# RefRea: Reference-Guided Reasoning with Meta-Cognition for Accurate Language Model Agents

**Yuxiang Mai[1,2,3], Qiyue Yin[1,2,3], Wancheng Ni[1,2,3,*], Jianwei Guo[4], Xiaogang Ouyang[4], Pei Xu[2,3], Kaiqi Huang[1,2,3,*]**

[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]CRISE, Institute of Automation, Chinese Academy of Sciences
[3]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[4]Intelligence Indeed
{maiyuxiang2020, wancheng.ni, pei.xu}@ia.ac.cn, {gaojianli,xinyi}@i-i.ai, {qyyin, kqhuang}@nlpr.ia.ac.cn,

## Abstract

In recent years, with the rapid development of large language models (LLMs), LLM-based agents have achieved remarkable progress across a wide range of tasks. However, reasoning inconsistencies in LLMs still significantly limit the performance of agents in complex decision-making scenarios. Cognitive science research suggests that individuals can benefit from observing others' explicit thinking processes to improve their strategy-making. Inspired by this mechanism, we propose Reference-guided Reasoning with meta-cognition (RefRea), a novel approach that enhances decision-making by introducing a reference language model to guide and calibrate the reasoning model's actions. RefRea enhances reasoning accuracy and stability by integrating a reference model and a meta-cognition module. The reference model relies solely on validated meta-cognition for consistent guidance, while the reasoning model interacts with the environment using both validated and exploratory meta-cognition. Guidance is provided by comparing the action similarity between the reference and reasoning models. This process is supported by the meta-cognition module, which generates summary knowledge by reflecting on action history and environmental feedback, leading to more adaptive and reliable behavior. We evaluate our algorithm in the text-based reasoning environment ScienceWorld. Experimental results demonstrate that RefRea outperforms state-of-the-art methods. Comprehensive ablation studies further highlight the effectiveness of both the reference model and the meta-cognition module.

## Introduction

In recent years, with the continuous advancement of LLMs, LLM-based agents have demonstrated strong potential in tasks such as robotic manipulation (Huang et al. 2023; Liu et al. 2024; Chu et al. 2023), autonomous navigation (Zhou, Hong, and Wu 2024; Doma, Arab, and Xiao 2024), and interactive dialogue systems (Andreas 2022; Bubeck et al. 2023). These agents typically leverage LLMs for task reasoning and decision-making, achieving remarkable performance across a variety of applications, primarily due to the models' advanced capabilities in language understanding and generation. However, as LLMs are inherently based on autoregressive language modeling (Brown et al. 2020) and lack explicit representations of world states and task structures (Vaswani et al. 2017), they are prone to issues such as reasoning inconsistency (Saparov and He 2022), which can severely compromise the agents' reliability and stability in complex decision-making scenarios.

To enhance the performance of LLM-based agents in complex tasks, a number of studies have attempted to guide the reasoning process of the models to improve their reasoning reliability. Chain-of-Thought (CoT) (Wei et al. 2022) prompts models to perform step-by-step reasoning, improving logical consistency. SayCan (Ahn et al. 2022) integrates high-level instructions with low-level feasibility by leveraging environmental feedback to filter viable actions. ReAct (Yao et al. 2023) interleaves "thought" and "action" steps during reasoning to dynamically adjust strategies. Reflexion (Shinn et al. 2023) introduces a reflection mechanism, encouraging models to summarize failed experiences to optimize future decisions. These approaches focus on guiding the reasoning process and leveraging feedback to refine strategies, and have made notable progress. However, they generally lack explicit verification mechanisms for the outputs of the language models themselves, making it difficult to effectively address core issues such as reasoning inconsistencies and factual errors induced by the models.

In real-world decision-making, humans often consult others when faced with complex problems, engaging in the exchange of thought processes to gain a more comprehensive perspective and make more accurate judgments. Cognitive science research has shown that such verbal protocols (Afflerbach 2001) not only help individuals understand others' reasoning paths but also encourage them to reflect on their own cognitive biases, thereby facilitating the refinement of their problem-solving strategies. The ability to monitor and regulate one's own cognitive processes, known as meta-cognition (Flavell 1979; Nelson 1990), plays a crucial role in human reasoning and decision-making. Inspired by these insights, we propose incorporating a reference model to calibrate the outputs of the primary reasoning model. Furthermore, we propose a language-based meta-cognition module that generates high-level strategic summaries from recent

*Corresponding authors
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

reasoning experiences and reuses them to guide future decisions, thereby improving decision-making accuracy.

In this paper, we propose Reference-guided Reasoning with meta-cognition, a novel dual-path reasoning framework designed to improve the performance of LLM-based agents by introducing a reference model that supervises and calibrates the outputs of a reasoning model. To support coordination between the two models, we design a language-based meta-cognition module that generates reflective summaries from historical environmental feedback and action trajectories. This meta-cognition captures high-level strategic information about prior reasoning behaviors and helps guide decision-making. The reference model generates reliable action suggestions based solely on previously validated meta-cognition, while the reasoning model integrates both the validated and newly generated meta-cognition to interact directly with the environment. By measuring the semantic similarity between the two models' actions, RefRea determines whether the newly generated meta-cognition should be retained. This mechanism enables behavior-level supervision and policy calibration, thereby addressing reasoning inconsistencies in LLM-based agents. We evaluate RefRea extensively on the ScienceWorld textual reasoning benchmark. Experimental results show that our method outperforms state-of-the-art baselines in performance and efficiency. Our contributions are summarized as follows:

- We propose RefRea, a dual-path framework that integrates a reasoning model with a reference model, improving consistency and accuracy in LLM-based agents through reference-guided supervision.
- We develop a language-based meta-cognition module that reflects on historical feedback and decision traces to generate high-level strategies, enabling long-term memory formation and supporting knowledge transfer between the reference and reasoning models.
- We conduct extensive experiments on the ScienceWorld benchmark and show that RefRea surpasses competitive baselines across diverse tasks. Comprehensive ablation studies further demonstrate the synergistic contributions of the reference model and the meta-cognition module in improving performance and efficiency.

## Related Works

### LLM Agents for Task Solving

With the rapid development of large language models (LLMs), a new generation of LLM-based agents has emerged, capable of solving complex, multi-step tasks in textual and interactive environments (Guo et al. 2024; Xi et al. 2025; Mou et al. 2024). These agents leverage the language modeling capacity of LLMs not only for understanding task instructions but also for generating sequences of reasoning and actions. A notable example is ReAct (Yao et al. 2023), which interleaves "thoughts" and "actions" in the agent's output, facilitating dynamic reasoning during interactions with the environment. ReAct demonstrates that reasoning steps generated before each action can improve task success by clarifying intent and preserving task focus. Other works, such as SayCan (Yao et al. 2023), integrate language planning with affordance models that evaluate the feasibility of executing actions in a given environment. SayCan grounds language-based intentions into executable behaviors, ensuring that agent plans are both logically coherent and physically realizable. To further improve reasoning adaptability, Reflexion (Shinn et al. 2023) introduces a reflective learning loop, where the agent generates a natural language summary of errors and uses this information to enhance future performance. However, Reflexion relies on a single LLM for both generating and processing reflections, lacking external validation or correction of the reasoning paths. This limitation may result in the reinforcement of biased reflections. In contrast, our approach integrates external validation and meta-cognitive feedback to ensure reasoning consistency and enhance decision-making accuracy.

### Behavioral Supervision

LLMs often suffer from inconsistencies and goal forgetting, especially in long-horizon or partially observable environments (Ji et al. 2023; Zhang et al. 2023). These challenges have driven growing interest in behavioral supervision. Constitutional AI (Bai et al. 2022) exemplifies this direction by using human-written principles and an evaluator model to assess helpfulness and harmlessness during fine-tuning. Likewise, RLAIF (Parisotto, Ba, and Salakhutdinov 2015; Lee et al. 2023) replaces human preference signals with AI-based evaluators to train reward models for instruction following. However, these methods mainly operate at the token or response level, limiting their effectiveness in supervising long-term strategies. Recent approaches such as Value Supervision (Peng et al. 2023) provide reward or critic-based feedback to align actions with task goals, but still focus on value signals rather than natural-language actions or planning traces. In contrast, our method integrates behavioral supervision with meta-cognitive feedback, enabling more consistent and aligned long-horizon reasoning.

### Memory-Augmented LLM Agents

Recent studies have explored integrating memory modules into LLM-based agents to support long-term reasoning and improved contextual understanding. Toolformer (Schick et al. 2023) augments LLMs with external tools and memory states for enhanced decision-making. MemGPT (Packer et al. 2024) introduces an operating system-like memory manager that enables structured reading and writing across interactions. DSPy (Khattab et al. 2023) offer frameworks for modular memory composition in agent systems. Other approaches such as RETRO (Borgeaud et al. 2022) and KNN-LM (Khandelwal et al. 2019) retrieve past examples to improve predictions. Episodic memory methods (Shinn et al. 2023; Xu et al. 2023) allow agents to reflect on past behaviors for continual adaptation. While these methods enhance scalability and retention, they often lack mechanisms to ensure that recalled memories are aligned with current reasoning goals. In contrast, our approach incorporates behavior-level consistency checking to ensure memory relevance and reliability during interactions with the environment, enhancing the ability to solve complex decision-making tasks.

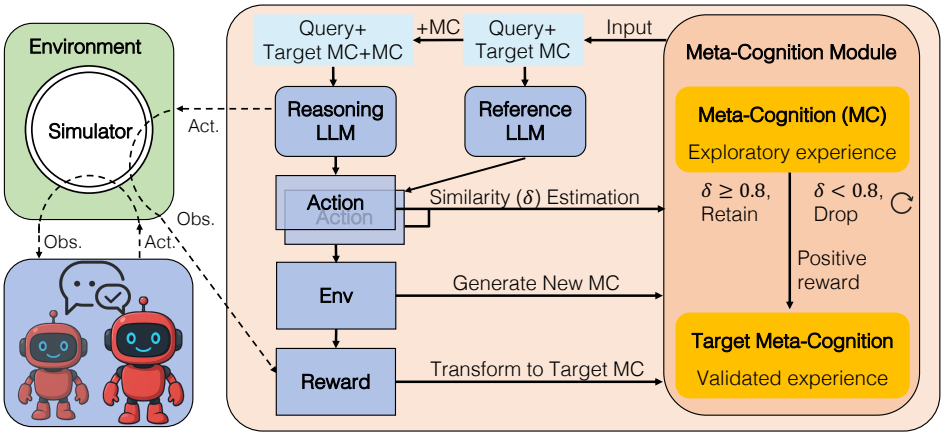


Figure 1: Framework of the proposed method, which includes a reasoning model, a reference model, and a meta-cognition module. It uses the reference model to correct the reasoning model's actions through meta-cognition, achieving higher accuracy and stability in reasoning.

## Method

We propose a dual-path reasoning optimization framework, RefRea (Reference-guided Reasoning with meta-cognition), which incorporates a reference model to supervise the behavior of a reasoning model. RefRea enhances the decision-making capability and reliability of large language model agents by maintaining a dynamic memory mechanism and enforcing behavioral consistency. The core components include a meta-cognition module, a reasoning model, a reference model, and a reference-guided update mechanism.

### Problem Formulation

At each time step $t$, the agent receives an observation $o_t$, maintains a history $H_t = (o_i, a_i)_{i=1}^{t}$, and generates an action $a_t$ based on the history and meta-cognition memory. The reasoning model explores the environment, while a reference model provides stable guidance. The action generation using the large language model $\pi_\theta$ follows:

$$a_t \sim \pi_\theta \left( \cdot \mid Q \left( H_t, \mathcal{M}^{\text{New}}, \mathcal{M}^{\text{Target}} \right) \right), \tag{1}$$

where $Q(\cdot)$ is a query template function. The objective is to learn a meta-cognition memory strategy $\mathcal{M}$ that guide the frozen LLM $\pi_\theta$ to act effectively and maximize cumulative rewards over a finite horizon $T$:

$$\max_{\mathcal{M}} \mathbb{E}_{a_t \sim \pi_\theta(\cdot \mid Q(H_t, \mathcal{M}))} \left[ \sum_{t=1}^{T} r_t \right] \tag{2}$$

### Meta-Cognition Module

To ensure efficient reasoning and long-term adaptability, RefRea maintains two types of memory: the exploratory meta-cognition memory $\mathcal{M}^{\text{New}}$ and the validated long-term memory $\mathcal{M}^{\text{Target}}$. These memory components are dynamically updated during interaction. Exploratory memory $\mathcal{M}^{\text{New}}$ is short-term and generated at fixed intervals based on the reasoning agent's current trajectory. It captures reflective summaries of recent behaviors, allowing the agent to propose tentative strategies and support ongoing reasoning. However, to ensure reliability, $\mathcal{M}^{\text{New}}$ undergoes a behavioral consistency check against the reference model's actions, preventing noisy or inconsistent guidance from influencing decision-making. Validated memory $\mathcal{M}^{\text{Target}}$ serves as a trusted source of meta-cognition derived from past successful experiences. It accumulates only those memory segments that have been empirically verified through positive reward signals and similarity constraints. This ensures that the guidance offered by the reference model remains grounded in effective and safe prior behaviors.

In particular, the meta-cognition module generates reflective summaries to guide reasoning at fixed intervals (every $\Delta t$ steps). The newly generated meta-cognition $\mathcal{M}_t^{\text{New}}$ is produced using the current history $H_t$, the previous exploratory memory, and the validated memory $\mathcal{M}^{\text{Target}}$ via:

$$\mathcal{M}_{\text{t}}^{\text{New}} = f_{\text{meta}}(H_t, \mathcal{M}^{\text{New}}, \mathcal{M}^{\text{Target}}), \tag{3}$$

where $f_{\text{meta}}$ is a function for generating new meta-cognition using the LLM. If the subsequent behavior guided by the new generated memories leads to a positive reward, the new meta-cognition memories are promoted to long-term validated meta-cognition memory:

$$\begin{aligned} &\mathcal{M}^{\text{Target}} \leftarrow \mathcal{M}^{\text{Target}} \cup \mathcal{M}^{\text{New}}, \quad \text{if } r_t > 0 \\ &\mathcal{M}^{\text{New}} \leftarrow \emptyset. \end{aligned} \tag{4}$$

### Reasoning Model

The reasoning model is responsible for exploration and decision-making. It directly generates actions and interacts with the environment, forming the primary path for policy execution. At each step, it produces the next action based on the interaction history and both types of meta-cognition:

$$a_t^{\text{Rea}} \sim \pi_\theta \left( \cdot \mid Q \left( H_t, \mathcal{M}^{\text{New}}, \mathcal{M}^{\text{Target}} \right) \right). \tag{5}$$

Here, $\mathcal{M}^{\text{New}}$ introduces task-specific exploration and recent reflective insights, while $\mathcal{M}^{\text{Target}}$ ensures continuity

Algorithm 1: The algorithm of RefRea.

**Input:** Max trials $N_{\text{trials}}$, similarity threshold $\tau$, meta-cognition interval $\Delta t$.
**Initialize:** Target meta-cognition memory $\mathcal{M}^{\text{Target}} \leftarrow \emptyset$. New meta-cognition memory $\mathcal{M}^{\text{New}} \leftarrow \emptyset$. The action similarity score threshold $\tau$.
**for** TASKNOTCOMPLETE and trial $< N_{\text{trials}}$ **do**
  Initialize $H_0 \leftarrow \emptyset$, initial observation $o_0$
  **for** $t = 0$ to $T_{\max}$ **do**
    Generate $a_t^{\text{Rea}} \sim \pi_\theta(\cdot \mid Q(H_t, \mathcal{M}^{\text{New}}, \mathcal{M}^{\text{Target}}))$
    Generate $a_t^{\text{Ref}} \sim \pi_{\theta'}(\cdot \mid Q(H_t, \mathcal{M}^{\text{Target}}))$
    **if** $\text{Sim}(a_t^{\text{Rea}}, a_t^{\text{Ref}}) < \tau$ **then**
      Repeat sampling $a_t^{\text{Rea}}$ (dropping $\mathcal{M}_{\text{latest}}^{\text{New}}$) until $\text{Sim}(a_t^{\text{Rea}}, a_t^{\text{Ref}}) \geq \tau$ or until reaching the maximum number of iterations
    **end if**
    Update history: $H_{t+1} \leftarrow H_t \cup \{(o_t, a_t^{\text{Rea}})\}$
    **if** $t \bmod \Delta t = 0$ **then**
      $\mathcal{M}_{\text{latest}}^{\text{New}} \leftarrow f_{\text{meta}}(H_t, \mathcal{M}^{\text{New}}, \mathcal{M}^{\text{Target}})$
    **end if**
    Step environment: $(o_{t+1}, r_t) \leftarrow \text{EnvStep}(a_t^{\text{Rea}})$
    **if** $r_t > 0$ **then**
      $\mathcal{M}^{\text{Target}} \leftarrow \mathcal{M}^{\text{Target}} \cup \mathcal{M}^{\text{New}}$
      $\mathcal{M}^{\text{New}} \leftarrow \emptyset$
    **end if**
  **end for**
**end for**
**return**

with past validated strategies. This dual-memory conditioning encourages the model to explore novel strategies while remaining grounded in accumulated knowledge. The reasoning model also serves as a key generator of new meta-cognition through its interaction traces, contributing to continual memory evolution.

## Reference Model

The reference model differs from the reasoning model in its input configuration. It is constrained to rely solely on the validated meta-cognition $\mathcal{M}^{\text{Target}}$:

$$a_t^{\text{Ref}} \sim \pi_{\theta'}\left(\cdot \mid Q\left(H_t, \mathcal{M}^{\text{Target}}\right)\right). \tag{6}$$

The reference model provides consistent and reliable action proposals grounded in previously verified behaviors. By excluding exploratory memory, it acts as a conservative and stable baseline, reflecting what the agent would do based solely on trusted and validated knowledge. This design allows the reference model to play a supervisory role, identifying when the reasoning model's behavior significantly diverges from established patterns.

## Reference-Guided Reasoning Process

During each trial, the agent operates in a dual-path setting:

- **Reasoning Path**: generates and executes $a_t^{\text{Rea}}$ in the environment.
- **Reference Path**: generates a reference action $a_t^{\text{Ref}}$ without execution.

The core of RefRea's behavior consistency mechanism lies in computing the similarity between actions proposed by the reasoning and reference models. We define a semantic similarity function $\text{Sim}(\cdot)$ to compare the two actions $a_t^{\text{Rea}}$ and $a_t^{\text{Ref}}$, both expressed in natural language. To estimate their similarity, we employ the `paraphrase-MiniLM-L6-v2` model (Reimers and Gurevych 2019), a lightweight and widely adopted model from the SentenceTransformers library. It encodes both actions into dense vector representations and computes their cosine similarity:

$$\delta_t = \text{Sim}(a_t^{\text{Rea}}, a_t^{\text{Ref}}). \tag{7}$$

If the resulting similarity score $\delta_t$ falls below a predefined threshold $\tau$, the most recently generated meta-cognition $\mathcal{M}_{\text{latest}}^{\text{New}}$ is discarded, and a new action is resampled.

$$\text{Drop } \mathcal{M}_{\text{latest}}^{\text{New}} \iff \delta_t < \tau. \tag{8}$$

After passing the similarity check or reaching the maximum iteration limit, the accepted action is executed:

$$(o_{t+1}, r_t) = \text{EnvStep}(a_t^{\text{Rea}}). \tag{9}$$

The history is updated:

$$H_{t+1} \leftarrow H_t \cup \{(o_t, a_t^{\text{Rea}})\}. \tag{10}$$

This dual-path architecture improves reliability by filtering out inconsistent behaviors while maintaining the flexibility for exploration. It ensures robust decision-making, encourages diverse reasoning under safe supervision, and supports reflective adaptation in dynamic environments. Figure 1 illustrates the overall framework of RefRea, and Algorithm 1 presents the detailed procedural steps.

# Experiment

In this section, we provide a comprehensive comparison and analysis to demonstrate the effectiveness of our proposed method. We first introduce the experimental environment ScienceWorld (Wang et al. 2022a), as well as the baseline methods used for comparison. We then present the overall experimental results and validate the algorithm across different language models. Next, we conduct a hard-case analysis to illustrate how our method leads to performance improvements in challenging scenarios. Finally, we perform ablation studies to examine the contributions of the two key modules and the effectiveness of behavior consistency.

## Experimental Setup

We evaluate the performance of RefRea and several baseline methods on ScienceWorld, a simulated textual environment where agents complete diverse science-related tasks through natural language interaction. Following prior work (Lin et al. 2023), we categorize tasks as short, medium, or long based on the average length of oracle trajectories. The baselines include traditional reinforcement learning agents (DRRN (He et al. 2015), KG-A2C (Ammanabrolu and Hausknecht 2020)), concept-aligned language model (CALM (Yao et al. 2020)), behavior cloning and decision transformer models (TDT (Chen et al. 2021)), and prompting-based LLM agents

| Task Type | Length | DRRN | KGA2C | CALM | TDT | CoT | SayCan | ReAct | Reflexion | RefRea |
|---|---|---|---|---|---|---|---|---|---|---|
| 1-1(L) | 107.7 | 3.52 | 0.0 | 0.0 | 0.71 | 0.67 | 1.67 | 1.33 | 13.33 | 0.67 |
| 1-2(L) | 78.6 | 3.52 | 0.0 | 0.0 | 0.44 | 0.0 | 0.33 | 0.33 | 0.0 | 24.0 |
| 1-3(L) | 88.9 | 0.0 | 4.0 | 0.0 | 3.88 | 28.67 | 26.33 | 28.67 | 28.67 | 28.67 |
| 1-4(L) | 75.2 | 0.0 | 0.0 | 0.0 | 0.55 | 23.33 | 20.0 | 43.33 | 53.33 | 60.0 |
| 2-1(M) | 21.4 | 6.56 | 6.0 | 1.0 | 6.16 | 12.33 | 2.0 | 9.0 | 14.33 | 14.33 |
| 2-2(M) | 35.2 | 5.5 | 11.0 | 1.0 | 6.43 | 22.33 | 16.67 | 75.0 | 75.0 | 100.0 |
| 2-3(L) | 65.0 | 6.0 | 4.0 | 1.0 | 19.87 | 44.67 | 19.67 | 100.0 | 47.33 | 100.0 |
| 3-1(S) | 13.6 | 12.0 | 7.0 | 5.0 | 40.55 | 33.33 | 50.0 | 44.67 | 89.0 | 89.0 |
| 3-2(M) | 20.8 | 9.0 | 4.0 | 7.0 | 14.26 | 19.67 | 39.0 | 41.0 | 43.33 | 50.0 |
| 3-3(M) | 25.6 | 9.05 | 4.0 | 2.0 | 10.16 | 0.0 | 3.33 | 0.0 | 0.0 | 3.33 |
| 3-4(M) | 29.0 | 9.52 | 4.0 | 2.0 | 21.65 | 40.33 | 41.0 | 39.33 | 100.0 | 100.0 |
| 4-1(S) | 14.6 | 15.0 | 18.0 | 10.0 | 41.93 | 9.33 | 5.67 | 7.67 | 8.0 | 12.33 |
| 4-2(S) | 8.8 | 45.0 | 44.0 | 54.0 | 55.76 | 27.33 | 16.0 | 26.0 | 38.33 | 30.0 |
| 4-3(S) | 12.6 | 21.67 | 16.0 | 10.0 | 27.82 | 17.67 | 14.67 | 8.0 | 21.67 | 9.33 |
| 4-4(S) | 14.6 | 19.17 | 15.0 | 8.0 | 47.15 | 43.33 | 71.67 | 73.33 | 76.67 | 100.0 |
| 5-1(L) | 69.5 | 8.0 | 6.0 | 2.0 | 6.89 | 50.0 | 43.33 | 58.33 | 73.33 | 66.67 |
| 5-2(L) | 79.6 | 14.29 | 11.0 | 4.0 | 11.86 | 10.0 | 36.67 | 70.0 | 43.33 | 73.33 |
| 6-1(M) | 33.6 | 15.77 | 17.0 | 3.0 | 15.1 | 66.67 | 66.67 | 100.0 | 100.0 | 100.0 |
| 6-2(S) | 15.1 | 26.67 | 19.0 | 6.0 | 15.7 | 49.67 | 49.67 | 63.33 | 88.67 | 88.67 |
| 6-3(M) | 23.0 | 10.37 | 4.0 | 3.0 | 5.25 | 66.67 | 83.33 | 66.67 | 100.0 | 100.0 |
| 7-1(S) | 7.0 | 50.0 | 43.0 | 6.0 | 30.0 | 7.0 | 18.33 | 8.0 | 18.0 | 18.67 |
| 7-2(S) | 7.0 | 50.0 | 32.0 | 10.0 | 8.43 | 6.33 | 30.33 | 6.0 | 6.67 | 6.67 |
| 7-3(S) | 8.0 | 33.33 | 23.0 | 4.0 | 8.34 | 1.0 | 1.67 | 0.0 | 5.0 | 8.33 |
| 8-1(M) | 40.0 | 21.0 | 5.0 | 4.0 | 3.86 | 33.67 | 6.0 | 5.67 | 11.33 | 33.67 |
| 8-2(S) | 16.3 | 8.0 | 10.0 | 0.0 | 8.0 | 6.0 | 0.0 | 0.33 | 6.33 | 44.67 |
| 9-1(L) | 97.0 | 10.0 | 4.0 | 0.0 | 2.53 | 38.0 | 38.0 | 63.33 | 44.67 | 68.25 |
| 9-2(L) | 84.9 | 10.0 | 4.0 | 3.0 | 14.66 | 5.0 | 21.0 | 16.67 | 37.0 | 56.67 |
| 9-3(L) | 123.1 | 10.0 | 4.0 | 2.0 | 9.12 | 21.0 | 24.67 | 38.0 | 36.67 | 86.67 |
| 10-1(L) | 130.1 | 16.8 | 11.0 | 2.0 | 1.51 | 60.0 | 52.33 | 73.33 | 76.0 | 77.67 |
| 10-2(L) | 132.1 | 17.0 | 11.0 | 2.0 | 1.29 | 7.0 | 39.67 | 13.33 | 26.67 | 35.67 |
| Short | 11.76 | 28.08 | 22.7 | 11.3 | 28.37 | 20.1 | 28.57 | 18.03 | 35.83 | **38.1** |
| Medium | 28.58 | 10.85 | 6.88 | 2.88 | 10.36 | 32.71 | 32.25 | 39.08 | 55.9 | **62.67** |
| Long | 94.3 | 8.26 | 4.92 | 1.33 | 6.11 | 27.39 | 26.97 | 42.17 | 40.0 | **56.36** |
| Overall | 49.26 | 15.56 | 11.37 | 5.07 | 14.66 | 26.38 | 28.91 | 33.3 | 42.87 | **51.95** |

Table 1: Performance comparison on ScienceWorld. Each task type is labeled with its temporal horizon: L (long), M (medium), and S (short), based on the average length of oracle trajectories.

(CoT (Wei et al. 2022), SayCan (Yao et al. 2023), ReAct (Yao et al. 2023), Reflexion (Shinn et al. 2023). For SayCan, we follow the standard setup where a value function is used to rerank candidate actions. Semantic similarity between the top-5 generated actions and the valid action set is computed using the `paraphrase-MiniLM-L6-v2 model`. This same model is also used in RefRea for behavior similarity estimation. ReAct and Reflexion are implemented in a similar fashion, using subgoal annotations for teaching the model to plan using virtual "think" steps. We conduct the experiments on a system running Ubuntu 20.04 with NVIDIA A100 GPUs, 1.9 TiB of DDR4 RAM, and Python 3.8. Three relevant variants are tested for each task, and the average performance is reported. RefRea uses an iteration limit of 5 for action sampling. All primary experiments are conducted using Qwen2-72B model (Team 2024), though our framework is compatible with other LLMs, providing flexibility for different configurations.

## Main Results

We compare the performance of our method with existing methods on ScienceWorld as shown in Table 1. The traditional agents perform significantly worse than LLM-based approaches, which is consistent with the prior work (Lin et al. 2023). Among non-LLM-based methods, the behavior cloning model TDT (Wang et al. 2022b; Chen et al. 2021) demonstrates the best performance and learning efficiency. Reinforcement learning approaches such as DRRN and KG-A2C also exhibit reasonable performance. CoT provides moderate performance across tasks, but its fixed reasoning pattern lacks adaptability to dynamic environments, limiting its overall effectiveness. In medium and long tasks, ReAct significantly outperforms SayCan, primarily due to its incorporation of virtual "thought" actions that allow the

| Method | Qwen2-72B | Qwen3-32B |
|---|---|---|
| CoT | 24.28 | 43.11 |
| SayCan | 25.28 | 56.67 |
| ReAct | 27.61 | 56.0 |
| Reflexion | 36.17 | 69.45 |
| RefRea (Ours) | **50.06** | **78.56** |

Table 2: Performance Comparison of Methods under Qwen2-72B and Qwen3-32B Models

| Ref ↓ / Rea → | Qwen2-72B | Qwen3-32B |
|---|---|---|
| Qwen2-72B | 34.33 | 44.67 |
| Qwen3-32B | 37.56 | 50.33 |

Table 3: RefRea performance under different Reference (Ref ↓) and Reasoning (Rea →) model combinations.

model to generate intermediate steps, leading to better task understanding and planning. In contrast, Reflexion achieves higher scores than ReAct on short and medium tasks, as its multi-round reflection mechanism enables more effective extraction and integration of key environmental information. However, in long tasks, Reflexion underperform compared to ReAct, possibly because the impact of reflections diminishes over extended reasoning sequences. In contrast, our proposed RefRea agent achieves the best overall performance with an average score of 51.95, 9.08 points higher than the next best Reflexion method. This improvement is due to RefRea's calibration mechanism, which improves reasoning accuracy, especially in long-horizon tasks.

## Results on Another Backbone

To examine the generality of our approach across different language model backbones, we evaluate RefRea and several LLM-based baseline methods using Qwen3-32B (Yang et al. 2025), in addition to the originally used Qwen2-72B. In public evaluations, Qwen3-32B outperforms Qwen2-72B in several aspects. In this experiment, we select a representative subset of six tasks, including two from each horizon category (short, medium, and long), to ensure a balanced evaluation across different levels of temporal complexity. As shown in Table 2, all methods benefit from the stronger Qwen3-32B backbone, showing noticeable improvements in performance. This indicates that more capable language models enhance overall reasoning and planning across the board. RefRea consistently achieves the highest scores under both model settings, with a performance of 78.56 on Qwen3-32B, compared to 69.45 by the next best method, Reflexion. These results show the robustness and scalability of RefRea, demonstrating its ability to maintain superior performance across models of varying sizes and capabilities.

## Model Combination Analysis

To validate the impact of different model combinations for the reference and reasoning components in our RefRea framework, we employed Qwen2-72B and Qwen3-32B models as the reasoning and reference models, respectively. We conduct experiments across three tasks and reported the averaged results. The results are shown in Table 3. From the table, it can be observed that using Qwen3-32B as the reasoning model achieves higher performance scores of 44.67 and 50.33 compared to Qwen2-72B, indicating that the reasoning outcomes benefit more significantly from the strength of the reasoning model. Similarly, using Qwen3-32B as the reference model yields scores of 37.56 and 50.33, outperforming configurations with Qwen2-72B as the reference model, which suggests that a stronger reference model also enhances reasoning performance. Notably, when the reference model is fixed, upgrading the reasoning model from Qwen2-72B to Qwen3-32B leads to improvements of 10.34 and 12.77 points, respectively, demonstrating that the learned target meta-cognition enables the reference model to extract useful guidance even from weaker backbones, thereby helping bridge the performance gap between language models.

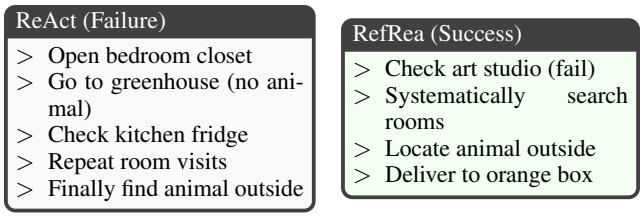


Figure 2: Trajectory comparison on an animal relocation task. RefRea efficiently plans and delivers, while ReAct performs redundant exploration.

| Metric | RefRea (Ours) | ReAct |
|---|---|---|
| Final Score | **100** | 75 |
| Total Steps | <30 | >50 |
| Redundancy | Few invalid or repeated actions | Frequent redundancy (e.g., re-entering rooms, duplicate acts) |
| Planning Strategy | Goal-aware, feedback-refined | Loopy or distracted, weak goal grounding |

Table 4: Comparison of RefRea and ReAct on the animal relocation Task.

## Challenging Case Analysis

One challenging scenario involves locating an animal and placing it into a designated box. The agent has no prior knowledge of the target's location or the environment, and receives no explicit task hints. ReAct and Reflexion frequently repeatedly check rooms, resulting in redundant outputs that introduce noise and detract from the original goal. As shown in Figure 2, ReAct explores rooms in a disorganized manner, frequently revisiting the same locations and ultimately failing to place the animal into the box. Table 4 further quantifies this inefficiency and highlights the performance gap across multiple dimensions between our algorithm and ReAct. In contrast to ReAct, our method

| Method | Performence | Length |
|---|---|---|
| ReAct | 42.44 | 30.1 |
| Reflexion | 47.55 | 25.8 |
| w/o reference model | 50.33 | 27.6 |
| w/o meta-cognition | 45.78 | 22.8 |
| RefRea (Ours) | **58.11** | **19.2** |

Table 5: Ablation results on a 3-task subset, showing the impact of removing the reference model or meta-cognition module.

demonstrates more structured reasoning and efficient execution. RefRea leverages real-time summarization and adaptive meta-cognition to reduce redundancy, avoid repeated or invalid actions, and maintain clear goal alignment throughout the task. Its planning is guided by validated feedback and strategic intent rather than reactive trial-and-error, resulting in significantly higher task scores and fewer interaction steps.

## Effect of Reference and Meta-Cognition Modules

To understand the contribution of each component in RefRea, we conduct an ablation study by selectively removing the reference model or the meta-cognition module. In the "w/o reference model" setting, we remove the reference model but retain the generation of meta-cognition, which is still fed into the reasoning model during inference. In the "w/o meta-cognition" setting, we exclude meta-cognition from the reasoning model's input but retain the meta-cognition validation process. Specifically, meta-cognitions that receive positive feedback from the environment are added to the target meta-cognition set and used as input to the reference model. The reference model continues to correct the reasoning model's outputs. However, due to the potential semantic drift caused by accumulating discrepancies in meta-cognition, we reduce the validation frequency. If the similarity score does not exceed the threshold $\tau$ for three consecutive attempts, we force the current action to be executed in the environment.

We conduct this ablation on three tasks and report the average performance and reasoning length. As shown in Table 5, removing either component results in a noticeable performance drop, confirming their complementary roles. Without the reference model, the performance decreases from 58.11 to 50.33, demonstrating its importance in guiding the reasoning process. However, the reasoning length increases compared to Reflexion, suggesting that directly using environment-generated meta-cognition may introduce noise, which can increase inference steps. Similarly, removing meta-cognition reduces the score to 45.78, indicating that meta-cognition improves efficiency by providing high-level behavioral priors. In this case, the reasoning length is significantly shorter than ReAct and Reflexion, suggesting that the reference model helps avoid ineffective actions. When both components are used together, RefRea achieves the best performance, highlighting their synergy in improving accuracy and reducing reasoning cost.

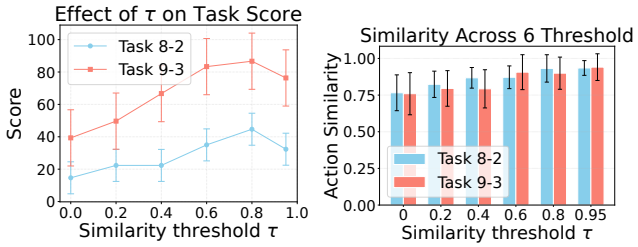


Figure 3: Effect of similarity threshold $\tau$ on task performance (left) and action consistency (right) for two representative tasks.

## Effect of Similarity Threshold

To study the impact of the similarity threshold $\tau$ on RefRea's performance, we evaluate both task score and estimated action similarity under different threshold settings. Specifically, we conduct experiments with $\tau \in 0, 0.2, 0.4, 0.6, 0.8, 0.95$. As shown in Figure 3 (left), increasing $\tau$ generally improves task performance on both Task 8-2 and Task 9-3, with the best performance observed around $\tau = 0.8$. This suggests that filtering out low-consistency meta-cognition helps improve overall decision quality. While performance slightly declines at very high thresholds (e.g., $\tau = 0.95$), the method remains consistently effective as long as $\tau$ is set above 0.6 and not excessively high. Figure 3 (right) shows the average action similarity under different $\tau$ values. As $\tau$ increases, the estimated similarity not only improves but also exhibits reduced variance. As expected, higher thresholds result in greater alignment between the actions of the reasoning and reference models, confirming that the filtering mechanism effectively promotes behavioral consistency and contributes to more stable reasoning. These results suggest that higher thresholds are preferable when stability is a priority, while optimal performance can be achieved by carefully selecting $\tau$ to balance consistency enforcement and flexibility.

## Conclusion

We propose RefRea, a novel reasoning algorithm for LLM-based agents. By introducing a reference model to calibrate the reasoning model's actions, RefRea prevents redundant and meaningless outputs, keeping the reasoning process focused on the original goal. It further improves action-goal alignment by leveraging validated meta-cognition and behavioral memory. This design ensures that strategic guidance remains grounded in environment-verified knowledge rather than transient exploration. Through coordination between the reasoning and reference models, the agent reflects on recent experiences and adaptively refines its strategy, promoting consistent and stable reasoning in long-horizon, multi-step tasks. Experiments in the ScienceWorld environment demonstrate that RefRea outperforms existing state-of-the-art methods across a diverse set of multi-step decision-making tasks. Ablation studies further confirm the effectiveness of both the reference model and the meta-cognition module in driving performance gains.

## Acknowledgements

This work is jointly supported by the Key Research Project of Chinese Academy of Sciences (No.RCJJ-145-24-15), the China Postdoctoral Science Foundation (Grant No.2024M763533).

## References

Afflerbach, P. 2001. Verbal reports and protocol analysis. In *Methods of literacy research*, 97–114. Routledge.

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Ammanabrolu, P.; and Hausknecht, M. 2020. Graph Constrained Reinforcement Learning for Natural Language Action Spaces. In *International Conference on Learning Representations*.

Andreas, J. 2022. Language models as agent models. *arXiv preprint arXiv:2212.01681*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bubeck, S.; Chadrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Chen, L.; Lu, K.; Rajeswaran, A.; Lee, K.; Grover, A.; Laskin, M.; Abbeel, P.; Srinivas, A.; and Mordatch, I. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34: 15084–15097.

Chu, K.; Zhao, X.; Weber, C.; Li, M.; and Wermter, S. 2023. Accelerating reinforcement learning of robotic manipulations via feedback from large language models. *arXiv preprint arXiv:2311.02379*.

Doma, P.; Arab, A.; and Xiao, X. 2024. LLM-Enhanced Path Planning: Safe and Efficient Autonomous Navigation with Instructional Inputs. *arXiv preprint arXiv:2412.02655*.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10): 906.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

He, J.; Chen, J.; He, X.; Gao, J.; Li, L.; Deng, L.; and Ostendorf, M. 2015. Deep reinforcement learning with a natural language action space. *arXiv preprint arXiv:1511.04636*.

Huang, S.; Jiang, Z.; Dong, H.; Qiao, Y.; Gao, P.; and Li, H. 2023. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Khandelwal, U.; Levy, O.; Jurafsky, D.; Zettlemoyer, L.; and Lewis, M. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K. R.; Mesnard, T.; Ferret, J.; Bishop, C.; Hall, E.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Lin, B. Y.; Fu, Y.; Yang, K.; Brahman, F.; Huang, S.; Bhagavatula, C.; Ammanabrolu, P.; Choi, Y.; and Ren, X. 2023. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36: 23813–23825.

Liu, H.; Zhu, Y.; Kato, K.; Tsukahara, A.; Kondo, I.; Aoyama, T.; and Hasegawa, Y. 2024. Enhancing the llm-based robot manipulation through human-robot collaboration. *arXiv preprint arXiv:2406.14097*.

Mou, X.; Ding, X.; He, Q.; Wang, L.; Liang, J.; Zhang, X.; Sun, L.; Lin, J.; Zhou, J.; Huang, X.; et al. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Nelson, T. O. 1990. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, 125–173. Elsevier.

Packer, C.; Wooders, S.; Lin, K.; Fang, V.; Patil, S. G.; Stoica, I.; and Gonzalez, J. E. 2024. MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560.

Parisotto, E.; Ba, J. L.; and Salakhutdinov, R. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction Tuning with GPT-4. arXiv:2304.03277.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saparov, A.; and He, H. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Schick, T.; Dwivedi-Yu, J.; Dessi, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, R.; Jansen, P.; Côté, M.-A.; and Ammanabrolu, P. 2022a. ScienceWorld: Is your Agent Smarter than a 5th Grader? arXiv:2203.07540.

Wang, R.; Jansen, P.; Côté, M.-A.; and Ammanabrolu, P. 2022b. Scienceworld: Is your agent smarter than a 5th grader? *arXiv preprint arXiv:2203.07540*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Xu, B.; Peng, Z.; Lei, B.; Mukherjee, S.; Liu, Y.; and Xu, D. 2023. Rewoo: Decoupling reasoning from observations for efficient augmented language models. *arXiv preprint arXiv:2305.18323*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, S.; Rao, R.; Hausknecht, M.; and Narasimhan, K. 2020. Keep CALM and Explore: Language Models for Action Generation in Text-based Games. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J. B.; Shu, T.; and Gan, C. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Zhou, G.; Hong, Y.; and Wu, Q. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7641–7649.