ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Why Cognitive Robotics Matters: Lessons from OntoAgent and LLM Deployment in HARMONIC for Safety-Critical Robot Teaming

Sanjay Oruganti[1], Sergei Nirenburg[1], Marjorie McShane[1], Jesse English[1],
Michael Roberts[1], Christian Arndt[1], Ramviyas Parasuraman[2], Luis Sentis[3]
[1]Rensselaer Polytechnic Institute, [2]University of Georgia, [3]University of Texas at Austin
Contact: sanjayovs@ieee.org
https://rpi-leia.github.io/WhyCognitiveRobotics/

***Abstract*—Robots operating alongside humans must recognize what they do not know before acting, diagnose problems from domain knowledge, and reason about action consequences. These capabilities are operational requirements, not optimization targets, and their absence produces silent and unrecoverable failures. We present a first-of-its-kind controlled comparison between OntoAgent, our content-centric cognitive architecture, and six LLMs spanning frontier and efficient tiers as drop-in replacements at the strategic layer of the same robotic system in HARMONIC. LLMs fail to verify their knowledge state before acting, even when given equivalent procedural knowledge. The deficit is architectural, not knowledge-based. Knowledge-grounded architectures must retain decision authority; LLMs contribute where their strengths apply.**

I. Motivation

Large language models are increasingly deployed as the strategic reasoning layer for robotic systems [1]–[3]. For conversational applications, stochastic errors are tolerable because humans remain in the loop to correct and regenerate. Physical embodiment removes that safety net entirely. In safety-critical human-robot teaming, a hallucinated fact becomes a wrong action, a wrong action becomes an unrecoverable failure, and that failure unfolds alongside humans who depend on the robot's judgment. A growing body of evidence documents systematic reasoning failures in LLMs that persist across model scale and prompting strategies [4], [5]. Whether LLMs can reliably provide the cognitive capabilities safety-critical settings demand has not been tested within a controlled embodied comparison. We identify three capabilities as critical: metacognitive self-monitoring, domain-grounded diagnosis, and consequence-based action selection. These are not emergent properties of scale but architectural commitments, provided by construction in cognitive architectures such as Soar [6], ACT-R/E [7], and OntoAgent [8], [10].

II. The HARMONIC Framework

HARMONIC is a dual-control cognitive-robotic architecture separating strategic (System 2) deliberative reasoning from tactical (System 1) reactive control [11] through a bidirectional interface. The strategic layer instantiates OntoAgent [8]–[10], whose reasoning operates over four interconnected knowledge resources: an ontological world model, procedural scripts and metascripts with explicit preconditions, episodic memory, and a continuously updated situation model. Prior to any action dispatch, OntoAgent inspects the situation model to verify preconditions. Unsatisfied preconditions trigger metascript activation, such as requesting information from a teammate. Diagnostic hypotheses are generated by traversing causal relations in the ontology, and action selection includes an actionability assessment before any command issues. A single verified execution trace fully characterizes system behavior, yielding the inspectability and traceability required for safety-critical deployment.

The tactical layer executes real-time motor control through Behavior Trees [12] and a shared blackboard, engaging skills from a modular library that includes state machines, classical controllers, learned policies, and vision-language-action models. Crucially, the strategic layer is interchangeable by design. Any reasoning system that processes timed perception frames and produces parameterized action commands can replace OntoAgent while the tactical infrastructure, perception pipeline, and task environment remain invariant. This modularity enables the controlled comparison reported here.

III. Experimental Design

We evaluate six LLMs as drop-in replacements for OntoAgent at the strategic layer: Claude Opus 4.6 and Haiku 4.5 (Anthropic), GPT-5.2 and GPT-5 Mini (OpenAI), Gemini 3 Pro and Gemini 3 Flash (Google). Each model operates through an LLMAgent module comprising a context manager, system prompt builder, LLM provider, and action parser that translates model outputs into the standardized command format. The evaluation scenario is a collaborative shipboard maintenance task in which the robot assists a mechanic in diagnosing an engine overheating issue and retrieving a replacement thermostat. The scenario imposes three cognitive demands that parallel the measurement targets: generating diagnostic hypotheses from domain knowledge, detecting missing information before executing a fetch plan, and selecting action primitives whose execution requirements match the deliberative-reactive timing constraints.

Each model runs five trials under two conditions, yielding $N = 60$ trials total. Under Internal Knowledge (IK), the LLM relies entirely on its pretrained knowledge.

Under Knowledge-Equalized (KE), the LLM additionally accesses a `FETCHPLAN` tool that retrieves the narrative rendering of OntoAgent's procedural scripts, specifying preconditions, diagnostic strategy, and expected action sequences. The KE condition separates the mechanism of reasoning from the availability of knowledge. If a deficit persists under KE, it cannot be attributed to missing domain knowledge.

IV. Results

OntoAgent achieves reference performance on all metrics, completing the task in 100% of trials with full precondition verification, domain-first diagnosis, and correct action selection.

Metacognitive Monitoring. Under IK, 100% of LLM trials dispatched a physical retrieval command before verifying preconditions. KE reduced this to 60% ($p < .001$), but improvement was concentrated in three of six models. Feature hallucination dropped from 100% to 57% ($p < .001$), yet two models showed no change despite accessing verification procedures.

Diagnostic Reasoning. Only 7% of IK trials exhibited domain-first diagnosis, with models treating the service log as the diagnostic framework rather than reasoning from causal knowledge. Under KE this reversed to 70% ($p < .001$, $|h| = 1.46$, the largest effect in the study). However, hallucinated facts were unaffected (IK: 1.4, KE: 1.6, $p = .41$), while expressed uncertainty rose from 43% to 93%. LLMs became verbally more cautious without becoming factually more accurate. Retrieval and reasoning were dissociable: Opus 4.6 queried the fetch procedure in 80% of KE trials but followed it in 0%, and queried the diagnose procedure in every KE trial yet followed it in none.

Action Consequence Reasoning. Correct action selection rose from 57% to 93% under KE ($p = .002$). Every wrong-action trial ($n = 15$) produced an unrecoverable cascade failure: behavioral loops (47%), hallucinated task completion (27%), stalls (20%), and backtrack-circling (7%). No model recovered from an incorrect action selection. Task completion rose from 47% under IK to 83% under KE, compared to 100% for OntoAgent.

V. Discussion

Three findings resist equalization entirely. First, hallucination is a generation-level property unaffected by retrieval. Second, the epistemic hedging dissociation shows LLMs mimicking calibration without the underlying mechanism. Third, scale does not predict reliability. Haiku 4.5, an efficient model, improved most on every metric, while Gemini 3 Pro, a frontier model, showed no improvement despite accessing the procedures. A system whose failure modes cannot be predicted or bounded cannot be certified where human safety depends on it.

Cascade failures instantiate the frame problem [13] at the embodied level. LLMs that selected `WAYPOINT` could not monitor perception frames fast enough to `STOP` the robot before passing the target. This is not a latency problem, it is an architectural limitation of systems that reason only at the boundaries of generation calls. OntoAgent avoids these failures through continuous precondition verification and actionability assessment before issuing any command.

A system that checks preconditions in 60% of trials is not 60% safe, it is unpredictably unsafe. Certification requires bounded, verifiable behavior, and LLM stochasticity, even at temperature zero, precludes it. Neuro-symbolic approaches such as LLM-Modulo [14] provide soundness guarantees through external critics but not metacognition or diagnostic reasoning. OntoAgent's full traceability, where every decision produces an inspectable transcript and every command traces to an ontological justification, is a prerequisite for accountability in safety-critical deployment.

VI. Conclusion and Future Work

Metacognition, domain-grounded diagnosis, and consequence-based action selection are architectural properties of knowledge-grounded systems, not emergent byproducts of LLM scaling or retrieval. In safety-critical embodied settings, decision authority must remain with cognitive architectures that provide these guarantees by construction. LLMs contribute where their strengths apply, including language-mediated interaction and in-context adaptation. We are extending HARMONIC through OntoAgentic AI, where OntoAgent orchestrates LLMs rather than being replaced by them, including leveraging LLMs to accelerate ontological knowledge acquisition.

Acknowledgment

This work was supported in part by ONR Grant #N00014-23-1-2060.

References

[1] M. Ahn et al., "Do as I can, not as I say: Grounding language in robotic affordances," *arXiv:2204.01691*, 2022.

[2] W. Huang et al., "Inner Monologue: Embodied reasoning through planning with language models," in *CoRL*, 2023.

[3] I. Singh et al., "ProgPrompt: Generating situated robot task plans using large language models," in *ICRA*, 2023.

[4] P. Song, P. Han, and N. Goodman, "A survey on large language model reasoning failures," in *AI for Math Workshop @ ICML*, 2025.

[5] M. Griot et al., "Large language models lack essential metacognition for reliable medical reasoning," *Nature Communications*, vol. 16, 2025.

[6] J. E. Laird, *The Soar Cognitive Architecture*. MIT Press, 2012.

[7] J. G. Trafton et al., "ACT-R/E: An embodied cognitive architecture for human-robot interaction," *J. Human-Robot Interaction*, vol. 2, no. 1, 2013.

[8] J. English and S. Nirenburg, "OntoAgent: Implementing content-centric cognitive models," in *Advances in Cognitive Systems*, 2020.

[9] M. McShane and S. Nirenburg, *Linguistics for the Age of AI*. MIT Press, 2021.

[10] M. McShane, S. Nirenburg, and J. English, *Agents in the Long Game of AI*. MIT Press, 2024.

[11] D. Kahneman, *Thinking, Fast and Slow*. Macmillan, 2011.

[12] M. Colledanchise and P. Ögren, *Behavior Trees in Robotics and AI*. CRC Press, 2018.

[13] J. McCarthy and P. J. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," in *Readings in AI*, 1981.

[14] S. Kambhampati et al., "LLMs can't plan, but can help planning in LLM-Modulo frameworks," in *ICML*, 2024.