Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/150/2025.22347

A Metacognitive Evaluation Framework for Embodied Intelligent Agents

Yuanhang Ling

Tomsk State University, Tomsk, Russia
voyaged.zero@gmail.com

Abstract: The study introduces a novel evaluation system designed to measure the metacognitive abilities of embodied agents. The system incorporates multiple metrics—including task success rate, self-monitoring accuracy (measured by AUC), error detection speed, and confidence calibration error—to provide a comprehensive assessment of an agent's internal monitoring and self-regulatory processes. Experiments were conducted in simulated environments (using Meta-World, etc.) and on a real robotic platform performing target grasping tasks. Two types of agents were compared: baseline agents relying solely on external feedback and agents enhanced with integrated metacognitive modules. The results demonstrate that agents with metacognitive capabilities consistently achieve higher performance, exhibit more precise self-monitoring, and respond more swiftly to unexpected events. This evaluation system serves as a robust tool for assessing metacognitive functions and offers promising implications for the development of more adaptable and reliable autonomous systems in dynamic environments, thus significantly enhancing overall system performance continuously.

Keywords: Metacognition, Embodied Agents, Self-Monitoring, Error Detection, Autonomous Systems

1. Introduction

Embodied agents refer to AI entities with physical or virtual bodies that interact with the environment via sensors and actuators [1]. Unlike disembodied intelligence, they face real-world complexity, positioning them as a key path toward artificial general intelligence (AGI). Metacognition—"thinking about thinking"—enables agents to monitor, evaluate, and regulate their cognitive processes, enhancing adaptability and autonomy. However, current AI systems often lack such flexibility when encountering novel scenarios [2]. While metacognitive abilities could allow embodied agents to self-monitor, adjust strategies, and improve learning efficiency, existing evaluation methods remain limited. Most focus on task outcomes, overlooking agents' self-awareness and lacking standardized assessment criteria [3]. To address this gap, this paper aims to establish a systematic evaluation framework for assessing metacognitive capabilities in embodied agents, identifying key metrics and methodologies to enhance the comparability and practical relevance of research in this field.

DOI: 10.54254/2755-2721/150/2025.22347

2. Related Work

2.1. Cognitive Models

In terms of cognitive architecture, many classical frameworks have begun integrating metacognitive mechanisms. For example, MIDCA (Metacognitive Integrated Dual-Cycle Architecture) employs a dual-loop system: the object-level handles routine planning and execution, while the meta-level monitors behavior, detects discrepancies, and guides adjustments [4]. Similarly, the Metacognitive Loop (MCL) architecture enables agents to detect mismatches between expected and actual outcomes, enhancing robustness in reinforcement learning contexts [5]. Other robotic control frameworks have incorporated self-error monitoring, allowing agents to evaluate their own actions for improved decision-making [6]. With the rise of deep learning, recent efforts have explored using large-scale pre-trained models to enhance metacognitive capabilities. For instance, vision-language models have been utilized to help robots interpret human-readable documentation for self-diagnosis and fault recovery [7-8]. In reinforcement learning, certain frameworks introduce competence awareness and strategy regulation, leading to improved performance in novel or unstructured tasks [9-10]. Beyond single-agent systems, research [11-12] has also shown that, in human-AI collaboration, expressing decision confidence improves team outcomes—highlighting the critical role of metacognition in both autonomous and collaborative settings.

2.2. Framework Needs

The absence of a unified metacognitive evaluation framework represents a significant gap in current research [13]. Although previous studies have demonstrated that metacognition can enhance the performance of embodied agents, the lack of standardized benchmarks hinders fair comparisons across methods and makes it difficult to identify weaknesses in existing mechanisms [14]. In real-world applications, evaluating an agent's self-monitoring and self-regulation is essential for determining its readiness for critical tasks. Yet, no standardized metacognitive evaluation method—comparable to the Turing Test—currently exists. To address this, we propose a systematic framework for assessing metacognitive capabilities in embodied agents. Our approach unifies prior insights, defines core metrics, and introduces standardized testing protocols. Key contributions include: (1) the first structured index system for agent metacognition; (2) a generalizable evaluation architecture; and (3) empirical validation demonstrating both the framework's utility and existing agent limitations.

3. Method

This study designs a metacognitive evaluation framework to quantify the performance of embodied agents in self-monitoring, self-regulation, and error detection.

3.1. Main Modules

- **Data Collection Module:** During task execution, this module records internal outputs—such as confidence scores, self-predictions, and error alerts—alongside environmental feedback, including task success rates and reward signals.
- **Evaluation and Analysis Module:** This module computes predefined metrics, including self-monitoring accuracy (evaluated using ROC curves and AUC of confidence predictions), error detection rate (the proportion of anomalies correctly identified by the agent), and self-regulation effectiveness (measured by improvements in task success rates before and after strategy adjustments).

DOI: 10.54254/2755-2721/150/2025.22347

- **Feedback Adjustment Module:** To further enhance agent performance, this module provides recommendations for improvement based on evaluation results, such as automatically adjusting parameters within the metacognitive module.

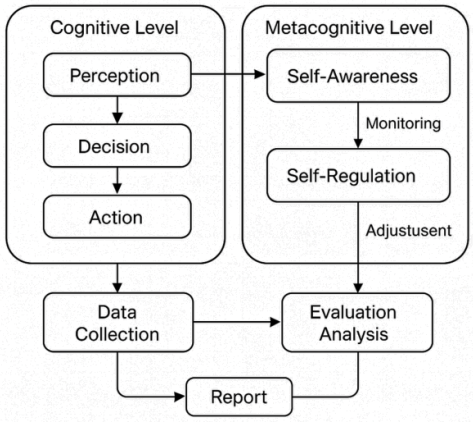

Figure 1: The metacognitive assessment architecture

This architecture draws inspiration from the monitor-control models in cognitive psychology and control theory, enabling agents not only to achieve high task performance but also to reflect on and adapt their decision-making processes in real time. Figure 1 shows the complete architecture diagram.

3.2. Metacognition Assessment System

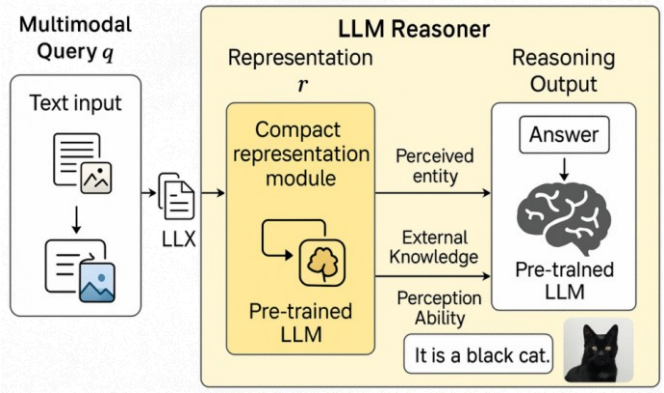

Figure 2: Framework

The figure 2 illustrates the detailed framework of the metacognition assessment system for embodied agents. It highlights the interaction between the perception-action cycle and the metacognitive layer, where the agent's internal assessments of confidence, error detection, and self-regulation are monitored and adjusted. The flow from sensory inputs through the cognitive processes to metacognitive feedback showcases the system's collaborative modules. This visualization emphasizes the dynamic feedback loop between the agent's performance and its metacognitive self-awareness and regulation. The system computes metacognitive calibration using metrics such as the Brier score:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{1}$$

DOI: 10.54254/2755-2721/150/2025.22347

where f_i is the agent's predicted confidence and o_i is the actual outcome, quantifying the accuracy of self-monitoring over time. Additionally, the system models self-regulation dynamics through an adaptive control function:

$$\pi_{t+1} = \pi_t + \eta \cdot \nabla_\pi \mathcal{L}_{meta}(\pi_t) \quad (2)$$

where $\pi_t$is the policy at time t, η is the learning rate, and $\mathscr{L}_{meta}$ represents the metacognitive loss guiding policy adjustment based on internal feedback.

4. Experiment

To validate the effectiveness of the proposed evaluation framework, experiments were conducted on two types of platforms.

4.1. Simulation Experiments

Tasks such as "open door" and "object relocation" were selected within the Meta-World and ALFWorld environments. For each task, anomalous scenarios were introduced by randomly altering object positions and adding sensor noise, simulating real-world unexpected disturbances.

Two types of agents were deployed for each task:

Metacognitive Agent: Integrated with self-monitoring and self-regulation modules, capable of evaluating its internal state, detecting errors, and adapting strategies in real time during task execution.

Baseline Agent: Lacking metacognitive modules, it relies solely on environmental feedback for decision-making.

To ensure statistical reliability, each task was repeated multiple times (e.g., 30 trials per task). Key recorded metrics included task success rate, self-monitoring accuracy (measured by AUC of ROC curves), error detection rate, and the performance gain attributed to self-regulation.

4.2. Real-World Robotic Experiments

A mobile autonomous robot was tasked with "target object grasping." The robot was required to autonomously navigate to a target zone and grasp an object, while variations in target location and obstacle interference were introduced to assess its adaptability under dynamic disturbances. As in the simulation experiments, both a metacognitive agent and a baseline agent were deployed. Task success rates and internal metacognitive metrics (e.g., confidence scores and error alerts) were recorded to evaluate the practical effectiveness of the metacognitive module in real-world conditions.

4.3. Key Metrics

- **Task Success Rate:** Proportion of tasks where the agent achieves the goal; reflects overall performance.
- **Self-Monitoring Accuracy (AUC):** Measures how well the agent predicts its own correctness; higher AUC means better self-assessment.
- **Error Detection Rate:** Percentage of failures or anomalies correctly identified by the agent; indicates awareness of unexpected events.
- **Self-Regulation Improvement:** Increase in success rate after strategy adjustment; reflects the agent's capacity to self-improve.

DOI: 10.54254/2755-2721/150/2025.22347

4.4. Experimental Results

4.4.1. Task Success and Performance Metrics

Table 1: Experimental Results Comparison

Task	Baseline Success Rate	Metacognitive Success Rate	Self-monitoring AUC	Error Detection Rate	Self-regulation Improvement Rate
Meta-World Door	64%	87%	0.81	23%	32%
ALFWorld Transportation	42%	68%	0.85	27%	47%
Real Robot Grasping	57%	76%	0.79	31%	38%

Table 1 compares the overall task success rate, self-monitoring accuracy (measured by AUC), error detection rate, and self-regulation improvement rate between the baseline and metacognitive agents across various tasks. The data clearly shows that the metacognitive agent consistently outperforms the baseline in all key performance areas.

4.4.2. Decision and Error Detection Delays

Table 2: Decision Delay and Error Detection Delay Statistics

Task	Baseline Decision Delay (ms)	Metacognitive Decision Delay (ms)	Baseline Error Detection Delay (ms)	Metacognitive Error Detection Delay (ms)
Meta-World Door	118	153	398	247
ALFWorld Transportation	127	161	447	282
Real Robot Grasping	205	237	502	319

Table 2 presents the average decision delay and error detection delay for both agents. Although the metacognitive agent incurs a slightly longer decision delay due to additional internal evaluations, it significantly reduces the error detection delay, enabling faster responses in abnormal conditions.

4.4.3. Confidence Calibration Error Comparison

Table 3: Confidence Calibration Error Statistics

Task	Baseline MCE	Metacognitive MCE
Meta-World Door	0.146	0.082
ALFWorld Transportation	0.173	0.109
Real Robot Grasping	0.198	0.123

Table 3 shows the average Mean Calibration Error (MCE) for both agents across different tasks. Lower MCE values indicate that the agent's confidence levels are more closely aligned with its actual performance. The metacognitive agent achieves substantially lower calibration errors, reinforcing the benefit of incorporating metacognitive processes.

4.5. Results Analysis

The evaluation system effectively differentiates between agents with and without metacognitive capabilities. In tasks such as door opening and transportation, metacognitive agents consistently achieved higher success rates and exhibited improved self-monitoring accuracy, as evidenced by increased AUC values. Additionally, these agents demonstrated significantly reduced error detection delays and lower confidence calibration errors compared to baseline agents. These metrics

DOI: 10.54254/2755-2721/150/2025.22347

collectively confirm that the evaluation system captures the essential aspects of metacognitive functioning, offering a comprehensive assessment of an agent's ability to monitor and adjust its own performance in dynamic environments.

5. Conclusion

The proposed evaluation system offers a robust framework for quantifying the metacognitive abilities of embodied agents. By integrating multiple performance metrics, it reliably distinguishes agents with enhanced self-monitoring, error detection, and self-regulation capabilities. The experimental results validate the system's effectiveness and highlight its value as a tool for advancing metacognitive functionalities in autonomous systems. This framework paves the way for further research into developing more adaptive and reliable embodied agents.

Acknowledgement

Support from Tomsk State University is gratefully acknowledged. The project benefited from insightful discussions and technical expertise provided by local experts and affiliated research centers. Their contributions played an essential role in the development and successful execution of this study, enhancing its overall impact.

References

[1] *Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W. and Lin, L. (2024) Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI. arXiv preprint. https://arxiv.org/abs/2407.06886*

[2] *Valiente, R. and Pilly, P.K. (2024) Metacognition for Unknown Situations and Environments (MUSE). arXiv preprint. https://arxiv.org/abs/2411.13537*

[3] *Leidner, D. (2024) Toward Robotic Metacognition: Redefining Self-Awareness in an Era of Vision-Language Models. In ICRA@ 40 40th Anniversary of the IEEE International Conference on Robotics and Automation.*

[4] *Cox, M., Alavi, Z., Dannenhauer, D., Eyorokon, V., Munoz-Avila, H. and Perlis, D. (2016) MIDCA: A Metacognitive, Integrated Dual-Cycle Architecture for Self-Regulated Autonomy. In Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).*

[5] *Haidarian, H., Dinalankara, W., Fults, S., Wilson, S., Perlis, D. and Anderson, M. (2010) The Metacognitive Loop: An Architecture for Building Robust Intelligent Systems. AAAI Fall Symposium - Technical Report.*

[6] *Rabiee, S. and Biswas, J. (2023) Introspective Perception for Mobile Robots. Artificial Intelligence, 324, 103999.*

[7] *Loftus, T.J., Tighe, P.J., Filiberto, A.C., Efron, P.A., Brakenridge, S.C., Mohr, A.M., Rashidi, P., Upchurch, G.R. and Bihorac, A. (2020) Artificial Intelligence and Surgical Decision-Making. JAMA Surgery, 155(2), 148–158.*

[8] *Bhattacharyya, R., Colombatto, C., Fleming, S., Posner, I. and Hawes, N. (2023) Investigating the Role of Metacognition for Joint Decision-Making in Human-Robot Collaboration.*

[9] *Didolkar, A., Goyal, A., Ke, N.R., Guo, S., Valko, M., Lillicrap, T., Jimenez Rezende, D., Bengio, Y., Mozer, M.C. and Arora, S. (2024) Metacognitive Capabilities of LLMs: An Exploration in Mathematical Problem Solving. Advances in Neural Information Processing Systems, 37, 19783–19812.*

[10] *Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, E.L., Zhang, R. et al. (2024) Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making. Advances in Neural Information Processing Systems, 37, 100428–100534.*

[11] *Kronsted, C., Kugele, S., Neemeh, Z.A., Ryan Jr, K.J. and Franklin, S. (2022) Embodied Intelligence: Smooth Coping in the Learning Intelligent Decision Agent Cognitive Architecture. Frontiers in Psychology, 13, 846931.*

[12] *Dorsch, J. (2022) Embodied Metacognition: How We Feel Our Hearts to Know Our Minds. The University of Edinburgh.*

[13] *Anil Meera, A. and Lanillos, P. (2024) Towards Metacognitive Robot Decision Making for Tool Selection. In Buckley, C.L., Cialfi, D., Lanillos, P., Ramstead, M., Sajid, N., Shimazaki, H., Verbelen, T. and Wisse, M. (Eds.), Active Inference, pp. 31–42. Cham: Springer Nature Switzerland.*

[14] *Davis, R.O., Park, T. and Vincent, J. (2022) A Meta-Analytic Review on Embodied Pedagogical Agent Design and Testing Formats. Journal of Educational Computing Research, 61(1), 30–67. https://doi.org/10.1177/0735633122 1100556*