A Survey of Reinforcement Learning for Large Language Models under Data Scarcity: Challenges and Solutions

Zhiyin Yu[1,2], **Yuchen Mou**[3], **Juncheng Yan**[4], **Junyu Luo**[1], **Chunchun Chen**[5], **Xing Wei**[5], **Yunhui Liu**[6], **Hongru Sun**[7], **Yuxing Zhang**[8], **Jun Xu**[4], **Yatao Bian**[3], **Ming Zhang**[1], **Wei Ye**[5], **Tieke He**[6], **Jie Yang**[7], **Guanjie Zheng**[8], **Zhonghai Wu**[1 †], **Bo Zhang**[2 †], **Lei Bai**[2], **Xiao Luo**[9]
[1]Peking University [2]Shanghai Artificial Intelligence Laboratory [3]National University of Singapore
[4]Nankai University [5]Tongji University [6]Nanjing University [7]University of Wollongong
[8]Shanghai Jiao Tong University [9]University of Wisconsin-Madison
zhiyinyu25@stu.pku.edu.cn, zhangbo@pjlab.org.cn, wuzh@pku.edu.cn, xiao.luo@wisc.edu

Abstract

Reinforcement learning (RL) has emerged as a powerful post-training paradigm for enhancing the reasoning capabilities of large language models (LLMs). However, reinforcement learning for LLMs faces substantial data scarcity challenges, including the limited availability of high-quality external supervision and the constrained volume of model-generated experience. These limitations make data-efficient reinforcement learning a critical research direction. In this survey, we present the first systematic review of reinforcement learning for LLMs under data scarcity. We propose a bottom-up hierarchical framework built around three complementary perspectives: the data-centric perspective, the training-centric perspective, and the framework-centric perspective. We develop a taxonomy of existing methods, summarize representative approaches in each category, and analyze their strengths and limitations. Our taxonomy aims to provide a clear conceptual foundation for understanding the design space of data-efficient RL for LLMs and to guide researchers working in this emerging area. We hope this survey offers a comprehensive roadmap for future research and inspires new directions toward more efficient and scalable reinforcement learning post-training for LLMs.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various domains, including mathematical reasoning (Cui et al., 2025; Guan et al., 2025), algorithmic programming (Li et al., 2024; Guo et al., 2024) and scientific research (Team et al., 2025a). As a prevailing and promising paradigm for post-training, reinforcement learning (RL) has shown strong potential to further enhance the reasoning abilities of LLMs.

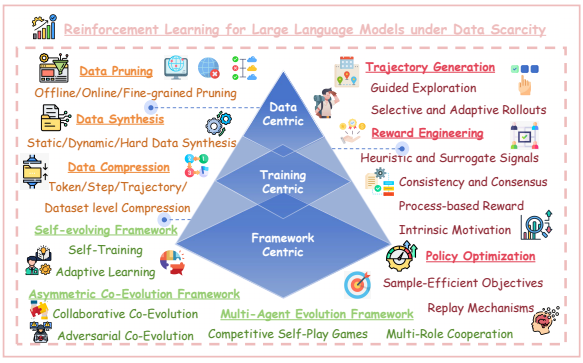

Figure 1: Overview of LLM-based reinforcement learning under data scarcity, illustrating data-, training-, and framework-centric perspectives.

Recent advancements such as DeepSeek-R1 (Guo et al., 2025) and OpenAI-o1 (OpenAI, 2024) indicate that RL-based post-training can elicit emergent behaviors including self-reflection (Zeng et al., 2025b), enabling LLMs to complete complex tasks (Team et al., 2025b; Hu et al., 2025).

However, data scarcity has become a critical bottleneck constraining effective RL for LLMs. This challenge manifests in two complementary forms. Specifically, external data scarcity refers to the limited availability of high-cost supervised signals, such as fine-grained human feedback (Wu et al., 2023a), preference data, expert annotations and step-by-step reasoning traces (Xia et al., 2025). By contrast, internal data scarcity arises from constraints on model-generated interactions, including the number of rollouts, trajectory lengths, and exploration budgets. As noted by Jones (2024), the AI revolution is running out of data. Simply scaling training with more data or computational resources often yields diminishing returns and may still fail to produce strong performance (Zuo et al., 2025). Recently, Silver and Sutton (2025) emphasized a shift toward "the era of experience", advocating a transition from heavy reliance on human supervision to approaches that allow models to evolve through experience. This perspective highlights

† Corresponding authors.
https://github.com/YuZhiyin/Data-Efficient-RL

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22823–22846
July 2-7, 2026

the importance of studying reinforcement learning for LLMs under data scarcity. This paper provides the **first systematic survey** on RL for LLMs under data scarcity, offering a unified framework that organizes the fragmented research landscape.

Existing research has explored various approaches to unlocking the potential of reinforcement learning for LLMs under data scarcity, including improving the utilization of limited human supervision (Fang et al., 2025b) and enhancing the efficiency of model-generated experience (Wang et al., 2025f). However, this field still lacks a systematic and unified survey. To bridge this gap, we review reinforcement learning under data scarcity from three complementary perspectives. Specifically, we introduce a conceptual framework consisting of data-centric, training-centric, and framework-centric perspectives (as illustrated in Figure 1). Based on this framework, we design a taxonomy to categorize existing methods, summarize core techniques, and identify promising directions for future exploration. We hope this survey can serve as a roadmap for researchers and inspire continued progress in data-efficient RL for LLMs.

Differences from Previous Surveys. Recently, several surveys have focused on related but orthogonal themes, including LLM and agentic RL (Zhang et al., 2025c,a), self-evolving agents (Tao et al., 2024; Fang et al., 2025a; ang Gao et al., 2026), and data-efficient post-training (Luo et al., 2025b). These surveys provide valuable insights but do not systematically study RL for LLMs under data scarcity. We address this gap with a bottom-up hierarchical framework and a unified analysis of existing methods through the lens of data scarcity.

2 Taxonomy

In this section, we present a taxonomy that categorizes reinforcement learning under data scarcity from three perspectives, as shown in Figure 2.

- ***Level 1: Data-Centric Perspective:*** *Optimizing the data itself to maximize usable information before, during, and after RL.* ❶ Data Pruning: Identifying the most informative subset from raw training data. ❷ Data Synthesis: Expanding or enriching the effective training distribution. ❸ Data Compression: Compressing tokens, reasoning steps, trajectories, or entire datasets while preserving essential information density.

- ***Level 2: Training-Centric Perspective:*** *Improving how RL generates trajectories, evaluates rewards, and updates policies under scarce data.* ❶ Trajectory Generation: Improving exploration efficiency to reduce wasted rollouts. ❷ Reward Engineering: Enhancing the quality and granularity of feedback signals when annotations are limited. ❸ Policy Optimization: Increasing update efficiency and stability under constrained trajectory budgets.

- ***Level 3: Framework-Centric Perspective:*** *Designing evolving RL frameworks that reduce dependence on external data.* ❶ Self-Evolving Frameworks: Single-model systems that continually refine themselves through self-generated data and self-rewarding mechanisms. ❷ Asymmetric Co-Evolution Frameworks: Dual-agent architectures where agents optimize cooperative or adversarial objectives. ❸ Multi-Agent Evolution Frameworks: Systems of interacting agents that produce multi-objective training signals through cooperative or competitive dynamics.

These three perspectives collectively form a comprehensive and structured approach to RL under data scarcity, moving from curating the data, to improving the efficiency of trajectory usage, and ultimately to constructing autonomous frameworks capable of continual self-improvement.

3 Data-Centric Perspective

3.1 Data Pruning

Data pruning shrinks the effective training pool and accelerates training by selectively retaining informative samples. As shown in Figure 3, we classify existing data pruning methods as: offline pruning, online pruning, and fine-grained pruning.

Offline Pruning. Heuristic pruning has been widely adopted across various LLM training paradigms (Stiennon et al., 2020; Rae et al., 2021; Touvron et al., 2023; Wang et al., 2023), where predefined rules are applied during data preprocessing from multiple perspectives, making it a critical component of data curation (Team et al., 2025b; Guo et al., 2025). LIMR (Li et al., 2025c) identifies high-impact prompts by measuring the alignment between each sample reward trajectory and the average learning curve. LearnAlign (Li et al., 2025b) uses a learnability-weighted gradient-alignment score to select reasoning samples that are both learnable and representative. EAS (Zhu et al., 2025) selects samples by integrating token-level predictive entropy along the generation trajectory.

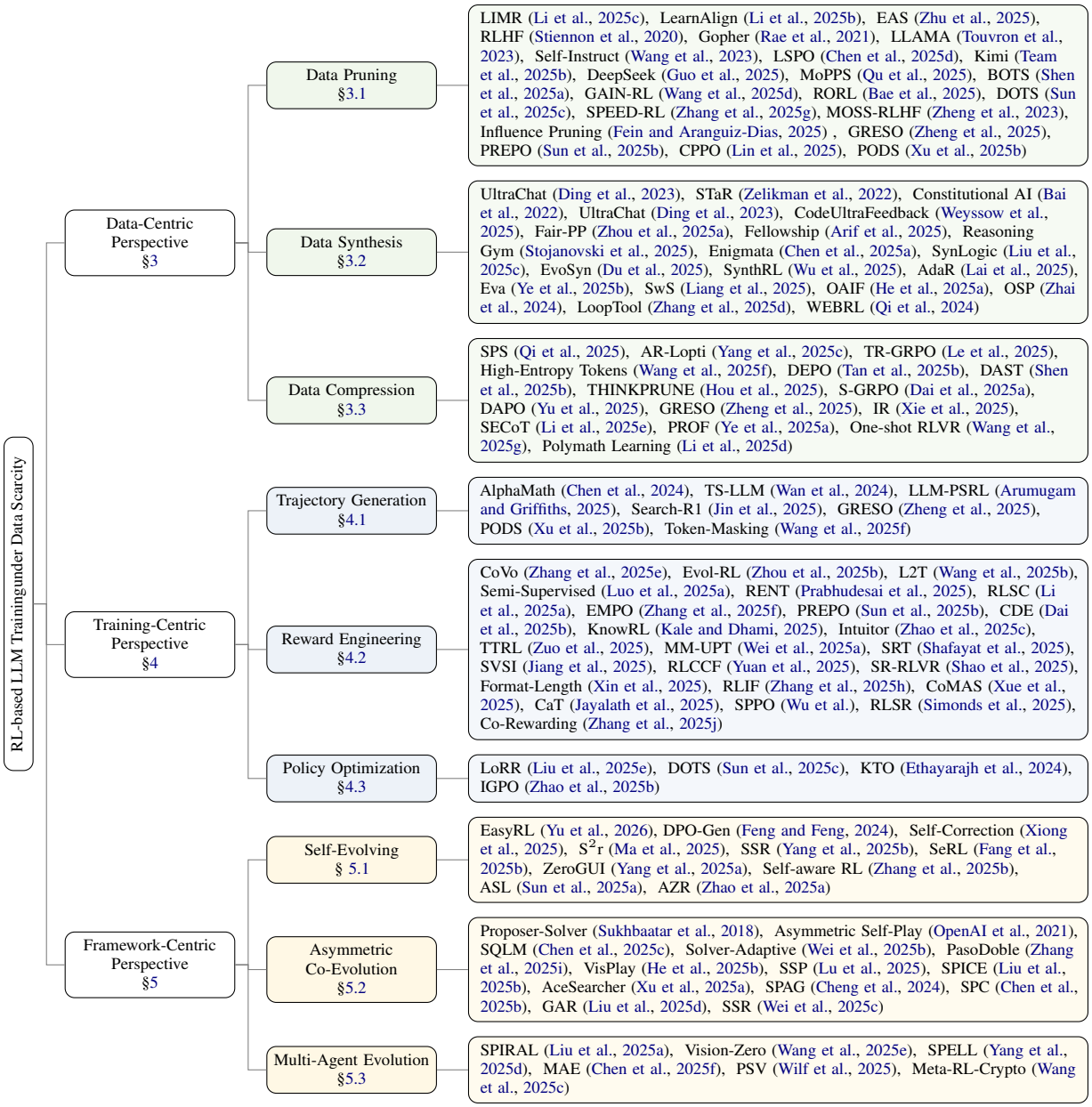

Figure 2: A taxonomy of RL-based LLM Training under Data Scarcity.

Online Pruning. Online pruning dynamically selects high-information-density samples for rollout. LSPO (Chen et al., 2025d) augments accuracy-based filtering with an additional length-based step. MMoPPS (Qu et al., 2025) estimates prompt difficulty as a latent success probability for adaptive prompt selection. BOTS (Shen et al., 2025a) further integrates explicit feedback from selected samples and implicit evidence from unselected samples to balance exploration and exploitation. GAIN-RL (Wang et al., 2025d) predicts the magnitude of gradient updates using the angular concentration of pre-filling hidden states and schedules an easy-to-hard curriculum. RORL (Bae et al., 2025) estimates pass rates from online rollouts to identify prompts of moderate difficulty, thereby focusing updates on samples with higher pass rate variance. DOTS (Sun et al., 2025c) uses attention-based similarity to propagate difficulty scores from a reference subset to the full training pool. SPEED-RL (Zhang et al., 2025g) allocates training budget to moderately difficult samples by estimating empirical pass rates on a few samples. Beyond explicit sample selection, Zheng et al. (2023) investigates implicit data filtering strategies in PPO training, including reward clipping, response deduplication, and advantage normalization.

Fine-grained Pruning. Fine-grained pruning reframes the problem toward precise pruning criteria and focuses selection on trajectories that are valuable to retain. Influence Pruning (Fein

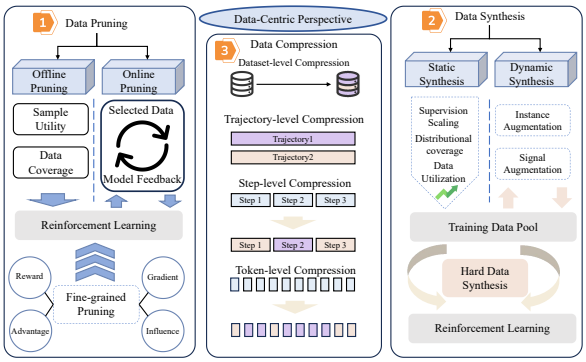

Figure 3: RL for LLMs in data-centric perspective.

and Aranguiz-Dias, 2025) approximates influence functions using a conjugate gradient solver to estimate each samples effect on validation loss. GRESO (Zheng et al., 2025) learns a reward-driven skipping policy to bypass prompts with zero reward variance, increasing the proportion of effective samples. PREPO (Sun et al., 2025b) leverages prompt perplexity to adopt an easy-to-hard curriculum when the available number of prompts is limited. CPPO (Lin et al., 2025) uses within-group advantages as a fine-grained learning signal to prune inefficient trajectories. PODS (Xu et al., 2025b) performs down-sampling to obtain a subset of trajectories with the largest reward variance.

3.2 Data Synthesis

Data synthesis can expand the scale of supervision and improve distributional coverage under a constrained sample budget. As shown in Figure 3, we classify existing methods as: static data synthesis before training, dynamic data synthesis during training, and hard data synthesis after training.

Static Data Synthesis. Static data synthesis (Ding et al., 2023; Zelikman et al., 2022) provides a solid data foundation (Guo et al., 2025) across different stages of LLM training. Constitutional AI (Bai et al., 2022), UltraFeedback (Cui et al., 2024), and CodeUltraFeedback (Weyssow et al., 2025) leverage strong LLMs to synthesize large scale preference and critique data with rich multi dimensional coverage. Meanwhile, Fair-PP (Zhou et al., 2025a) and Fellowship of the LLMs (Arif et al., 2025) further incorporate rule based constraints or multi agent collaboration. Reasoning Gym (Stojanovski et al., 2025), Enigmata (Chen et al., 2025a), SynLogic (Liu et al., 2025c), and EvoSyn (Du et al., 2025) scale the synthesis of verifiable reasoning data within offline pipelines by designing diverse interaction schemes between generators and verifiers. Meanwhile, SynthRL (Wu et al., 2025) and AdaR (Lai et al., 2025) focus on generating logically equivalent variant problems, with an emphasis on broader coverage of the training distribution and improved generalization.

Dynamic Data Synthesis. Dynamic data synthesis strengthens LLM training by continuously generating and augmenting data within the learning loop and can operate at different levels, such as instances and signals. OSP (Zhai et al., 2024) optimizes the policy using self-generated soft preference advantages, turning offline preference optimization into an online alignment pipeline while OAIF (He et al., 2025a) uses an online LLM annotator as a reward model to directly label on-policy response pairs.

Hard Data Synthesis. SwS (Liang et al., 2025) identifies the weakness problems that exhibit consistently low accuracy and degrading performance across epochs, and generates new training problems targeted at these weaknesses. LoopTool (Zhang et al., 2025d) identifies and corrects label errors after each RL phase. EVA (Ye et al., 2025b) finds the most useful prompts based on reward signals and generates variants of these prompts for subsequent training. WEBRL (Qi et al., 2024) automatically generates new tasks from a model's failed interactions and continuously incorporates these tasks into subsequent reinforcement learning stages.

3.3 Data Compression

Data compression aims to minimize both data and computational costs in reinforcement learning. As illustrated in Figure 3, existing data compression methods can be categorized into four levels: token, step, trajectory, and dataset.

Token-level Compression. Shallow Preference Signals (Qi et al., 2025) finds that many tokens are ineffective for reinforcement learning, indicating token-level compression is feasible. AR-Lopti (Yang et al., 2025c) reweighting or isolating low-probability tokens during training. TR-GRPO (Le et al., 2025) downweights low-probability tokens while placing greater emphasis on high-probability ones during gradient computation. High-Entropy Minority Tokens (Wang et al., 2025f) proposes updating the policy using only high-entropy tokens, which play a dominant role in determining reasoning branches and policy updates, while masking gradients from low-entropy tokens. DEPO (Tan et al., 2025b) separates reasoning tokens into efficient and inefficient segments, downweights the advantages of inefficient tokens, combined with difficulty-aware length penalties

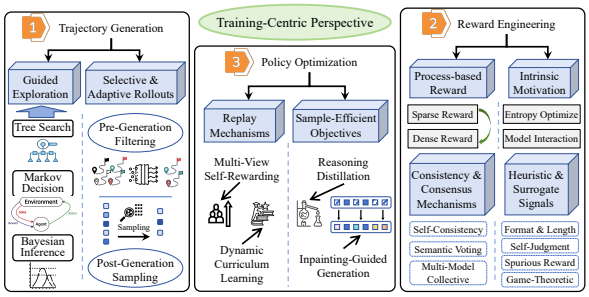

Figure 4: RL for LLMs in training-centric perspective.

and advantage clipping for stable optimization.

Step-level Compression. DAST (Shen et al., 2025b) proposes a difficulty-aware reward mechanism based on a Token Length Budget, enabling models to automatically shorten reasoning for simple problems while preserving sufficient CoT for complex ones. THINKPRUNE (Hou et al., 2025) encouraging models to actively prune redundant reasoning steps and progressively learn more efficient reasoning structures through iterative budget tightening. S-GRPO (Dai et al., 2025a) introduces a serial grouping strategy with decaying rewards, guiding models to recognize when intermediate reasoning is already sufficient and perform early exit along the reasoning trajectory. Interleaved Reasoning (Xie et al., 2025) trains models via reinforcement learning to alternate between generating intermediate answers and reasoning steps, thereby providing denser and more verifiable intermediate reward signals. Step-Entropy-based CoT Compression (Li et al., 2025e) shows that many low-entropy reasoning steps are redundant, and trains models to actively skip such steps during generation.

Trajectory-level Compression. DAPO (Yu et al., 2025) oversamples prompts during training and filters out entire groups of trajectories that produce zero gradients, retaining only trajectories that yield non-zero gradient signals for optimization. PROF (Ye et al., 2025a) performs trajectory-level ranking and filtering after rollout, retaining only those complete trajectories whose process rewards are consistent with outcome rewards.

Dataset-level Compresion. One-shot RLVR (Wang et al., 2025g) shows that the effective RLVR training set can be reduced to a single example while still achieving performance comparable to RLVR trained on datasets containing 7.5k samples. Polymath learning (Li et al., 2025d) restricts the RL training set to a single example, selecting or synthesizing a high-information-density sample by explicitly covering salient reasoning skills.

Key takeaways

- ***Model-Dependent Valuation.*** Data value estimation depends on model capability, and overemphasizing high-contribution data may weaken long-tail learning.
- ***Adaptive Data Governance.*** Future work should jointly and adaptively schedule data pruning, compression, and synthesis.

4 Training-Centric Perspective

From the perspective of training, researchers have addressed data dependence through three principal avenues: trajectory generation, reward engineering, and policy optimization (see Figure 4).

4.1 Trajectory Generation

Policy updates rely on numerous trajectory samples. To avoid the unaffordable consumption of computational resources and time, researchers have shifted from unstructured random exploration toward guided exploration and selective rollouts.

Guided Exploration. Guided exploration seeks to improve sample efficiency through structured sampling. A major breakthrough in this area integrates planning algorithms such as Monte Carlo Tree Search (MCTS) into the LLM decoding process (Chen et al., 2024; Wan et al., 2024), leveraging the structural advantages of search trees to increase the probability of generating valid samples. LLM-PSRL (Arumugam and Griffiths, 2025) adopts a Bayesian posterior sampling approach through prompting engineering. It first samples a hypothesis from the posterior distribution, then takes optimal actions based on that hypothesis. A more advanced exploration direction uses external search tools as part of the exploration process. Search-R1 (Jin et al., 2025) enables models to learn when to proactively initiate search queries in knowledge-blind regions. Such dynamic interleaving of search and reasoning trajectories encodes richer information-processing logic.

Selective and Adaptive Rollouts. Even with guided exploration, generating massive trajectories remains computationally expensive. Efficient rollout strategies (Zheng et al., 2025) have therefore become essential. GRESO (Zheng et al., 2025) exemplifies *pre-generation* filtering by maintaining historical states for each prompt and computing a skipping probability. Prompts predicted to yield low-information rollouts are directly by-

passed. PODS (Xu et al., 2025b), a *post-generation* sampling method, addresses the computational asymmetry between generation and update phases. Token-level masking methods (Wang et al., 2025f) identify high-entropy tokens as the true forking points that determine reasoning path directions. By computing policy gradients only for high-entropy tokens, models not only maintain performance but can even achieve improvements.

4.2 Reward Engineering

Evaluating trajectory quality is a central RL challenge. Sparse rewards complicate credit assignment, motivating automated process-based rewards and intrinsic motivation signals.

Process-based Reward. Process rewards aim to provide dense rewards for each reasoning step. In the absence of human-annotated process data, researchers leverage statistical properties of model outputs to synthesize process rewards. CoVo (Zhang et al., 2025e) observes that correct reasoning paths typically converge to the same answer, while incorrect paths diverge chaotically. It exploits consistency and volatility patterns to achieve a self-rewarding mechanism. Evol-RL (Zhou et al., 2025b) designs a novelty-promoting mechanism that rewards samples whose reasoning paths are semantically distant from existing paths yet still arrive at correct answers. L2T (Wang et al., 2025b) incorporates a compression penalty term that penalizes steps with low information gain but high token consumption. This reward design compels the model to learn efficient thinking, achieving maximum confidence improvement with minimal reasoning steps.

Intrinsic Motivation. When external feedback is unavailable, agents must rely on intrinsic motivation. Entropy-based metrics (Luo et al., 2025a) have become central to measuring intrinsic motivation. However, the community holds divergent views on how to leverage entropy. *Entropy minimization* perspective (Agarwal et al., 2025) argues that pretrained models already contain most required knowledge, and errors primarily stem from decoding uncertainty. Therefore, RL should minimize entropy to make models more confident (Prabhudesai et al., 2025; Li et al., 2025a; Zhang et al., 2025f). However, *Entropy maximization* perspective counters that solely minimizing entropy leads to premature convergence and overfitting. To push beyond capability boundaries, relative entropy must be maximized. PREPO (Sun et al., 2025b) and CDE (Dai et al., 2025b) reward rollouts with higher relative entropy (greater divergence from the old policy) to encourage exploration of new strategies. Beyond entropy, inter-model interactions also serve as important intrinsic rewards. SCoRe (Kumar et al., 2025) and ReviewRL (Zeng et al., 2025a) train models through multi-round and multi-agent RL to provide more reliable rewards.

Consistency and Consensus Mechanisms. In addition to entropy-driven exploration and multi-agent interaction, another powerful source of internal signals emerges from the self-consistency of the model's own response. For instance, KnowRL (Kale and Dhami, 2025) incorporates introspection and consensus-based rewarding mechanisms to allow models to self-improve their boundary awareness with the internally-generated data. Intuitor (Zhao et al., 2025c) utilizes the confidence of the model itself as the sole reward signal, realizing fully unsupervised learning. Majority voting reward methods (Zuo et al., 2025; Wei et al., 2025a; Shafayat et al., 2025) estimate reward signals through majority voting, allowing models to self-evolve during test-time inference. SVSI (Jiang et al., 2025) leverages lightweight sentence embeddings to measure semantic similarity among generated responses, thereby constructing preference pairs for model training. RLCCF (Yuan et al., 2025) also uses self-consistency to weight the votes from multiple heterogeneous LLMs.

Heuristic and Surrogate Signals. Since consensus-based rewards require multiple deployments for each query, researchers turn to lighter heuristic rewards from individual responses. (Shao et al., 2025) uncovers that RL can improve mathematical reasoning in specific models even with spurious rewards. It shows that RL training activates reasoning patterns already learned during pretraining. (Xin et al., 2025) suggests using format correctness and response length as surrogate reward signals. Researchers also find that the efficacy of internal feedback-based RL depends on the models initial policy entropy (Zhang et al., 2025h). CoMAS (Xue et al., 2025) provide a co-evolution framework for multi-agent systems using intrinsic rewards, while (Zhang et al., 2025j) proposes a co-rewarding framework to provide complementary supervision on the data side and model side. CaT (Jayalath et al., 2025) synthesizes a reference answer from the models parallel inference-time rollouts and converts it into reference-free rewards. SPPO (Wu et al.) frames LLM alignment as a two-

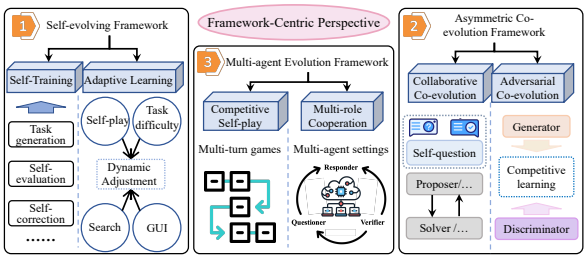

Figure 5: RL for LLMs in framework perspective.

player constant-sum game and iteratively updates policies to approximate the Nash equilibrium.

4.3 Policy Optimization

The effectiveness of policy optimization determines whether the model can efficiently acquire reliable reasoning capabilities from generated trajectories and engineered rewards.

Replay Mechanisms. To stabilize training and enhance sample efficiency, some methods revisit experience replay to adapt it to the unique structure of LLM reasoning trajectories. (Liu et al., 2025e) designs a multi-view self-rewarding mechanism to construct cross-validated intrinsic reward signals across diverse reasoning paths or problem formulations. (Sun et al., 2025c) presents a curriculum learning strategy with dynamic difficulty adjustment in reinforcement learning, enabling models to progressively advance from simple to complex problems in mathematical reasoning.

Sample-Efficient Objectives. Another pipeline focuses on redesigning the learning objective to maximize knowledge extraction from self-generated data. KTO (Ethayarajh et al., 2024) develops a distillation mechanism that leverages self-generated intermediate reasoning steps as supervision signals during training to enhance LLMs' reasoning abilities. IGPO (Zhao et al., 2025b) introduces an inpainting-like controllable generation mechanism into diffusion language models to enable guidance and alignment of generated content.

Key takeaways

- ***Mining Intrinsic Experience.*** The paradigm shifts from external supervision to mining intrinsic experiential value via guided exploration and endogenous signals.
- ***Efficiency versus Noise.*** A core challenge is trading off sampling cost and signal-to-noise ratio of self-generated rewards, while preventing reward-hacking collapse.

5 Framework-Centric Perspective

From a framework-centric perspective, RL under data scarcity leverages self-feedback to enable continuous evolution. We categorize methods into three paradigms: self-evolving, asymmetric co-evolution, and multi-agent evolution (see Figure 5).

5.1 Self-evolving Framework

Self-evolving utilizes a single model acting as both generator and evaluator, effectively closing the learning loop without external supervision.

Self-Training. Self-training methods evolve through iterative task generation and self-evaluation. Yu et al. (2026) enables self-evolving by progressively leveraging easy labeled data and harder unlabeled data via pseudo-labeling and RL. Feng and Feng (2024) combines Direct Preference Optimization with self-generated trajectories to reduce reliance on human labels. Similarly, self-correction frameworks (Xiong et al., 2025) and self-verification mechanisms like S^2R (Ma et al., 2025) reinforce learning at both outcome and process levels. Recent works further refine this via self-judging rewards for machine translation (Yang et al., 2025b), adaptive reward interpolation in RLER (Tan et al., 2025a), and bootstrapping via self-instruction in SeRL (Fang et al., 2025b).

Adaptive Learning. These methods focus on dynamic strategy adjustment w.r.t environmental changes. ZeroGUI (Yang et al., 2025a) automates task generation for GUI agents to eliminate handcrafted evaluations. Self-aware RL (Zhang et al., 2025b) enables capability prediction to proactively request data. ASL (Sun et al., 2025a) unifies task generation and execution in search environments. Similarly, AZR (Zhao et al., 2025a) uses self-play to autonomously generate and solve training tasks.

5.2 Asymmetric Co-Evolution Framework

Asymmetric co-evolution typically involves two distinct agents to enhance learning through cooperative or adversarial interactions.

Collaborative Co-Evolution. This approach often pairs a proposer with a solver (Sukhbaatar et al., 2018; OpenAI et al., 2021), which utilizes intrinsic motivation for automatic curriculum generation. This dynamic extends to self-questioning (Chen et al., 2025c), difficulty adjustment (Wei et al., 2025b), and visual reasoning (He et al., 2025b). Recent methods push the capability frontier by generating challenging queries via a Challenger (Liu

et al., 2025b) or unifying roles for search (Lu et al., 2025; Xu et al., 2025a). PasoDoble (Zhang et al., 2025i) further stabilizes training by decoupling updates between the proposer and solver.

Adversarial Co-Evolution. Adversarial frameworks often pit a generator against a discriminator. SPAG (Cheng et al., 2024) utilizes games like Adversarial Taboo to enforce information-reserved constraints. Verification capabilities are refined in SPC (Chen et al., 2025b) via a "sneaky generator" designed to fool a critic, and in GAR (Liu et al., 2025d) through a discriminator providing dense logical rewards. SSR (Wei et al., 2025c) employs an injection-repair loop to construct a code repair curriculum via bug generation.

5.3 Multi-Agent Evolution Framework

This paradigm generalizes self-play beyond binary interactions, utilizing complex game dynamics or specialized roles to internalize evaluation.

Competitive Self-Play Games. Competitive self-play allows LLMs to autonomously acquire reasoning skills by competing against themselves in structured, zero-sum environments. SPIRAL (Liu et al., 2025a) uses multi-turn games (e.g., poker) to incentivize systematic reasoning. Similarly, Vision-Zero (Wang et al., 2025e) adapts competitive logic to visual domains via "Who is the Spy" games, achieving SOTA results in visual reasoning.

Multi-Role Cooperation. These models employ specialized roles (e.g., Proposer, Solver, Verifier) to address complex tasks. SPELL (Yang et al., 2025d) and MAE (Chen et al., 2025f) utilize triplet structures to handle long-context and general reasoning. Others leverage formal verification signals (Wilf et al., 2025) or hierarchical actor-judge architectures (Wang et al., 2025c) to refine both policies and evaluation criteria within a closed loop.

Key takeaways

- ***Trade-offs.*** Self-evolving frameworks prioritize efficiency, while multi-agent methods trade cost for deeper reasoning.
- ***De-biasing.*** Interactions in co-evolution introduce external signals, breaking "echo chambers" and reducing self-delusion.
- ***Autogenous Curricula.*** These frameworks replace static datasets with dynamic generation, where task difficulty adaptively matches the model's evolving capability.

6 Challenges and Future Directions

▷ ***Reliability of Internal Rewards.*** LLM RL under data scarcity relies heavily on internal rewards (Zuo et al., 2025) such as consistency, entropy and heuristic signals. However, these signals are often noisy and susceptible to reward hacking (Shafayat et al., 2025) or model collapse (Shumailov et al., 2024), making reliable credit assignments challenging. Future research should explore robust process-based signals (Zhang et al., 2025e), and hybrid reward designs (Zhou et al., 2025b) that remain stable even under scarce or noisy feedback.

▷ ***Generalization to Unverifiable and Open-Ended Tasks.*** Most existing RL methods for LLMs under data scarcity focus on verifiable domains such as mathematics or coding (Liu et al., 2025b; Wei et al., 2025b) and are highly domain-dependent (Zhang et al., 2025i). Future work should address unverifiable or subjective tasks (e.g., creative writing, open-ended dialogue (Huang et al., 2025), scientific discovery), real-world solving (e.g., world modeling and embodied agents (Zhao et al., 2025a)), and generalization to out-of-distribution tasks.

▷ ***Safety Risks in Self-Play Frameworks.*** Although self-play frameworks can generate data that reduce reliance on human-crafted tasks, they also introduce significant safety risks (Wang et al., 2024, 2025a). For example, Llama-3.1-8B may exhibit uh-oh moments in chain-of-thought reasoning (Zhao et al., 2025a), and unsupervised self-evolution can propagate or even amplify biases from initial seed data (Fang et al., 2025b). Under data-scarce conditions, models may further develop spurious generalization patterns. To mitigate these risks, future work should incorporate online filtering (Chen et al., 2025c) and safety-aware training mechanisms to ensure reliable self-play processes.

7 Conclusion

We present a review of reinforcement learning for LLMs under data scarcity and introduce a bottom-up hierarchical taxonomy that categorizes existing approaches from data, training, and framework perspectives. Our survey reveals that addressing data scarcity goes beyond scaling supervised signals or interactive experience, but instead requires external data processing, effective internal data utilization during RL training, and evolving frameworks. We hope this survey provides a solid foundation for future research on data-efficient RL for LLMs.

8 Limitations

While this survey provides a first unified framework for reinforcement learning under data scarcity, it has limitations. Due to the rapid development of the field, the proposed framework requires timely updates to comprehensively cover emerging methods. We hope that this survey serves as inspiration for both theoretical and practical advancements in data-efficient reinforcement learning.

9 Ethical Statement

We follow the ACL Code of Ethics and maintain high standards of research integrity throughout this survey. As a literature review, our work does not involve human subjects, human annotation, or the collection of private or sensitive data. All referenced datasets, benchmarks, and models discussed in this survey are drawn from publicly available sources, and we cite them in accordance with their original usage and licensing conditions.

We aim to present a balanced and transparent analysis of existing methods, carefully summarizing their contributions, limitations, and potential risks. No conflicts of interest or sponsorship biases have been identified. We remain committed to addressing any ethical considerations raised during the review process and to promoting responsible research on data-efficient reinforcement learning for large language models.

10 Acknowledgments

The work of Zhiyin Yu, Bo Zhang, and Lei Bai was supported by the Shanghai Artificial Intelligence Laboratory. The authors thank the anonymous reviewers for their valuable comments and suggestions.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Huan ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 8 others. 2026. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Preprint*, arXiv:2507.21046.

Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2025. The fellowship of the LLMs: Multi-model workflows for synthetic preference optimization dataset generation. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 30–45, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Dilip Arumugam and Thomas L Griffiths. 2025. Toward efficient exploration by large language model agents. In *The Exploration in AI Today Workshop at ICML 2025*.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. 2025. Online difficulty filtering for reasoning oriented reinforcement learning. *CoRR*, abs/2504.03380.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024. Alphamath almost zero: process supervision without process. *Advances in Neural Information Processing Systems*, 37:27689–27724.

Jiangjie Chen, Qianyu He, Siyu Yuan, Aili Chen, Zhicheng Cai, Weinan Dai, Hongli Yu, Qiying Yu, Xuefeng Li, Jiaze Chen, Hao Zhou, and Mingxuan Wang. 2025a. Enigmata: Scaling logical reasoning in large language models with synthetic verifiable puzzles. *CoRR*, abs/2505.19914.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K. Wong. 2025b. SPC: Evolving self-play critic via adversarial games for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lili Chen, Mihir Prabhudesai, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025c. Self-questioning language models. *Preprint*, arXiv:2508.03682.

Weizhe Chen, Sven Koenig, and Bistra Dilkina. 2025d. LSPO: length-aware dynamic sampling for policy optimization in LLM reasoning. *CoRR*, abs/2510.01459.

Xingwu Chen, Tianle Li, and Difan Zou. 2025e. Reshaping reasoning in llms: A theoretical analysis of rl training dynamics through pattern selection. *Preprint*, arXiv:2506.04695.

Yixing Chen, Yiding Wang, Siqi Zhu, Haofei Yu, Tao Feng, Muhan Zhang, Mostofa Patwary, and Jiaxuan You. 2025f. Multi-agent evolve: Llm self-improve through co-evolution. *Preprint*, arXiv:2510.23595.

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, nan du, and Xiaolong Li. 2024. Self-playing adversarial language game enhances LLM reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, and 6 others. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *Preprint*, arXiv:2310.12773.

Muzhi Dai, Chenxu Yang, and Qingyi Si. 2025a. S-grpo: Early exit via reinforcement learning in reasoning models. *arXiv preprint arXiv:2505.07686*.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025b. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.

He Du, Bowen Li, Aijun Yang, Siyang He, Qipeng Guo, and Dacheng Tao. 2025. Evosyn: Generalizable evolutionary data synthesis for verifiable learning. *CoRR*, abs/2510.17928.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. 2025a. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *Preprint*, arXiv:2508.07407.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. 2025b. SeRL: Self-play reinforcement learning for large language models with limited data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Daniel Fein and Gabriela Aranguiz-Dias. 2025. Influence functions for preference dataset pruning. *CoRR*, abs/2507.14344.

Shuang Feng and Grace Feng. 2024. An extremely data-efficient and generative llm-based reinforcement learning agent for recommenders. *Preprint*, arXiv:2408.16032.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small LLMs can master math reasoning with self-evolved deep thinking. In *Forty-second International Conference on Machine Learning*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *Preprint*, arXiv:2308.08998.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *Preprint*, arXiv:2401.14196.

Li He, He Zhao, Stephen Wan, Dadong Wang, Lina Yao, and Tongliang Liu. 2025a. Direct advantage regression: Aligning llms with online AI reward. *CoRR*, abs/2504.14177.

Qiang He and Setareh Maghsudi. 2026. Pareto multi-objective alignment forălanguage models. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 257–272, Cham. Springer Nature Switzerland.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. 2025b. Visplay: Self-evolving vision-language models from images. *Preprint*, arXiv:2511.15661.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. 2025. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning. *arXiv preprint arXiv:2504.01296*.

Sam Houliston, Alizée Pace, Alexander Immer, and Gunnar Rätsch. 2024. Uncertainty-penalized direct preference optimization. *Preprint*, arXiv:2410.20187.

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. *Preprint*, arXiv:2503.24290.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. *Preprint*, arXiv:2508.05004.

Dulhan Jayalath, Shashwat Goel, Thomas Foster, Parag Jain, Suchin Gururangan, Cheng Zhang, Anirudh Goyal, and Alan Schelten. 2025. Compute as teacher: Turning inference compute into reference-free supervision. *arXiv preprint arXiv:2509.14234*.

Kaixuan Ji, Jiafan He, and Quanquan Gu. 2024. Reinforcement learning from human feedback with active queries. *Trans. Mach. Learn. Res.*, 2025.

Chunyang Jiang, Yonggang Zhang, Yiyang Cai, Chi-Min Chan, Yulong Liu, Mingming Chen, Wei Xue, and Yike Guo. 2025. Semantic voting: A self-evaluation-free approach for efficient llm self-improvement on unverifiable open-ended tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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*.

N Jones. 2024. The ai revolution is running out of data. *What can researchers do*.

Sahil Kale and Devendra Singh Dhami. 2025. Knowrl: Teaching language models to know what they know. *arXiv preprint arXiv:2510.11407*.

Bryce Kan, Wei Yang, Emily Nguyen, Ganghui Yi, Bowen Yi, Chenxiao Yu, and Yan Liu. 2026. De-conflating preference and qualification: Constrained dual-perspective reasoning for job recommendation with large language models. *Preprint*, arXiv:2602.03097.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2025. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

Zhejian Lai, Xiang Geng, Zhijun Wang, Yang Bai, Jiahuan Li, Rongxiang Weng, Jingang Wang, Xuezhi Cao, Xunliang Cai, and Shujian Huang. 2025. Making mathematical reasoning adaptive. *CoRR*, abs/2510.04617.

Tue Le, Nghi D. Q. Bui, Linh Ngo Van, and Trung Le. 2025. Sharpness-controlled group relative policy optimization with token-level probability shaping. *Preprint*, arXiv:2511.00066.

Bowen Li, Wenhan Wu, Ziwei Tang, Lin Shi, John Yang, Jinyang Li, Shunyu Yao, Chen Qian, Binyuan Hui, Qicheng Zhang, Zhiyin Yu, He Du, Ping Yang, Dahua Lin, Chao Peng, and Kai Chen. 2024. Prompting large language models to tackle the full software development lifecycle: A case study. *Preprint*, arXiv:2403.08604.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. 2025a. Confidence is all you need: Few-shot rl fine-tuning of language models. *arXiv preprint arXiv:2506.06395*.

Shikun Li, Shipeng Li, Zhiqin Yang, Xinghua Zhang, Gaode Chen, Xiaobo Xia, Hengyu Liu, and Zhe Peng. 2025b. Learnalign: Reasoning data selection for reinforcement learning in large language models based on improved gradient alignment. *CoRR*, abs/2506.11480.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025c. LIMR: less is more for RL scaling. *CoRR*, abs/2502.11886.

Yiyuan Li, Zhen Huang, Yanan Wu, Weixun Wang, Xuefeng Li, Yijia Luo, Pengfei Liu, Wenbo Su, and Bo Zheng. 2025d. One sample to rule them all: Extreme data efficiency in RL scaling. In *NeurIPS 2025 Workshop on Efficient Reasoning*.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. 2025e. Compressing chain-of-thought in llms via step entropy. *Preprint*, arXiv:2508.03346.

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. *CoRR*, abs/2506.08989.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. 2025. CPPO: accelerating the training of group relative policy optimization-based reasoning models. *CoRR*, abs/2503.22342.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. 2025a. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *Preprint*, arXiv:2506.24119.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. 2025b. Spice: Self-play in corpus environments improves reasoning. *Preprint*, arXiv:2510.24684.

Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, Yunan Huang, Mozhi Zhang, Pengyu Zhao, Junjie Yan, and Junxian He. 2025c. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *CoRR*, abs/2505.19641.

Qihao Liu, Luoxin Ye, Wufei Ma, Yu-Cheng Chou, and Alan Yuille. 2025d. Generative adversarial reasoner: Enhancing llm reasoning with adversarial reinforcement learning. *Preprint*, arXiv:2512.16917.

Zichuan Liu, Jinyu Wang, Lei Song, and Jiang Bian. 2025e. Sample-efficient llm optimization with reset replay. *arXiv preprint arXiv:2508.06412*.

Hongliang Lu, Yuhang Wen, Pengyu Cheng, Ruijin Ding, Haotian Xu, Jiaqi Guo, Chutian Wang, Haonan Chen, Xiaoxi Jiang, and Guanjun Jiang. 2025. Search self-play: Pushing the frontier of agent capability without supervision. *Preprint*, arXiv:2510.18821.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025a. Semi-supervised fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2795–2808.

Junyu Luo, Bohan Wu, Xiao Luo, Zhiping Xiao, Yiqiao Jin, Rong-Cheng Tu, Nan Yin, Yifan Wang, Jingyang Yuan, Wei Ju, and Ming Zhang. 2025b. A survey on efficient large language model training: From data-centric perspectives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30904–30920, Vienna, Austria. Association for Computational Linguistics.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. 2025. S^2r: Teaching llms to self-verify and self-correct via reinforcement learning. *Preprint*, arXiv:2502.12853.

Luckeciano C. Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. 2024. Deep bayesian active learning for preference modeling in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 118052–118085. Curran Associates, Inc.

OpenAI. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI OpenAI, Matthias Plappert, Raul Sampedro, Tao Xu, Ilge Akkaya, Vineet Kosaraju, Peter Welinder, Ruben D'Sa, Arthur Petron, Henrique P. d. O. Pinto, Alex Paino, Hyeonwoo Noh, Lilian Weng, Qiming Yuan, Casey Chu, and Wojciech Zaremba. 2021. Asymmetric self-play for automatic goal discovery in robotic manipulation. *Preprint*, arXiv:2101.04882.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Xuan Qi, Jiahao Qiu, Xinzhe Juan, Yue Wu, and Mengdi Wang. 2025. Shallow preference signals: Large language model aligns even better with truncated data? *arXiv preprint arXiv:2505.17122*.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Xinyue Yang, Jiadai Sun, Yu Yang, Shuntian Yao, Tianjie Zhang, and 1 others. 2024. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. *arXiv preprint arXiv:2411.02337*.

Yun Qu, Qi Wang, Yixiu Mao, Vincent Tao Hu, Björn Ommer, and Xiangyang Ji. 2025. Can prompt difficulty be online predicted for accelerating RL finetuning of reasoning models? *CoRR*, abs/2507.04632.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, and 61 others. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can large reasoning models self-train? *Preprint*, arXiv:2505.21444.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, and 1 others. 2025. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*.

Qianli Shen, Daoyuan Chen, Yilun Huang, Zhenqing Ling, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025a. BOTS: A unified framework for bayesian online task selection in LLM reinforcement finetuning. *CoRR*, abs/2510.26374.

Yi Shen, Jian Zhang, Jieyun Huang, Shuming Shi, Wenjing Zhang, Jiangze Yan, Ning Wang, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025b. Dast: Difficulty-adaptive slow-thinking for large reasoning models. *arXiv preprint arXiv:2503.04472*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

David Silver and Richard S. Sutton. 2025. Welcome to the era of experience. Technical report, Google AI.

Toby Simonds, Kevin Lopez, Akira Yoshiyama, and Dominique Garmier. 2025. Self rewarding self improving. *arXiv preprint arXiv:2505.08827*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. 2025. REASONING GYM: reasoning environments for reinforcement learning with verifiable rewards. *CoRR*, abs/2505.24760.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. 2018. Intrinsic motivation and automatic curricula via asymmetric self-play. In *International Conference on Learning Representations*.

Wangtao Sun, Xiang Cheng, Jialin Fan, Yao Xu, Xing Yu, Shizhu He, Jun Zhao, and Kang Liu. 2025a. Towards agentic self-learning llms in search environment. *Preprint*, arXiv:2510.14253.

Yan Sun, Jia Guo, Stanley Kok, Zihao Wang, Zujie Wen, and Zhiqiang Zhang. 2025b. Efficient reinforcement learning for large language models with intrinsic exploration. *CoRR*, abs/2511.00794.

Yifan Sun, Jingyan Shen, Yibin Wang, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, and Huan Zhang. 2025c. Improving data efficiency for LLM reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. *CoRR*, abs/2506.05316.

Chuyi Tan, Peiwen Yuan, Xinglin Wang, Yiwei Li, Shaoxiong Feng, Yueqi Zhang, Jiayi Shi, Ji Zhang, Boyuan Pan, Yao Hu, and Kan Li. 2025a. Diagnosing and mitigating system bias in self-rewarding rl. *Preprint*, arXiv:2510.08977.

Zezhong Tan, Hang Gao, Xinhong Ma, Feng Zhang, and Ziqiang Dong. 2025b. Towards flash thinking via decoupled advantage policy optimization. *arXiv preprint arXiv:2510.15374*.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *Preprint*, arXiv:2404.14387.

InternAgent Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Runmin Ma, Yusong Hu, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, Zhilong Wang, Jinyao Liu, Tianshuo Peng, Peng Ye, Dongzhan Zhou, Shufei Zhang, Xiaosong Wang, and 7 others. 2025a. Internagent: When agent becomes the scientist – building closed-loop system from hypothesis to verification. *Preprint*, arXiv:2505.16938.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 77 others. 2025b. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus Mcaleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training. In *International Conference on Machine Learning*, pages 49890–49920. PMLR.

Huanqian Wang, Yang Yue, Rui Lu, Jingxin Shi, Andrew Zhao, Shenzhi Wang, Shiji Song, and Gao Huang. 2025a. Model surgery: Modulating LLM's behavior via simple parameter editing. 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)*, pages 6337–6357, Albuquerque, New Mexico. Association for Computational Linguistics.

Jingyao Wang, Wenwen Qiang, Zeen Song, Changwen Zheng, and Hui Xiong. 2025b. Learning to think: Information-theoretic reinforcement fine-tuning for llms. *arXiv preprint arXiv:2505.10425*.

Junqiao Wang, Zhaoyang Guan, Guanyu Liu, Tianze Xia, Xianzhi Li, Shuo Yin, Xinyuan Song, Chuhan Cheng, Tianyu Shi, and Alex Lee. 2025c. Meta-learning reinforcement learning for crypto-return prediction. *Preprint*, arXiv:2509.09751.

Qinsi Wang, Jinghan Ke, Hancheng Ye, Yueqian Lin, Yuzhe Fu, Jianyi Zhang, Kurt Keutzer, Chenfeng Xu, and Yiran Chen. 2025d. Angles don't lie: Unlocking training-efficient RL through the model's own signals. *CoRR*, abs/2506.02281.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. 2025e. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play. *Preprint*, arXiv:2509.25541.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2024. Boosting LLM agents with recursive contemplation for effective deception handling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9909–9953, Bangkok, Thailand. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025f. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shuai Wang, Yaoming Yang, Bingdong Li, Hao Hao, and Aimin Zhou. 2026. Ib-grpo: Aligning llm-based learning path recommendation with educational objectives via indicator-based group relative policy optimization. *Preprint*, arXiv:2601.14686.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025g. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. 2025a. First sft, second rl, third upt: Continual improving multimodal llm reasoning via unsupervised post-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yongxian Wei, Yilin Zhao, Li Shen, Xinrui Chen, Runxi Cheng, Sinan Du, Hao Yu, Gang Liu, Jiahong Yan, Chun Yuan, and Dian Li. 2025b. Learning to pose problems: Reasoning-driven and solver-adaptive data synthesis for large reasoning models. *Preprint*, arXiv:2511.09907.

Yuxiang Wei, Zhiqing Sun, Emily McMilin, Jonas Gehring, David Zhang, Gabriel Synnaeve, Daniel Fried, Lingming Zhang, and Sida Wang. 2025c. Toward training superintelligent software agents through self-play swe-rl. *Preprint*, arXiv:2512.18552.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. 2025. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

Alex Wilf, Pranjal Aggarwal, Bryan Parno, Daniel Fried, Louis-Philippe Morency, Paul Pu Liang, and Sean Welleck. 2025. Propose, solve, verify: Self-play through formal verification. *Preprint*, arXiv:2512.18160.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023a. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, volume 36, pages 59008–59033. Curran Associates, Inc.

Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan, and Michael Qizhe Shieh. 2025. Synthrl: Scaling visual reasoning with verifiable data synthesis. *CoRR*, abs/2506.02096.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2025. Beyond chain-of-thought: A survey of chain-of-X paradigms for LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10795–10809, Abu Dhabi, UAE. Association for Computational Linguistics.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. 2025. Interleaved reasoning for large language models via reinforcement learning. *arXiv preprint arXiv:2505.19640*.

Rihui Xin, Han Liu, Zecheng Wang, Yupeng Zhang, Dianbo Sui, Xiaolin Hu, and Bingning Wang. 2025. Surrogate signals from format and length: Reinforcement learning for solving mathematical problems without ground truth answers. *arXiv preprint arXiv:2505.19439*.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. 2025. Self-rewarding correction for mathematical reasoning. *Preprint*, arXiv:2502.19613.

Ran Xu, Yuchen Zhuang, Zihan Dong, Ruiyu Wang, Yue Yu, Joyce C. Ho, Linjun Zhang, Haoyu Wang, Wenqi Shi, and Carl Yang. 2025a. Acesearcher: Bootstrapping reasoning and search for LLMs via reinforced self-play. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yixuan Even Xu, Yash Savani, Fei Fang, and Zico Kolter. 2025b. Not all rollouts are useful: Down-sampling rollouts in LLM reinforcement learning. *CoRR*, abs/2504.13818.

Xiangyuan Xue, Yifan Zhou, Guibin Zhang, Zaibin Zhang, Yijiang Li, Chen Zhang, Zhenfei Yin, Philip

Torr, Wanli Ouyang, and Lei Bai. 2025. Comas: Co-evolving multi-agent systems via interaction rewards. *arXiv preprint arXiv:2510.08529*.

Chenyu Yang, Shiqian Su, Shi Liu, Xuan Dong, Yue Yu, Weijie Su, Xuehui Wang, Zhaoyang Liu, Jinguo Zhu, Hao Li, Wenhai Wang, Yu Qiao, Xizhou Zhu, and Jifeng Dai. 2025a. Zerogui: Automating online gui learning at zero human cost. *Preprint*, arXiv:2505.23762.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2025b. Ssr-zero: Simple self-rewarding reinforcement learning for machine translation. *Preprint*, arXiv:2505.16637.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. 2025c. Do not let low-probability tokens over-dominate in rl for llms.

Ziyi Yang, Weizhou Shen, Ruijun Chen, Chenliang Li, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. 2025d. Spell: Self-play reinforcement learning for evolving long-context language models. *Preprint*, arXiv:2509.23863.

Chenlu Ye, Zhou Yu, Ziji Zhang, Hao Chen, Narayanan Sadagopan, Jing Huang, Tong Zhang, and Anurag Beniwal. 2025a. Beyond correctness: Harmonizing process and outcome rewards through rl training. *arXiv preprint arXiv:2509.03403*.

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V. Le, Qijun Tan, and Yuan Liu. 2025b. Reward-guided prompt evolving in reinforcement learning for llms. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhiyin Yu, Bo Zhang, Qibin Hou, Zhonghai Wu, Xiao Luo, and Lei Bai. 2026. Easy samples are all you need: Self-evolving llms via data-efficient reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*.

Wenzhen Yuan, Shengji Tang, Weihao Lin, Jiacheng Ruan, Ganqu Cui, Bo Zhang, Tao Chen, Ting Liu, Yuzhuo Fu, Peng Ye, and 1 others. 2025. Wisdom of the crowd: Reinforcement learning from coevolutionary collective feedback. *arXiv preprint arXiv:2508.12338*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, and 1 others. 2025a. Reviewrl: Towards automated scientific review with rl. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16942–16954.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun MA, and Junxian He. 2025b. SimpleRL-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*.

Yuanzhao Zhai, Zhuo Zhang, Kele Xu, Hanyang Peng, Yue Yu, Dawei Feng, Cheng Yang, Bo Ding, and Huaimin Wang. 2024. Online self-preferring language models. *CoRR*, abs/2405.14103.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Francisco Piedrahita-Velez, Yue Liao, Hongru Wang, and 6 others. 2025a. The landscape of agentic reinforcement learning for llms: A survey. *Preprint*, arXiv:2509.02547.

Hangfan Zhang, Siyuan Xu, Zhimeng Guo, Huaisheng Zhu, Shicheng Liu, Xinrun Wang, Qiaosheng Zhang, Yang Chen, Peng Ye, Lei Bai, and Shuyue Hu. 2025b. The path of self-evolving large language models: Achieving data-efficient learning via intrinsic feedback. *Preprint*, arXiv:2510.02752.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, and 20 others. 2025c. A survey of reinforcement learning for large reasoning models. *Preprint*, arXiv:2509.08827.

Kangning Zhang, Wenxiang Jiao, Kounianhua Du, Yuan Lu, Weiwen Liu, Weinan Zhang, Lei Zhang, and Yong Yu. 2025d. Looptool: Closing the data-training loop for robust llm tool calls. *Preprint*, arXiv:2511.09148.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. 2025e. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *Preprint*, arXiv:2506.08745.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025f. Right question is already half the answer: Fully unsupervisedd LLM reasoning incentization. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*.

Ruiqi Zhang, Daman Arora, Song Mei, and Andrea Zanette. 2025g. SPEED-RL: faster training of reasoning models via online curriculum learning. *CoRR*, abs/2506.09016.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025h. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*.

Zhengxin Zhang, Chengyu Huang, Aochong Oliver Li, and Claire Cardie. 2025i. Better llm reasoning via dual-play. *Preprint*, arXiv:2511.11881.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. 2025j. Co-rewarding: Stable self-supervised rl for eliciting reasoning in large language models. *arXiv preprint arXiv:2508.00410*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025a. Absolute zero: Reinforced self-play reasoning with zero data. *Preprint*, arXiv:2505.03335.

Siyan Zhao, Mengchen Liu, Jing Huang, Miao Liu, Chenyu Wang, Bo Liu, Yuandong Tian, Guan Pang, Sean Bell, Aditya Grover, and 1 others. 2025b. Inpainting-guided policy optimization for diffusion large language models. *arXiv preprint arXiv:2509.10396*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025c. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.

Yang Zhao, Hepeng Wang, Xiao Ding, Yangou Ouyang, Bibo Cai, Kai Xiong, Jinglong Gao, Zhouhao Sun, Li Du, Bing Qin, and Ting Liu. 2026. Maestro: Meta-learning adaptive estimation of scalarization trade-offs for reward optimization. *Preprint*, arXiv:2601.07208.

Yang Zhao, Kai Xiong, Xiao Ding, Li Du, YangouOuyang, Zhouhao Sun, Jiannan Guan, Wenbin Zhang, Bin Liu, Dong Hu, Bing Qin, and Ting Liu. 2025d. Ufo-rl: Uncertainty-focused optimization for efficient reinforcement learning data selection. *Preprint*, arXiv:2505.12457.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. 2025. Act only when it pays: Efficient reinforcement learning for LLM reasoning via selective rollouts. *CoRR*, abs/2506.02177.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, and 8 others. 2023. Secrets of rlhf in large language models part i: Ppo. *Preprint*, arXiv:2307.04964.

Qi Zhou, Jie Zhang, Dongxia Wang, Qiang Liu, Tianlin Li, Jin Song Dong, Wenhai Wang, and Qing Guo. 2025a. Fair-pp: A synthetic dataset for aligning LLM with personalized preferences of social equity. *CoRR*, abs/2505.11861.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025b. Evolving language models without labels: Majority drives selection, novelty promotes variation. *Preprint*, arXiv:2509.15194.

Yongfu Zhu, Lin Sun, Guangxiang Zhao, Weihong Lin, and Xiangzheng Zhang. 2025. Uncertainty under the curve: A sequence-level entropy area metric for reasoning LLM. *CoRR*, abs/2508.20384.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. TTRL: Test-time reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

A Statistics

To show the research momentum in data-efficient reinforcement learning for large reasoning models (LLMs), we conduct a statistical analysis of the publication year of all surveyed papers. As shown in Figure 6, the field has experienced an extraordinary growth trajectory. Early works are scarce, with 1 publication in 2018 and 2020 each, followed by 2 papers in both 2021 and 2022, and 2 papers in 2023. The field witnessed a modest increase to 8 papers in 2024, before experiencing an explosive surge to 109 publications by December 2025. This remarkable acceleration highlights the field's transition from an emerging research direction to a mainstream paradigm in LLM development.

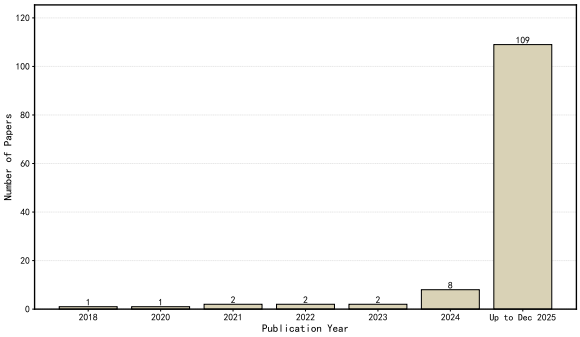

Figure 6: Distribution on publication year of surveyed papers.

Furthermore, we perform a word frequency analysis on the titles of all surveyed papers. Figure 7 reveals the core methodological and conceptual themes in current research. Prominently, *Reasoning* and *Learning* emerge as the most central terms, which reflect the field's primary focus on enhancing reasoning capabilities through learning mechanisms. The significant presence of *Reinforcement*, *Training*, *Data* and *Model* underscores the importance of reinforcement learning based training strategies. Additionally, key concepts such as *Self*, *Evolving*, and *Agent* highlight the growing emphasis on self-improvement and agentic frameworks, and terms such as *Verifiable* and *Rewarding* also indicate the field's focus on reward engineering. This visualization demonstrates the rapid advances that data-efficient RL for large reasoning models has seen in recent years, laying a foundation for future progress in LLM capabilities.

Figure 7: Word cloud of research paper titles.

B Acknowledgment of AI Assistance in Writing and Revision

We only use large language models for grammar checking and language refinement, and we strictly adhere to the ACL policy on AI writing assistance. All ideas and technical content in this paper are original contributions of the authors.

C Further Analysis

C.1 Discussion on Data Pruning

Uncertainty-Aware Pruning. Uncertainty-aware pruning uses epistemic uncertainty to identify samples where the models knowledge is most insufficient, concentrating the training budget on informative data. UFO-RL (Zhao et al., 2025d) computes a single-pass confidence score from average token-level log-probabilities and selects samples within the models zone of proximal development, matching full-data performance with only 10% of data. ReST (Gulcehre et al., 2023) iteratively grows the training set by generating candidates from the current policy and filtering them through a reward threshold, coupling data curation with policy improvement. BAL-PM (Melo et al., 2024) combines ensemble-based epistemic uncertainty of the reward model with a prompt-distribution entropy term for active preference data acquisition, reducing required labels by 3368%. ADPO (Ji et al., 2024) formulates RLHF as a contextual dueling bandit and introduces an uncertainty-aware query criterion that requests human feedback only when the reward gap is poorly estimated, halving query complexity while matching DPO performance. Uncertainty-Penalized DPO (Houliston et al., 2024) attenuates the DPO loss gradient for preference pairs with high reward uncertainty, down-weighting ambiguous or mislabeled samples.

C.2 Discussion on Reward Engineering

Theoretical Perspectives on Internal Feedback. Recent work has begun to provide theoretical grounding for why internal feedback signals can substitute for external rewards. Zhang et al. (2025h) formally investigate Reinforcement Learning from Internal Feedback, showing that token-level entropy minimization, trajectory-level entropy minimization, and self-certainty maximization are partially equivalent optimization objectives under mild assumptions. CoVo (Zhang et al., 2025e) provides a complementary theoretical perspective by reinterpreting its consistency-volatility reward as a variational inference objective, treating reasoning trajectories as latent variables. This formulation grounds the self-rewarding mechanism in both reasoning paths and final answers, offering a principled explanation for why correct responses exhibit convergent trajectory patterns while incorrect ones diverge.

From a broader perspective, Shao et al. (2025) present a striking empirical finding with theoretical implications: RLVR can elicit strong mathematical reasoning even with spurious rewards on certain model families. Chen et al. (2025e) further support this point by a two-stage mathematical framework modeling reasoning as $q \rightarrow r \rightarrow a$ (question-reason-answer). For RLVR, they prove convergence to the reasoning pattern with the highest success rate. For RLIF, their analysis explains both the initial performance gains and eventual degradation observed empirically. Agarwal et al. (2025) further demonstrate the unreasonable effectiveness of entropy minimization: using negative entropy as the sole RL reward, their EM-RL achieves performance comparable to supervised GRPO and RLOO on challenging reasoning benchmarks. These theoretical and empirical results collectively suggest a key insight that under data scarcity, RL may function less as a mechanism for acquiring new capabilities and more as a redistribution mechanism that sharpens the policy over reasoning patterns already learned during pretraining.

Adaptive Multi-Objective Trade-offs. Under data scarcity, static reward weighting fails without sufficient validation data for tuning. Effective reward engineering thus requires denser signals and adaptive mechanisms to balance conflicting objectives. FINE-GRAINED RLHF (Wu et al., 2023b) replaces holistic feedback with segment-level, multi-type rewards, yielding denser supervision and higher sample efficiency. Safe-RLHF (Dai et al., 2023) decouples helpfulness and harmlessness into reward/cost models, using Lagrangian constraints for dynamic balancing without hyperparameter searches. MAESTRO (Zhao et al., 2026) treats scalarization as a dynamic latent policy, co-evolving weights via group-relative meta-signals to eliminate manual tuning. Extending this paradigm, PAMA (He and Maghsudi, 2026) transforms multi-objective RLHF into a convex optimization with closed-form Pareto convergence at $O(n)$ complexity. JobRec (Kan et al., 2026) applies similar Lagrangian constraints to de-conflate candidate preference and employer qualification under data scarcity. IB-GRPO (Wang et al., 2026) employs ϵ-dominance indicators to compute group-relative advantages across multiple objectives, avoiding manual scalarization while leveraging hybrid expert demonstrations. Collectively, these methods shift reward engineering from static design toward adaptive, data-efficient mechanisms for navigating conflicting objectives without extensive validation data.

D Literature Review Summary

To provide an overview of the literature examined in this study, we present a detailed summary table in this appendix covering all discussed works in our analysis. Each entry includes six columns : **Title** (the complete publication name); **Section** and **Subsection** (place in the three-level taxonomy); **Year** (the publication date); **Venue** (the publication outlet) and **Link** (URL to the original source of paper), as shown in Table 1.

Table 1: Summary of Referenced Papers

Title	Section	Subsection	Year	Venue	Link
Uncertainty Under the Curve: A Sequence-Level Entropy Area Metric for Reasoning LLM	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
LLaMA: Open and Efficient Foundation Language Models	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
Learning to summarize from human feedback	Level 1: Data-Centric Perspective	Data Pruning	2020	NeurIPS	link
Scaling Language Models: Methods, Analysis & Insights from Training Gopher	Level 1: Data-Centric Perspective	Data Pruning	2021	Arxiv	link
Self-Instruct: Aligning Language Models with Self-Generated Instructions	Level 1: Data-Centric Perspective	Data Pruning	2023	ACL	link
LIMR: Less is More for RL Scaling	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
LearnAlign: Reasoning Data Selection for Reinforcement Learning in Large Language Models Based on Improved Gradient Alignment	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
Unified Data Selection for LLM Reasoning	Level 1: Data-Centric Perspective	Data Pruning	2025	ICLR	link
LSPO: Length-aware Dynamic Sampling for Policy Optimization in LLM Reasoning	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
Can Prompt Difficulty be Online Predicted for Accelerating RL Finetuning of Reasoning Models?	Level 1: Data-Centric Perspective	Data Pruning	2025	KDD	link
BOTS: A Unified Framework for Bayesian Online Task Selection in LLM Reinforcement Finetuning	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
Angles Don't Lie: Unlocking Training?Efficient RL Through the Model's Own Signals	Level 1: Data-Centric Perspective	Data Pruning	2025	NeurIPS	link
Online Difficulty Filtering for Reasoning Oriented Reinforcement Learning	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
SPEED-RL: Faster Training of Reasoning Models via Online Curriculum Learning	Level 1: Data-Centric Perspective	Data Pruning	2025	ICML Workshop	link
Secrets of RLHF in Large Language Models Part I: PPO	Level 1: Data-Centric Perspective	Data Pruning	2023	NeurIPS Workshop	link
Influence Functions for Preference Dataset Pruning	Level 1: Data-Centric Perspective	Data Pruning	2025	NeurIPS Workshop	link
CPPO: Accelerating the Training of Group Relative Policy Optimization-Based Reasoning Models	Level 1: Data-Centric Perspective	Data Pruning	2025	NeurIPS	link
Not All Rollouts are Useful: Down-Sampling Rollouts in LLM Reinforcement Learning	Level 1: Data-Centric Perspective	Data Pruning	2025	Arxiv	link
Enhancing Chat Language Models by Scaling High-quality Instructional Conversations	Level 1: Data-Centric Perspective	Data Synthesis	2023	EMNLP	link
STaR: Bootstrapping Reasoning With Reasoning	Level 1: Data-Centric Perspective	Data Synthesis	2022	NeurIPS	link
Constitutional AI: Harmlessness from AI Feedback	Level 1: Data-Centric Perspective	Data Synthesis	2022	Arxiv	link

Continued on next page

Table 1 – Continued

Title	Section	Subsection	Year	Venue	Link
ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback	Level 1: Data-Centric Perspective	Data Synthesis	2024	ICML	link
CodeUltraFeedback: An LLM-as-a-Judge Dataset for Aligning Large Language Models to Coding Preferences	Level 1: Data-Centric Perspective	Data Synthesis	2025	ACM Transactions on Software Engineering and Methodology	link
Fair-PP: A Synthetic Dataset for Aligning LLM with Personalized Preferences of Social Equity	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
The Fellowship of the LLMs: Multi-Model Workflows for Synthetic Preference Optimization Dataset Generation	Level 1: Data-Centric Perspective	Data Synthesis	2025	GEM	link
Reasoning Gym: Reasoning Environments for Reinforcement Learning with Verifiable Rewards	Level 1: Data-Centric Perspective	Data Synthesis	2025	NeurIPS	link
Enigmata: Scaling Logical Reasoning in Large Language Models with Synthetic Verifiable Puzzles	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
SynLogic: Synthesizing Verifiable Reasoning Data at Scale for Learning Logical Reasoning and Beyond	Level 1: Data-Centric Perspective	Data Synthesis	2025	NeurIPS	link
EvoSyn: Generalizable Evolutionary Data Synthesis for Verifiable Learning	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
SynthRL: Scaling Visual Reasoning with Verifiable Data Synthesis	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
Making Mathematical Reasoning Adaptive	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
SwS: Self-aware Weakness-driven Problem Synthesis in Reinforcement Learning for LLM Reasoning	Level 1: Data-Centric Perspective	Data Synthesis	2025	NeurIPS	link
Reward-Guided Prompt Evolving in Reinforcement Learning for LLMs	Level 1: Data-Centric Perspective	Data Synthesis	2025	ICML	link
Direct Advantage Regression: Aligning LLMs with Online AI Reward	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
Online Self-Preferring Language Models	Level 1: Data-Centric Perspective	Data Synthesis	2024	Arxiv	link
LoopTool: Closing the DataTraining Loop for Robust LLM Tool Calls	Level 1: Data-Centric Perspective	Data Synthesis	2025	Arxiv	link
WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning	Level 1: Data-Centric Perspective	Data Synthesis	2025	ICLR	link
Shallow Preference Signals: Large Language Model Aligns Even Better with Truncated Data?	Level 1: Data-Centric Perspective	Data Compression	2025	GEM	link
Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link
Dapo: An open-source llm reinforcement learning system at scale	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link
Beyond Correctness: Harmonizing Process and Outcome Rewards through RL Training	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link
Reinforcement Learning for Reasoning in Large Language Models with One Training Example	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link

Continued on next page

Table 1 – Continued

Title	Section	Subsection	Year	Venue	Link
Sharpness-Controlled Group Relative Policy Optimization with Token-Level Probability Shaping	Level 1: Data-Centric Perspective	Data compression	2025	Arxiv	link
One Sample to Rule Them All: Extreme Data Efficiency in RL Scaling	Level 1: Data-Centric Perspective	Data Compression	2025	Arxiv	link
Do Not Let Low-Probability Tokens Over-Dominate in RL for LLMs	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link
Token-Regulated Group Relative Policy Optimization for Stable Reinforcement Learning in Large Language Models	Level 1: Data-Centric Perspective	Data Compression	2025	Arxiv	link
Towards Flash Thinking via Decoupled Advantage Policy Optimization	Level 1: Data-Centric Perspective	Data Compression	2025	Arxiv	link
DAST: Difficulty-Adaptive Slow-Thinking for Large Reasoning Models	Level 1: Data-Centric Perspective	Data Compression	2025	EMNLP	link
ThinkPrune: Pruning Long Chain-of-Thought of LLMs via Reinforcement Learning	Level 1: Data-Centric Perspective	Data Compression	2025	TMLR	link
S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models	Level 1: Data-Centric Perspective	Data Compression	2025	NeurIPS	link
Interleaved Reasoning for Large Language Models via Reinforcement Learning	Level 1: Data-Centric Perspective	Data Compression	2025	Arxiv	link
Compressing Chain-of-Thought in LLMs via Step Entropy	Level 1: Data-Centric Perspective	Data Compression	2025	Arxiv	link
Sample-efficient LLM Optimization with Reset Replay	Level 2: Training-Centric Perspective	Policy Optimization	2025	Arxiv	link
Inpainting-Guided Policy Optimization for Diffusion Large Language Models	Level 2: Training-Centric Perspective	Policy Optimization	2025	Arxiv	link
KTO: Model alignment as prospect theoretic optimization	Level 2: Training-Centric Perspective	Policy Optimization	2024	ICML	link
Can Large Reasoning Models Self-Train?	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Test-Time Reinforcement Learning	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Semantic Voting: A Self-Evaluation-Free Approach for Efficient LLM Self-Improvement on Unverifiable Open-ended Tasks	Level 2: Training-Centric Perspective	Reward Engineering	2025	NeurIPS	link
Wisdom of the Crowd: Reinforcement Learning from Coevolutionary Collective Feedback	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
First SFT, Second RL, Third UPT: Continual Improving Multi-Modal LLM Reasoning via Unsupervised Post-Training	Level 2: Training-Centric Perspective	Reward Engineering	2025	NeurIPS	link
Learning to Reason without External Rewards	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
KnowRL: Teaching Language Models to Know What They Know	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Self Rewarding Self Improving	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link

Continued on next page

Table 1 – Continued

Title	Section	Subsection	Year	Venue	Link
Spurious rewards: Rethinking training signals in rlvr	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Surrogate Signals from Format and Length: Reinforcement Learning for Solving Mathematical Problems without Ground Truth Answers	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
No Free Lunch: Rethinking Internal Feedback for LLM Reasoning	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
CoMAS: Co-Evolving Multi-Agent Systems via Interaction Rewards	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Compute as teacher: Turning inference compute into reference-free supervision	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Self-Play Preference Optimization for Language Model Alignment	Level 2: Training-Centric Perspective	Reward Engineering	2025	ICLR	link
Right Question is Already Half the Answer: Fully Unsupervised LLM Reasoning Incentivization	Level 2: Training-Centric Perspective	Reward Engineering	2025	NeurIPS	link
Consistent Paths Lead to Truth: Self-Rewarding Reinforcement Learning for LLM Reasoning	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Evolving language models without labels: Majority drives selection, novelty promotes variation	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Learning to think: Information-theoretic reinforcement fine-tuning for llms	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Semi-supervised Fine-tuning for Large Language Models	Level 2: Training-Centric Perspective	Reward Engineering	2025	*ACL	link
Maximizing Confidence Alone Improves Reasoning	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Confidence Is All You Need: Few-Shot RL Fine-Tuning of Language Models	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
The unreasonable effectiveness of entropy minimization in llm reasoning	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models	Level 2: Training-Centric Perspective	Reward Engineering	2025	Arxiv	link
Training Language Models to Self-Correct via Reinforcement Learning	Level 2: Training-Centric Perspective	Reward Engineering	2025	ICLR	link
ReviewRL: Towards Automated Scientific Review with RL	Level 2: Training-Centric Perspective	Reward Engineering	2025	*ACL	link
AI-powered peer review needs human supervision	Level 2: Training-Centric Perspective	Reward Engineering	2025	Journal of Information, Communication and Ethics in Society	link
AlphaZero-Like Tree-Search can Guide Large Language Model Decoding and Training	Level 2: Training-Centric Perspective	Trajectory Generation	2024	ICML	link
Alphamath almost zero: process supervision without process	Level 2: Training-Centric Perspective	Trajectory Generation	2024	NeurIPS	link
Toward Efficient Exploration by Large Language Model Agents	Level 2: Training-Centric Perspective	Trajectory Generation	2025	ICML Workshop	link

Continued on next page

Table 1 – Continued

Title	Section	Subsection	Year	Venue	Link
Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning	Level 2: Training-Centric Perspective	Trajectory Generation	2025	Arxiv	link
Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning	Level 2: Training-Centric Perspective	Trajectory Generation	2025	Arxiv	link
Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2018	ICLR	link
Asymmetric self-play for automatic goal discovery in robotic manipulation	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2021	Arxiv	link
Self-Questioning Language Models	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
Learning to Pose Problems: Reasoning-Driven and Solver-Adaptive Data Synthesis for Large Reasoning Models	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
Better LLM Reasoning via Dual-Play	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
VisPlay: Self-Evolving Vision-Language Models from Images	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
Search Self-play: Pushing the Frontier of Agent Capability without Supervision	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
SPICE: Self-Play In Corpus Environments Improves Reasoning	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	Arxiv	link
AceSearcher: Bootstrapping Reasoning and Search for LLMs via Reinforced Self-Play	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution	2025	NeurIPS	link
ZeroGUI: Automating Online GUI Learning at Zero Human Cost	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
The Path of Self-Evolving Large Language Models: Achieving Data-Efficient Learning via Intrinsic Feedback	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
Towards Agentic Self-Learning LLMs in Search Environment	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
Absolute Zero: Reinforced Self-play Reasoning with Zero Data	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
An Extremely Data-efficient and Generative LLM-based Reinforcement Learning Agent for Recommenders	Level 3: Framework-Centric Perspective	Self-evolving Framework	2024	Arxiv	link
Self-rewarding correction for mathematical reasoning	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
S2R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
SSR-Zero: Simple Self-Rewarding Reinforcement Learning for Machine Translation	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
Diagnosing and Mitigating System Bias in Self-Rewarding RL	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
SeRL: Self-Play Reinforcement Learning for Large Language Models with Limited Data	Level 3: Framework-Centric Perspective	Self-evolving Framework	2025	Arxiv	link
SPC: Evolving Self-Play Critic via Adversarial Games for LLM Reasoning	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution Frameworks	2025	NeurIPS	link

Continued on next page

Table 1 – Continued

Title	Section	Subsection	Year	Venue	Link
Generative Adversarial Reasoner: Enhancing LLM Reasoning with Adversarial Reinforcement Learning	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution Frameworks	2025	Arxiv	link
Toward Training Superintelligent Software Agents through Self-Play SWE-RL	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution Frameworks	2025	Arxiv	link
Self-playing Adversarial Language Game Enhances LLM Reasoning	Level 3: Framework-Centric Perspective	Asymmetric Co-Evolution Frameworks	2025	Arxiv	link
SPIRAL: Self-Play on Zero-Sum Games Incentivizes Reasoning via Multi-Agent Multi-Turn Reinforcement Learning	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
Vision-Zero: Scalable VLM Self-Improvement via Strategic Gamified Self-Play	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
Propose, Solve, Verify: Self-Play Through Formal Verification	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
SPELL: Self-Play Reinforcement Learning for evolving Long-Context Language Models	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
Multi-Agent Evolve: LLM Self-Improve through Co-evolution	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
Meta-Learning Reinforcement Learning for Crypto-Return Prediction	Level 3: Framework-Centric Perspective	Multi-Agent Evolution Frameworks	2025	Arxiv	link
Towards Understanding Self-play for LLM Reasoning	Level 3: Framework-Centric Perspective	Discussion	2025	Arxiv	link
OpenAI o1 System Card	Introduction	Background	2024	Arxiv	link
SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild	Introduction	Background	2025	COLM	link
Kimi k1.5: Scaling Reinforcement Learning with LLMs	Introduction	Background	2025	Arxiv	link
Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model	Introduction	Background	2025	NeurIPS	link
A survey of reinforcement learning for large reasoning models	Survey	Survey	2025	Arxiv	link
The Landscape of Agentic Reinforcement Learning for LLMs: A Survey	Survey	Survey	2025	Arxiv	link
A Survey on Efficient Large Language Model Training From Data-centric Perspectives	Survey	Survey	2025	ACL	link
A Survey on Self-Evolution of Large Language Models	Survey	Survey	2024	Arxiv	link
A Comprehensive Survey of Self-Evolving AI Agents: A New Paradigm Bridging Foundation Models and Lifelong Agentic Systems	Survey	Survey	2025	Arxiv	link
A Survey of Self-Evolving Agents: What, When, How, and Where to Evolve on the Path to Artificial Super Intelligence	Survey	Survey	2026	Arxiv	link