Grounding Open-Domain Knowledge from LLMs to Real-World Reinforcement Learning Tasks: A Survey

Haiyan Yin[1] , **Hangwei Qian**[1†] , **Yaxin Shi**[1†] , **Ivor Tsang**[1,2] , **Yew-Soon Ong**[1,2]
[1]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[2]College of Computing and Data Science, Nanyang Technological University (NTU), Singapore
{yin_haiyan, qian_hangwei, shi_yaxin, ivor_tsang}@cfar.a-star.edu.sg, asysong@ntu.edu.sg

Abstract

Grounding open-domain knowledge from large language models (LLMs) into real-world reinforcement learning (RL) tasks represents a transformative frontier in developing intelligent agents capable of advanced reasoning, adaptive planning, and robust decision-making in dynamic environments. In this paper, we introduce the *LLM-RL Grounding Taxonomy*, a systematic framework that categorizes emerging methods for integrating LLMs into RL systems by bridging their open-domain knowledge and reasoning capabilities with the task-specific dynamics, constraints, and objectives inherent to real-world RL environments. This taxonomy encompasses both training-free approaches, which leverage the zero-shot and few-shot generalization capabilities of LLMs without fine-tuning, and fine-tuning paradigms that adapt LLMs to environment-specific tasks for improved performance. We critically analyze these methodologies, highlight practical examples of effective knowledge grounding, and examine the challenges of alignment, generalization, and real-world deployment. Our work not only illustrates the potential of LLM-RL agents for enhanced decision-making, but also offers actionable insights for advancing the design of next-generation RL systems that integrate open-domain knowledge with adaptive learning.

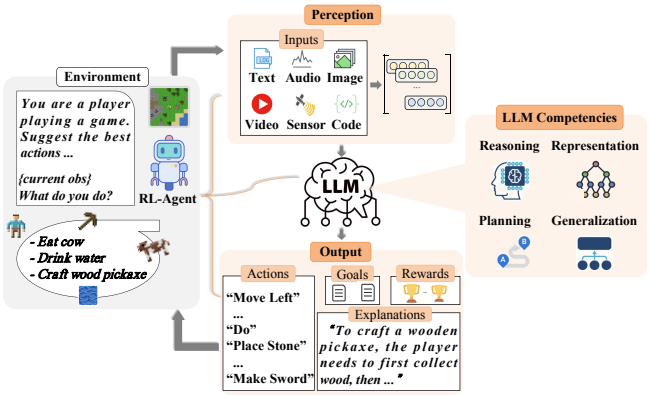

Figure 1: **Overview of the LLM-RL Grounding Taxonomy.** This framework illustrates how large language models support RL agents through multimodal perception, structured reasoning, adaptive planning, and task generalization, enabling decision-making in complex, dynamic environments.

1 Introduction

The integration of large language models (LLMs) with reinforcement learning (RL) represents a transformative milestone in the development of intelligent agents. By leveraging LLMs' powerful capabilities in reasoning, generalization, contextual understanding, and their rich priors over world knowledge, RL agents can overcome persistent challenges such as sample inefficiency, poor generalization, and brittle task-specific behavior in complex environments. This synergy equips agents with greater foresight, semantic grounding, and adaptability, opening new possibilities for robust decision-making across domains such as robotics, healthcare, and multi-agent systems.

At the core of this integration lies the concept of **grounding**, a critical process that aligns the broad, open-domain knowledge embedded within LLMs with the structured, goal-directed nature of RL tasks. Grounding involves the translation of high-level, abstract knowledge into actionable insights that are consistent with the dynamic state-action-reward paradigms of RL. This encompasses several key dimensions: *what to ground* (e.g., commonsense reasoning, procedural knowledge, and semantic priors), *how to ground* (through mechanisms like prompting, retrieval, or fine-tuning), and *why grounding matters* (to enhance decision efficiency, task generalization, and policy robustness). Effective grounding ensures that RL agents not only react to environmental feedback but also proactively reason, plan, and adapt in contextually meaningful ways.

Recent advances in LLM-powered agents, exemplified by GPT-4 [OpenAI, 2023], DeepSeek [DeepSeek, 2023], Gemini [DeepMind, 2024] and LLAMA [Touvron *et al.*, 2023], highlight their growing potential in tackling complex, real-world tasks by providing strong priors over language, knowl-

† Hangwei Qian and Yaxin Shi are the corresponding authors.

edge, and commonsense reasoning. These capabilities position LLMs as promising complements to RL, which traditionally depends on learning through direct interaction. Yet, integrating LLMs into RL pipelines remains fundamentally challenging. The knowledge encoded in LLMs, shaped by pretraining on broad, general-purpose data, often misaligns with the structured, interactive demands of RL environments, leading to hallucinations, overconfidence, or brittle policies. While RL agents are designed to adapt continuously through feedback, most LLMs operate in static inference settings and lack mechanisms for contextual adaptation. Bridging this gap requires a principled understanding of grounding: how to translate abstract knowledge into behavior that is sensitive to environmental dynamics, responsive to interaction, and robust across varied tasks. Developing such grounding strategies is essential to unlocking the potential of LLM and RL integration.

To this end, we propose the **LLM-RL Grounding Taxonomy**, a principled framework that organizes and critically examines the methodologies used to align LLM capabilities with RL objectives. This survey is based on a targeted review of approximately 40 recent papers, primarily published between 2022 and 2024 in top-tier AI venues such as NeurIPS, ICML, and ICLR. The selection emphasizes high-quality studies that explicitly integrate LLMs into RL systems through grounding mechanisms. The taxonomy categorizes these approaches into two principal paradigms: *training-free grounding* and *fine-tuning-based grounding*. Training-free methods leverage techniques such as structured prompting, chain-of-thought reasoning, and retrieval-augmented inference to guide LLM behavior without modifying model weights. In contrast, fine-tuning-based strategies adapt LLMs through parameter updates, feedback-driven optimization, and modular architectural design. This taxonomy clarifies the emerging design space, surfaces key techniques and trade-offs, and offers a practical guide for researchers and practitioners developing grounded LLM-RL agents.

Furthermore, this survey highlights three cross-cutting directions that are central to advancing LLM-RL systems. First, *multimodal grounding* aims to connect language models with perceptual inputs such as vision and audio, enabling agents to interpret and act within richer sensory environments. A central challenge is aligning symbolic reasoning with continuous, high-dimensional observations. Second, *hierarchical reasoning* supports planning and control across both temporal and semantic scales, bridging abstract objectives with low-level execution through structures like subgoals or modular policies. Third, *adaptive grounding* focuses on mechanisms that enable agents to dynamically adjust to feedback and evolving environments. Progress across these areas will also require evaluation frameworks that go beyond task completion to assess generalization, robustness, and the effectiveness of grounding itself.

1.1 Objectives of This Survey

This survey aims to offer a focused synthesis of recent methods for grounding pretrained LLMs to enhance the decision-making capabilities of RL agents. Rather than providing exhaustive coverage, we emphasize conceptual clarity and practical taxonomy to support both understanding and application. Beyond categorization, we present a critical analysis of training-free and fine-tuning-based paradigms, identifying key algorithms alongside the underlying design principles and architectural patterns that drive their success across diverse RL environments. Through the proposed *LLM-RL Grounding Taxonomy*, we offer a strategic framework that maps the evolving landscape, reveals synergies between grounding strategies, and serves as a practical reference for researchers. We extract actionable insights and classify state-of-the-art techniques that connect LLMs' generalized knowledge with task-specific demands. Additionally, we highlight core challenges, such as the brittleness of LLM reasoning in dynamic environments and the inefficiencies of scaling fine-tuned models, while identifying opportunities for developing adaptive, scalable agents. Unlike prior surveys on general LLM-based agents, which typically emphasize planning, reasoning, or static tool use, our focus is specifically on agents that learn and adapt through reinforcement feedback in interactive, dynamic environments. Ultimately, this survey aims to advance the integration of language understanding and RL, laying the groundwork for versatile, context-aware intelligent systems capable of robust real-world performance.

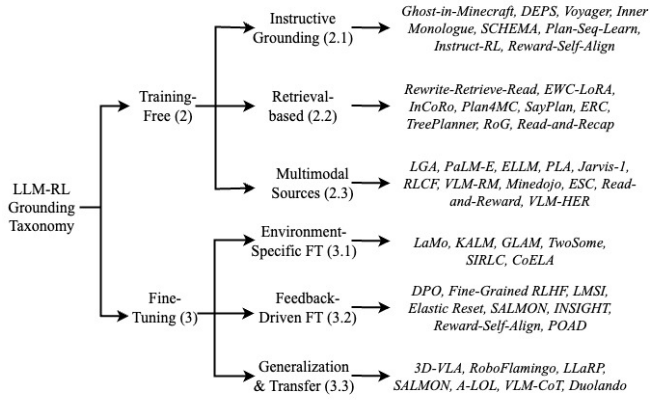

Figure 2: **LLM-RL Grounding Taxonomy**: categorizing grounding paradigms for LLM-RL agents into *training-free* and *fine-tuning (FT)*-based approaches, along with their respective subcategories and representative models.

2 Training-free Grounding Paradigms

Training-free grounding paradigms for LLM-RL agents leverage the general reasoning abilities, contextual understanding, and vast prior knowledge embedded in LLMs to enable seamless integration into RL tasks. By bypassing the need for additional parameter updates, these methods minimize computational complexity and simplify implementation, making them a practical and efficient choice for diverse decision-making scenarios.

2.1 Instructive Grounding

Instructive grounding guides LLM-RL agents using explicit instructional signals, including natural language prompts,

demonstrations, and task-specific templates. These signals serve to structure the agent's behavior, aligning high-level reasoning with desired outcomes without modifying the LLM's parameters.

LLMs, enriched with extensive pretrained knowledge from complex decision-making problem domains, can be strategically guided through prompting to translate their capabilities into grounded, task-relevant behavior. *Ghost-in-Minecraft* [Zhu *et al.*, 2023] employs structured instruction templates that define action interfaces, illustrate queries, specify response formats, and incorporate error-enriched interaction guidelines to ground LLM for task execution in sparse-reward environments. *DEPS* [Wang *et al.*, 2023b] dynamically generates agent-in-the-loop programs, using a self-explanation module for error correction and a trainable goal selector to rank and prioritize sub-goals for effective task decomposition. *Voyager* [Wang *et al.*, 2024] employs GPT-4 to autonomously generate executable programs, maintaining a dynamic curriculum of skills while actively collecting bug messages from the game for self-correction. These approaches demonstrate how carefully designed prompts, curriculum structures, and self-corrective feedback loops can enable robust grounding in sparse-reward environments.

Instructive grounding can also be utilized to teach LLMs essential RL dynamics and outcomes, effectively aligning their reasoning capabilities with task execution. *Inner Monologue* [Huang *et al.*, 2022] integrates the planning capabilities of LLMs with robotic control policies via textual instructions that incorporate real-time environmental feedback. This approach allows the agent to iteratively refine its actions by evaluating observed outcomes, seamlessly linking high-level reasoning with low-level control. Similarly, *SCHEMA* [Niu *et al.*, 2024] employs LLMs to generate step representations by describing state changes at each step, achieved through sophisticated chain-of-thought prompting, enabling a deeper understanding of task dynamics. *Plan-Seq-Learn* [Dalal *et al.*, 2024] designs structured prompts to guide LLMs in breaking down complex robotic control tasks into manageable sequential sub-goals. The model generates step-by-step action plans that link subgoals to objectives through chain-of-thought reasoning, and iteratively refines them using environment feedback to adapt to task-specific dynamics.

Beyond guiding action selection, instructional signals have also been used to shape reward functions, aligning agent behavior more directly with human intent. Recent work extends instructive grounding beyond task guidance, using language to shape internal objectives and reward functions. *Instruct-RL* [Hu and Sadigh, 2023] uses pretrained LLMs to generate a prior policy conditioned on the human instruction and leverages it to regularize the RL objective, steering policy learning toward behavior aligned with human preferences, effectively grounding the agent's behavior in explicitly human provided directives. *Reward-Self-Align* [Zeng *et al.*, 2024] employs LLMs to autonomously generate and refine reward functions via iterative self-alignment. By analyzing its outputs and iteratively adjusting the reward signals, the LLM creates task-specific feedback loops, directly grounding the reward design in structured prompts. This enables the agent to refine its actions dynamically, with the LLM continuously optimizing the reward function to encourage behaviors that achieve successful task execution.

Discussions Instructive grounding is most effective when tasks admit language-based decompositions and when pretrained knowledge can reduce the need for environment-driven exploration. It is particularly well-suited to sparse-reward and long-horizon settings, where structured prompts and task priors help scaffold behavior early in training. However, success depends critically on the clarity and completeness of instructions, the agent's ability to ground language in situated perception and dynamics, and the robustness of feedback loops over time. Common patterns across methods include hierarchical prompt design, subgoal generation, and self-refinement through interaction. While enabling fast iteration without fine-tuning, it remains limited in scenarios requiring fine-grained perception or real-time adaptation beyond what static language can express.

2.2 Retrieval-based Grounding

Retrieval-based grounding enhances LLM-RL agents by injecting external, task-relevant knowledge at inference time. By dynamically querying sources such as document corpora, demonstration traces, or structured representations, these methods allow agents to incorporate context that cannot be memorized or inferred from pretraining alone. This capability is particularly beneficial for tasks that involve sparse rewards, long horizons, or shifting domain knowledge.

A common strategy is to retrieve exemplars or environment traces to ground decision-making in specific contexts. For example, *Rewrite-Retrieve-Read* [Ma *et al.*, 2023] leverages web search to retrieve contextual information during query rewriting, ensuring the responses are better aligned with task requirements. *EWC-LoRA* [Xiang *et al.*, 2023] incorporates 2-10 exemplars in prompts, comprising instructions, question contexts, and answers, to facilitate effective in-context learning. In robotics, *InCoRo* [Zhu *et al.*, 2024] provides demonstration traces to the robotic manipulation system for imitation learning, enabling zero-shot generalization to new tasks. *Plan4MC* [Yuan *et al.*, 2023] constructs skill categorization in Minecraft, covering actions such as finding, manipulating, and crafting, to instruct LLMs to extract the relationship between skills to construct a skill graph beforehand.

A prominent subclass of retrieval-based methods incorporates structured representations, such as knowledge graphs and 3D scene graphs, to scaffold high-level reasoning and policy planning. *SayPlan* [Rana *et al.*, 2023] leverages hierarchical 3D scene graphs to perform iterative replanning in complex household environments. *Embodied Robotic Control (ERC)* [Qi *et al.*, 2024] integrates knowledge graphs with LLMs to enforce safety constraints in service robotics. *TreePlanner* [Hu *et al.*, 2024] reframes task planning with LLMs into three distinct phases: plan sampling, action tree construction, and grounded decision-making, to enable LLMs to perform top-down control. Beyond planning, other graph-based methods emphasize knowledge-guided reasoning and distillation, using structured representations to transfer policies and abstract complex tasks. *Reasoning on Graphs (RoG)* [Luo *et al.*, 2024a] implements a planning-retrieval-reasoning pipeline, where relation paths grounded in knowl-

edge graphs serve as faithful intermediate plans, supporting multi-step reasoning and compositional knowledge transfer.

Discussions Retrieval-based grounding is effective when essential knowledge lies outside the LLM's pretrained scope. It supports generalization in long-horizon tasks by injecting contextual or structured information at inference time, alleviating the need for extensive environment-driven exploration. Graph-based methods, in particular, enable interpretable reasoning and safety-aware planning through symbolic constraints. However, performance depends critically on retrieval quality, latency, and the agent's capacity to interpret retrieved content in context. These methods are less suited for real-time or reactive settings, where tight feedback loops or high-frequency decisions are required.

2.3 Grounding from Multimodal Data

Grounding from multimodal sources involves integrating information from diverse data modalities, such as text, images, and video, to enhance contextual understanding and task relevance, particularly in environments where single-modality inputs fall short. These approaches use multimodal LLMs to align perceptual inputs with language-conditioned objectives, enabling dynamic grounding of observations into task-relevant decisions.

One primary direction for multimodal grounding is to enhance task representation. *LGA* [Peng *et al.*, 2024] investigates how language can be leveraged to create state abstractions that streamline decision-making in RL agents. By utilizing pretrained LLMs, LGA automatically constructs state abstraction functions tailored to new, unseen tasks. The approach then trains an imitation policy using a smaller set of demonstrations, operating on the generalized abstract states, thereby enhancing the efficiency and adaptability of the RL agent. *PaLM-E* [Driess *et al.*, 2023] integrates multiple sensory inputs—such as vision, language, and embodied actions—into a unified model that can understand and respond to complex tasks in real-world environments. This alignment is learned by training the model to correlate visual observations with language-conditioned objectives and action sequences. RL is used to optimize the model's policy, with task success serving as feedback to refine multimodal representations. *ELLM* [Du *et al.*, 2023] encourages exploration by rewarding agents for achieving LLM-specified goals, using a state captioner to bridge visual observations and language instructions. *PLA* [Gao *et al.*, 2024] aligns multi-domain images based on textual prompts to facilitate zero-shot policy transfer. In contrast, *Jarvis-1* [Wang *et al.*, 2023a] is designed as a multi-modal open-world agent focused on scalability and adaptability in complex environments like Minecraft. Unlike PaLM-E's tightly integrated structure, Jarvis-1 employs a modular architecture where VLMs are decoupled from the planning and control modules. This design allows for more flexible task adaptation, with the vision-language component primarily responsible for semantic understanding, while task planning is handled by specialized models optimized for open-ended exploration. While PaLM-E is well-suited for end-to-end robotic control, Jarvis-1 is more adaptable in dynamic, unstructured environments, benefiting from its modularity and long-horizon memory design.

Multimodal information can be harnessed to ground the decision-making through constructing informative reward functions. *RLCF* [Zhao *et al.*, 2024] proposes an RL framework with CLIP reward feedback for test time adaptation for VLM models in zero-shot classification problems. The authors propose a novel CLIP reward which samples low-entropy predictions from multiple views as test-time samples for reward maximization. *VLM-RM* [Rocamonde *et al.*, 2024] introduces a CLIP-based model to generate zero-shot rewards, enabling MuJoCo humanoid agents to learn complex tasks without the need for manually specified reward functions. The approach utilizes a simple, single-sentence text prompt to describe the desired task, with minimal prompt engineering, while the VLM generates correlation scores that serve as reward signals for training. *Minedojo* [Fan *et al.*, 2022] leverages internet-scale multimodal Minecraft knowledge to pre-train a VLM model named MineCLIP to compute the correlation between a language goal string and visual RGB frames, which can further serve as a reward function to train RL agents. *ESC* [Zhou *et al.*, 2023] proposes goal-conditioned exploration with soft commonsense constraints to transfer commonsense knowledge in pre-trained models to open-world object navigation without any navigation experience or other training on the visual environment. ESC models the commonsense constraints into navigation actions with soft logic predicates for efficient object navigation in the Habitat and RoboTHOR navigation challenges. *Read-and-Reward* [Wu *et al.*, 2023a] simultaneously performs vision captioning on Atari game frames and QA extraction and reasoning module to learn to play Atari from the user manual.

Furthermore, multimodal grounding has been explored through retrospective learning and hierarchical decomposition. *VLM-HER* [Sumers *et al.*, 2023] utilizes pre-trained VLMs to integrate visual and textual data for enhancing agent learning. By combining VLMs with hindsight experience replay (HER), the approach retroactively generates language descriptions of agent trajectories based on visual and task-specific observations. These relabeled trajectories enable the agent to learn multiple tasks along different dimensions, effectively grounding its behavior in multimodal data. This approach not only leverages the strengths of vision-language models for semantic understanding but also enhances data efficiency by reusing past experiences in a more informative way. By transforming raw trajectories into structured language, VLM-HER facilitates more interpretable and transferable policy learning across diverse tasks.

Discussions Multimodal grounding enables RL agents to incorporate signals from language, vision, and environmental context, producing decisions that are more situationally aware and semantically aligned. This integration supports expressive task representations, adaptive reward shaping, and transparent reasoning, especially valuable in open-ended or real-world environments. It also allows agents to disambiguate complex scenarios by cross-referencing multiple modalities, enhancing robustness in noisy or partially observable settings. Furthermore, leveraging pretrained multimodal models can improve sample efficiency by injecting semantic priors into perception and action loops. As multimodal systems continue

to evolve, ensuring alignment across modalities and preserving interpretability will be critical for real-world deployment.

3 Fine-Tuning-Based Grounding Paradigms

Fine-tuning-based grounding paradigms adapt pre-trained LLMs to specific RL tasks by updating model parameters with task-specific data and feedback. This approach enhances task alignment and adaptability, enabling improved performance in dynamic and complex scenarios at the cost of additional computational overhead and design complexity.

3.1 Environment-Specific Fine-Tuning

Environment-specific fine-tuning tailors LLMs to the dynamics of a particular RL setting by modifying model architectures or leveraging parameter-efficient adaptation techniques. These methods aims to bridge the gap between general-purpose language pretraining and the grounded, interaction-heavy requirements of RL tasks.

One approach involves architectural modifications to enhance adaptability in complex domains. *LaMo* [Siyao *et al.*, 2024] replaces standard linear projections with MLPs, improving representation learning in offline RL. It also uses LoRA (Low-Rank Adaptation) to fine-tune just 0.7% of the parameters, achieving computational efficiency while maintaining generalization. However, limiting updates to a small subset of parameters can constrain the model's ability to internalize nuanced task dynamics—particularly in long-horizon or rapidly changing environments. This highlights a key trade-off in parameter-efficient tuning: preserving general knowledge versus fully adapting to new domains.

A related challenge is catastrophic forgetting, where fine-tuning on a specific task erodes prior capabilities. This issue undermines the robustness of LLM-RL agents in lifelong or multi-environment settings, where continual retention of diverse competencies is critical. Other methods address the need to bridge language-based reasoning with low-level environmental signals. *KALM* [Pang *et al.*, 2024b] adapts LLMs for robotic control by incorporating MLP layers that translate textual goals into executable trajectories. This enables alignment between linguistic reasoning and numeric observations. Despite its strong performance in domains like CLEVR-Robot and Meta-World, KALM's reliance on dense fine-tuning of multimodal layers raises scalability concerns, especially when expanding to new domains. This motivates the need for adaptive strategies that adjust tuning depth based on task complexity and modality.

In embodied and interactive settings, fine-tuning can enhance task grounding through policy optimization. *GLAM* [Carta *et al.*, 2023] augments LLMs with a value head and uses PPO to align predictions with RL rewards. *TwoSome* [Tan *et al.*, 2024] addresses action-length bias by combining LoRA fine-tuning with word normalization, improving performance in long-horizon tasks like VirtualHome. *POAD* [Wen *et al.*, 2024] proposes Policy Optimization with Action Decomposition (POAD), which tokenizes actions and applies fine-grained credit assignment to each component. This formulation enables more expressive gradient signals and enhances learning efficiency in structured environments like Overcooked.

Modular task designs offer another direction for scalable fine-tuning. *CoELA* [Zhang *et al.*, 2024] embeds pretrained LLMs in a modular multi-agent framework, using LoRA to adapt specific components while freezing general reasoning layers. This approach supports both linguistic generalization and domain-specific control, enabling decentralized coordination in cooperative environments such as TDW-MAT and C-WAH. Modular designs strike a balance between adaptability and interpretability, allowing agents to reason, plan, and communicate effectively in dynamic multi-agent settings.

Discussions Environment-specific fine-tuning enables tight coupling between LLM representations and RL task dynamics, often yielding high performance in static or in-distribution settings. Yet this comes at the cost of scalability, particularly in multi-task or continually evolving environments. While parameter-efficient methods like LoRA reduce overhead, they often trade off adaptation depth and long-term flexibility. A promising direction is modular grounding: freezing core language reasoning while attaching lightweight adapters for fast, task-specific alignment, preserving generality without sacrificing adaptability.

3.2 Feedback-Driven Fine-Tuning

Feedback-driven fine-tuning adapts LLMs to RL tasks by incorporating signals such as preferences, rewards, critiques, or self-assessments. It iteratively adjusts model behavior using feedback rather than static labels, allowing the model to better align with task-specific objectives in dynamic environments.

A prominent direction is preference-based fine-tuning, where LLMs are trained to prefer certain outputs over alternatives. For instance, *DPO* [Rafailov *et al.*, 2023] simplifies alignment by directly optimizing a preference-aware loss, avoiding explicit reward modeling or reinforcement learning loops. This yields efficient training pipelines while maintaining strong alignment in tasks like summarization and dialogue. *Fine-Grained RLHF* [Wu *et al.*, 2023b] extends this framework with multi-dimensional feedback signals (e.g., fluency, coherence, informativeness), enabling LLMs to learn nuanced task-aligned behaviors via fine-tuned reward shaping. Such approaches offer flexibility but require careful signal design to avoid reinforcing superficial cues.

Another important class of methods explore self-generated feedback. *SIRLC* [Pang *et al.*, 2024a] uses diverse reasoning paths to produce high-confidence answers, which are then used to fine-tune the model via supervised learning. *Elastic Reset* [Noukhovitch *et al.*, 2023] introduces a stabilization technique, periodically reverting the model to a moving average of its parameters to preserve generalization while still optimizing for task-specific rewards. These techniques highlight the promise of internal feedback but still face limitations in feedback quality and signal calibration.

Feedback can also be structured through symbolic principles or alignment rules. *SALMON* [Sun *et al.*, 2024] generates synthetic preferences using human-aligned heuristics to shape a reward model for policy optimization. Complementing this approach, *INSIGHT* [Luo *et al.*, 2024b] combines symbolic reasoning with reward feedback, enabling the model to produce interpretable decisions and explanations. These methods exemplify how grounding can be en-

hanced through principled, iterative, and explanation-aware feedback, enabling more interpretable and trustworthy RL agents.

Other works explore dynamic and multi-level feedback for more granular control. *SIRLC* [Pang *et al.*, 2024a] generates reward signals from its own output quality and updates its policy accordingly. *ArCHer* [Zhou *et al.*, 2024] introduces multi-level reinforcement signals, at both token and utterance levels, to fine-tune long-horizon dialogue agents. *Thought Cloning* [Hu and Clune, 2023] captures human decision processes by training agents on synchronized demonstrations of both actions and underlying reasoning, improving alignment with human-like behavior and thought patterns. These approaches reflect a broader trend toward interactive, cognitively aligned fine-tuning, where models adapt not only to task feedback but to process-level guidance that reflects human-like decision flows.

Discussions Feedback-driven fine-tuning enables iterative alignment between LLM behavior and task-specific RL objectives, offering adaptability where static supervision falls short. Its flexibility makes it especially suited for complex or underspecified settings, but also exposes models to risks like feedback loops, misalignment, or reward hacking. While preference learning and self-evaluation provide lightweight alternatives to full supervision, their effectiveness hinges on signal quality and calibration. Symbolic priors and structured feedback offer a path forward, enabling more stable and interpretable fine-tuning with clearer grounding dynamics.

3.3 Task Generalization and Transfer

Task generalization and transfer under fine-tuning-based grounding aim to equip LLMs with the flexibility to operate beyond the training distribution, adapting to new tasks, instructions, and environments through targeted parameter updates. This capability is crucial for LLM-RL agents intended to function in open-ended, multi-task settings with minimal supervision. A representative example is *3D-VLA* [Zhen *et al.*, 2024], which fine-tunes a vision-language-action model for 3D embodied environments. It aligns multimodal inputs—language, perception, and action—through task-specific interaction tokens and a generative diffusion planner, enabling spatial reasoning and policy generalization across previously unseen 3D manipulation scenarios. In a complementary direction, *RoboFlamingo* [Li *et al.*, 2024] selectively fine-tunes task-specific modules (e.g., policy heads and cross-modal attention layers) while freezing the core pretrained vision-language backbone. This modular tuning strategy supports effective transfer to novel language-conditioned manipulation tasks, achieving strong generalization without overfitting to specific environments. Similarly, *LLaRP* [Szot *et al.*, 2024] adopts a reinforcement learning framework that fine-tunes peripheral modules while keeping the central LLM frozen. This setup preserves general-purpose reasoning while allowing efficient adaptation to novel instructions, tasks, and paraphrased command variations in embodied contexts.

Beyond modularization, several approaches leverage principle-driven reward mechanisms to improve transfer. *SALMON* [Sun *et al.*, 2024] incorporates human-aligned heuristics into a synthetic preference model, which then guides iterative fine-tuning of LLMs via reinforcement learning. This facilitates alignment with abstract goals and promotes consistency across tasks. *A-LOL* [Baheti *et al.*, 2024] introduces advantage-weighted learning in offline RL to optimize sequence-level rewards, supporting adaptation across a broad spectrum of language tasks. Zhai et al. [Zhai *et al.*, 2024] further combine chain-of-thought prompting with RL fine-tuning of vision-language models, improving intermediate reasoning and enabling generalization to previously unseen multi-step tasks. By embedding structural priors and reasoning scaffolds into the training process, these approaches expand transfer capabilities while reducing the fragility often associated with conventional fine-tuning.

Discussions The ability to generalize across tasks and environments remains a central bottleneck for LLM-RL agents. While freezing pretrained LLM backbones preserves broad reasoning skills, it limits fine-grained adaptation to specific control and interaction patterns. Emerging strategies that fine-tune peripheral modules or modular policy heads show promise in striking a balance between reuse and flexibility. By decoupling abstract semantic reasoning from grounded decision-making, these methods enable more efficient, task-specific adaptation with minimal interference to general capabilities. Moving forward, advances in structured memory, compositional skill libraries, and meta-adaptation mechanisms will be key to achieving robust, low-shot generalization in open-ended, dynamic environments.

4 Challenges and Opportunities

Grounding LLMs in RL tasks holds immense potential to revolutionize intelligent decision-making across diverse domains, yet realizing this vision demands overcoming several deep and interrelated challenges:

- *Contextual Adaptation to Environment-Specific Dynamics*: While LLMs are pretrained on broad open-domain corpora, real-world RL environments exhibit domain-specific dynamics, temporal dependencies, and task constraints that shift over time. Aligning LLM predictions with such evolving state-action structures remains a core challenge, particularly under non-stationarity and distributional drift. Effective adaptation requires mechanisms like environment-conditioned prompting, online fine-tuning, and continual knowledge distillation that can respond to task variation without sacrificing prior competence.
- *Mitigating Hallucination and Misalignment Risks*: LLMs are prone to hallucinations and generating overconfident yet incorrect outputs, which can severely compromise the safety and reliability of RL agents in critical tasks. Verifying, quantifying, and mitigating such risks requires robust grounding mechanisms, uncertainty-aware decision models, and fail-safe fallback strategies to prevent cascading errors in sequential decision-making.
- *Evaluating Effectiveness and Reliability*: Traditional RL evaluation metrics focus on reward performance, which may not fully capture the nuanced contributions of LLM-augmented reasoning. Developing comprehensive evaluation frameworks that assess consistency, assumptions,

Model	Domain	Tasks	Mod	FT	HR	EMB	MTL	Backbone
Ghost-in-Minecraft [Zhu *et al.*, 2023]	Games	Minecraft	T	×	✓	✓	✓	GPT-3
DEPS [Wang *et al.*, 2023b]	Games	Minecraft	T	×	✓	✓	✓	GPT-3
Voyager [Wang *et al.*, 2024]	Games	Minecraft	T	×	✓	✓	×	GPT-4
SCHEMA [Niu *et al.*, 2024]	Video	CrossTask, COIN, NIV	T+V	×	✓	×	✓	GPT3.5
Plan-Seq-Learn [Dalal *et al.*, 2024]	Robotics	MetaWorld, Kitchen, Robosuite, Obstructed Suite	T+V	×	✓	×	✓	GPT-4
InstructRL [Hu and Sadigh, 2023]	Games	Hanabi, Say-Select	T	×	×	×	✓	GPT-3.5
Plan4MC [Yuan *et al.*, 2023]	Games	Minecraft	T+V	×	✓	✓	✓	GPT-3.5
SayPlan [Rana *et al.*, 2023]	Robotics	Home, Office	T+V	×	✓	✓	✓	GPT-3.5
ELLM [Du *et al.*, 2023]	Games	Crafter; Housekeep	T	×	×	×	×	GPT-3
VLM-RM [Rocamonde *et al.*, 2024]	Robotics	Humanoid-v4	T+V	×	×	×	×	CLIP
Read and Recap [Wu *et al.*, 2023a]	Games	Atari	T+V	×	×	×	×	RoBERTa+Macaw
INSIGHT [Luo *et al.*, 2024b]	Neuro-Symbolic	Atari	T+V	✓	✓	×	×	GPT-4
PaLM-E [Driess *et al.*, 2023]	Robotics	TAMP	T+V	✓	✓	✓	✓	PaLM
SALMON [Sun *et al.*, 2024]	Assistant	OpenAssistant	T	✓	×	×	×	Llama2-70B
Duolando [Siyao *et al.*, 2024]	Motion	Duet Dance 100	T	✓	×	×	✓	minGPT
KALM [Pang *et al.*, 2024b]	Robotics	CLEVR-Robot	T	✓	×	×	✓	Llama2-7B
GLAM [Carta *et al.*, 2023]	Games	BabyAI	T	✓	×	×	✓	T5-Large
TWOSOME [Tan *et al.*, 2024]	Embodied	Overcooked, VirtualHome	T	✓	×	✓	×	Llama-7B
POAD [Wen *et al.*, 2024]	Embodied	Overcooked, VirtualHome, DataSciCoding	T	✓	×	✓	×	Llama2-7B
CoELA [Zhang *et al.*, 2024]	Embodied	C-WAH, TDW-MAT	T+V	✓	✓	✓	✓	Co/Llama2
DPO [Rafailov *et al.*, 2023]	Alignment	NLP tasks	T	✓	×	×	✓	GPT-2/J
Fine-Grained RLHF [Wu *et al.*, 2023b]	Alignment	NLP tasks	T	✓	×	×	✓	GPT-2/T5-L
SIRLC [Pang *et al.*, 2024a]	Alignment	NLP tasks	T	✓	×	×	✓	FLAN-T5
Elastic-Reset [Noukhovitch *et al.*, 2023]	Alignment	NLP tasks	T	✓	×	×	✓	GPT-2/Llama2-7B
ArCHer [Zhou *et al.*, 2024]	Language Agent	Multi-turn tasks	T	✓	✓	×	✓	GPT-2/RoBERTa
Thought Cloning [Hu and Clune, 2023]	Games	BabyAI (BossLevel)	T	✓	✓	×	×	BabyAI LM
3D-VLA [Zhen *et al.*, 2024]	Robotics	RLBench, CALVIN	T+V	✓	✓	✓	✓	BLIP2+T5$_{XL}$
RoboFlamingo [Li *et al.*, 2024]	Robotics	RLBench, CALVIN	T+V	✓	×	✓	✓	OpenFlamingo
LLaRP [Szot *et al.*, 2024]	Robotics	ALFRED	T+V	✓	×	✓	✓	Llama2-7/13B
A-LoL [Baheti *et al.*, 2024]	Alignment	NLP Tasks	T	✓	×	×	✓	Llama-7B
LLaVA-RL [Zhai *et al.*, 2024]	Robotics	ALFRED	T+V	✓	✓	✓	✓	LLaVA-1.6

Table 1: **Grounding LLMs in RL: A Taxonomy of Approaches.** This table categorizes representative methods for grounding pretrained LLMs within reinforcement learning pipelines, structured under six core dimensions: (1) **Model**: the method or system name; (2) **Domain**: the application area (e.g., games, robotics, alignment); (3) **Tasks**: the benchmark environments or task suites; (4) **Mod**: input modality, with **T** for text and **V** for vision; (5) **FT**: whether the LLM is fine-tuned during grounding; (6) **HR**: use of hierarchical policy structures; (7) **EMB**: whether the LLM is embodied within an agent interacting with an environment; (8) **MTL**: support for multitask learning and generalization; and (9) **Backbone**: the underlying LLM architecture. The taxonomy highlights diverse grounding strategies, revealing emerging trends such as multimodal integration, hierarchical control, and fine-tuning for decision-making in embodied agents.

reasoning accuracy, policy robustness, and generalization is essential for reliable grounding. This includes benchmarks for multi-task learning, out-of-distribution performance, and longitudinal adaptability.

- *Data Efficiency and Computational Scalability*: Fine-tuning large models for RL is resource-intensive, especially when paired with RL's high sample complexity. Yet, the scalability challenge extends beyond training: real-world deployment introduces tight constraints on inference latency, memory usage, and system integration, particularly in robotics and embedded settings. These practical bottlenecks demand grounding strategies that are not only computationally efficient during learning but also lightweight and responsive at inference time.

- *Trust, Interpretability, and Ethical Alignment*: The integration of LLMs into RL pipelines introduces new capabilities, but also amplifies the need for interpretability, trust, and ethical safeguards. In high-stakes domains, these are not ancillary concerns; they are prerequisites. Interpretability must go beyond surface-level explanations and be woven into the learning and decision-making fabric of the agent. Ethical alignment, likewise, must be op-

erationalized at the level of reward modeling, preference elicitation, and interactive feedback, not treated as an afterthought. Recent advances in causal reasoning, policy summarization, and natural language rationalization offer practical avenues for making agent behavior transparent and scrutinizable. Coupled with human-in-the-loop oversight and normative constraints, these techniques are essential to ensure that LLM-RL systems are not only performant but also accountable and aligned with societal values.

While these challenges are significant, they also present a unique opportunity to redefine the boundaries of reinforcement learning and catalyze transformative progress across both research and real-world applications:

- *Dynamic Knowledge Adaptation for Real-Time Decision-Making*: LLMs can serve as dynamic knowledge modules that adapt to real-time changes in the environment, enabling RL agents to continuously update their policies without exhaustive retraining. This facilitates robust decision-making in non-stationary and high-variance environments, such as autonomous driving and real-time strategy games.
- *Augmenting Exploration with Knowledge-Driven Priors*: Traditional RL agents rely on random exploration, which can be inefficient. LLMs can provide structured priors based on accumulated knowledge, guiding agents toward more promising state-action spaces. This knowledge-driven exploration accelerates learning in sparse-reward settings and complex environments.
- *Cross-Domain Transfer and Generalization*: The broad generalization capabilities of LLMs enable RL agents to transfer knowledge across domains with minimal adaptation. This paves the way for universal RL agents that can operate effectively in diverse environments, from healthcare simulations to industrial robotics, without domain-specific retraining.
- *Human-AI Collaborative Learning*: By leveraging LLMs for natural language interaction, RL agents can seamlessly incorporate human feedback into their learning loops. This human-AI collaboration enhances the alignment of agent behavior with human values and improves safety in critical tasks through intuitive guidance.

5 Conclusion

This survey presents a comprehensive synthesis of grounding methodologies for integrating large language models (LLMs) into reinforcement learning (RL) systems. It covers both training-free approaches, including instructive prompting, retrieval-augmented reasoning, and multimodal integration, and fine-tuning-based strategies such as environment-specific adaptation, feedback-driven refinement, and multi-task transfer. The taxonomy in Table 1 organizes grounding methods into coherent paradigms, clarifying how LLM capabilities in reasoning, abstraction, and generalization can be harnessed to enhance RL agents' adaptability in dynamic environments. While recent progress is encouraging, significant challenges remain in scaling grounding methods, balancing generalization with task specificity, and integrating symbolic reasoning with embodied, multimodal interaction. Addressing these open questions, especially through safe and data-efficient feedback-driven fine-tuning, will be critical for advancing toward context-aware agents capable of effective reasoning, planning, and real-world decision-making.

Ethical Statement

This paper is a survey of previously published research and does not involve original experimentation, data collection, human participants, or animal subjects. Accordingly, it does not raise any direct ethical concerns.

Acknowledgments

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, Career Development Fund (CDF) of the Agency for Science, Technology and Research (A*STAR) (No: C233312007), and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the National Research Foundation, Singapore, and Infocomm Media Development Authority.

References

[Baheti *et al.*, 2024] Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark O. Riedl. Leftover lunch: Advantage-based offline reinforcement learning for language models. In *ICLR*, 2024.

[Carta *et al.*, 2023] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In *ICML*, volume 202, pages 3676–3713, 2023.

[Dalal *et al.*, 2024] Murtaza Dalal, Tarun Chiruvolu, Devendra Singh Chaplot, and Ruslan Salakhutdinov. Plan-seq-learn: Language model guided RL for solving long horizon robotics tasks. In *ICLR*, 2024.

[DeepMind, 2024] Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across modalities. *arXiv preprint arXiv:2403.05530*, 2024.

[DeepSeek, 2023] DeepSeek. Deepseek-v2: Bridging chat and completion with mixture of experts. *arXiv preprint arXiv:2311.17035*, 2023.

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *ICML*, 2023.

[Du *et al.*, 2023] Yuqing Du, Olivia Watkins, Zihan Wang, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. In *ICML*, 2023.

[Fan *et al.*, 2022] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Neurips*, 2022.

[Gao *et al.*, 2024] Haihan Gao, Rui Zhang, Qi Yi, Hantao Yao, Haochen Li, Jiaming Guo, Shaohui Peng, Yunkai Gao, QiCheng Wang, Xing Hu, Yuanbo Wen, Zihao Zhang, Zidong Du, Ling Li, Qi Guo, and Yunji Chen. Prompt-based visual alignment for zero-shot policy transfer. *CoRR*, abs/2406.03250, 2024.

[Hu and Clune, 2023] Shengran Hu and Jeff Clune. Thought cloning: Learning to think while acting by imitating human thinking. In *Neurips*, 2023.

[Hu and Sadigh, 2023] Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-ai coordination. In *ICML*, 2023.

[Hu *et al.*, 2024] Mengkang Hu, Yao Mu, Xinmiao Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. Tree-planner: Efficient close-loop task planning with large language models. In *ICLR*, 2024.

[Huang *et al.*, 2022] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *CoRL*, 2022.

[Li *et al.*, 2024] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *ICLR*, 2024.

[Luo *et al.*, 2024a] Linhao Luo, Yuan-Fang Li, Reza Haf, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR*, 2024.

[Luo *et al.*, 2024b] Lirui Luo, Guoxi Zhang, Hongming Xu, Yaodong Yang, Cong Fang, and Qing Li. End-to-end neuro-symbolic reinforcement learning with textual explanations. In *ICML*, 2024.

[Ma *et al.*, 2023] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *CoRR*, abs/2305.14283, 2023.

[Niu *et al.*, 2024] Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. SCHEMA: state changes matter for procedure planning in instructional videos. In *ICLR*, 2024.

[Noukhovitch *et al.*, 2023] Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C. Courville. Language model alignment with elastic reset. In *Neurips*, 2023.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Pang *et al.*, 2024a] Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. Language model self-improvement by reinforcement learning contemplation. In *ICLR*, 2024.

[Pang *et al.*, 2024b] Jing-Cheng Pang, Si-Hang Yang, Kaiyuan Li, Jiaji Zhang, Xiong-Hui Chen, Nan Tang, and Yang Yu. Knowledgeable agents by offline reinforcement learning from large language model rollouts. In *Neurips*, 2024.

[Peng *et al.*, 2024] Andi Peng, Ilia Sucholutsky, Belinda Z. Li, Theodore R. Sumers, Thomas L. Griffiths, Jacob Andreas, and Julie Shah. Learning with language-guided state abstractions. In *ICLR*, 2024.

[Qi *et al.*, 2024] Yong Qi, Gabriel Kyebambo, Siyuan Xie, Wei Shen, Shenghui Wang, Bitao Xie, Bin He, Zhipeng Wang, and Shuo Jiang. Safety control of service robots with llms and embodied knowledge graphs. *CoRR*, abs/2405.17846, 2024.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

[Rana *et al.*, 2023] Krishan Rana, Jesse Haviland, Sourav Garg, and Niko Sünderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot planning. In *CoRL*, 2023.

[Rocamonde *et al.*, 2024] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *ICLR*, 2024.

[Siyao *et al.*, 2024] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

[Sumers *et al.*, 2023] Theodore R. Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling internet-scale vision-language models into embodied agents. In *ICML*, 2023.

[Sun *et al.*, 2024] Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. SALMON: self-alignment with instructable reward models. In *ICLR*, 2024.

[Szot *et al.*, 2024] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R. Devon Hjelm, and Alexander T. Toshev. Large language models as generalizable policies for embodied tasks. In *ICLR*, 2024.

[Tan *et al.*, 2024] Weihao Tan, Wentao Zhang, Shanqi Liu, Longtao Zheng, Xinrun Wang, and Bo An. True knowledge comes from practice: Aligning large language

models with embodied environments via reinforcement learning. In *ICLR*, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang *et al.*, 2023a] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. JARVIS-1: open-world multi-task agents with memory-augmented multimodal language models. *CoRR*, abs/2311.05997, 2023.

[Wang *et al.*, 2023b] Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, abs/2302.01560, 2023.

[Wang *et al.*, 2024] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *TMLR*, 2024.

[Wen *et al.*, 2024] Muning Wen, Ziyu Wan, Weinan Zhang, Jun Wang, and Ying Wen. Reinforcing language agents via policy optimization with action decomposition. In *Neurips*, 2024.

[Wu *et al.*, 2023a] Yue Wu, Yewen Fan, Paul Pu Liang, Amos Azaria, Yuanzhi Li, and Tom Mitchell. Read and reap the rewards: Learning to play atari with the help of instruction manuals. In *NeurIPS*, 2023.

[Wu *et al.*, 2023b] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *NeurIPS*, 2023.

[Xiang *et al.*, 2023] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. In *NeurIPS*, 2023.

[Yuan *et al.*, 2023] Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *CoRR*, abs/2303.16563, 2023.

[Zeng *et al.*, 2024] Yuwei Zeng, Yao Mu, and Lin Shao. Learning reward for robot skills using large language models via self-alignment. *CoRR*, 2405.07162, 2024.

[Zhai *et al.*, 2024] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *CoRR*, abs/2405.10292, 2024.

[Zhang *et al.*, 2024] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, and Chuang Gan. Building cooperative embodied agents modularly with large language models. In *ICLR*, 2024.

[Zhao *et al.*, 2024] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *ICLR*, 2024.

[Zhen *et al.*, 2024] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *ICML*, 2024.

[Zhou *et al.*, 2023] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. ESC: exploration with soft commonsense constraints for zero-shot object navigation. In *ICML*, volume 202, pages 42829–42842, 2023.

[Zhou *et al.*, 2024] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn RL. In *ICML*, 2024.

[Zhu *et al.*, 2023] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *CoRR*, 2305.17144, 2023.

[Zhu *et al.*, 2024] Jiaqiang Ye Zhu, Carla Gomez Cano, David Vázquez Bermudez, and Michal Drozdzal. Incoro: In-context learning for robotics control with feedback loops. *CoRR*, abs/2402.05188, 2024.