Large Language Models in Human–Robot Collaboration with Cognitive Validation

Nived E K
LTVE22CS148

Department of Computer Science And Engineering
College of Engineering Trivandrum

Guided by Prof. Neethi Mohan

Table of Contents

Introduction

- LLMs (like GPT) are capable of interpreting and generating human-like text.
- Applied in HRC for flexible instruction understanding and action planning.
- Challenge: LLMs may generate hallucinated content - plausible but factually incorrect.
- Robotics needs **fact-based validation** before execution.

Motivation

- Ensure **safety and reliability** in collaborative robots.
- Avoid **blind trust** in LLM outputs.
- Combine **neural language understanding** with **symbolic verification**.

Problem Statement

To develop an LLM-powered HRC framework that detects and mitigates **context-induced hallucinations** by integrating a **cognitive validation mechanism** before execution.

Literature Survey

Title	Author(s)	Date	Summary	Research Gap
An Empirical Study on Hallucinations in Embodied Agents	Trishna Chakraborty et al.	2025	Evaluates hallucination rates in LLM agents; finds triggers and inconsistencies.	More mitigation, wider agent studies needed.
Human-in-the-loop Multi-Robot Collaboration Framework	Zhaoxing Li et al.	2025	Hybrid LLM+human task allocation; reduces hallucinations; handles diverse robots.	Scalability and large-fleet integration not addressed.
Hallucination Study in LLM-Agents	Yan Zhang	2025	Communication strategies for multi-LLM context/intention.	Empirical studies and comm. overhead understudied.
Working together: A review on safe human-robot collaboration in industrial environments	Sandra Robla-Gómez et al.	2017	early Human-Robot Collaboration (HRC) research, before the use of Large Language Models (LLMs), prioritized physical safety and interaction.	early research prioritized physical safety, it neglected conversational collaboration, creating a need for more intuitive, human-like communication interfaces..

Objectives

- Embed LLMs for natural language interpretation in robots.
- Design a symbolic **validation layer** to detect hallucinations.
- Improve **execution safety** and **task accuracy**.
- Evaluate in simulated collaborative environments.

Methodology Overview

- **Step 1:** User gives instruction via text or speech.
- **Step 2:** LLM parses and generates an action plan.
- **Step 3:** Cognitive Validator checks against internal knowledge graph.
- **Step 4:** Valid plans get executed by robot.

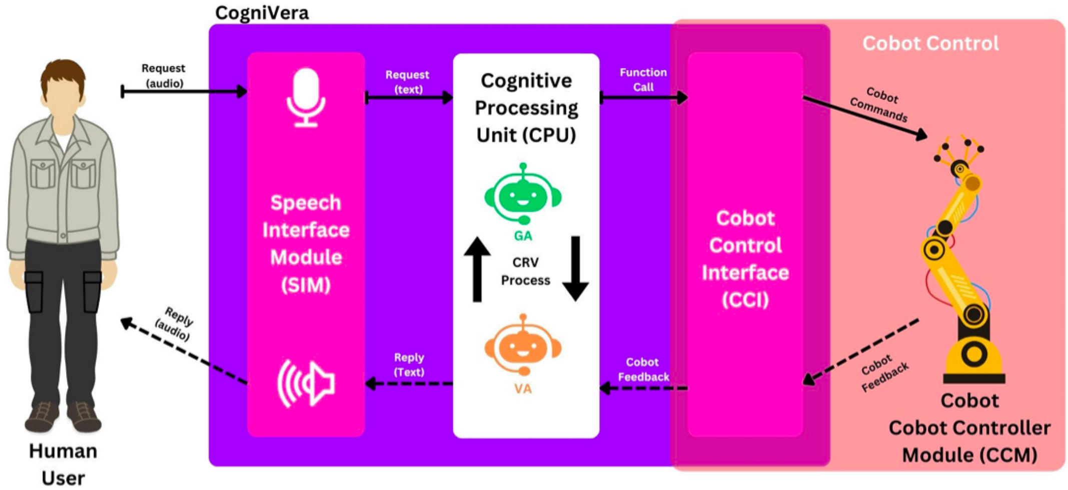
System Architecture

- **Input Interface** – Accepts natural language.
- **LLM Module** – Generates candidate actions.
- **Validator Module** – Symbolically verifies validity.
- **Executor** – Performs physical tasks if validated.

Cognitive Validation Framework

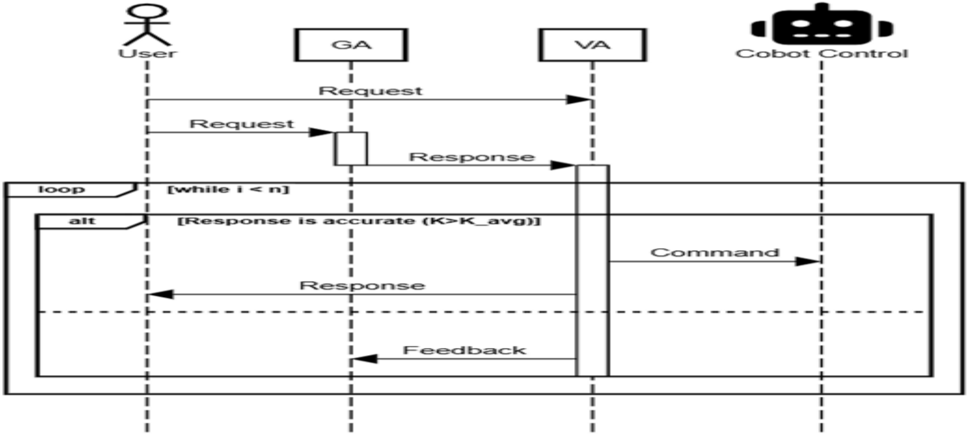
- Uses a **symbolic knowledge base** to check factual consistency.
- Handles object affordances, known entities, and logic.
- Flags hallucinated commands like "carry a fridge with one hand".

System Overview



The CogniVera framework for human-robot collaborative tasks using dual-agent LLM-based

System Overview (contd.)



CogniVera framework workflow.

Experimental Setup

- Simulated domestic robot environment.
- Tasks: Fetching, table setting, object manipulation.
- Compared baseline LLM vs LLM + Validator.

Results

- Achieved **99% error detection** compared to 50% in conventional systems.
- Completed **96.6% of collaborative tasks** versus 40% without validation.
- Slight increase in processing time (**5.47s vs 3.26s**) is offset by improved reliability.

Discussions

- Enhances trust and explainability in robotic systems.
- Prevents risky execution from flawed instructions.
- Cognitive layer enables "common sense" filtering.

Challenges

- Building and maintaining symbolic knowledge bases.
- Balancing accuracy vs latency.
- Extending to open-world tasks.

Conclusion

- Merging LLMs with symbolic reasoning improves HRC safety.
- Cognitive validation prevents hallucinated commands.
- Shows promise for trustworthy, autonomous robot collaborators.

Future Work

- Automate knowledge acquisition.
- Combine multimodal inputs (vision + speech).
- Real robot deployment and long-term learning.

References

[1] N. Ranasinghe, W. M. Mohammed, K. Stefanidis, and J. L. Schoonderwoerd, "Large language models in human–robot collaboration with cognitive validation against context-induced hallucinations," *IEEE Access*, 2025.

[2] T. Chakraborty, A. Roy, S. Dey and M. Banerjee, "An empirical study on hallucinations in embodied agents," in *Proc. IEEE*, 2025.

[3] Z. Li, W. Wu, Y. Wang, and H. Zhang, "Human-in-the-loop multi-robot collaboration framework," *IEEE Trans. on Systems, Man, and Cybernetics*, 2025.

[4] Y. Zhang, "Hallucination study in LLM-agents," *IEEE Early Access*, 2025.