Novelty Assessment Report

Paper: Meta-RL Induces Exploration in Language Agents
PDF URL: https://openreview.net/pdf?id=4GiBscHW1k
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Reinforcement learning (RL) has enabled the training of Large Language Model (LLM) agents to interact with the environment and to solve multi-turn longhorizon tasks. However, the RL-trained agents often struggle in tasks that require active exploration and fail to efficiently adapt from trial-and-error experiences. In this paper, we present LaMer, a general Meta-RL framework that enables LLM agents to actively explore and learn from the environment feedback at test time. LaMer consists of two key components: (i) a cross-episode training framework to encourage exploration and long term rewards optimization; and (ii) in-context policy adaptation via reflection, allowing the agent to adapt their policy from task feedback signal without gradient update. Experiments across diverse environments show that LaMer significantly improves performance over RL baselines, with 11\%, 14\%, and 19\% performance gains on Sokoban, MineSweeper and Webshop, respectively. Moreover, LaMer also demonstrates better generalization to more challenging or previously unseen tasks compared to the RL-trained agents. Overall, our results demonstrate that meta-reinforcement learning provides a principled approach to induce exploration in language agents, enabling more robust adaptation to novel environments through learned exploration strategies.

Core Task Landscape

This paper addresses: **Inducing Exploration in Language Agents through Meta-Reinforcement Learning**

A total of **14 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Meta-RL Frameworks for Language Agent Exploration**
- **Meta-Learning Exploration Strategies and Curiosity Mechanisms**
- **LLM-Guided Meta-RL for Recommendation and Content Discovery**
- **Offline Meta-RL with Natural Language Supervision**
- **Meta-Learning for Balancing Imitation and Exploration**
- **Reflective Memory and Reusable Meta-Policies for LLM Agents**
- **Meta-RL in Specialized Application Domains**

Complete Taxonomy Tree

- Inducing Exploration in Language Agents through Meta-Reinforcement Learning Survey Taxonomy
- Meta-RL Frameworks for Language Agent Exploration
  - Cross-Episode Meta-RL with In-Context Adaptation ★ (2 papers)
  - [0] Meta-RL Induces Exploration in Language Agents (Anon et al., 2026) View paper
  - [1] Can large language models explore in-context? (Dylan Foster, 2024) View paper
  - Meta-RL for Multi-Skill Mastery and Generalization (2 papers)
  - [8] Meta-reinforcement learning for mastering multiple skills and generalizing across environments in text-based games (Zhenjie Zhao, 2021) View paper
  - [14] Logits and Labyrinths: Using Meta RL to Play Text Based Adventure Games (J Arifov, n.d.) View paper
  - Emergent Exploration through Exploitation-Only Objectives (1 papers)
  - [10] Exploitation Is All You Need... for Exploration (Roberts, 2025) View paper
- Meta-Learning Exploration Strategies and Curiosity Mechanisms
  - Meta-Learned Exploration Policies with Decision Transformers (1 papers)
  - [9] Meta-Learning Exploration Strategies with Decision Transformers (Welch, 2025) View paper
  - Meta-Learning Curiosity Algorithms via Program Search (1 papers)
  - [11] Meta-learning curiosity algorithms (Alet, 2020) View paper
- LLM-Guided Meta-RL for Recommendation and Content Discovery
  - Self-Evolving Meta-RL with LLM Policy Adaptation (2 papers)
  - [2] MetaEvo-Rec: Self-Evolving Meta-Reinforcement Learning Recommendation with Large-Language-Model Guided Policy Adaptation (Alamdari, 2025) View paper
  - [6] Self-Reflective Multi-Agent Reinforcement Architecture for Autonomous Recommendation Policy Evolution (Zare, 2025) View paper
- Offline Meta-RL with Natural Language Supervision (1 papers)
  - [3] Text-to-Decision Agent: Offline Meta-Reinforcement Learning from Natural Language Supervision (Zhang ShiLin, 2025) View paper
- Meta-Learning for Balancing Imitation and Exploration (1 papers)
  - [5] AMFT: Aligning LLM Reasoners by Meta-Learning the Optimal Imitation-Exploration Balance (HE Lixuan, 2025) View paper

- Reflective Memory and Reusable Meta-Policies for LLM Agents (1 papers)
  - [12] Meta-Policy Reflexion: Reusable Reflective Memory and Rule Admissibility for Resource-Efficient LLM Agent (Wu ChunLong, 2025) View paper
- Meta-RL in Specialized Application Domains
  - Meta-RL for Natural Language Generation (1 papers)
  - [4] Metaex-gan: Meta exploration to improve natural language generation via generative adversarial networks (Yun-Yen Chuang, 2023) View paper
  - Meta-RL for Atmospheric Guidance and Control (1 papers)
  - [7] Transformer-based Robust Feedback Guidance for Atmospheric Powered Landing (Jacopo Carradori, 2025) View paper
  - Dialogue as Implicit Meta-RL for Complex Non-Verifiable Domains (1 papers)
  - [13] Self-evolving expertise in complex non-verifiable subject domains: dialogue as implicit meta-RL (Bailey, 2025) View paper

Narrative

Core task: inducing exploration in language agents through meta-reinforcement learning. The field structure reflects diverse approaches to enabling language-based agents to explore effectively across tasks. At the highest level, the taxonomy distinguishes between frameworks that directly integrate meta-RL with language agent architectures, methods that meta-learn exploration strategies or curiosity mechanisms, and specialized applications such as recommendation systems or domain-specific deployments. Some branches focus on offline meta-RL with natural language supervision, while others address the balance between imitation and exploration or develop reflective memory systems that allow agents to reuse meta-policies. Representative works span from foundational meta-RL text game environments like Meta-RL Text Games[8] to more recent efforts such as MetaEvo-Rec[2] for content discovery and Text-to-Decision Agent[3] for offline decision-making with language supervision.

A particularly active line of work centers on cross-episode meta-RL with in-context adaptation, where agents leverage large language models to rapidly adjust exploration behavior based on accumulated experience within and across episodes. Meta-RL Language Exploration[0] exemplifies this direction by combining meta-reinforcement learning with in-context learning to induce exploratory behavior in language agents. This approach contrasts with LLM In-Context Exploration[1], which similarly exploits in-context mechanisms but may differ in how meta-level policies are structured or updated. Meanwhile, methods like Meta-Learning Curiosity[11] and Exploitation for Exploration[10] emphasize learning intrinsic motivation signals, and Meta-Policy Reflexion[12] integrates reflective memory to enable reusable exploration strategies. The original paper sits within the cross-episode adaptation cluster, sharing the emphasis on in-context learning with LLM In-Context Exploration[1] while potentially offering distinct meta-RL formulations or exploration incentives that differentiate its contribution from closely related contemporaries.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Can large language models explore in-context?

Authors: Dylan Foster, Keegan Harris, Akshay Krishnamurthy, Aleksandrs Slivkins, Cyril Zhang | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ resulted in satisfactory exploratory behavior: GPT-4 with chainâ¦ not result in robust exploratory behavior, including those with â¦ with a Bayesian meta-reinforcement learning perspective [64]â¦

Relationship Analysis

Both papers belong to the Cross-Episode Meta-RL with In-Context Adaptation category, focusing on meta-reinforcement learning frameworks that enable language agents to explore and adapt policies in-context. While the original paper (LAMER) proposes a training framework that combines cross-episode RL with self-reflection for policy adaptation across multiple episodes, this candidate paper investigates whether existing pre-trained LLMs can natively explore in-context within multi-armed bandit environments without training interventions. The key difference is that LAMER actively trains agents using meta-RL objectives to induce exploration, whereas this paper evaluates the inherent exploration capabilities of frozen LLMs through prompt engineering alone.

Contributions Analysis

Overall novelty summary. The paper introduces LaMer, a meta-reinforcement learning framework that combines cross-episode training with in-context policy adaptation via reflection to improve exploration in LLM agents. According to the taxonomy, this work resides in the 'Cross-Episode Meta-RL with In-Context Adaptation' leaf, which contains only two papers total: the original submission and one sibling work (LLM In-Context Exploration). This positioning suggests the paper targets a relatively sparse but emerging research direction within the broader meta-RL for language agents landscape, which encompasses fourteen papers across multiple branches addressing exploration, curiosity mechanisms, and specialized applications.

The taxonomy reveals that neighboring research directions include meta-learned exploration policies using decision transformers, curiosity algorithm search via program synthesis, and reflective memory systems for reusable meta-policies. The original paper's leaf explicitly excludes methods lacking cross-episode training or in-context adaptation mechanisms, distinguishing it from branches focused on emergent exploration through exploitation-only objectives or offline meta-RL with natural language supervision. This structural context indicates the work bridges meta-RL training paradigms with LLM in-context learning capabilities, occupying a niche between traditional meta-RL frameworks and pure prompt-based adaptation approaches that do not employ cross-episode optimization.

Among the three contributions analyzed, the literature search examined twenty-seven candidates total, identifying refutable prior work for each component. The core LaMer framework (ten candidates examined, one refutable) and cross-episode training with trajectory discounting (seven candidates, one refutable) each show limited overlap within the search scope. In-context policy adaptation via self-reflection (ten candidates, two refutable) appears to have more substantial prior work among the examined papers. These statistics reflect a targeted semantic search rather than exhaustive coverage, suggesting that while some overlap exists in the examined subset, the specific combination of cross-episode meta-RL with reflection-based adaptation may represent a novel integration within the limited candidate pool.

Based on the analysis of thirty candidates from top-K semantic search, the work appears to occupy a sparsely populated research direction with one closely related sibling paper in its taxonomy leaf. The contribution-level statistics indicate varying degrees of prior work across components, with reflection-based adaptation showing more overlap than the cross-episode training mechanism. However, the limited search scope means this assessment captures only a snapshot of the most semantically similar work, not a comprehensive field survey, and the novelty evaluation remains contingent on this bounded literature examination.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: LAMER: A Meta-RL framework for training LLM agents

Description: The authors propose LAMER, a meta-reinforcement learning framework designed to train large language model agents. This framework enables agents to balance exploration and exploitation across multiple episodes, allowing them to actively explore environments and adapt their policies at test time without gradient updates.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Instructrag: Leveraging retrieval-augmented generation on instruction graphs for llm-based task planning

URL: View paper

Brief Assessment

InstructRAG[17] uses meta-reinforcement learning within a multi-agent framework specifically for task planning with retrieval-augmented generation, not for general LLM agent training with exploration-exploitation balance across episodes. The frameworks serve different purposes and architectural designs.

---

2. Multimodal Agentic AI Architecture for High Frequency Trading Using Reinforcement Learning and Temporal Graph Encoders

URL: View paper

Brief Assessment

Multimodal Trading Agent[23] applies RL to financial trading with temporal graph encoders, not a general meta-RL framework for LLM agents with test-time adaptation capabilities.

---

3. Efficient meta reinforcement learning for preference-based fast adaptation

URL: View paper

Brief Assessment

Efficient Meta-RL Preferences[19] focuses on preference-based fast adaptation in traditional RL environments (MuJoCo tasks), not on training large language model agents for multi-turn interactive tasks.

---

4. MetaEvo-Rec: Self-Evolving Meta-Reinforcement Learning Recommendation with Large-Language-Model Guided Policy Adaptation

URL: View paper

Brief Assessment

MetaEvo-Rec[2] focuses on recommendation systems using meta-RL for policy adaptation, not general LLM agent training frameworks for interactive environments. The candidate's domain (recommendation) and application context differ fundamentally from LAMER's multi-turn interactive task environments.

---

5. Meta-Learning Online Adaptation of Language Models

URL: View paper

Brief Assessment

Meta-Learning Online Adaptation[16] focuses on meta-learning token importance weights for online fine-tuning of language models on document streams, not on training interactive agents with multi-episode reinforcement learning for exploration and exploitation in environments.

---

6. Optimizing test-time compute via meta reinforcement fine-tuning

URL: View paper

Prior Art Analysis

Meta-RL Test-Time[15] demonstrates that meta-reinforcement learning frameworks for training language model agents with test-time adaptation existed prior to LAMER. Both papers formalize the problem of optimizing test-time compute as a meta-RL problem where the agent learns to balance exploration and exploitation across multiple episodes. Meta-RL Test-Time[15] explicitly segments output streams into episodes, uses dense rewards based on progress throughout thinking traces, and enables in-context policy adaptation without gradient updates - all core features claimed as novel in LAMER.

Evidence

Evidence 1 - **Rationale**: Both papers formalize test-time compute optimization as a meta-RL problem with multi-episode structures. Meta-RL Test-Time[15] explicitly segments output streams into episodes for test-time adaptation, demonstrating this approach existed before LAMER. - **Original**: we present lamer, a general meta-rl framework that enables llm agents to actively explore and learn from the environment feedback at test time. lamerconsists of two key components:(i)a cross-episode training framework to encourage exploration and long-term rewards optimization; and(ii)in-context pol... - **Candidate**: we formalize the problem of optimizing test-time compute as a meta-reinforcement learning (rl) problem, which provides a principled perspective on spending test-time compute. this perspective enables us to view the long output stream from the llm as consisting of several episodes run at test time an...

Evidence 2 - **Rationale**: Both frameworks use episode-based segmentation with feedback mechanisms for in-context adaptation. Meta-RL Test-Time[15] uses dense rewards based on progress across episodes, while LAMER uses reflection, but the underlying meta-RL structure with episode-based adaptation is present in both. - **Original**: in lamer, we propose a self-reflection based strategy(shinn et al., 2023) toadapt the policy in-context(brown et al., 2020; laskin et al., 2023). specifically, after each episode finishes, we prompt the agent to generate the textual reflection on the previous attempt, providing specific feedback and... - **Candidate**: mrt uses dense rewards based on progress throughout the thinking trace (segmented into 'episodes') to improve final performance and test-time efficiency. standard fine-tuning only trains models with outcome rewards at the end, thus reinforcing several traces that make subpar progress but somehow suc...

Evidence 3 - **Rationale**: Both papers describe multi-episode meta-RL frameworks for LLM training that balance exploration and exploitation. Meta-RL Test-Time[15] explicitly segments outputs into episodes and frames the problem as learning explore-exploit tradeoffs, demonstrating this framework existed prior to LAMER. - **Original**: building upon this, we propose lamer(llmagent withmeta-rl), a general meta-rl framework for llm agent training. lamercontains two important design factors. first, unlike standard single-episode rl, lameris designed around a multi-episode structure to train the agent to solve the problem through tria... - **Candidate**: to build our approach, we segment the output stream from an llm on a given problem into multiple episodes (figure 2). if we were to only care about (a) efficiency, then the llm should only learn to exploitanddirectlyoutputthefinalanswerwithout spending too many episodes. on the other hand, if the ll...

Evidence 4 - **Rationale**: This pair directly refutes LAMER's novelty claim. Meta-RL Test-Time[15] formalized and implemented a meta-RL framework for LLM training before LAMER, contradicting the claim that LAMER is 'the first time a meta-rl framework is used for llm agent training.' - **Original**: to the best of our knowledge, this is the first time a meta-rl framework is used for llm agent training. - **Candidate**: we formalize the problem of optimizing test-time compute as a meta-reinforcement learning (rl) problem, which provides a principled perspective on spending test-time compute.

---

7. Meta-Learning Reinforcement Learning for Crypto-Return Prediction

URL: View paper

Brief Assessment

Meta-RL Crypto Returns[18] applies meta-learning RL to cryptocurrency trading with financial reward signals, not to training general LLM agents for interactive multi-turn tasks with environment exploration.

---

8. Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning

URL: View paper

Brief Assessment

Meta-RL Lifelong Context[20] focuses on meta-RL for general RL tasks with distributional shift robustness, not specifically on training LLM agents for multi-turn language-based tasks. The candidate uses transformers for Bayesian inference in traditional RL environments, while LAMER addresses exploration-exploitation in language agents with self-reflection mechanisms.

---

9. Sample-Efficient Online Learning in LM Agents via Hindsight Trajectory Rewriting

URL: View paper

Brief Assessment

Hindsight Trajectory Rewriting[21] focuses on hindsight experience replay for online learning from failed trajectories, not meta-RL across multiple episodes for exploration-exploitation balance. The candidate addresses trajectory rewriting and counterfactual generation, while LAMER addresses cross-episode training and in-context policy adaptation.

---

10. Adapting Like Humans: A Metacognitive Agent with Test-time Reasoning

URL: View paper

Brief Assessment

Metacognitive Test-Time Agent[22] focuses on vision-language models for Atari games with metacognitive reasoning and memory systems, not on meta-RL frameworks for training LLM agents in text-based interactive environments.

Contribution 2: Cross-episode training framework with trajectory discount factor

Description: The authors introduce a training scheme that optimizes rewards across multiple sequential episodes rather than single episodes. This approach uses a trajectory discount factor to assign credit across episodes, encouraging the agent to explore in early episodes and exploit gathered information in later ones.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. When Does Reward Drive Exploration?

URL: View paper

Brief Assessment

Reward Drives Exploration[31] focuses on conditions for emergent exploration in meta-RL (recurring structure, memory, credit assignment) rather than proposing a cross-episode training framework with trajectory discount factors as a novel contribution.

---

2. Bachelor's Thesis Submitted in 2025

URL: View paper

Brief Assessment

Bachelor Thesis[28] focuses on agricultural robot navigation using TD-MPC2 for single-task crop-following, not multi-episode meta-RL frameworks with cross-episode credit assignment for exploration-exploitation trade-offs across sequential episodes.

---

3. Efficient Cross-Episode Meta-RL

URL: View paper

Prior Art Analysis

Efficient Cross-Episode Meta-RL[24] demonstrates that cross-episode training frameworks were proposed prior to the original paper. The candidate paper introduces a hierarchical transformer architecture that processes sequences of transitions across multiple episodes to learn task representations, using both intra-episode and cross-episode information. While the original paper uses a trajectory discount factor γ_traj for credit assignment across episodes, the candidate achieves similar cross-episode learning through its Cross-Episodic Transformer (CET) architecture that processes episode representations without requiring explicit discount factors. Both approaches optimize rewards across multiple sequential episodes rather than single episodes, though through different mechanisms.

Evidence

Evidence 1 - **Rationale**: Both papers propose training frameworks that utilize multiple episodes sequentially. The candidate's ECET processes cross-episode information through its architecture, while the original uses explicit trajectory discounting. - **Original**: cross-episode training framework.in the training of lamer, each trial consists ofnepisodes sequentially generated by the agent: t= (τ (0), τ(1), . . . , τ(n-1)),whereτ (n) ∼π (n) θ (·), n∈[0, n-1] - **Candidate**: we introduce efficient cross-episodic transformers (ecet), a new algorithm for online meta-reinforcement learning that addresses the challenge of enabling reinforcement learning agents to perform effectively in previously unseen tasks. we demonstrate how past episodes serve as a rich source of in-co...

Evidence 2 - **Rationale**: Both papers explicitly optimize across multiple episodes. The original uses a mathematical formulation with discount factors, while the candidate uses transformer architecture to capture cross-episode dynamics. - **Original**: to enhance the exploration and maximize the long-term reward, in lamerframework we define the discounted returng (n) t across the episodesoftas: g(n) t =g (n) t|{z} within-the-episode + n-1x m=n+1 γm-n traj g(m) | {z } cross-episode - **Candidate**: our approach enhances the adaptability of rl agents across diverse sets of tasks by leveraging both intra-episode and cross-episode experiences, providing a more comprehensive learning process. our learned algorithm is capable of outperforming the previous state-of-the-art and providing more efficie...

Evidence 3 - **Rationale**: While the original paper uses a trajectory discount factor to balance exploration and exploitation across episodes, the candidate achieves similar cross-episode credit assignment through its CET architecture that processes episode representations. - **Original**: here,γ traj is an important factor for the trade-off betweenexplorationandexploitation. ideally, smallγ traj biases the objective towards early episodes and will lead to rapid exploitation to solve the problem. in comparison, a largerγtraj emphasizes long-horizon return and therefore encourages more... - **Candidate**: the cross-episodic transformer (cet) computes a representation that transforms a sequence of t transitions from the e-th episode into a vector representation ze ∈ rk as formalized below: ze = iet (χe 1, . . . , χe t) , iet : (s x a x r x {0, 1})t → rk, ze ∈ rk

---

4. Policy gradient

URL: View paper

Brief Assessment

Policy Gradient[27] discusses standard discount factors for expected return within episodes, not cross-episode trajectory discount factors for multi-episode meta-RL training as proposed in the original paper.

---

5. Delayed Geometric Discounts: An Alternative Criterion for Reinforcement Learning

URL: View paper

Brief Assessment

Delayed Geometric Discounts[26] focuses on generalizing geometric discounts within single episodes using delayed discount factors, not on cross-episode training frameworks that optimize across multiple sequential episodes with trajectory-level discounting.

---

6. Motivated optimal developmental learning for sequential tasks without using rigid time-discounts

URL: View paper

Brief Assessment

Motivated Developmental Learning[29] focuses on emergent neural representations for sequential tasks with biologically-inspired reinforcement (serotonin/dopamine systems), not cross-episode meta-RL training frameworks. The candidate explicitly avoids time-discount approaches used in the original paper's trajectory discount factor.

---

7. Reciprocal reward influence encourages cooperation from self-interested agents

URL: View paper

Brief Assessment

Reciprocal Reward Influence[25] focuses on opponent shaping through reciprocal rewards within single episodes in multi-agent settings, not on cross-episode training frameworks for single-agent exploration and exploitation.

Contribution 3: In-context policy adaptation via self-reflection

Description: The authors develop a mechanism where the agent generates textual reflections after each episode to summarize experiences and adjust strategy. This enables policy adaptation through context modification rather than parameter updates, naturally leveraging the in-context learning capabilities of large language models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Reflexion: Language agents with verbal reinforcement learning

URL: View paper

Prior Art Analysis

Reflexion[32] demonstrates that in-context policy adaptation via self-reflection was already proposed and implemented before the original paper. Reflexion[32] explicitly describes a mechanism where agents generate textual reflections after episodes to summarize experiences and adjust strategy, enabling policy adaptation through context modification rather than parameter updates. This directly refutes the novelty claim of the original paper's contribution, as Reflexion[32] published this approach in 2023 at NeurIPS, predating the original submission.

Evidence

Evidence 1 - **Rationale**: Both papers describe the same core mechanism: agents reflect on task feedback and use this reflection to improve decision-making in subsequent trials without parameter updates. - **Original**: in-context policy adaptation via reflection, allowing the agent to adapt their policy from task feedback signal without gradient update - **Candidate**: reflexion agents verbally reflect on task feedback signals, then maintain their own reflective text in an episodic memory buffer to induce better decision-making in subsequent trials

Evidence 2 - **Rationale**: Reflexion[32] explicitly states it reinforces agents without updating weights, using verbal reflection on task feedback stored in episodic memory - the exact approach claimed as novel by the original paper. - **Original**: in-context policy adaptation via reflection, allowing the agent to adapt their policy from task feedback signal without gradient update - **Candidate**: we propose reflexion, a novel framework to reinforce language agents not by updating weights, but instead through linguistic feedback. concretely, reflexion agents verbally reflect on task feedback signals, then maintain their own reflective text in an episodic memory buffer to induce better decisio...

---

2. REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments

URL: View paper

Brief Assessment

REGENT[39] focuses on retrieval-augmented in-context learning without self-reflection mechanisms. The candidate uses retrieved state-action pairs for adaptation, not textual reflections generated by the agent to summarize experiences.

---

3. PromptPilot: Autonomous Prompt Optimization via Genetic Particle Filtering and Dynamic Exploration

URL: View paper

Brief Assessment

PromptPilot[38] focuses on prompt optimization through genetic particle filtering for improving LLM outputs, not on multi-episode RL agent training with self-reflection for policy adaptation. The candidate addresses a different problem domain (prompt engineering) rather than agentic reinforcement learning with cross-episode learning.

---

4. Language agent tree search unifies reasoning acting and planning in language models

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

5. Evotest: Evolutionary test-time learning for self-improving agentic systems

URL: View paper

Brief Assessment

EvoTest[36] focuses on evolving entire agent configurations (prompts, memory, hyperparameters, tool-use) between episodes in text-based games, rather than the original paper's in-context reflection mechanism for RL agents. The candidate uses reflection as one component within a broader evolutionary framework, not as the primary adaptation mechanism.

---

6. Self-Adapting Financial Agents: Evolution Through Feedback-Driven Meta-Prompt Optimization

URL: View paper

Brief Assessment

Self-Adapting Financial Agents[41] focuses on meta-prompt optimization for financial trading agents, not general RL frameworks for multi-turn interactive tasks. The candidate's feedback-driven meta-prompt optimization operates in a financial domain with different objectives than the original paper's multi-episode exploration-exploitation framework.

---

7. Agentic context engineering: Evolving contexts for self-improving language models

URL: View paper

Prior Art Analysis

Agentic Context Engineering[34] demonstrates prior work on in-context policy adaptation using self-reflection without parameter updates. The candidate paper explicitly describes a reflector component that 'distills concrete insights from successes and errors' and integrates them into context updates, enabling adaptation through context modification rather than weight updates. This mechanism closely parallels the original paper's self-reflection approach for policy adaptation, where agents generate textual reflections to adjust strategy through context rather than gradient updates.

Evidence

Evidence 1 - **Rationale**: Both papers describe mechanisms where agents adapt their policy through reflection-based context modification without gradient updates. The candidate's reflector component serves the same function as the original's self-reflection mechanism. - **Original**: in-context policy adaptation via reflection, allowing the agent to adapt their policy from task feedback signal without gradient update - **Candidate**: ace introduces a structured division of labor across three roles (figure 4): thegenerator, which produces reasoning trajectories; thereflector, which distills concrete insights from successes and errors; and thecurator, which integrates these insights into structured context updates

Evidence 2 - **Rationale**: Both papers describe generating textual reflections or natural language feedback after episodes/attempts to guide subsequent behavior through context modification, implementing in-context adaptation without parameter updates. - **Original**: we propose a self-reflection based strategy(shinn et al., 2023) toadapt the policy in-context(brown et al., 2020; laskin et al., 2023). specifically, after each episode finishes, we prompt the agent to generate the textual reflection on the previous attempt, providing specific feedback and plan to g... - **Candidate**: context adaptation (or context engineering) refers to methods that improve model behavior by constructing or modifying inputs to an llm, rather than altering its weights. the current state of the art leveragesnatural language feedback[4, 40, 54]. in this paradigm, a language model inspects the curr...

Evidence 3 - **Rationale**: Both approaches update policy through context modification that accumulates historical information and reflections, enabling in-context learning without weight updates. - **Original**: The policy is therefore updated through modifying the context,π (n) θ (·) =πθ(·|h(n)) whereh (n) denotes the inter-episode memory that contains both the history trajectories and reflections - **Candidate**: instead of condensing knowledge into terse summaries or static instructions, ace treats contexts as evolving playbooks that continuously accumulate, refine, and organize strategies over time

---

8. â¦ GÃ¶del Machine: A Commentary on Novelty and Implications From formal proofs to empirical evolution: re-energizing self-improving AI with the Darwin GÃ¶del â¦

URL: View paper

Brief Assessment

Darwin Gödel Machine[40] focuses on evolutionary optimization of agent structures rather than in-context policy adaptation through self-reflection. The candidate discusses gradient-free optimization and evolution of action-to-reasoning structures, which differs from the original paper's mechanism of generating textual reflections after episodes to modify context for policy adaptation.

---

9. Improving LLM Agent Planning with In-Context Learning via Atomic Fact Augmentation and Lookahead Search

URL: View paper

Brief Assessment

Atomic Fact Lookahead[37] focuses on atomic fact extraction and lookahead search for planning, not on self-reflection mechanisms for policy adaptation. The candidate's fact-based approach differs fundamentally from the original's reflection-based episodic learning framework.

---

10. DORA: Dynamic Optimization Prompt for Continuous Reflection of LLM-based Agent

URL: View paper

Brief Assessment

DORA[35] focuses on optimizing reflection prompts through Bayesian optimization to address 'early stop reflection' in agent frameworks, rather than proposing the fundamental mechanism of in-context policy adaptation via self-reflection itself. The original paper's contribution is the meta-RL framework that enables policy adaptation through reflection across episodes without gradient updates.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Meta-RL Induces Exploration in Language Agents View paper
- [1] Can large language models explore in-context? View paper
- [2] MetaEvo-Rec: Self-Evolving Meta-Reinforcement Learning Recommendation with Large-Language-Model Guided Policy Adaptation View paper
- [3] Text-to-Decision Agent: Offline Meta-Reinforcement Learning from Natural Language Supervision View paper
- [4] Metaex-gan: Meta exploration to improve natural language generation via generative adversarial networks View paper
- [5] AMFT: Aligning LLM Reasoners by Meta-Learning the Optimal Imitation-Exploration Balance View paper
- [6] Self-Reflective Multi-Agent Reinforcement Architecture for Autonomous Recommendation Policy Evolution View paper
- [7] Transformer-based Robust Feedback Guidance for Atmospheric Powered Landing View paper
- [8] Meta-reinforcement learning for mastering multiple skills and generalizing across environments in text-based games View paper
- [9] Meta-Learning Exploration Strategies with Decision Transformers View paper
- [10] Exploitation Is All You Need... for Exploration View paper
- [11] Meta-learning curiosity algorithms View paper
- [12] Meta-Policy Reflexion: Reusable Reflective Memory and Rule Admissibility for Resource-Efficient LLM Agent View paper

- [13] Self-evolving expertise in complex non-verifiable subject domains: dialogue as implicit meta-RL View paper
- [14] Logits and Labyrinths: Using Meta RL to Play Text Based Adventure Games View paper
- [15] Optimizing test-time compute via meta reinforcement fine-tuning View paper
- [16] Meta-Learning Online Adaptation of Language Models View paper
- [17] Instructrag: Leveraging retrieval-augmented generation on instruction graphs for llm-based task planning View paper
- [18] Meta-Learning Reinforcement Learning for Crypto-Return Prediction View paper
- [19] Efficient meta reinforcement learning for preference-based fast adaptation View paper
- [20] Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning View paper
- [21] Sample-Efficient Online Learning in LM Agents via Hindsight Trajectory Rewriting View paper
- [22] Adapting Like Humans: A Metacognitive Agent with Test-time Reasoning View paper
- [23] Multimodal Agentic AI Architecture for High Frequency Trading Using Reinforcement Learning and Temporal Graph Encoders View paper
- [24] Efficient Cross-Episode Meta-RL View paper
- [25] Reciprocal reward influence encourages cooperation from self-interested agents View paper
- [26] Delayed Geometric Discounts: An Alternative Criterion for Reinforcement Learning View paper
- [27] Policy gradient View paper
- [28] Bachelor's Thesis Submitted in 2025 View paper
- [29] Motivated optimal developmental learning for sequential tasks without using rigid time-discounts View paper
- [30] Learning to Explore with In-Context Policy for Fast Peer Adaptation View paper
- [31] When Does Reward Drive Exploration? View paper
- [32] Reflexion: Language agents with verbal reinforcement learning View paper
- [33] Language agent tree search unifies reasoning acting and planning in language models View paper
- [34] Agentic context engineering: Evolving contexts for self-improving language models View paper
- [35] DORA: Dynamic Optimization Prompt for Continuous Reflection of LLM-based Agent View paper
- [36] Evotest: Evolutionary test-time learning for self-improving agentic systems View paper
- [37] Improving LLM Agent Planning with In-Context Learning via Atomic Fact Augmentation and Lookahead Search View paper
- [38] PromptPilot: Autonomous Prompt Optimization via Genetic Particle Filtering and Dynamic Exploration View paper
- [39] REGENT: A Retrieval-Augmented Generalist Agent That Can Act In-Context in New Environments View paper
- [40] â GÃ¶del Machine: A Commentary on Novelty and Implications From formal proofs to empirical evolution: re-energizing self-improving AI with the Darwin GÃ¶del â View paper
- [41] Self-Adapting Financial Agents: Evolution Through Feedback-Driven Meta-Prompt Optimization View paper