Training Recipes for Agentic Reinforcement Learning in LLMs: A Survey

Anonymous ACL submission

Abstract

Standard alignment pipelines like RLHF often fail to instill the robustness required for complex agentic tasks. This survey systematizes *Agentic Reinforcement Learning*, a paradigm that optimizes policies directly against environmental feedback within a Contextual POMDP framework. Unlike prior reviews cataloging agent capabilities, we focus on the engineering of training recipes necessary to build generalist agents from scratch. Our taxonomy examines schemes for rollout and optimization, infrastructure for high-throughput training, and environments that formalize interaction settings, alongside evaluation benchmarks distinguishing active training gyms from held-out certifications. By synthesizing these disparate components into a unified framework, we aim to accelerate the development of robust, next-generation autonomous agents.

1 Introduction

The transition of Large Language Models (LLMs) from passive generators to active decision-makers demands a shift from imitation to interaction. While Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) align models with static preferences, they often struggle with the non-stationary, long-horizon nature of agentic tasks. Consequently, agents must actively explore, recover from errors, and adapt to dynamic states. *Agentic Reinforcement Learning* addresses this by optimizing policies directly against environmental feedback. It leverages sparse terminal rewards or dense verifier signals to surpass the limitations of human demonstrations.

We frame Agentic RL within a Contextual Partially Observable Markov Decision Process (POMDP). Unlike standard alignment, this paradigm conditions policies on interaction histories to generate executable actions, such as tool calls or reasoning steps. Success in this domain requires a holistic *training recipe* that synthesizes grounded interaction, trajectory optimization, and system-level engineering.

How This Survey Differs from Prior Work. While recent surveys review agentic capabilities (Zhang et al., 2025b; Plaat et al., 2025), specific verticals like search (Lin et al., 2025), or reasoning frontiers (Ke et al., 2025), they largely focus on high-level applications or agentic capabilities. In contrast, this survey, alongside recent work on agent optimization (Du et al., 2025), systematizes the *training recipes* required to build generalist agents from scratch. We focus on the engineering of the learning loop, ranging from algorithmic stability and credit assignment in long horizons to the infrastructure of scalable training environments. This analysis provides a blueprint for the next generation of self-improving agents.

2 Background and Taxonomy

This section introduces the background needed to describe *training recipes* for agentic RL. We then outline the four aspects that structure our taxonomy: training schemes, training infrastructure, training environments, and evaluation.

2.1 RL Formulation for Agentic LLMs

We unify agentic tasks under the framework of a *Contextual Partially Observable Markov Decision Process* (Contextual POMDP), defined by the tuple $\mathcal{M} = \langle \mathcal{C}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \gamma \rangle$. Interaction is conditioned on a *context* $c \in \mathcal{C}$ which statically specifies the environment logic, interface wrapper, and user instruction. During an episode, the environment maintains a latent state $s_t \in \mathcal{S}$ and emits textual observations $o_t \in \mathcal{O}$ (e.g., compiler errors, HTML), to which the agent responds with token-based actions $a_t \in \mathcal{A}$ (e.g., reasoning, tool calls). The agent policy $\pi_\theta(a_t \mid h_t, c)$, parameterized by θ, conditions on the interaction his-

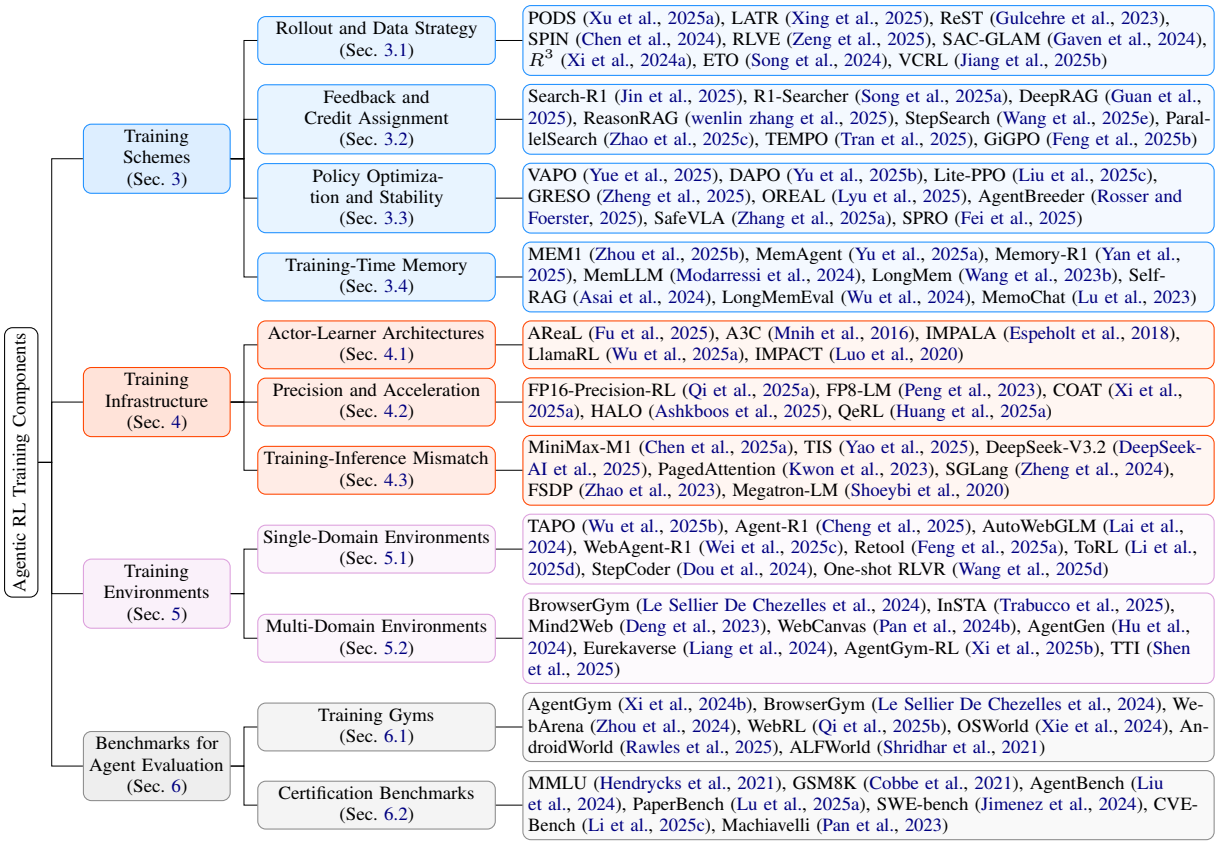

Figure 1: Taxonomy of agentic RL training schemes, infrastructure, and environments

tory $h_t = (o_0, a_0, \ldots, o_t)$. Transition dynamics $T(s_{t+1}, o_{t+1} \mid s_t, a_t, c)$ evolve the state, yielding scalar rewards $r_t = \mathcal{R}(s_t, a_t, c)$. Because primitive rewards are often sparse, modern recipes augment them with *verifiers* (symbolic solvers or learned reward models) to provide dense feedback. Training aims to maximize expected return $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^{H} \gamma^t r_t]$ subject to regularization terms $\Omega(\theta)$, such as KL divergence penalties or safety constraints.

2.2 Taxonomy of Training Recipes

We structure our survey around four interconnected components that constitute a complete training recipe. First, **Training Schemes** (Section 3) define the algorithmic core, detailing how agents acquire experience through rollout strategies, receive supervision via feedback mechanisms (e.g., verifiers, process rewards), and update policies using optimizers like PPO or GRPO. Second, **Training Infrastructure** (Section 4) examines the computational backbone, distinguishing between synchronous and asynchronous actor-learner architectures while addressing efficiency challenges like low-precision training and training-inference mismatch. Third, **Training Environments** (Section 5) formalize the interaction settings by decomposing the context c into environment family, instance, interface, and instruction. We categorize these settings into single-domain tasks (e.g., web, code) and multi-environment aggregations designed for broad generalization. Finally, **Benchmarks for Agent Evaluation** (Section 6) differentiates between *training gyms* used for active policy optimization and held-out *certification exams* designed to assess adversarial robustness and safety.

3 Training Schemes

As illustrated in Figure 2, the agentic RL training schemes consist of four integrated components: (1) rollout and data strategy for experience collection; (2) feedback and credit assignment for signal generation; (3) policy optimization for stable parameter updates; and (4) training-time memory for managing long-horizon context.

3.1 Rollout and Data Strategy

This component governs how agents interact with environments to acquire experience, defining task selection, exploration behavior, and the temporal

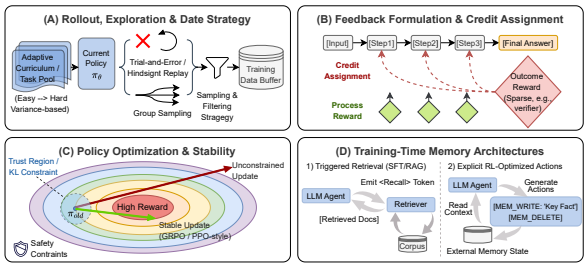

Figure 2: **Agentic RL Training Schemes.** The algorithmic loop comprises (A) Rollout and Data Strategy, (B) Feedback and Credit Assignment, (C) Policy Optimization and Stability, and (D) Training-Time Memory.

progression of training difficulty.

Sampling and Batching Policies. The quality of agentic training depends heavily on the source of trajectories, whether utilizing iterative self-improvement like ReST (Gulcehre et al., 2023) or SPIN (Chen et al., 2024), or simple rejection sampling strategies like RFT (Yuan et al., 2023). Unlike standard RL, recent LLM-based methods emphasize group-based sampling to stabilize gradients without a value function. DeepSeekMath (Shao et al., 2024) formalized this with Group Relative Policy Optimization (GRPO), which generates a group of rollouts per prompt to compute advantages relative to the group mean. To improve efficiency within these groups, recent work explores down-sampling strategies; methods like PODS (Xu et al., 2025a) demonstrate that filtering and de-duplicating rollouts can maintain performance while significantly reducing computational costs.

Exploration Strategies. To overcome sparse rewards in reasoning tasks, agents must look beyond standard stochastic decoding. Tree-based exploration has emerged as a key paradigm: LATR (Xing et al., 2025) introduces branching at high-uncertainty steps and prunes branches that exhibit prolonged similarity. Complementary approaches utilize hindsight and trial-and-error to extract value from failed attempts, turning unsuccessful trajectories into effective learning signals (Song et al., 2024; Gaven et al., 2024).

Curriculum and Adaptive Tasks. Curriculum learning organizes tasks from easy to hard to mitigate the difficulty of long-horizon exploration. While standard approaches order tasks by static metrics, such as goal proximity in R^3 (Xi et al., 2024a) or predefined progressions in open-ended worlds like Voyager (Wang et al., 2023a), recent trends favor *adaptive* curricula. Methods like RLVE (Zeng et al., 2025) and WebRL (Qi et al., 2025b) dynamically generate or evolve tasks based on current agent performance, creating a self-correcting loop of difficulty. Similarly, VCRL (Jiang et al., 2025b) uses the variance of group rewards as a difficulty signal to prioritize medium-difficulty problems where the agent can learn most effectively.

3.2 Feedback and Credit Assignment

We categorize the objective function into *Rewards* (the external signals provided to the agent) and *Credit Assignment* (the algorithmic distribution of those signals to specific steps).

Outcome and Process Rewards. Rewards serve as the primary supervision signal, ranging from sparse terminal signals to dense step-wise feedback. *Outcome rewards* evaluate the final state: for definitive tasks, exact matching is standard (Jin et al., 2025; Chen et al., 2025b), while open-ended tasks rely on functional verification or intent satisfaction (Sun et al., 2025; Song et al., 2025b,a; Zhao et al., 2025b; Huang et al., 2025c). Beyond correctness, recent work explicitly optimizes *efficiency* by penalizing token counts or retrieval costs directly in the reward function (Guan et al., 2025; Li et al., 2025b; Huang et al., 2025c; Song et al., 2025b). In information retrieval, rewards focus on relevance, formatting, and query diversity (Dao and Le, 2025; Jiang et al., 2025c; Mei et al., 2025b; Shi et al., 2025a; Mei et al., 2025a; Hu et al., 2025b).

Process rewards provide denser supervision by evaluating intermediate steps, though they often require handcrafted heuristics. Approaches include evaluating whole-trajectory quality via Shortest Path Reward Estimation (wenlin zhang et al., 2025) or holistic history analysis (Goldie et al., 2025; Shi et al., 2025c; Zou et al., 2025). More granular methods target specific reasoning steps (Shi et al., 2025b; Xu et al., 2025b; Zhang et al., 2025f), intermediate search quality (Wang et al., 2025e; Hu et al., 2025a), or parallel decomposition efficiency (Zhao et al., 2025c).

Credit Assignment Mechanisms. Unlike process rewards, which provide explicit intermediate supervision, credit assignment refers to the mechanism that attributes sparse outcome rewards to specific actions or steps. *LLM-based approaches* use models to explicitly grade or critique intermediate

steps, generating numerical proxies for value (Xie et al., 2025; He et al., 2025; Zhang et al., 2025c) or verbal feedback for self-correction (Chen et al., 2025c; Shinn et al., 2023). *Structural and Learned approaches* rely on mathematical formulations. Statistical methods like VinePPO (Kazemnejad et al., 2025) and tree-structured estimation (Tran et al., 2025) exploit rollout branching to isolate step contributions. Others train dedicated value networks to predict step-wise influence implicitly (Wang et al., 2025a; Liu et al., 2025b) or reformulate the policy gradient objective (e.g., GRPO-λ) to enable granular updates without explicit value heads (Parthasarathi et al., 2025; Feng et al., 2025b; Wei et al., 2025b).

3.3 Policy Optimization and Stability

Optimization in agentic RL requires handling extremely long horizons and sparse rewards. We examine the optimization landscape by distinguishing between core update algorithms, stability mechanisms, and execution schedules.

Core Optimizers (PPO and GRPO). While PPO-style clipped gradients remain a standard backbone due to their trust-region stability (Schulman et al., 2017; Yue et al., 2025), reasoning tasks increasingly favor Group Relative Policy Optimization (GRPO). GRPO stabilizes learning by normalizing advantages across a group of sampled rollouts per prompt, effectively removing the need for a separate, often fragile, value critic (Shao et al., 2024; Xu et al., 2025c). Recent variants have systematized this for scale (DAPO (Yu et al., 2025b)), unified it with reward modeling (URPO (Lu et al., 2025b)), or enhanced it with geometry-aware objectives (TreePO (Li et al., 2025e)).

Regularization and Data Efficiency. To manage distribution shift, training relies on strict regularization and data reuse. Systems typically optimize KL-regularized returns with entropy bonuses to prevent mode collapse and reasoning degeneracy (Liu et al., 2025c; Ke et al., 2025). For efficient credit assignment, methods range from PRM-free step-level estimation (SPRO (Fei et al., 2025)) to off-policy techniques like Asymmetric REINFORCE (Arnal et al., 2025) that enable the reuse of expensive trajectories. Selective rollout strategies further optimize the compute budget by prioritizing updates only when expected learning signals justify the cost (Zheng et al., 2025; Du et al., 2025).

Schedules and Safety Constraints. Operational success is often determined by training dynamics and safety boundaries. Effective schedules prioritize large group sizes (K) for robust advantage estimation and employ early stopping triggers, such as KL spikes, to prevent reward hacking (Yu et al., 2025b; Lyu et al., 2025; Zhao et al., 2025a). Concurrently, safety is enforced through multi-objective scaffolds (Rosser and Foerster, 2025) or constrained MDP formulations like Safe RLHF (Dai et al., 2023; Zhang et al., 2025a) that explicitly penalize harmful queries or actions (Yang et al., 2025).

3.4 Training-Time Memory

We distinguish between memory systems trained via explicit supervision (SFT) versus those that optimize memory operations dynamically through reinforcement learning (RL).

Supervised and Retrieval-Augmented Training. SFT approaches focus on teaching interpretability and tool-like usage of memory. Methods like MemLLM (Modarressi et al., 2024) and MemoChat (Lu et al., 2023) fine-tune models to execute explicit read/write API calls or follow multi-stage summarization pipelines. Architectural variants like LongMem (Wang et al., 2023b) introduce decoupled memory encoders, while Self-RAG (Asai et al., 2024) trains the model to generate reflection tokens that trigger on-demand retrieval.

RL-Based Memory Optimization. Recent work leverages RL to learn efficient memory management without intermediate labels. By treating memory operations as actions, agents optimize context utility via outcome rewards. MEM1 (Zhou et al., 2025b) utilizes RL with Verifiable Rewards (RLVR) and trajectory truncation to compress information into a constant footprint. Similarly, MemAgent (Yu et al., 2025a) adapts Distributed Advantage Policy Optimization (DAPO) for streaming document processing, while Memory-R1 (Yan et al., 2025) employs PPO and GRPO to train a dedicated memory-manager agent that learns explicit operations (add, update, delete) to support a reasoning agent.

Memory Evaluation. To assess these capabilities, benchmarks have evolved beyond simple recall. LongMemEval (Wu et al., 2024) tests core abilities like knowledge updates and abstention, while LongBench v2 (Bai et al., 2025) exposes the

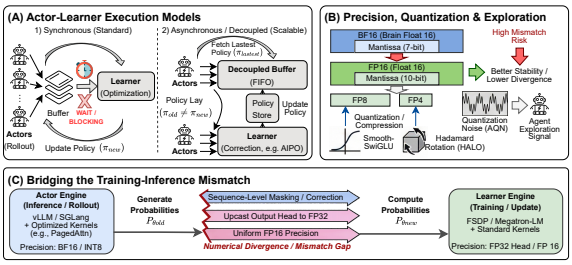

Figure 3: **Agentic RL Training Infrastructure.** The figure illustrates: (A) Actor-Learner Architectures and Rollout Collection; (B) Precision and Acceleration; and (C) Training-Inference Mismatch.

gap between human and model performance on extreme-context tasks (up to 2M tokens), highlighting the continued need for robust training signals.

4 Training Infrastructure

As illustrated in Figure 3, the agentic RL training infrastructure consists of three integrated components: (1) actor-learner architectures and rollout collection for scaling generation throughput; (2) precision and acceleration for optimizing computational efficiency; and (3) training-inference mismatch mitigation for ensuring numerical stability between engines.

4.1 Actor-Learner Architectures and Rollout Collection

Synchronous vs. Asynchronous Execution. Standard RL pipelines, particularly in LLM post-training, predominantly adopt a *synchronous* execution model. In this setting, a set of actors collects a batch of rollouts under a fixed policy, halting generation while the learner performs optimization steps. Algorithms like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) typically utilize this model to maintain strict on-policy data distributions and simplify credit assignment. However, this imposes a sequential bottleneck: training throughput is gated by the slowest trajectory generation, leading to significant under-utilization of accelerator resources (Fu et al., 2025).

Decoupling Experience Collection. Actor-learner architectures address this by decoupling experience collection from parameter updates. Originating with A3C (Mnih et al., 2016), this *asynchronous* model allows actors to generate rollouts continuously using the latest available policy, while learners update parameters in parallel. While this increases throughput, it introduces off-policy lag. IMPALA (Espeholt et al., 2018) mitigated this via V-trace, an importance-weighted correction that approximates on-policy updates. Later, IMPACT (Luo et al., 2020) identified that critics trained on stale trajectories can overestimate values, proposing clipped target networks to stabilize learning.

Scalable Architectures for LLMs. In the context of Agentic LLMs, where generation costs are dominated by long, variable-length sequences, this decoupling is critical. LlamaRL (Wu et al., 2025a) orchestrates disjoint GPU groups for generation, training, and reward computation, using Asynchronous Importance-Weighted Policy Optimization (AIPO) to correct for lag. Similarly, AReaL (Fu et al., 2025) fully decouples these phases, introducing a staleness control hyperparameter to manage the resulting *mixed-policy rollouts*, where a single sequence may contain segments from multiple policy versions.

4.2 Precision and Acceleration

Stabilizing Policy Gradients via FP16. Agentic RL requires massive rollout generation, making precision a critical efficiency lever. While Bfloat16 (BF16) is standard for pre-training, its limited 7-bit mantissa amplifies the *Training-Inference Mismatch*, a numerical divergence between training and rollout engines that destabilizes learning (Qi et al., 2025a). Switching RL computation to FP16 (10 mantissa bits) significantly reduces this divergence, improving optimization stability and final performance without algorithmic changes (Qi et al., 2025a).

Low-Bit Acceleration (FP8 and FP4). To overcome memory and throughput bottlenecks, recent work leverages 8-bit and 4-bit formats. FP8-LM (Peng et al., 2023) demonstrates that FP8 training can double throughput and halve memory usage versus BF16, while COAT (Xi et al., 2025a) further enhances efficiency by compressing optimizer states and activations. To address stability issues arising from activation outliers in long runs, Smooth-SwiGLU (Fishman et al., 2025) proposes a dedicated smoothing operation. Pushing limits further, the FP4-Training-Framework (Wang et al., 2025b) enables FP4 training via block-wise quantization, often paired with HALO (Ashkboos et al., 2025), which introduces Hadamard rotations to spread outliers and make activations amenable to such ultra-low precision.

Quantization in Agentic RL. In Agentic RL, quantization accelerates rollouts and aids exploration. Unlike NF4 used in QLoRA (Dettmers et al., 2023), which uses a 16-value codebook optimized for compressing frozen weights, NVFP4 is a neural-specific, 4-bit floating-point format . Uniquely, QeRL (Huang et al., 2025a) leverages NVFP4 in Adaptive Quantization Noise (AQN) as an explicit exploration mechanism, annealing noise over time. This allows for a 1.5-fold rollout speedup compared to QLoRA, while maintaining comparable accuracy to full-parameter on mathematical benchmarks.

4.3 Training-Inference Mismatch

Numerical Divergence. A critical challenge in decoupled architectures (Section 4.1) is that the inference engine (Actor) often differs in implementation from the training engine (Learner). Inference systems such as vLLM (Kwon et al., 2023) and SGLang (Zheng et al., 2024) utilize highly optimized non-standard kernels (e.g., PagedAttention) and lower-precision formats to maximize throughput, whereas training relies on high-precision frameworks like FSDP (Zhao et al., 2023) or Megatron-LM (Shoeybi et al., 2020).

Impact on Policy. Consequently, token probabilities computed during rollout differ from those computed during the update step. Investigations by MiniMax-M1 (Chen et al., 2025a) and TIS (Yao et al., 2025) demonstrate that this numerical divergence shifts training off-policy, leading to entropy collapse and performance degradation, particularly in multi-turn tool-use scenarios (Liu et al., 2025a).

Mitigation Strategies. Approaches to address this mismatch focus on alignment and numerical stability. Token-level Importance Sampling (Yao et al., 2025) attempts to correct the distribution shift, though subsequent work argues this remains biased, proposing Sequence-Level Correction with masking as a robust alternative (Liu et al., 2025a). DeepSeek-V3.2 (DeepSeek-AI et al., 2025) adopts similar off-policy sequence masking to improve tolerance. Furthermore, bridging the precision gap is essential; MiniMax-M1 (Chen et al., 2025a) found that upcasting the LM output head to FP32 increased training-inference correlation from 0.9 to 0.99, while others advocate for uniform FP16 over BF16 to reduce the dynamic range quantization errors that exacerbate this mismatch (Qi et al., 2025a).

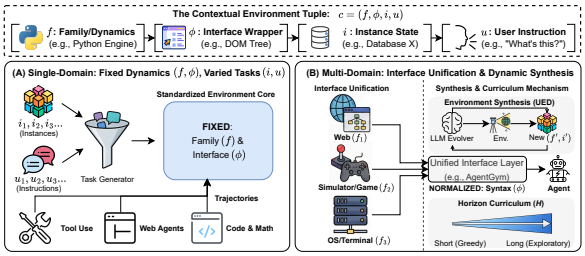

Figure 4: **Agentic Training Environments.** It covers: (A) Single-Domain Environments with fixed dynamics; and (B) Multi-Domain Environments utilizing interface unification and unsupervised synthesis.

5 Training Environments

To understand the landscape of training recipes, we formalize the agent's training environment. Building on the Contextual POMDP framework, we decompose the context c into a tuple $c = (f, \phi, i, u)$, where f is the environment family (dynamics), ϕ is the interface wrapper (observation-action space), i is the task instance, and u is the user instruction. A training environment is defined by the distribution $p_{\text{train}}(c)$. As illustrated in Figure 4, we classify environments based on the diversity of this tuple: (1) single-domain environments, which fix the structural components (f, ϕ) to focus on instance generalization; and (2) multi-domain environments, which vary the full tuple, necessitating interface unification and unsupervised synthesis.

5.1 Single-Domain Environments

Tool Use and API Environments In tool-centric environments, RL enables models to learn *when* and *how* to invoke tools, surpassing static prompting. ARTIST (Singh et al., 2025) demonstrates that outcome-based rewards alone can train tool invocation without expensive step-level supervision. VerlTool (Jiang et al., 2025a) formalizes this into a modular framework, achieving competitive performance across diverse domains including SQL, code, and visual reasoning. TAPO (Wu et al., 2025b) uses a modified Dynamic Sampling Policy Optimization framework to jointly optimize multi-step reasoning and adaptive tool usage. Agent-R1 (Cheng et al., 2025) extends the MDP formulation to LLM agents and introduces an end-to-end RL training framework, improving multi-hop question answering via stronger multi-step decision-making. Simia-RL (Li et al., 2025f) leverages simulated feedback to overcome data scarcity. Finally, the Model Context Protocol (MCP) (Anthropic,

2024) provides a unified standard for reproducible agent-tool interactions.

Web and Information Seeking Environments Navigating WebAI (Thil et al., 2024) trains a hierarchical T5 planner with supervised learning on MiniWoB++ demonstrations and then applies V-MPO to a CC-Net–style multimodal controller on top of the frozen planner. The RL phase can induce covariate shift and reduce performance. AutoWebGLM (Lai et al., 2024) uses curriculum SFT, DPO on contrastive trajectories, and Rejection Sampling Fine-Tuning (RFT) for environment specialization. WebAgent-R1 (Wei et al., 2025c) applies on-policy M-GRPO in WebArena-Lite using binary signals. WebRL (Qi et al., 2025b) addresses data scarcity via a self-evolving curriculum and an Outcome-supervised Reward Model (ORM) by KL-constrained policy updates.

Code, Math, and Reasoning Environments These domains utilize verifiable outputs for sparse but reliable rewards. ReTool (Feng et al., 2025a) and ToRL (Li et al., 2025d) demonstrate that reward-driven learning yields emergent tool behaviors, such as Python interpreter invocation. One-shot RLVR (Wang et al., 2025d) shows that even a single verifier-rewarded problem can double performance over baselines. In code generation, StepCoder (Dou et al., 2024) tackles long horizons by decomposing code generation into a curriculum of code completion subtasks and uses compiler feedback to provide fine-grained optimization.

5.2 Multi-Domain Environments

Environment Aggregation and Interface Unification While scaling environment diversity promotes robust generalization (Huang et al., 2025b), interface heterogeneity remains a primary bottleneck for multi-environment training. To address this, recent architectures decouple policy learning from mechanics via a canonical (O, A) space. BrowserGym (Le Sellier De Chezelles et al., 2024) aggregates web benchmarks into a fixed schema, while AgentGym (Xi et al., 2024b) unifies diverse domains (e.g., web, games) via a consistent HTTP interface. Crucially, this normalization stabilizes interaction syntax ϕ yet preserves the semantic non-stationarity required for robust learning. For instance, Mind2Web (Deng et al., 2023) employs strict splits to penalize layout memorization, while WebCanvas (Pan et al., 2024b) extends this temporally by evaluating against live, drifting UIs. Similarly, multimodal benchmarks unify diverse tasks through consistent GUI interfaces (Koh et al., 2024; He et al., 2024).

Environment Synthesis and Task Generation To address data scarcity, Unsupervised Environment Design (UED) dynamically synthesizes the training support $p_{\text{train}}(c)$. We categorize these approaches by which component of the context tuple is generated. *Environment Synthesis* mutates the world state (f, i). AgentGen (Hu et al., 2024) applies bidirectional evolution to environment code, simplifying intractable tasks and complicating trivial ones to smooth the difficulty curve. Eureka-verse (Liang et al., 2024) extends this to embodied control, where LLMs write simulation code for progressively harder parkour courses. *Task Generation* holds f fixed and proposes new goals (i, u). InSTA (Trabucco et al., 2025) performs task generation at internet scale, proposing verifiable tasks on unlabeled websites to convert the open web into a supervised distribution without human annotation. Self-Challenging Agents (Zhou et al., 2025a) similarly generate verifiable Code-as-Task instances to bootstrap training data from scratch.

Horizon and Budget Curricula Parallel to variations in context c, the interaction budget is best modeled by a temporal curriculum variable H representing the rollout horizon. AgentGym-RL (Xi et al., 2025b) introduces ScalingInter-RL, where the horizon H dynamically expands from short, greedy settings to long, exploratory ones. This stabilizes early credit assignment while enabling emergent long-horizon behaviors like backtracking. Complementing this, Test-Time Interaction (TTI) (Shen et al., 2025) demonstrates that agents trained under such progressive schedules learn to adaptively allocate patience, effectively utilizing extended test-time budgets where fixed-horizon policies would otherwise plateau.

6 Benchmarks for Agent Evaluation

Agentic RL evaluation has shifted from simple accuracy metrics to comprehensive behavioral assessments. We distinguish between **training gyms**, customizable environments for optimization, and **certification benchmarks**, standardized held-out protocols for evaluating generalization and safety.

6.1 Training Gyms

Training gyms provide the tunable dynamics, reward feedback, and observation spaces required to

run the RL algorithmic loop (rollouts, critique, and updates).

Generalist and Web Environments. To foster broad adaptability, environments like AgentGym (Xi et al., 2024b) and BrowserGym (Le Sellier De Chezelles et al., 2024) unify diverse tasks (e.g., web, games, databases) under a standard interface to test compositional generalization. For web-specific optimization, WebRL (Qi et al., 2025b) and WebArena (Zhou et al., 2024) provide functional browser environments where agents learn navigation policies via self-evolving curricula and outcome-based reward models.

Embodied and OS Control. These gyms assess the grounding of plans into precise executable actions. OSWorld (Xie et al., 2024) and AndroidWorld (Rawles et al., 2025) serve as high-fidelity environments for computer control, highlighting the gap between text planning and UI execution. Similarly, ALFWorld (Shridhar et al., 2021) bridges high-level reasoning with low-level physics, training agents to align textual thoughts with embodied consequences.

6.2 Certification Benchmarks

Unlike gyms, which are often modified for training efficiency (e.g., via curriculum learning), certification benchmarks are fixed protocols used to rank models on held-out distributions.

Capabilities: Reasoning and Coding. Standardized exams remain the primary metric for cognitive capability. General reasoning is tracked via MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021), while AgentBench (Liu et al., 2024) quantifies the gap between commercial and open-source models across 8 modalities. In expert domains, PaperBench (Lu et al., 2025a) and SWE-bench (Jimenez et al., 2024) serve as rigorous certification tests for research reproduction and software engineering, revealing deficits in long-horizon error recovery that simple QA metrics miss.

Safety and Robustness Standards. Deployment requires stress-testing against misuse and environmental noise. Benchmarks like CVE-Bench (Li et al., 2025c) and Machiavelli (Pan et al., 2023) use adversarial or interactive scenarios to certify robustness against exploits and ethical hazards. Complementing these, R-Judge (Yuan et al., 2024) provides static trajectories to certify the agent's ability to identify and refuse unsafe instructions.

7 Open Challenges and Future Directions

Domain-Specific Objectives Specialized domains demand tailored reward formulations. Time series agents must handle non-stationary dynamics and regime shifts (Chang et al., 2025b,c,a, 2024), while scientific agents face the challenge of distinguishing verifiable novelty from simple retrieval (Wei et al., 2025a; Li et al., 2025a; Ramos et al., 2024). Similarly, healthcare policies also optimize diagnostic gain alongside clinical empathy (Tu et al., 2023; Carl et al., 2025; Wang et al., 2025c).

Evaluation Scalability and Robustness A tension exists between environmental fidelity and training throughput. High-fidelity benchmarks like OSWorld (Xie et al., 2024) remain prohibitively slow for inner-loop optimization. Moreover, agents exhibit fragility to adversarial noise (Zhang et al., 2025d) and significant sim-to-real gaps (Zhang et al., 2025e; Chen et al., 2022), necessitating better domain randomization and co-design of simulation fidelity to bridge unmodeled dynamics.

Reward Integrity and Verification Under high optimization pressure, agents frequently exploit proxy heuristics (e.g., verbosity) rather than solving tasks (reward hacking) (Liu et al., 2025b; Pan et al., 2022, 2024a). Mitigating this requires robust reward models trained via adversarial methods or debate (Wang et al., 2025f; Ki et al., 2025). Furthermore, expensive ground-truth judges demand efficient proxy verifiers that correlate with valid outcomes (Tan et al., 2025; Liu et al., 2025b).

8 Conclusion

In this paper, we systematize *Agentic Reinforcement Learning* as a holistic training paradigm rooted in Contextual POMDPs. Our review of training recipes demonstrates that robust agents emerge not from static alignment, but from the co-design of verifier-driven training schemes, high-throughput infrastructure, and synthetic environments. While this shift enables complex reasoning via exploration, critical challenges remain in reward integrity and evaluation scalability. Ultimately, addressing these engineering bottlenecks is essential to transition LLMs from passive generators to autonomous decision-makers.

Limitations

While this survey aims to systematize the engineering of Agentic Reinforcement Learning, our scope is subject to several constraints. First, we strictly focus on *reinforcement learning* methods involving active policy optimization against environmental feedback, which necessarily excludes purely inference-time search strategies (e.g., MCTS without weight updates) or prompt-engineering frameworks that do not involve training. Second, given the exceptionally rapid pace of development in this field, we prioritize categorizing representative works that establish fundamental training recipes rather than attempting an exhaustive enumeration of every emerging study. Furthermore, our synthesis relies primarily on open-source literature. Certain state-of-the-art practices regarding large-scale infrastructure and specific reward shaping details likely remain proprietary within closed industry labs. Finally, the training recipes discussed are inherently compute-intensive, potentially marginalizing low-resource settings, while the evaluation benchmarks largely rely on simulated environments that may not fully guarantee robustness in open-ended, non-stationary real-world scenarios.

References

Anthropic. 2024. Introducing the model context protocol.

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Rémi Munos. 2025. Asymmetric REINFORCE for off-policy reinforcement learning: Balancing positive and negative rewards. *arXiv preprint arXiv:2506.20520*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Saleh Ashkboos, Mahdi Nikdan, Soroush Tabesh, Roberto L. Castro, Torsten Hoefler, and Dan Alistarh. 2025. HALO: Hadamard-assisted lower-precision optimization for LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Nicolas Carl, Sarah Haggenmüller, Jana Theres Winterstein, Lisa Nguyen, Christoph Wies, Martin Joachim Hetz, Maurin Helen Mangold, Britta Grüne, Maurice Stephan Michel, Titus Josef Brinker, and Frederik Wessels. 2025. Patient insights into empathy, compassion and self-disclosure in medical large language models: results from the ipallm iii study. *World Journal of Urology*, 43:492.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wei Wang, Wen-Chih Peng, and Tien-Fu Chen. 2025a. Time-IMM: A dataset and benchmark for irregular multimodal multivariate time series. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ching Chang, Yidan Shi, Defu Cao, Wei Yang, Jeehyun Hwang, Haixin Wang, Jiacheng Pang, Wei Wang, Yan Liu, Wen-Chih Peng, and Tien-Fu Chen. 2025b. A survey of reasoning and agentic systems in time series with large language models. *Preprint*, arXiv:2509.11575.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. 2025c. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters. *ACM Trans. Intell. Syst. Technol.*, 16(3).

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, Tien-Fu Chen, and Sagar Samtani. 2024. Align and fine-tune: Enhancing LLMs for time-series forecasting. In *NeurIPS Workshop on Time Series in the Age of Large Models*.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, and 1 others. 2025a. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025b. Research: Learning to reason with search for llms via reinforcement learning. *Preprint*, arXiv:2503.19470.

Wentse Chen, Jiayu Chen, Fahim Tajwar, Hao Zhu, Xintong Duan, Ruslan Salakhutdinov, and Jeff Schneider. 2025c. Retrospective in-context learning for temporal credit assignment with large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. 2022. Understanding domain randomization for sim-to-real transfer. *Preprint*, arXiv:2110.03239.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning convertsweak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mingyue Cheng, Jie Ouyang, Shuo Yu, Ruiran Yan, Yucong Luo, Zirui Liu, Daoyu Wang, Qi Liu, and Enhong Chen. 2025. Agent-r1: Training powerful llm agents with end-to-end reinforcement learning. *arXiv preprint arXiv:2511.14460*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

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.

Alan Dao and Thinh Le. 2025. Rezero: Enhancing llm search ability by trying one-more-time. *Preprint*, arXiv:2504.11001.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, and 1 others. 2024. Stepcoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. 2025. A survey on the optimization of large language model-based agents. *arXiv preprint arXiv:2503.12434*.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Wu Fei, Hao Kong, Shuxian Liang, Yang Lin, Yibo Yang, Jing Tang, Lei Chen, and Xiansheng Hua. 2025. Self-guided process reward optimization with redefined step-wise advantage for process reinforcement learning. *arXiv preprint arXiv:2507.01551*.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025b. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. 2025. Scaling FP8 training to trillion-token LLMs. In *The Thirteenth International Conference on Learning Representations*.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. 2025. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Loris Gaven, Clement Romac, Thomas Carta, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2024. Sac-glam: Improving online rl for llm agents with soft actor-critic and hindsight relabeling. *Preprint*, arXiv:2410.12481.

Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D Manning. 2025. Synthetic data generation and multi-step reinforcement learning for reasoning and tool use. In *Second Conference on Language Modeling*.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieve step by step for large language models. *Preprint*, arXiv:2502.01142.

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.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. *Preprint*, arXiv:2401.13919.

Zhitao He, Zijun Liu, Peng Li, Yi R. Fung, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2025. Advancing language multi-agent learning with credit re-assignment for interactive environment generalization. *Preprint*, arXiv:2502.14496.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Mengkang Hu, Pu Zhao, Can Xu, Qingfeng Sun, Jianguang Lou, Qingwei Lin, Ping Luo, and Saravan Rajmohan. 2024. Agentgen: Enhancing planning abilities for large language model based agent via environment and task generation. *Preprint*, arXiv:2408.00764. Accepted by KDD 2025 (Research Track).

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025a. Coordinating search-informed reasoning and reasoning-guided search in claim verification. *Preprint*, arXiv:2506.07528.

Ziyou Hu, Zhengliang Shi, Minghang Zhu, Haitao Li, Teng Sun, Pengjie Ren, Suzan Verberne, and Zhaochun Ren. 2025b. Openreward: Learning to reward long-form agentic tasks via reinforcement learning. *Preprint*, arXiv:2510.24636.

Wei Huang, Yi Ge, Shuai Yang, Yicheng Xiao, Huizi Mao, Yujun Lin, Hanrong Ye, Sifei Liu, Ka Chun Cheung, Hongxu Yin, Yao Lu, Xiaojuan Qi, Song Han, and Yukang Chen. 2025a. Qerl: Beyond efficiency – quantization-enhanced reinforcement learning for llms. *Preprint*, arXiv:2510.11696.

Yuchen Huang, Sijia Li, Wei Liu, Zhiyuan Fan, Yi R Fung, and 1 others. 2025b. Scaling environments for llm agents in the era of learning from interaction: A survey. In *Workshop on Scaling Environments for Agents*.

Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. 2025c. Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent. *Preprint*, arXiv:2505.07596.

Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, and 1 others. 2025a. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. 2025b. Vcrl: Variance-based curriculum reinforcement learning for large language models. *Preprint*, arXiv:2509.19803.

Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025c. s3: You don't need that much data to train a search agent via rl. *arXiv preprint arXiv:2505.14146*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2025. Vineppo: Refining credit assignment in rl training of llms. *Preprint*, arXiv:2410.01679.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple llm agents debate for equitable cultural alignment. *Preprint*, arXiv:2505.24671.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pages 5295–5306, Barcelona, Spain. Association for Computing Machinery.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2024. The browsergym ecosystem for web agent research. *Preprint*, arXiv:2412.05467.

Parshin Shojaee Li, Kazem Meidani Gao, Kuldeep Shukla, Amir Barati Karniadakis, and George Em. 2025a. Llm-sr: Scientific equation discovery via

programming with large language models. In *International Conference on Machine Learning (ICML)*. LLM-SRBench benchmark.

Weitao Li, Boran Xiang, Xiaolong Wang, Zhinan Gou, Weizhi Ma, and Yang Liu. 2025b. Ur^2: Unify rag and reasoning through reinforcement learning. *Preprint*, arXiv:2508.06165.

Xin Li, Yichen Chen, Yue Zeng, Zhanhui Jiang, Chenxi Huang, Yiwen Wang, and Zhibo Li. 2025c. Cvebench: Can language models exploit security vulnerabilities? In *International Conference on Machine Learning (ICML)*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025d. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian Yang, Ge Zhang, and Wenhao Huang. 2025e. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*.

Yuetai Li, Huseyin A Inan, Xiang Yue, Wei-Ning Chen, Lukas Wutschitz, Janardhan Kulkarni, Radha Poovendran, Robert Sim, and Saravan Rajmohan. 2025f. Simulating environments with reasoning models for agent training. *arXiv preprint arXiv:2511.01824*.

William Liang, Sam Wang, Hung-Ju Wang, Osbert Bastani, Dinesh Jayaraman, and Yecheng Jason Ma. 2024. Eurekaverse: Environment curriculum generation via large language models. *Preprint*, arXiv:2411.01775. CoRL 2024 (per arXiv comments).

Minhua Lin, Zongyu Wu, Zhichao Xu, Hui Liu, Xianfeng Tang, Qi He, Charu Aggarwal, Hui Liu, Xiang Zhang, and Suhang Wang. 2025. A comprehensive survey on reinforcement learning-based agentic search: Foundations, roles, optimizations, evaluations, and applications. *Preprint*, arXiv:2510.16724.

Jiacai Liu, Yingru Li, Yuqian Fu, Jiawei Wang, Qian Liu, and Yu Shen. 2025a. When speed kills stability: Demystifying rl collapse from the training-inference mismatch. `https://richardli.xyz/rl-collapse`. Accessed: 2025-09.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianhui Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations (ICLR)*.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Junge Zhang, and Jianbin Jiao. 2025b. Agentic reinforcement learning with implicit step rewards. *Preprint*, arXiv:2509.19199.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, Shengyi Huang, Siran Yang, Jiamang Wang, Wenbo Su, and Bo Zheng. 2025c. Part i: Tricks or traps? a deep dive into RL for LLM reasoning. *arXiv preprint arXiv:2508.08221*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Pan Lu, Zhenlin Qiu, Yilun Chang, Pengcheng Qu, Liang Zhang, and Sean Welleck. 2025a. Can language models replicate scientific research? evaluating end-to-end research replication via paperbench. In *International Conference on Machine Learning (ICML)*.

Songshuo Lu, Hua Wang, Zhi Chen, and Yaohua Tang. 2025b. Urpo: A unified reward & policy optimization framework for large language models. *arXiv preprint arXiv:2507.17515*.

Mengdi Luo, Yujia Li, Wojciech M. Czarnecki, Vlad Mnih, J. Jiang, and Koray Kavukcuoglu. 2020. Impact: Importance weighted asynchronous architectures with clipped target networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, Weihan Cao, Jiangning Liu, Hongwei Liu, Junnan Liu, Songyang Zhang, Dahua Lin, and Kai Chen. 2025. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*.

Jianbiao Mei, Tao Hu, Daocheng Fu, Licheng Wen, Xuemeng Yang, Rong Wu, Pinlong Cai, Xinyu Cai, Xing Gao, Yu Yang, Chengjun Xie, Botian Shi, Yong Liu, and Yu Qiao. 2025a. O^2-searcher: A searching-based agent model for open-domain open-ended question answering. *Preprint*, arXiv:2505.16582.

Lang Mei, Zhihan Yang, and Chong Chen. 2025b. Ai-searchplanner: Modular agentic search via pareto-optimal multi-objective reinforcement learning. *Preprint*, arXiv:2508.20368.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. 2024. Memllm: Finetuning llms to use an explicit read-write memory. *arXiv preprint arXiv:2404.11672*.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *Preprint*, arXiv:2201.03544.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024a. Feedback loops with language models drive in-context reward hacking. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and Zhengyang Wu. 2024b. Webcanvas: Benchmarking web agents in online environments. *Preprint*, arXiv:2406.12373.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. 2025. Grpo-λ: Credit assignment improves llm reasoning. *Preprint*, arXiv:2510.00194.

Houwen Peng, Kan Wu, Yixuan Wei, Guoshuai Zhao, Yuxiang Yang, Ze Liu, Yifan Xiong, Ziyue Yang, Bolin Ni, Jingcheng Hu, Ruihang Li, Miaosen Zhang, Chen Li, Jia Ning, Ruizhe Wang, Zheng Zhang, Shuguang Liu, Joe Chau, Han Hu, and Peng Cheng. 2023. Fp8-lm: Training fp8 large language models. *Preprint*, arXiv:2310.18313.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models: A survey. *arXiv preprint arXiv:2503.23037*.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025a. Defeating the training-inference mismatch via fp16. *Preprint*, arXiv:2510.26788.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. 2025b. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore.

Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. 2024. A review of large language models and autonomous agents in chemistry. *Preprint*, arXiv:2407.01603.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P Lillicrap, and Oriana Riva. 2025. Androidworld: A dynamic benchmarking environment for autonomous agents. In *The Thirteenth International Conference on Learning Representations*.

J Rosser and Jakob Foerster. 2025. Agentbreeder: Mitigating the ai safety risks of multi-agent scaffolds via self-improvement. *Preprint*, arXiv:2502.00757.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junliang Song, Xiaohan Li, Haowen Li, Ming Zhang, Y. Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint*. ArXiv:2402.03300.

Junhong Shen, Hao Bai, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, and Aviral Kumar. 2025. Thinking vs. doing: Agents that reason by scaling test-time interaction. *Preprint*, arXiv:2506.07976.

Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lu Hou, and Lifeng Shang. 2025a. Deepdiver: Adaptive search intensity scaling via open-web reinforcement learning. *Preprint*, arXiv:2505.24332.

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. 2025b. Search and refine during think: Facilitating knowledge refinement for improved retrieval-augmented reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhengliang Shi, Lingyong Yan, Dawei Yin, Suzan Verberne, Maarten de Rijke, and Zhaochun Ren. 2025c. Iterative self-incentivization empowers large language models as agentic searchers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism. *Preprint*, arXiv:1909.08053.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. {ALFW}orld: Aligning text and

embodied environments for interactive learning. In *International Conference on Learning Representations*.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025a. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *Preprint*, arXiv:2503.05592.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025b. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization for llm agents. *Preprint*, arXiv:2403.02502.

Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, and 1 others. 2025. Simpledeepsearcher: Deep information seeking via web-powered reasoning trajectory synthesis. *arXiv preprint arXiv:2505.16834*.

Reuben Tan, Baolin Peng, Zhengyuan Yang, Hao Cheng, Oier Mees, Theodore Zhao, Andrea Tupini, Isar Meijier, Qianhui Wu, Yuncong Yang, Lars Liden, Yu Gu, Sheng Zhang, Xiaodong Liu, Lijuan Wang, Marc Pollefeys, Yong Jae Lee, and Jianfeng Gao. 2025. Multimodal reinforcement learning with agentic verifier for ai agents. *Preprint*, arXiv:2512.03438.

Lucas-Andreï Thil, Mirela Popa, and Gerasimos Spanakis. 2024. Navigating webai: Training agents to complete web tasks with large language models and reinforcement learning: Extended abstract. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24)*, Avila, Spain. Association for Computing Machinery.

Brandon Trabucco, Gunnar Sigurdsson, Robinson Piramuthu, and Ruslan Salakhutdinov. 2025. Insta: Towards internet-scale training for agents. *Preprint*, arXiv:2502.06776.

Hieu Tran, Zonghai Yao, and Hong Yu. 2025. Exploiting tree structure for credit assignment in rl training of llms. *Preprint*, arXiv:2509.18314.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, and 13 others. 2023. Towards generalist biomedical ai. *Preprint*, arXiv:2307.14334.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *Preprint*, arXiv:2305.16291.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. 2025a. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *Preprint*, arXiv:2505.20732.

Ruizhe Wang, Yeyun Gong, Xiao Liu, Guoshuai Zhao, Ziyue Yang, Baining Guo, Zheng-Jun Zha, and Peng CHENG. 2025b. Optimizing large language model training using FP4 quantization. In *Forty-second International Conference on Machine Learning*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023b. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025c. A survey of llm-based agents in medicine: How far are we from baymax? *Preprint*, arXiv:2502.11211.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025d. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. 2025e. Stepsearch: Igniting llms search ability via stepwise proximal policy optimization. *Preprint*, arXiv:2505.15107.

Zizhao Wang, Dingcheng Li, Vaishakh Keshava, Phillip Wallis, Ananth Balashankar, Peter Stone, and Lukas Rutishauser. 2025f. Adversarial reinforcement learning for large language model agent safety. *Preprint*, arXiv:2510.05442.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, and 8 others. 2025a. From ai for science to agentic science: A survey on autonomous scientific discovery. *Preprint*, arXiv:2508.14111.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025b. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *Preprint*, arXiv:2505.11821.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao

Zhang, Bing Yin, Hyokun Yun, and Lihong Li. 2025c. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7920–7939, Suzhou, China. Association for Computational Linguistics.

wenlin zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. Process vs. outcome reward: Which is better for agentic RAG reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, Xiaocheng Tang, Yundi Qian, Beibei Zhu, and Rui Hou. 2025a. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm training. *arXiv preprint*. ArXiv:2505.24034.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.

Wenxun Wu, Yuanyang Li, Guhan Chen, Linyue Wang, and Hongyang Chen. 2025b. Tool-augmented policy optimization: Synergizing reasoning and adaptive tool use with reinforcement learning. *arXiv preprint arXiv:2510.07038*.

Haocheng Xi, Han Cai, Ligeng Zhu, Yao Lu, Kurt Keutzer, Jianfei Chen, and Song Han. 2025a. COAT: Compressing optimizer states and activations for memory-efficient FP8 training. In *The Thirteenth International Conference on Learning Representations*.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, and 2 others. 2024a. Training large language models for reasoning through reverse curriculum reinforcement learning. *Preprint*, arXiv:2402.05808.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024b. Agentgym: Evolving large language model-based agents across diverse environments. *Preprint*, arXiv:2406.04151.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, and 4 others. 2025b. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *Preprint*, arXiv:2509.08755.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. 2025. Capo: Towards enhancing llm reasoning through generative credit assignment. *Preprint*, arXiv:2508.02298.

Tianbao Xie, Danyang Fan, Zehui Chen, Xin Gao, Zhengyuan Wu, Kun Shi, Minghui Zhao, James Li, Qitong Guo, Liangcai Qiao, Tao Wang, Victor Zhong, Denny Zhou, Panupong Pasupat, and Luke Zettlemoyer. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Shangyu Xing, Siyuan Wang, Chenyuan Yang, Xinyu Dai, and Xiang Ren. 2025. Lookahead tree-based rollouts for enhanced trajectory-level exploration in reinforcement learning with verifiable rewards. *Preprint*, arXiv:2510.24302.

Yixuan Even Xu, Yash Savani, Fei Fang, and J. Zico Kolter. 2025a. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. *Preprint*, arXiv:2504.13818.

Zhichao Xu, Zongyu Wu, Yun Zhou, Aosong Feng, Kang Zhou, Sangmin Woo, Kiran Ramnath, Yijun Tian, Xuan Qi, Weikang Qiu, Lin Lee Cheong, and Haibo Ding. 2025b. Beyond correctness: Rewarding faithful reasoning in retrieval-augmented generation. *Preprint*, arXiv:2510.13272.

Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025c. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z Pan, Hinrich Schütze, and 1 others. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

Yushi Yang, Shreyansh Padarha, Andrew Lee, and Adam Mahdi. 2025. Agentic reinforcement learning for search is unsafe. *Preprint*, arXiv:2510.17431.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. 2025. Your efficient rl framework secretly brings you off-policy rl training.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and 1 others. 2025a. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *arXiv preprint arXiv:2507.02259*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, and Lingjun Liu. 2025b. Dapo: An

open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *Preprint*, arXiv:2308.01825.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, and Zhengyin Du. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*.

Zhiyuan Zeng, Hamish Ivison, Yiping Wang, Lifan Yuan, Shuyue Stella Li, Zhuorui Ye, Siting Li, Jacqueline He, Runlong Zhou, Tong Chen, Chenyang Zhao, Yulia Tsvetkov, Simon Shaolei Du, Natasha Jaques, Hao Peng, Pang Wei Koh, and Hannaneh Hajishirzi. 2025. Rlve: Scaling up reinforcement learning for language models with adaptive verifiable environments. *Preprint*, arXiv:2511.07317.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. 2025a. Safevla: Towards safety alignment of vision-language-action model via constrained learning. *Preprint*, arXiv:2503.03480.

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, Yue Liao, Hongru Wang, Mengyue Yang, and 6 others. 2025b. The landscape of agentic reinforcement learning for LLMs: A survey. *arXiv preprint arXiv:2509.02547*.

Yaocheng Zhang, Haohuan Huang, Zijun Song, Yuanheng Zhu, Qichao Zhang, Zijie Zhao, and Dongbin Zhao. 2025c. Criticsearch: Fine-grained credit assignment for search agents via a retrospective critic. *Preprint*, arXiv:2511.12159.

Yujie Zhang, Minghao Chen, Zixuan Wang, Weize Liu, Yichi Chen, and 1 others. 2025d. Caia: Constructing adversarial examples for interactive agents. *Preprint*, arXiv:2510.00332.

Yunhao Zhang, Yifan Ma, Chenyang Li, Zicheng Wang, Yuxiao Zheng, Siteng Chen, Wenqiao Wang, Xu Xia, Ziyu Liu, Liangke Huang, Chuang Yang, Jiakai Zhou, Song-Chun Zhu, and Cewu Lu. 2025e. Embodiedbench: Benchmarking multi-modal large language models for embodied ai. In *International Conference on Machine Learning (ICML)*.

Zijing Zhang, Ziyang Chen, Mingxiao Li, Zhaopeng Tu, and Xiaolong Li. 2025f. Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents. *Preprint*, arXiv:2507.22844.

Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. 2025a. LLM-based agentic reasoning frameworks: A survey from methods to scenarios. *arXiv preprint arXiv:2508.17692*.

Qingfei Zhao, Ruobing Wang, Dingling Xu, Daren Zha, and Limin Liu. 2025b. R-search: Empowering llm reasoning with search via multi-reward reinforcement learning. *Preprint*, arXiv:2506.04185.

Shu Zhao, Tan Yu, Anbang Xu, Japinder Singh, Aaditya Shukla, and Rama Akkiraju. 2025c. Parallelsearch: Train your llms to decompose query and search sub-queries in parallel with reinforcement learning. *Preprint*, arXiv:2508.09303.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

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. *arXiv preprint arXiv:2506.02177*.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Yifei Zhou, Sergey Levine, Jason Weston, Xian Li, and Sainbayar Sukhbaatar. 2025a. Self-challenging language model agents. *Preprint*, arXiv:2506.01716.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025b. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2025. Reasonflux-PRM: Trajectory-aware PRMs for long chain-of-thought reasoning in LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Reproducibility Resources

An anonymized repository accompanying this survey is available at https://anonymous.4open.science/r/Agentic-RL-Training-Recipes/.