and multi-controller paradigms with a hierarchical interface, introduces a 3D-HybridEngine for parameter re-sharding, and applies automatic device mapping to optimize flexibility and resource utilization. TRLX [53] supported a wide range of distributed training paradigms, including data parallelism, model sharding, tensor parallelism, sequence parallelism, and pipeline parallelism. RL4LMs [138] is designed for optimizing language generators via reinforcement learning. The library implements online policy optimization algorithms and is compatible with any encoder or encoder-decoder model from the HuggingFace Transformers library [188], while supporting arbitrary reward functions. Colossal-AI was among the first to open-source a complete RLHF pipeline, i.e., ColossalChat [214], which included supervised data collection, supervised fine-tuning, reward model training, and reinforcement learning fine-tuning. DeepSpeed-Chat [212] integrates multiple optimization techniques for both training and inference into a unified framework. OpenRLHF [62] is built using Ray [97], vLLM, DeepSpeed [212], and HuggingFace Transformers [188], offering high resource efficiency and support for multiple training strategies. TRL [163] is designed for post-training foundation models using techniques such as SFT, PPO, and DPO. Built on top of the Transformers ecosystem [188], it supported various model architectures and modalities and scales across diverse hardware environments. Wang et al. [180] proposed the RAGEN system, which enhances large language models' reasoning and decision-making capabilities in multi-turn interaction environments through the StarPO reinforcement learning framework. Fu et al. [42] introduced AReaL, a fully asynchronous reinforcement learning system that completely decouples the generation and training processes. In AReaL, rollout workers continuously generate new outputs without waiting, while training workers update the model immediately upon collecting a batch of data. ROLL [171] is a library designed to simplify reinforcement learning for large language models. It addresses the challenges faced by technologists, product developers, and algorithm researchers in managing multi-model, multi-stage training workflows. Nemo RL [1] is a scalable and efficient post-training library capable of supporting models ranging from small to over 100 billion parameters and training environments from a single GPU to thousands. LlamaRL [189] is a PyTorch-based distributed asynchronous reinforcement learning framework that enables efficient training of large-scale language models (ranging from 8B to 405B parameters), achieving significant speedups while maintaining strong performance. Yao et al. [210] proposed Flash-LLM-RL, a package that patches vLLM to support model quantization with parameter updates. FlashRL [106] introduced Truncated Importance Sampling (TIS) to mitigate the gap between rollout and training, enabling the use of quantized rollouts without sacrificing downstream performance. DistFlow [182] introduced a fully distributed reinforcement learning training framework that addresses the common single-controller bottleneck in large-scale language model post-training. It employed a multi-controller architecture and user-defined DAG-based task pipelines to achieve decentralized management of both data and computation.

# 7 Open Discussion

### 7.1 Research Challenges

While reinforcement learning has undeniably enhanced LLM alignment and reasoning, several fundamental challenges continue to hinder its full potential.

7.1.1 Scalability and Training Stability. At the system level, large-scale RL on LLMs remains compute-intensive and sometimes unstable. Fine-tuning billion-parameter, high-action-space models demands vast resources and careful Manuscript submitted to ACM

hyperparameter control; even with distributed frameworks like VeRL [150], achieving stable large-scale convergence is non-trivial. Misspecified rewards or poorly managed dynamics can cause policy collapse or divergence [137, 216]. Tooling is fragmented—libraries vary in interfaces and scope, complicating pipeline integration [53, 212]. More efficient algorithms and robust, unified open-source frameworks are still needed to make RL both accessible and reliable at scale.

- 7.1.2 Reward Design and Credit Assignment. From a methodological perspective, challenges center on reward design, credit assignment, and exploration. Outcome-only rewards bias learning toward obvious, high-probability reasoning and overlook complex or unconventional solutions [190, 218]. Richer signals, such as step-level dense feedback [222], and entropy-based or diversity-promoting rewards [28, 168], show promise but remain immature; balancing exploration with efficient convergence is unresolved. Long-horizon credit assignment is especially difficult when rewards arrive only after lengthy reasoning [96], motivating new reward schemes and algorithms.
- 7.1.3 Theoretical Understanding and Reliability. At a theoretical and analytical level, we lack a clear account of generalization and stability in RL-trained LLMs. It remains open whether RL genuinely yields new reasoning abilities or simply amplifies pre-trained patterns [190, 218, 235]. Misconfigured optimization can degrade calibration or core knowledge [83, 84]. We need criteria for when RL helps versus hurts, and techniques, e.g., reward model regularization or conservative updates to curb instability. Deeper theory and lifecycle-wide interpretability studies are limited but essential for safer, more effective RL.
- 7.1.4 Application-Level Challenges of Agentic LLMs. At the application level, integrating LLMs with agentic and tool use via RL presents both exciting opportunities and unresolved difficulties. Recent work has started to treat LLMs as autonomous agents [37, 39, 132, 165] that can plan, act, and interact with external tools or environments to accomplish complex goals. Reinforcement learning is a natural fit for training such agentic LLMs because it provides a feedback loop for trial-and-error learning in interactive scenarios. However, scaling this idea up surfaces challenges in efficiency, safety, and controllability. Training an LLM agent through environment interactions (e.g., simulated tool APIs [94], web browsing[30], or games [96]) is extremely resource-intensive, as it requires running the expensive model many times to explore different action sequences. Ensuring the safety of agentic behavior is even more critical: an RL-driven agent might discover strategies that technically maximize reward while violating user intent or ethical normswang [41, 121, 122, 168]. Unlike constrained single-turn text generation, an autonomous LLM agent could take a sequence of harmful or undesirable actions if its reward function is misspecified [180]. Therefore, developing safe RL techniques for LLM-based agents is an urgent area of research. Additionally, current approaches often lack a memory or planning mechanism to handle very long interaction sequences, making it hard for agents to perform tasks requiring long-term planning or to recover from mistakes [115, 159, 215]. Integrating external memory combined with RL is a promising direction to address long-horizon agency, but it remains largely unexplored. Correspondingly, significant challenges remain with respect to datasets and evaluation benchmarks. Current studies often rely on bespoke datasets or task-specific benchmarks, making it difficult to systematically compare RL fine-tuning methods and to separate genuine improvements from task-specific gains. Although initial attempts such as Polymath [176], Humanity's Last Exam [133] provide more rigorous tests for advanced reasoning, the broader field still lacks standardized, communitywide benchmarks and unified metrics. Developing such resources remains essential for establishing a solid theoretical and empirical foundation for RL-augmented LLMs.

#### 7.2 Future Trends

- 7.2.1 Evolving Learning Paradigms. Looking ahead, we anticipate several key research trends to shape the future of RL-enhanced LLMs. First, there will be a push toward richer and more nuanced reward modeling. Rather than relying solely on outcome-based reward signals, future work is expected to incorporate process-level supervision and intermediate rewards that evaluate the quality of reasoning steps, justification logic, or adherence to constraints throughout the generation process [209, 222]. Such process-oriented rewards would help address the long-horizon credit assignment problem and encourage the model to develop more transparent and verifiable reasoning paths. Second, we foresee tighter integration of RL with structured reasoning paradigms and knowledge representations [117]. By embedding logical or graph-structured inductive biases into the RL process, models may learn reasoning strategies that transfer more robustly to new tasks, as opposed to the relatively unstructured trial-and-error approach currently prevalent. For example, an LLM might use RL to learn how to traverse and update a knowledge graph or to plan sequences of tool calls, thereby gaining a form of systematic reasoning that purely neural approaches struggle with.
- 7.2.2 Expanding Application Frontiers. Finally, the scope of RL applications for LLMs will continue to broaden, driving further innovations. We expect significant growth in multimodal reasoning tasks where LLMs augmented with vision, audio, or other modalities use RL to coordinate between modalities and achieve complex goals [45, 68] as well as in specialized domains like scientific research assistants, formal theorem proving, or decision-support systems. Each of these new domains will bring its own challenges, likely necessitating customized reward functions and safety considerations. The introduction of more comprehensive benchmarks and competitions targeting these scenarios will spur methodological advances by highlighting the limitations of current techniques. Meanwhile, given that agentic systems are inherently well-suited to the RL training paradigm and hold broad application prospects, RL-enhanced Agentic LLMs are undoubtedly an emerging and future technological trend.
- 7.2.3 Toward a Virtuous Research Cycle. In combination, these trends indicate a shift beyond today's relatively conservative fine-tuning toward a paradigm of using RL to train more adaptive, robust, and safe LLMs. The long-term vision is that reinforcement learning, underpinned by strong theoretical insights and practical tools, will enable LLMs to not only align with human values but also continuously improve their reasoning through experience, ultimately inching closer to systems that can learn how to reason in a human-like, self-correcting manner. Each challenge outlined above also represents an opportunity: by overcoming issues of stability, reward design, theoretical understanding, and evaluation, the field can unlock the next wave of progress in large-scale intelligent systems. Each emerging solution, in turn, feeds into a virtuous cycle—better tools and benchmarks lead to more rigorous research, which yields more capable and aligned models, which then require new evaluation standards—pushing the frontier of what RL-enhanced LLMs can achieve. When considered together, these trends signify a shift toward more comprehensive, structured, and diversified RL training for LLMs. Each challenge represents both a limitation and an opportunity: resolving scalability, reward design, theory gaps, application challenges of agentic LLMs, and evaluation will unlock new reasoning frontiers. Crucially, solutions reinforce each other better tools and benchmarks lead to stronger models, which in turn necessitate improved evaluation. This self-reinforcing cycle will drive the evolution of more aligned, generalizable, and safe RL-enhanced LLMs.

#### 8 Conclusion

This survey presents a comprehensive review of reinforcement learning for large language models, organized around the full training lifecycle from pre-training to alignment and reasoning. Particular emphasis is given to RLVR technology,

Manuscript submitted to ACM

which represents a promising direction for incorporating objective and reliable optimization signals. In addition, the survey consolidates datasets, benchmarks, and open-source frameworks, providing a structured reference for both evaluation and practical implementation. By integrating these perspectives, the survey delivers a lifecycle-based synthesis that highlights both methodological advances and supporting resources, serving as a state-of-the-art reference for future research in RL-enhanced LLMs.

### References

- [1] 2025. NeMo RL: A Scalable and Efficient Post-Training Library. https://github.com/NVIDIA-NeMo/RL. GitHub repository.
- [2] Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. arXiv preprint arXiv:2503.04697 (2025).
- [3] Anthropic. 2025. Claude Sonnet 4. https://www.anthropic.com/claude/sonnet
- [4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861 (2021).
- [5] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR. 4447–4455.
- [6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022).
- [7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022).
- [8] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. 2025. Thinking machines: A survey of llm based reasoning strategies. arXiv preprint arXiv:2503.10814 (2025).
- [9] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Nusa Dua, Bali, 675–718. doi:10.18653/v1/2023.ijcnlp-main.45
- [10] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ<sup>2</sup>-Bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv preprint arXiv:2506.07982 (2025).
- [11] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 610–623.
- [12] Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, et al. 2025. Reasoning language models: A blueprint. arXiv preprint arXiv:2501.11223 (2025).
- [13] Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought Anchors: Which LLM Reasoning Steps Matter? arXiv preprint arXiv:2506.19143 (2025).
- [14] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).
- [15] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. 2024. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. IEEE Transactions on Neural Networks and Learning Systems (2024)
- [16] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. Comput. Surveys (2024).
- [17] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. 2025. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. arXiv preprint arXiv:2506.13585 (2025).
- [18] Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Yuefeng Huang, et al. 2025. ACEBench: Who Wins the Match Point in Tool Learning? arXiv e-prints (2025), arXiv-2501.
- [19] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. arXiv preprint arXiv:2504.11468 (2025).
- [20] Jiawei Chen, Dingkang Yang, Yue Jiang, Mingcheng Li, Jinjie Wei, Xiaolu Hou, and Lihua Zhang. 2024. Efficiency in Focus: LayerNorm as a Catalyst for Fine-tuning Medical Visual Language Models. In Proceedings of the 32nd ACM International Conference on Multimedia. 3122–3130.
- [21] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024. Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024).

- [22] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023. TheoremQA: A Theorem-driven Question Answering Dataset. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, 7889–7901. doi:10.18653/v1/2023.emnlp-main.489
- [23] Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. 2025. Rm-r1: Reward modeling as reasoning. arXiv preprint arXiv:2505.02387 (2025).
- [24] Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. 2025. Compile scene graphs with reinforcement learning. arXiv preprint arXiv:2504.13617 (2025).
- [25] Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, et al. 2025.
  Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective. arXiv preprint arXiv:2506.14965 (2025).
- [26] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021).
- [27] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025).
- [28] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. 2025. The entropy mechanism of reinforcement learning for reasoning language models. arXiv preprint arXiv:2505.22617 (2025).
- [29] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. 2025. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. arXiv preprint arXiv:2503.07065 (2025).
- [30] Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025.
  Tool-Star: Empowering LLM-Brained Multi-Tool Reasoner via Reinforcement Learning. arXiv preprint arXiv:2505.16410 (2025).
- [31] Qingxiu Dong, Li Dong, Yao Tang, Tianzhu Ye, Yutao Sun, Zhifang Sui, and Furu Wei. 2025. Reinforcement Pre-Training. arXiv preprint arXiv:2506.08007 (2025).
- [32] Shihan Dou, Yan Liu, Haoxiang Jia, Enyu Zhou, Limao Xiong, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, Zhiheng Xi, Yuhao Zhou, Tao Ji, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. StepCoder: Improving Code Generation with Reinforcement Learning from Compiler Feedback. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, 4571–4585. doi:10.18653/v1/2024.acl-long.251
- [33] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. arXiv preprint arXiv:2502.14739 (2025).
- [34] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In Forty-first International Conference on Machine Learning.
- [35] Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025. SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward. arXiv preprint arXiv:2505.17018 (2025).
- [36] Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: Llm learns when to think. arXiv preprint arXiv:2505.13379 (2025).
- [37] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025. Retool: Reinforcement learning for strategic tool use in llms. arXiv preprint arXiv:2504.11536 (2025).
- [38] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025).
- [39] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. Group-in-group policy optimization for llm agent training. arXiv preprint arXiv:2505.10978 (2025).
- [40] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt. Advances in neural information processing systems 36 (2023), 27699–27744.
- [41] Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. 2025. Reward shaping to mitigate reward hacking in rlhf. arXiv preprint arXiv:2502.18770 (2025).
- [42] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, et al. 2025. AReaL: A Large-Scale Asynchronous Reinforcement Learning System for Language Reasoning. arXiv preprint arXiv:2505.24298 (2025).
- [43] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 3356–3369. doi:10.18653/v1/2020.findings-emnlp.301
- [44] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are We Done with MMLU?. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Albuquerque, New Mexico, 5069–5096. doi:10.18653/v1/2025.naacl-long.262
- [45] Dibya Ghosh and Sergey Levine. 2025. Visual Pre-Training on Unlabeled Images using Reinforcement Learning. arXiv preprint arXiv:2506.11967 (2025).

[46] Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D Manning. 2025. Synthetic data generation & multi-step rl for reasoning & tool use. arXiv preprint arXiv:2504.04736 (2025).

- [47] Yandong Guan, Xilin Wang, Xingxi Ming, Jing Zhang, Dong Xu, and Qian Yu. 2025. CAD-Coder: Text-to-CAD Generation with Chain-of-Thought and Geometric Reward. arXiv preprint arXiv:2505.19713 (2025).
- [48] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025).
- [49] Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Reward reasoning model. arXiv preprint arXiv:2505.14674 (2025).
- [50] Yiduo Guo, Zhen Guo, Chuanwei Huang, Zi-Ang Wang, Zekai Zhang, Haofei Yu, Huishuai Zhang, and Yikang Shen. 2025. Synthetic Data RL: Task Definition Is All You Need. arXiv preprint arXiv:2505.17063 (2025).
- [51] Qianyue Hao, Sibo Li, Jian Yuan, and Yong Li. 2025. Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning. arXiv preprint arXiv:2505.14140 (2025).
- [52] Hado Hasselt. 2010. Double Q-learning. Advances in neural information processing systems 23 (2010).
- [53] Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. 2023. trlX: A framework for large scale reinforcement learning from human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 8578–8595.
- [54] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, 3828–3850. doi:10.18653/v1/2024.acl-long.211
- [55] Shenghua He, Tian Xia, Xuan Zhou, and Hui Wei. 2025. Response-Level Rewards Are All You Need for Online Reinforcement Learning in LLMs: A Mathematical Perspective. arXiv preprint arXiv:2506.02553 (2025).
- [56] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021. Measuring Coding Challenge Competence With APPS. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper\_files/paper/2021/file/c24cd76e1ce41366a4bbe8a49b02a028-Paper-round2.pdf
- [57] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In ICLR. OpenReview.net.
- [58] Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, 11170–11189. doi:10.18653/v1/2024.emnlp-main.626
- [59] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. 2025. GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv preprint arXiv:2507.01006 (2025).
- [60] Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. TreeRL: LLM Reinforcement Learning with On-Policy Tree Search. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, 12355–12369. doi:10.18653/v1/2025.acl-long.604
- [61] Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. arXiv preprint arXiv:2501.03262 (2025).
- [62] Jian Hu, Xibin Wu, Zilin Zhu, Weixun Wang, Dehao Zhang, Yu Cao, et al. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. arXiv preprint arXiv:2405.11143 (2024).
- [63] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290 (2025).
- [64] Yang Hu, Xingyu Zhang, Xueji Fang, Zhiyang Chen, Xiao Wang, Huatian Zhang, and Guojun Qi. 2025. SLOT: Sample-specific Language Model Optimization at Test-time. arXiv preprint arXiv:2505.12392 (2025).
- [65] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Representation Learning*, Vol. 2024. 32808–32824. https://proceedings.iclr.cc/paper\_files/paper/2024/file/8b4add8b0aa8749d80a34ca5d941c355-Paper-Conference.pdf
- [66] Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. 2025. Boosting mllm reasoning with text-debiased hint-grpo. arXiv preprint arXiv:2503.23905 (2025).
- [67] Ting Huang, Zeyu Zhang, and Hao Tang. 2025. 3D-R1: Enhancing Reasoning in 3D VLMs for Unified Scene Understanding. arXiv preprint arXiv:2507.23478 (2025).
- [68] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025).
- [69] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024).
- [70] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024).

- [71] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. arXiv preprint arXiv:2403.07974 (2024).
- [72] Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. 2025. A survey on progress in llm alignment from the perspective of reward design. arXiv preprint arXiv:2505.02666 (2025).
- [73] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025).
- [74] Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025. Think only when you need with large hybrid-reasoning models. arXiv preprint arXiv:2505.14631 (2025).
- [75] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024. A survey on human preference learning for large language models. arXiv preprint arXiv:2406.11191 (2024).
- [76] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv:2503.09516 (2025).
- [77] Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. arXiv preprint arXiv:2502.18581 (2025).
- [78] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925 [cs.LG] https://arxiv.org/abs/2312.14925
- [79] Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. 2025. Big-bench extra hard. arXiv preprint arXiv:2502.19187 (2025).
- [80] Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, et al. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. arXiv preprint arXiv:2504.09037 (2025).
- [81] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. arXiv preprint arXiv:2103.14659 (2021).
- [82] Vijay Konda and John Tsitsiklis. 1999. Actor-critic algorithms. Advances in neural information processing systems 12 (1999).
- [83] Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. Advances in Neural Information Processing Systems 35 (2022), 16203–16220.
- [84] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *ICLR*. OpenReview.net.
- [85] 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. 2025. Llm post-training: A deep dive into reasoning large language models. arXiv preprint arXiv:2502.21321 (2025)
- [86] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward Design with Language Models. In ICLR. OpenReview.net.
- [87] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124 (2024).
- [88] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. (2023).
- [89] Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. 2025. Large Vision-Language Models as Emotion Recognizers in Context Awareness. In Asian Conference on Machine Learning. PMLR, 111–126.
- [90] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. Advances in neural information processing systems 35 (2022), 3843–3857.
- [91] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. 2025. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. arXiv preprint arXiv:2503.16188 (2025).
- [92] Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. 2025. Confidence Is All You Need: Few-Shot RL Fine-Tuning of Language Models. arXiv preprint arXiv:2506.06395 (2025).
- [93] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. arXiv preprint arXiv:2406.11939 (2024).
- [94] Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Torl: Scaling tool-integrated rl. arXiv preprint arXiv:2503.23383 (2025).
- [95] Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. 2025. Generalist Reward Models: Found Inside Large Language Models. arXiv preprint arXiv:2506.23235 (2025).
- [96] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, Ser-Nam Lim, and Hengshuang Zhao. 2025. LARM: Large Auto-Regressive Model for Long-Horizon Embodied Intelligence. In Forty-second International Conference on Machine Learning. https://openreview.net/forum?id=zcx7jqUZg5
- [97] Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. 2018. RLlib: Abstractions for distributed reinforcement learning. In *International conference on machine learning*. PMLR, 3053–3062.
- [98] Guannan Liang and Qianqian Tong. 2025. LLM-Powered AI Agent Systems and Their Applications in Industry. arXiv preprint arXiv:2505.16120

[99] Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yang Wang, Hengyuan Zhang, Yelong Shen, Ying Nian Wu, and Weizhu Chen. 2025. SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2506.08989 (2025).

- [100] Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. 2025. Improved visual-spatial reasoning via r1-zero-like training. arXiv preprint arXiv:2504.00883 (2025).
- [101] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. In Forty-second International Conference on Machine Learning. https://openreview.net/forum?id=sTAJ9QyA61
- [102] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chengdang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024).
- [103] Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. 2025. SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn Reinforcement Learning. arXiv preprint arXiv:2506.24119 (2025).
- [104] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models. arXiv preprint arXiv:2505.21523 (2025).
- [105] 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. Advances in Neural Information Processing Systems 36 (2023), 21558–21572.
- [106] Liyuan Liu, Feng Yao, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. 2025. FlashRL: 8Bit Rollouts, Full Power RL. https://fengyao.notion.site/flash-rl
- [107] Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. 2025. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. arXiv preprint arXiv:2505.24864 (2025).
- [108] Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. 2025.
  X-reasoner: Towards generalizable reasoning across modalities and domains. arXiv preprint arXiv:2505.03981 (2025).
- [109] Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024. AlignBench: Benchmarking Chinese Alignment of Large Language Models. Association for Computational Linguistics, Bangkok, Thailand, 11621–11640. doi:10.18653/v1/2024.acl-long.624
- [110] Yihao Liu, Shuocheng Li, Lang Cao, Yuhang Xie, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. 2025. SuperRL: Reinforcement Learning with Supervision to Boost Language Model Reasoning. arXiv preprint arXiv:2506.01096 (2025).
- [111] Yizhou Liu, Jingwei Wei, Zizhi Chen, Minghao Han, Xukun Zhang, Keliang Liu, and Lihua Zhang. 2025. Breaking Reward Collapse: Adaptive Reinforcement for Open-ended Medical Reasoning with Enhanced Semantic Discrimination. arXiv preprint arXiv:2508.12957 (2025).
- [112] Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. 2025. There may not be an amoment in r1-zero-like training—a pilot study.
- [113] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025).
- [114] Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling. arXiv preprint arXiv:2504.02495 (2025).
- [115] Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2025. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory. arXiv preprint arXiv:2508.09736 (2025).
- [116] Chenwei Lou, Zewei Sun, Xinnian Liang, Meng Qu, Wei Shen, Wenqi Wang, Yuntao Li, Qingping Yang, and Shuangzhi Wu. 2025. AdaCoT: Pareto-Optimal Adaptive Chain-of-Thought Triggering via Reinforcement Learning. arXiv preprint arXiv:2505.11896 (2025).
- [117] Haoran Luo, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, Luu Anh Tuan, et al. 2025. Graph-R1: Towards Agentic GraphRAG Framework via End-to-end Reinforcement Learning. arXiv preprint arXiv:2507.21892 (2025).
- [118] Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. arXiv preprint arXiv:2504.09858 (2025).
- [119] Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. 2025. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. arXiv preprint arXiv:2503.12797 (2025).
- [120] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models. arXiv preprint arXiv:2310.12931 (2023).
- [121] Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. Advances in Neural Information Processing Systems 37 (2024), 134387–134429.
- [122] Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. 2025. The energy loss phenomenon in rlhf: A new perspective on mitigating reward hacking. arXiv preprint arXiv:2501.19358 (2025).
- [123] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In ICLR. OpenReview.net.
- [124] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. nature 518, 7540 (2015), 529–533.
- [125] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393 (2025).
- [126] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

- [127] OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical Report. OpenAI. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf System Card officially released by OpenAI on April 16, 2025.
- [128] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. 2025. SpaceR: Reinforcing MLLMs in Video Spatial Reasoning. arXiv preprint arXiv:2504.01805 (2025).
- [129] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems 35 (2022), 27730–27744.
- [130] Samuel J Paech. 2025. EQ-Bench Creative Writing Benchmark v3. https://github.com/EQ-bench/creative-writing-bench.
- [131] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. arXiv preprint arXiv:2402.13228 (2024).
- [132] Feng Peiyuan, Yichen He, Guanhua Huang, Yuan Lin, Hanchong Zhang, Yuchen Zhang, and Hang Li. 2024. Agile: A novel reinforcement learning framework of llm agents. Advances in Neural Information Processing Systems 37 (2024), 5244–5284.
- [133] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. arXiv preprint arXiv:2501.14249 (2025).
- [134] Moschoula Pternea, Prerna Singh, Abir Chakraborty, Yagna Oruganti, Mirco Milletari, Sayli Bapat, and Kebei Jiang. 2024. The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models. Journal of Artificial Intelligence Research 80 (2024), 1525–1573.
- [135] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. 2025. VLN-R1: Vision-Language Navigation via Reinforcement Fine-Tuning. arXiv preprint arXiv:2506.17221 (2025).
- [136] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jianxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115
- [137] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems 36 (2023), 53728–53741.
- [138] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is Reinforcement Learning (Not) for Natural Language Processing: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. In ICLR. OpenReview.net.
- [139] Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. 2025. Magistral. arXiv preprint arXiv:2506.10910 (2025).
- [140] 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 First Conference on Language Modeling.
- [141] Gavin A Rummery and Mahesan Niranjan. 1994. On-line Q-learning using connectionist systems. Vol. 37. University of Cambridge, Department of Engineering Cambridge, UK.
- [142] Soumya Rani Samineni, Durgesh Kalwar, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. 2025. RL in Name Only? Analyzing the Structural Assumptions in RL post-training for LLMs. arXiv preprint arXiv:2505.13697 (2025).
- [143] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In International conference on machine learning. PMLR, 1889–1897.
- [144] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
- [145] Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, et al. 2025. Rethinking reflection in pre-training. arXiv preprint arXiv:2504.04022 (2025).
- [146] Chenyang Shao, Xinyang Liu, Yutang Lin, Fengli Xu, and Yong Li. 2025. Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router. arXiv preprint arXiv:2506.05901 (2025).
- [147] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024).
- [148] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025).
- [149] Hui Shen, Taiqiang Wu, Qi Han, Yunta Hsieh, Jizhou Wang, Yuyue Zhang, Yuxin Cheng, Zijian Hao, Yuansheng Ni, Xin Wang, et al. 2025. PhyX: Does Your Model Have the "Wits" for Physical Reasoning? arXiv preprint arXiv:2505.15929 (2025).
- [150] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In Proceedings of the Twentieth European Conference on Computer Systems. 1279–1297.
- [151] Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2506.06941 (2025).
- [152] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. arXiv preprint

- arXiv:2501.17399 (2025).
- [153] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on machine learning research (2023).
- [154] Saksham Sahai Srivastava and Vaneet Aggarwal. 2025. A Technical Survey of Reinforcement Learning Techniques for Large Language Models. arXiv preprint arXiv:2507.04136 (2025).
- [155] Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. 2025.
  Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025).
- [156] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems 12 (1999).
- [157] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. arXiv preprint arXiv:2401.04056 (2024).
- [158] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503 (2021).
- [159] Zhen Tan, Jun Yan, I Hsu, Rujun Han, Zifeng Wang, Long T Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. arXiv preprint arXiv:2503.08026 (2025).
- [160] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599 (2025).
- [161] Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-R1: Chain-of-Tool-Thought for Ultra-Long Egocentric Video Reasoning. arXiv preprint arXiv:2506.13654 (2025).
- [162] Carel van Niekerk, Renato Vukovic, Benjamin Matthias Ruppik, Hsien chin Lin, and Milica Gašić. 2025. Post-Training Large Language Models via Reinforcement Learning from Self-Feedback. arXiv:2507.21931 [cs.CL] https://arxiv.org/abs/2507.21931
- [163] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl.
- [164] Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, et al. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. arXiv preprint arXiv:2506.01713 (2025).
- [165] Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. 2025. SPA-RL: Reinforcing LLM Agents via Stepwise Progress Attribution. arXiv preprint arXiv:2505.20732 (2025).
- [166] Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, et al. 2025. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning. arXiv preprint arXiv:2504.11354 (2025).
- [167] Jiacong Wang, Zijiang Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. 2025. VGR: Visual Grounded Reasoning. arXiv preprint arXiv:2506.11991 (2025).
- [168] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. arXiv preprint arXiv:2506.01939 (2025).
- [169] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. 2024. Reinforcement learning enhanced llms: A survey. arXiv preprint arXiv:2412.10400 (2024).
- [170] Tevin Wang and Chenyan Xiong. 2025. AutoRule: Reasoning Chain-of-thought Extracted Rule-based Rewards Improve Preference Learning. arXiv preprint arXiv:2506.15651 (2025).
- [171] Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. 2025.
  Reinforcement Learning Optimization for Large-Scale Learning: An Efficient and User-Friendly Scaling Library. arXiv preprint arXiv:2506.06122 (2025).
- [172] Xubin Wang, Jianfei Wu, Yichen Yuan, Deyu Cai, Mingzhe Li, and Weijia Jia. 2024. Demonstration selection for in-context learning via reinforcement learning. arXiv preprint arXiv:2412.03966 (2024).
- [173] Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, et al. 2025. ViCrit: A Verifiable Reinforcement Learning Proxy Task for Visual Perception in VLMs. arXiv preprint arXiv:2506.10128 (2025).
- [174] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. arXiv preprint arXiv:2505.03318 (2025).
- [175] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems 37 (2024), 95266–95290.
- [176] Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. 2025.
  Polymath: Evaluating mathematical reasoning in multilingual contexts. arXiv preprint arXiv:2504.18428 (2025).
- [177] Yunhao Wang, Yuhao Zhang, Tinghao Yu, Can Xu, Feng Zhang, and Fengzong Lian. 2025. Adaptive Deep Reasoning: Triggering Deep Thinking When Needed. arXiv preprint arXiv:2505.20101 (2025).
- [178] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. arXiv preprint arXiv:2407.16216 (2024).

- [179] Zhexu Wang, Yiping Liu, Yejie Wang, Wenyang He, Bofei Gao, Muxi Diao, Yanxu Chen, Kelin Fu, Flood Sung, Zhilin Yang, et al. 2025. OJBench: A Competition Level Code Benchmark For Large Language Models. arXiv preprint arXiv:2506.16395 (2025).
- [180] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. 2025. Ragen: Understanding self-evolution in Ilm agents via multi-turn reinforcement learning. arXiv preprint arXiv:2504.20073 (2025).
- [181] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025. Octothinker: Mid-training incentivizes reinforcement learning scaling. arXiv preprint arXiv:2506.20512 (2025)
- [182] Zhixin Wang, Tianyi Zhou, Liming Liu, Ao Li, Jiarui Hu, Dian Yang, Jinlong Hou, Siyuan Feng, Yuan Cheng, and Yuan Qi. 2025. DistFlow: A Fully Distributed RL Framework for Scalable and Efficient LLM Post-Training. arXiv preprint arXiv:2507.13833 (2025).
- [183] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. Machine learning 8, 3 (1992), 279-292.
- [184] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. arXiv preprint arXiv:2411.04368 (2024).
- [185] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 (2021).
- [186] Colin White et al. 2025. LiveBench: A Challenging, Contamination-Free LLM Benchmark. In *The Thirteenth International Conference on Learning Representations*.
- [187] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.
- [188] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 38–45.
- [189] Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, et al. 2025.
  Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm trainin. arXiv preprint arXiv:2505.24034 (2025).
- [190] Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. 2025. The Invisible Leash: Why RLVR May Not Escape Its Origin. arXiv preprint arXiv:2507.14843 (2025).
- [191] Haoyuan Wu, Xueyi Chen, Rui Ming, Jilong Gao, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. ToTRL: Unlock LLM Tree-of-Thoughts Reasoning Potential through Puzzles Solving. arXiv preprint arXiv:2505.12717 (2025).
- [192] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. \[ \beta-DPO: Direct Preference Optimization with Dynamic \beta. In \[ Advances in Neural Information Processing Systems, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 129944–129966. \[ \https://proceedings.neurips.cc/paper\_files/paper/2024/file/ea888178abdb6fc233226d12321d754f-Paper-Conference.pdf
- [193] Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Yanwei Fu, Qin Liu, et al. 2025. Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination. arXiv preprint arXiv:2507.10532 (2025).
- [194] xAI. 2025. Grok 3 Beta The Age of Reasoning Agents. https://x.ai/news/grok-3. Accessed: 2025-08-25.
- [195] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. 2024. Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. In ICLR. OpenReview.net.
- [196] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2023. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. arXiv preprint arXiv:2312.11456 (2023).
- [197] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. arXiv preprint arXiv:2501.09686 (2025).
- [198] Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025. KDRL: Post-Training Reasoning LLMs via Unified Knowledge Distillation and Reinforcement Learning. arXiv preprint arXiv:2506.02208 (2025).
- [199] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. 2025.
  Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. arXiv preprint arXiv:2504.15279 (2025).
- [200] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. 2025. Visual Planning: Let's Think Only with Images. arXiv preprint arXiv:2505.11409 (2025).
- [201] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. 2025. DanceGRPO: Unleashing GRPO on Visual Generation. arXiv preprint arXiv:2505.07818 (2025).
- [202] 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. arXiv preprint arXiv:2504.14945 (2025).
- [203] Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. arXiv preprint arXiv:2508.19828 (2025).
- [204] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025.
  Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025).
- [205] Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. 2024.
  Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. Advances in Neural Information Processing Systems 37

- (2024), 138632-138662,
- [206] Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2025. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39. 25606–25614.
- [207] John Yang, Kilian Leret, Carlos E Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. 2025. Swe-smith: Scaling data for software engineering agents. arXiv preprint arXiv:2504.21798 (2025).
- [208] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025.
  R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025).
- [209] Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. 2025. TreeRPO: Tree Relative Policy Optimization. arXiv preprint arXiv:2506.05183 (2025).
- [210] Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. 2025. Your Efficient RL Framework Secretly Brings You Off-Policy RL Training. https://fengyao.notion.site/off-policy-rl
- [211] Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, et al. 2025. MMReason: An Open-Ended Multi-Modal Multi-Step Reasoning Benchmark for MLLMs Toward AGL. arXiv preprint arXiv:2506.23563 (2025).
- [212] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. arXiv preprint arXiv:2308.01320 (2023).
- [213] Jingyang Yi, Jiazheng Wang, and Sida Li. 2025. Shorterbetter: Guiding reasoning models to find optimal inference length for efficient reasoning. arXiv preprint arXiv:2504.21370 (2025).
- [214] Yang You. 2023. "Colossalchat: An open-source solution for cloning chatgpt with a complete rlhf pipeline.
- [215] Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. 2025. MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent. arXiv preprint arXiv:2507.02259 (2025).
- [216] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025).
- [217] Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. 2025.
  RewardAnything: Generalizable Principle-Following Reward Models. arXiv preprint arXiv:2506.03637 (2025).
- [218] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? arXiv preprint arXiv:2504.13837 (2025).
- [219] Kaiwen Zha, Zhengqi Gao, Maohao Shen, Zhang-Wei Hong, Duane S Boning, and Dina Katabi. 2025. RL Tango: Reinforcing Generator and Verifier Together for Language Reasoning. arXiv preprint arXiv:2505.15034 (2025).
- [220] Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, et al. 2025. GThinker: Towards General Multimodal Reasoning via Cue-Guided Rethinking. arXiv preprint arXiv:2506.01078 (2025).
- [221] Junyu Zhang, Runpei Dong, Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai, Jitendra Malik, Saurabh Gupta, et al. 2025. AlphaOne: Reasoning Models Thinking Slow and Fast at Test Time. arXiv preprint arXiv:2505.24863 (2025).
- [222] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937 (2025).
- [223] Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025. Adaptthink: Reasoning models can learn when to think. arXiv preprint arXiv:2505.13417 (2025).
- [224] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. 2025. A Survey of Reinforcement Learning for Large Reasoning Models. arXiv preprint arXiv:2509.08827 (2025).
- [225] Xiaoyun Zhang, Jingqing Ruan, Xing Ma, Yawen Zhu, Haodong Zhao, Hao Li, Jiansong Chen, Ke Zeng, and Xunliang Cai. 2025. When to continue thinking: Adaptive thinking mode switching for efficient reasoning. arXiv preprint arXiv:2505.15400 (2025).
- [226] Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. 2025. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. arXiv preprint arXiv:2504.14286 (2025).
- [227] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. Tinyllava-video-r1: Towards smaller lmms for video reasoning. arXiv preprint arXiv:2504.09641 (2025).
- [228] Xiaoqing Zhang, Huabin Zheng, Ang Lv, Yuhan Liu, Zirui Song, Xiuying Chen, Rui Yan, and Flood Sung. 2025. Divide-Fuse-Conquer: Eliciting" Aha Moments" in Multi-Scenario Games. arXiv preprint arXiv:2505.16401 (2025).
- [229] Yudi Zhang, Lu Wang, Meng Fang, Yali Du, Chenghua Huang, Jun Wang, Qingwei Lin, Mykola Pechenizkiy, Dongmei Zhang, Saravan Rajmohan, et al. 2025. Distill Not Only Data but Also Rewards: Can Smaller Language Models Surpass Larger Ones? arXiv preprint arXiv:2502.19557 (2025).
- [230] Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, Kun Qian, Yuxin Tang, Ran Xue, Houyu Zhang, Qingjun Cui, Yufan Guo, and Dakuo Wang. 2025. Shop-R1: Rewarding LLMs to Simulate Human Behavior in Online Shopping via Reinforcement Learning. arXiv:2507.17842 [cs.CL] https://arxiv.org/abs/2507.17842
- [231] Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025. No Free Lunch: Rethinking Internal Feedback for LLM Reasoning. arXiv preprint arXiv:2506.17219 (2025).
- [232] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. Absolute zero: Reinforced self-play reasoning with zero data. arXiv preprint arXiv:2505.03335 (2025).

- [233] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. 2025. Embodied-R: Collaborative Framework for Activating Embodied Spatial Reasoning in Foundation Models via Reinforcement Learning. arXiv preprint arXiv:2504.12680 (2025).
- [234] Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian, Biqing Qi, Xiu Li, et al. 2025. Genprm: Scaling test-time compute of process reward models via generative reasoning. arXiv preprint arXiv:2504.00891 (2025).
- [235] Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. arXiv preprint arXiv:2504.07912 (2025).
- [236] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025. Learning to reason without external rewards. arXiv preprint arXiv:2505.19590 (2025).
- [237] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. 2025. Group Sequence Policy Optimization. arXiv preprint arXiv:2507.18071 (2025).
- [238] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. arXiv preprint arXiv:2505.14362 (2025).
- [239] Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. arXiv preprint arXiv:2504.12328 (2025).
- [240] 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 (2025).
- [241] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. arXiv preprint arXiv:2503.05132 (2025).
- [242] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911 (2023).
- [243] Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. 2025. Reinforcing General Reasoning without Verifiers. arXiv preprint arXiv:2505.21493 (2025).
- [244] Jason Zhu and Hongyu Li. 2025. Towards Concise and Adaptive Thinking in Large Reasoning Models: A Survey. arXiv preprint arXiv:2507.09662 (2025).
- [245] Ke Zhu, Yu Wang, Jiangjiang Liu, Qunyi Xie, Shanshan Liu, and Gang Zhang. 2025. On Data Synthesis and Post-training for Visual Abstract Reasoning. arXiv preprint arXiv:2504.01324 (2025).
- [246] Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. 2025. VAU-R1: Advancing Video Anomaly Understanding via Reinforcement Fine-Tuning. arXiv preprint arXiv:2505.23504 (2025).
- [247] 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 (2025).
- [248] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in LLM reasoning. arXiv preprint arXiv:2506.01347 (2025).
- [249] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. 2025. Ttrl: Test-time reinforcement learning. arXiv preprint arXiv:2504.16084 (2025).