MASTER'S THESIS

EMBODIED ARTIFICIAL COGNITION: EVALUATING SELF-AWARENESS IN MULTIMODAL LARGE LANGUAGE MODELS WITH ROBOTIC SENSORY INTEGRATION

SEPTEMBER **2025**

Iñaki Luciano Dellibarda Varela

MASTER'S THESIS
TO OBTAIN A MASTER'S DEGREE IN AUTOMATION AND ROBOTICS

MASTER'S THESIS SUPERVISOR:
Dr. Eduardo Rocon de Lima
Dr. Manuel Cebrián Ramos

INDUSTRIALES
ETSII | UPM

POLITÉCNICA

UNIVERSIDAD
POLITÉCNICA
DE MADRID

Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Industriales

Master's Degree in Automation and Robotics

Masters's Thesis

Embodied Artificial Cognition: Evaluating Self-Awareness in Multimodal Large Language Models with Robotic Sensory Integration

Autor: **Iñaki Luciano Dellibarda Varela**

Director 1: **Dr. Eduardo Rocon de Lima**

Director 2: **Dr. Manuel Cebrián Ramos**

September, 2025

Familiae et amicis meis, omnia mihi.

Acknowledgments

With this work, I am bringing to a close a very beautiful stage of my life, my university years. However, I will remain a student for the rest of my life, because if this journey has taught me anything, it is that the most beautiful thing in this world is learning. For this reason, I would like to thank all the people who have accompanied me on this journey, starting with all my classmates and professors with whom I have traveled this path during four years of undergraduate study and one year of master's study; you have been great guides.

This work would not have been possible without the invaluable support of Eduardo Rocon, who has been a mentor to me over the last few years, welcoming me into his group, trusting me, and giving me opportunities for which I will always be grateful. To Manuel Cebrián, my other great mentor, I am very grateful for the countless hours of conversations we have had this past year, whether about work, sports, or metaphysical reflections; you have helped me expand my mind and not set limits for myself. I am also grateful to the rest of my colleagues at CAR, who motivate me every day to continue working on what I am most passionate about, especially Pablo Romero, Gabriel Delgado, Diego Torricelli, Nacho Serrano, Lola del Castillo, and Álvaro Gutiérrez. You are the most intelligent people I know, and this work has been possible thanks to you.

Cristina García, my eternal partner, has always been there to accompany me during countless hours of work and celebration; you make me happier than anyone else in this world and I feel lucky every day to be by your side.

And finally, to my parents, Alejandra and Fito. One paragraph is not enough to thank you for everything you have done for me. Thank you for teaching me to be curious, for showing me that I can achieve anything I set my mind to, and for picking me up every time I fell. I am who I am thanks to you; my successes are yours. Thank you for giving your all for me; I can look at myself in the mirror with pride when I see you. I love you with all my heart.

Abstract

Self-Awareness — the capacity of an individual to represent and understand itself as the subject of experience and action — is sustained as the foundation of intelligence and autonomous behavior. The most recent advances in AI have reached human-like performance in tasks that integrate multimodal information, especially in large language models (LLMs), which has raised interest in the embodiment capabilities of AI agents in non-human platforms such as robots.

For centuries, different fields of study, from philosophy to neuroscience, have devoted significant efforts to the definition and characterization of Self-Awareness. In the present study, the capabilities of a LLM to develop Self-Awareness are analyzed when embedded in an autonomous mobile robot, relying solely on sensorimotor experience.

By integrating a multimodal LLM into an autonomous mobile robot, we test its capacity to achieve artificial Self-Awareness. We find that the system demonstrates solid environmental awareness, self-recognition, and predictive awareness, which allows it to infer its robotic nature and movement characteristics. Structural Equation Modeling (SEM) reveals how sensory integration influences different dimensions of Self-Awareness and its coordination with past–present memory, as well as the hierarchical internal associations that drive self-identification. Moreover, through SEM we identify similarities between the cognitive constructs developed by the system and the human brain structures responsible for Self-Awareness.

Ablation tests of sensory inputs identify critical modalities for each dimension, demonstrate compensatory interactions between sensors, and confirm the essential role of structured episodic memory in coherent reasoning. These findings show that, given adequate sensory information about the world and itself, multimodal LLMs exhibit emergent Self-Awareness, opening the door to embodied artificial cognitive systems.

Keywords: Self-Awareness, Multimodal Large Language Models (MM-LLMs), Structural Equation Modeling (SEM), Ablation Test, Multi-Dimensional Awareness, Cognitive

Robotics & Past-Present Memory.

UNESCO Codes:

1203 - Computer Science

- 1203.04 - Artificial Intelligence
- 1203.17 - Informatics

1207 - Operational Research

- 1207.09 - Control Systems

1209 - Statistics

- 1209.05 - Analysis and Design of Experiments
- 1209.09 - Multivariate Analysis

2490 - Neurosciences

- 2490.01 - Neurophysiology

3399 - Robotics

- 3399.01 - Cognitive Robotics

6106 - Experimental Psychology

- 6106.01 - Brain Activity
- 6106.06 - Memory Processes
- 6106.09 - Perception Processes
- 6106.12 - Sensory Processes

Resumen

Autoconsciencia — la capacidad del individuo para representarse y entenderse a sí mismo como sujeto de la experiencia y la acción — se sustenta como base de la inteligencia y del comportamiento autónomo. Los avances más recientes en IA han alcanzado un rendimiento similar al humano en tareas que integran información multimodal, especialmente en modelos amplios de lenguaje (LLMs), lo que ha suscitado interés en las capacidades de encarnación de los agentes de IA en plataformas no humanas, como los robots.

Durante siglos, diferentes campos de estudio, desde la filosofía hasta la neurociencia, han dedicado numerosos esfuerzos a la definición y caracterización de la Autoconsciencia. En el presente estudio se analizan las capacidades de un LLM para desarrollar Autoconsciencia al embeberse en un robot móvil autónomo, simplemente mediante la experiencia sensorio-motora.

Al integrar un LLM multimodal en un robot móvil autónomo, probamos su capacidad para alcanzar Autoconsciencia artificial. Descubrimos que el sistema muestra una sólida conciencia del entorno, autorreconocimiento y conciencia predictiva, lo que le permite inferir su naturaleza robótica y sus características de movimiento. El modelado de ecuaciones estructurales (SEM) revela cómo la integración sensorial influye en distintas dimensiones de la Autoconsciencia y su coordinación con la memoria pasado-presente, así como las asociaciones internas jerárquicas que impulsan la autoidentificación. Además, mediante SEM podemos encontrar similitudes entre los constructos cognitivos que desarrolla el sistema y las estructuras humanas del cerebro responsables de la Autoconsciencia.

Las pruebas de ablación de las entradas sensoriales identifican modalidades críticas para cada dimensión, demuestran interacciones compensatorias entre los sensores y confirman el papel esencial de la memoria estructurada y episódica en el razonamiento coherente. Estos hallazgos muestran que, dada la información sensorial adecuada sobre el mundo y sobre sí mismo, los LLM multimodales exhiben una Autoconsciencia emergente, lo que abre la puerta a sistemas cognitivos artificiales encarnados.

Palabras clave: Autoconsciencia, Modelos de Lenguaje Multimodales de Gran Escala (MM-LLMs), Modelado de Ecuaciones Estructurales (SEM), Prueba de Ablación, Conciencia Multidimensional, Robótica Cognitiva y Memoria Pasado-Presente.

Códigos UNESCO:

1203 - Informática

- 1203.04 - Inteligencia Artificial
- 1203.17 - Informática

1207 - Investigación Operativa

- 1207.09 - Sistemas de Control

1209 - Estadística

- 1209.05 - Análisis y Diseño de Experimentos
- 1209.09 - Análisis Multivariante

2490 - Neurociencias

- 2490.01 - Neurofisiología

3399 - Robótica

- 3399.01 - Robótica Cognitiva

6106 - Psicología Experimental

- 6106.01 - Actividad Cerebral
- 6106.06 - Procesos de Memoria
- 6106.09 - Procesos de Percepción
- 6106.12 - Procesos Sensoriales

Index

List of Figures

List of Tables

Acronyms

AG: Angular Gyrus
AI: Artificial Intelligence
AIC: Anterior Insular Cortex
AMY: Amygdala
API: Application Programming Interface
CAR: Center for Automation and Robotics
Cau: Caudate
CFI: Comparative Fit Index
CLIP: Contrastive Language-Image Pretraining (CLIP)
CNN: Convolutional Neural Network
CSIC: Spanish National Research Council
DMN: Default Mode Network
dmPFC: Dorsal Medial Prefrontal Cortex
FA: Factor Analysis
GMR: Giant Magnetoresistance
IMU: Inertial Measurement Unit
JSON: JavaScript Object Notation
LLM: Large Language Model
MCP: Model Context Protocol
MLE: Maximum Likelihood Estimation
MLR: Multiple Linear Regression
MM-LLMs: MultiModal Large Language Models
MTG: Middle Temporal Gyrus
MTL: Medial Temporal Lobe
NLP: Natural Language Processing
OECD: Organization for Economic Co-operation and Development
PCC: Posterior Cingulate Cortex

RLHF: Reinforcement Learning from Human Feedback
RMSEA: Root Mean Square Error of Approximation
RNN: Recurrent Neural Network
ROS: Robot Operating System
RSP: Responsible Scaling Policy
SEM: Structural Equation Modeling
SLAM: Simultaneous Localization and Mapping
SDG: Sustainable Development Goals (SDG)
Thal: Thalamus
TII: Technology Innovation Institute
TLI: Tucker-Lewis Index
USPTO: United States Patent and Trademark Office
ViT: Vision Transformer
vmPFC: Ventral Medial Prefrontal Cortex
VQA: Visual Question Answering

Chapter 1

Introduction and Motivation

In 2017, the field of Artificial Intelligence (AI) begins a revolution that profoundly impacts society. A team of researchers from Google published the paper *Attention Is All You Need* [1], introducing the Transformer architecture. In the following years, this technology enables the emergence of Large Language Models (LLMs) as powerful tools. Trained on massive text corpora to learn statistical patterns of language, LLMs are capable of understanding and generating human-like text [2], [3] and even recreating human patterns of behavior and reasoning [4]–[6].

As a result, in recent years thousands of papers are published yearly, exploring the capabilities of LLMs across multiple domains [5], [7], [8], conducting sociological experiments [9]–[12], evaluating their strengths and limitations with complex benchmarks [5], [13]–[16], and applying them to embodiment in robotics [17]–[21]. In this latter field, most studies focus on assessing the ability of LLMs to act as an artificial brain for controlling robot movements and actions.

In the present study, we focus on another dimension of robotic embodiment through LLMs: analyzing their capacity to understand themselves and their surrounding world purely from sensory inputs, without any prior context or external information, and studying their potential to develop artificial Self-Awareness.

Looking at the direction the world is heading, where autonomous machines and AI are becoming inherent elements of everyday life, it is crucial to understand the capabilities and limitations of these embodied systems, both for safety and psychological considerations. This study takes a first step in that direction by testing and analyzing the current cognitive and perceptual capacities of LLMs, evaluating how they perceive themselves, how they interpret their surroundings, and how they interact with the world.

1.1 Background and Context

Philosophers since antiquity have considered Self-Awareness essential to cognition, most famously articulated by Descartes in his dictum Cogito, ergo sum—"I think, therefore I am" (Descartes, 1637). In psychology, self-recognition is traditionally assessed through the mirror test, introduced by Gallup in 1970, revealing that certain animals, including primates, dolphins, and birds, possess a basic form of Self-Awareness [22]. Neuroscientifically, Self-Awareness emerges from intricate neural interactions involving the prefrontal and insular cortices, integrating bodily sensations and introspection [23].

However, in AI the question of Self-Awareness remains unresolved and is often conflated with the ability to mimic human-like behavior, as exemplified by the classic Turing Test [24]. While the Turing Test gauges behavioral indistinguishability from humans, genuine Self-Awareness requires internal recognition of oneself as distinct from the environment—an attribute yet to be thoroughly explored in artificial systems [25], [26].

Advances in machine learning and neural networks, often inspired by neurobiological principles, enable artificial intelligence systems embedded in complex hardware to achieve success rates once believed exclusive to biological brains—and in some cases, even surpass them [27]–[29].

LLMs rapidly advance, demonstrating human-like or superior performance in complex cognitive tasks such as language comprehension, reasoning, multimodal perception and the interpretation of subtle discourse phenomena like irony and faux pas [30]–[34]. The evolution into multimodal LLMs (MM-LLMs), which integrate text, vision and other sensory modalities, marks a pivotal step toward human-level artificial intelligence [35]–[38] and opens the door to machines replicating aspects of Self-Awareness.

Yet, most research integrating LLMs and robots has focused on command-based interactions—robots executing human-issued instructions—without investigating whether these models can autonomously interpret their own sensory experiences to develop an internal sense of self [17], [39]–[42]. Here, we address precisely this unexplored frontier: Can a MM-LLM, embedded in an initially unknown robotic entity, develop Self-Awareness purely from multimodal sensorimotor data acquired during active exploration?

Defining Self-Awareness in humans is a complex and non-trivial task. This challenge becomes even greater when attempting to identify or quantify Self-Awareness in an artificial intelligence given system, such as a robot. In this study, to simplify the problem, we divide Self-Awareness into three key dimensions: (i) Environmental Awareness, the ability to perceive and interpret surroundings through multimodal sensory inputs; (ii) Individual Awareness, the capacity to infer one's own physical structure and characteristics; and (iii) Predictive Awareness, the refinement of self-perception through the integration of past experiences and sensor data.

Using Gemini 2.0 Flash [38], [43], we evaluated an MM-LLM embedded in an omni-

directional mobile robot, systematically examining these three core aspects of artificial Self-Awareness. By addressing these dimensions, our study investigates the potential of MM-LLMs to autonomously generate a coherent sense of self through direct interaction with their environment, advancing the frontier toward genuinely Self-Aware artificial systems.

1.2 Structure of the Thesis

The master's thesis is organized into nine chapters and eight appendices:

- **Chapter 1:** introduction and motivation behind this study. Brief introduction to the current state of AI, the need for this research and the the objectives pursued.
- **Chapter 2:** state-of-the-art of the present study. It: analyzes the evolution of AI that has led to the current state, focusing on LLMs; applications of LLMs in robotics; definition of Self-Awareness according to different fields of study; and cognitive robotic systems.
- **Chapter 3:** Definition of the objectives to be achieved in this study.
- **Chapter 4:** description of the technical methodology implemented in the study: robotics and sensor technology, MM-LLM used, system architecture, and establishment of the experimental framework.
- **Chapter 5:** presentation and discussion of the results obtained through experimental techniques (autonomous navigation studies and ablation tests), and through a complex statistical analysis.
- **Chapter 6:** presentation of the conclusions of the study.
- **Chapter 7:** description the future work.
- **Chapter 8:** study of ethical, legal, and professional implications and responsibilities.
- **Bibliography:** used in this TFM.
- **Appendix A:** timeline of the project.
- **Appendix B:** budget of the project.
- **Appendix C:** analysis of the environmental and social implications and contribution to the sustainable development goals associated with the study.

- **Appendix D:** detailed explanation of the criteria for searching for information and bibliographic data on the state-of-the-art of LLMs in robotics.
- **Appendix E:** definition of the system prompt used during the experimental process.
- **Appendix F:** rubrics used for evaluating the system.
- **Appendix G:** git repository with access to data, code, and materials supporting the findings of this study.
- **Appendix H:** summary of the replication of the experiments carried out for this study, but using manual navigation instead of autonomous robot navigation.

Chapter 2

State of the Art

We present a comprehensive overview of the state of the art in AI and LLMs, followed by a multidimensional analysis of Self-Awareness and consciousness from technical, psychological, and philosophical perspectives. Next, we survey recent advances in cognitive robotic systems and conclude with prior studies that directly inform the present research.

2.1 AI and LLMs

Since the term Artificial Intelligence was coined in 1955 in the proposal for the Dartmouth Summer Research Project [44], the discipline has experienced alternating "AI summers" and "AI winters" [45]—periods of intense optimism, breakthroughs, and investment, followed by phases of disillusionment, stagnation, and reduced funding.

The first sparks of practical artificial intelligence appear in 1958 with Rosenblatt's perceptron [46], a single-layer neural network capable of learning to classify inputs by adjusting connection weights. However, its technical limitations—most notably its inability to solve linearly inseparable problems, as highlighted in the Lighthill report [47]—trigger the first AI winter [45], as confidence in neural approaches collapses.

In 1986, Hinton and collaborators revive the discipline with the backpropagation algorithm [48], enabling the training of multi-layer neural networks and laying the foundations for modern deep learning [49]. This method uses gradient descent to adjust weights across multiple layers, allowing networks to extract increasingly abstract representations from data. For his impressive contribution to the field, in 2024, Hinton, in the company of Hopfield, won the Nobel Prize in Physics for foundational discoveries and inventions that enable machine learning with artificial neural networks [50].

From the 1990s onwards, AI undergoes cycles of progress and setbacks, but in 2017 the introduction of the Transformer architecture redefines the state of the art of the field [1]. Transformers replace recurrence with self-attention mechanisms, enabling models to

process entire sequences in parallel while capturing long-range dependencies with high efficiency. Since their introduction, the field of AI experiences exponential growth (see Figure 2.1), driving the development of Large Language Models and powering a new generation of multimodal systems (MM-LLM).

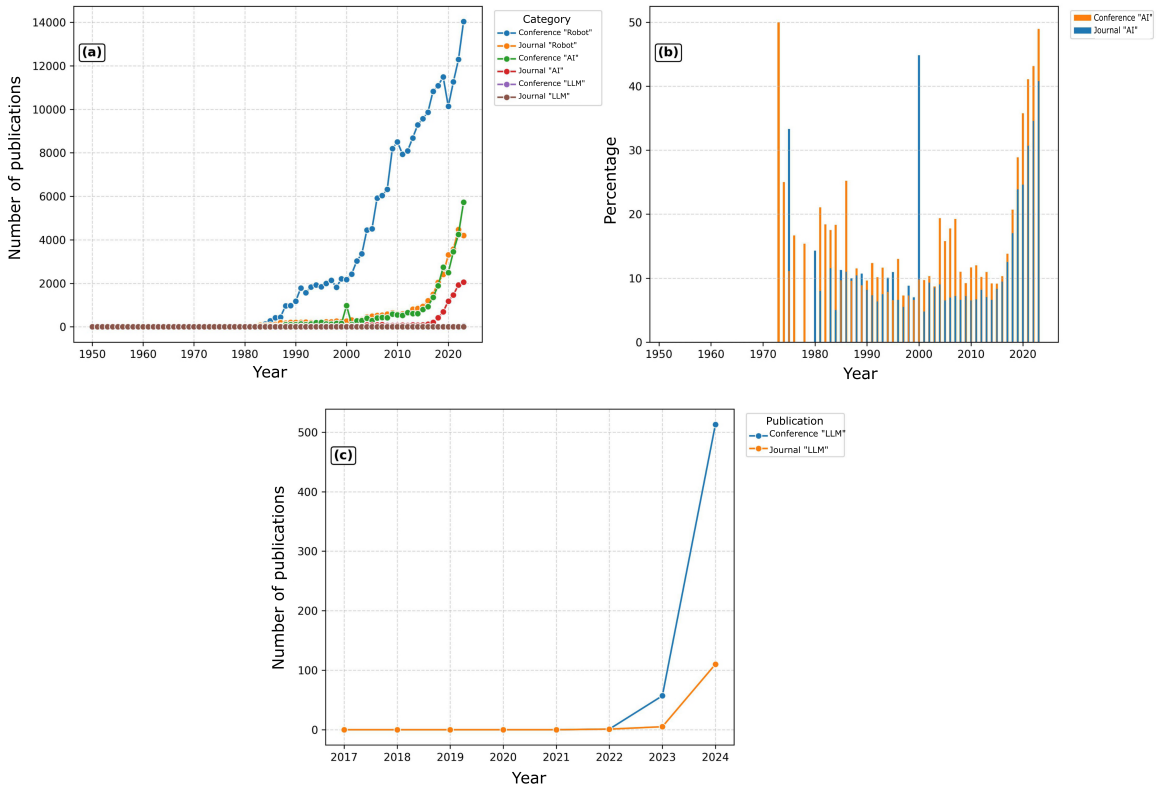

Figure 2.1. Evolution of scientific publications in the field of robotics and AI. (a) Annual number of publications related to robotics and AI in scientific conferences and journals. (b) Percentage of AI-related publications with respect to the total number of robotics publications, distinguishing between conferences and journals. (c) Evolution of the number of publications on LLMs since 2017. The detailed process of obtaining the bibliometric data and processing is in Appendix D.

Figure 2.1 a) shows the number of publications in both conferences and journals related to AI, robotics, and LLMs. AI exhibits a steady and gradual growth from its origins until the mid-to-late 2010s, when the introduction of the Transformer architecture and, consequently, LLMs sparks a profound surge in interest and capabilities, driven by advances in computational power and novel algorithms. This shift is reflected in the sharp increase in publications. The upward trend continues, with no signs of stabilization in the near term.

This growth extends to numerous fields, with robotics research among the most affected (see Figure 2.1 b). Before the Transformer era, fewer than 10% of robotics publications were AI-related. Since then, the share has risen steadily each year, reaching almost 50% of conference publications and surpassing 40% of journal publications in 2023.

Focusing specifically on LLM-related studies, their surge occurs slightly later than in other AI domains. Although the first publications appear in 2017, their widespread use and the development of applications do not accelerate until 2022. This inflection coincides with the launch of ChatGPT in November 2022, which introduces LLMs to the broader technical and non-technical public [51]. The recent and sustained increase in the number of publications across AI, LLMs, and robotics fields highlights the impact, growing interest, and potential of these technologies.

2.1.1 Architectural Foundations of Transformers

The Transformer architecture (see Figure 2.2 a) represents a radical breakthrough in the field of AI by replacing traditional recurrent and convolutional mechanisms with a model based exclusively on the self-attention mechanism. This mechanism enables the model to process entire sequences of data by tokenizing the inputs, converting each token into a numerical vector (*embedding*), and computing similarity relationships between all tokens in the sequence. Through these relationships, the model identifies which parts of the input are most relevant for generating each element of the output, allowing for a rich and flexible contextual representation. This approach has proven particularly effective in tasks such as language modeling, machine translation, and semantic understanding [1]. The self-attention based architecture offers computational efficiency and scalability, enabling the training of models of unprecedented size, with billions of parameters [52], [53].

Due to its potential and versatility, in the last years different variants of the Transformers architecture has emerged, allowing multimodal information processing. One of the most relevant is the Vision Transformer (ViT) architecture [54]. The ViT (see Figure 2.2 b) build on the original Transformer architecture, adapting self-attention for image processing. Instead of operating directly on pixels or using convolutional kernels, ViT divides an image into fixed-size patches, flattens them, and projects each patch into an embedding vector, which is then treated as a token in a sequence. ViT captures long-range dependencies between image regions, overcoming some of the locality and inductive bias limitations of Convolutional Neural Networks (CNNs) [54].

Similar to the ViT, the Speech-Transformers are specialized and adapted for the audio domain and speech recognition. Their main advantage over traditional audio processing methods lies in the ability to process audio in parallel chunks through self-attention, rather than relying on recurrent sequence-to-sequence architectures. They use feature embeddings extracted from spectrograms or similar representations as input, treating them as a sequence of tokens analogous to text tokenization or patch embeddings in vision. Positional encodings are integrated to preserve the temporal structure of speech, compensating for the loss of implicit order provided by Recurrent Neural Networks (RNNs) [55].

As a synthesis of the architectures discussed so far, multimodal Transformers have

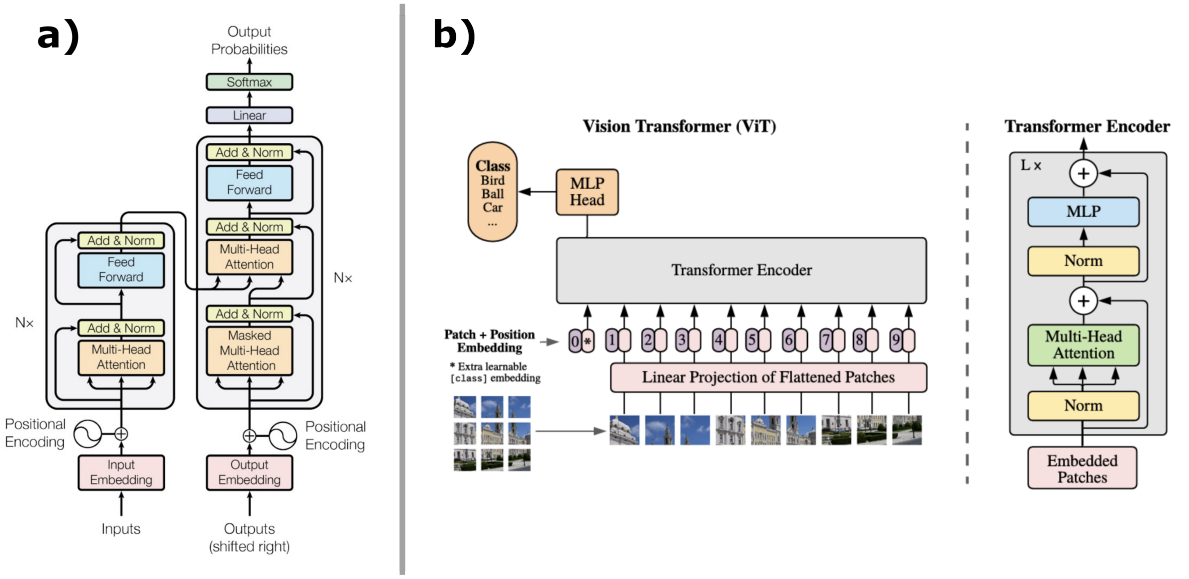

Figure 2.2. (a) Transformer architecture. Figure extracted from [1]; (b) Vision Transformers (ViT). Figure extracted from [54]

been developed. These models combine image and text processing, as in Contrastive Language–Image Pretraining (CLIP), which integrates a visual encoder and a textual encoder [56] trained on millions of image–text pairs collected from the internet. Other approaches extend this idea by incorporating video or audio processing into the same framework, such as Flamingo [57] and Kosmos [58].

2.1.2 LLMs and MM-LLMs

The rapid evolution, scalability, parallelism, cross-domain adaptability, and multimodal integration of Transformers have laid the foundation for the development of Large Language Models (LLMs) and Multimodal Large Language Models (MM-LLMs). These are advanced Natural Language Processing (NLP) systems trained at massive scale, with hundreds of billions of parameters, on diverse multimodal data such as text, images, and videos. LLMs were initially specialized in text comprehension and generation, but have since been extended to process and integrate multiple modalities, giving rise to MM-LLMs.

The great potential of this technology has led to the emergence of major companies, such as OpenAI and Anthropic, and has driven technology giants, including Meta, X, and Google, to invest heavily in leading the market, positioning these models as a key technol-

ogy for the future. This competitive race for market leadership and the development of state-of-the-art systems has resulted in the release of highly capable models, both closed- and open-weights. While some models provide full public access to their weights, enabling unrestricted research and deployment, most remain proprietary but offer powerful APIs for developers and the general public. The state-of-the-art models are:

Closed-weight commercial models: these models are the proprietary assets of an enterprise, accessible exclusively through an API where users pay for token usage. Their parameters (weights) are neither public nor modifiable. The developing company retains full control over the model's architecture, development, and application. Key examples of these models include:

- **OpenAI**: Its current most powerful model, GPT-5, surpasses other commercial models in several widely used benchmarks, achieving unprecedented scores such as 74.9% on SWE-bench Verified [59] for coding tasks, and 96.7% on the τ2-bench [60] agent evaluation. It is characterized by advanced complex reasoning [61], multimodal integration, and long-context understanding. Its API supports inputs of up to 272,000 tokens (traditionally, one token is approximately equivalent to one syllable), representing a significant increase over previous versions and currently leading the field [35], [62].

- **Anthropic:** Its most powerful reasoning model, Claude Opus 4.1, achieves outstanding results, with a 74.5% performance on SWE-bench Verified [59], placing it on par with models like GPT-5. One of Anthropic's primary differentiators from prior LLMs lies in its profound focus on user utility and safety. The company introduced Constitutional AI, a set of guiding principles that govern model training, and employs a Responsible Scaling Policy (RSP) to evaluate various risks, from chemical weapons to ethical principles, establishing a security scope that exceeds traditional frameworks [63].

  Anthropic is also responsible for developing the Model Context Protocol (MCP), a widely accepted standardized interface. This protocol is designed to facilitate seamless interaction between AI models and external tools and resources, thereby breaking down data silos and enhancing interoperability across diverse systems [64].

- **Google DeepMind:** this specialized team at Google focuses on advanced AI research, particularly in the fields of NLP and LLM-derived tools. Among their most outstanding contributions are: the Gemini family of models, led by its most powerful variants, Gemini 2.5 Flash (optimized for exceptional speed and efficiency) and Gemini 2.5 Pro (designed for deep reasoning) [43]; the open-weight Gemma models, which are distilled from their more powerful counterparts [65]; Genie 3, an advanced real-time video generator that creates immersive virtual worlds [66]; MedGemini, a

model specialized in medical knowledge [67]; and Gemini Robotics, which enables robot control via natural language [17].

- **Other Models:** notable models include Grok, developed by **xAI**, which is specifically optimized for social media interaction and real-time information processing [68]. Additionally, **Microsoft** Copilot [69] is designed to augment productivity through its direct integration with the Microsoft suite. It is important to note that Copilot is not a foundational model; rather, it is a sophisticated orchestration engine that relies on foundational models like GPT to function.

Open-Weight Commercial Models: these models make their weights publicly available and modifiable, granting users the ability to fine-tune them. This allows for specialization in a certain area of knowledge or a particular domain of actuation. Key examples of these models include:

- **Meta:** The company develops the Llama family of models, which follows an open-source strategy and is highly popular among developers due to the ease of fine-tuning. These model modifications provide complex support, such as: helping medical professionals more quickly and precisely diagnose bacterial infections and recommend the appropriate antibiotics [70]; processing information and offering live assistance for customer support agents [71]; or providing improved, knowledge-based financial assistance, among other applications [72].

- **Mistral AI**: is a prominent European AI company that has quickly gained recognition for its innovative approach to developing powerful, efficient, and open-weight models. Their philosophy focuses on releasing highly performant models that are small enough to be easily deployed and fine-tuned [73].

- **DeepSeek**: is a Chinese enterprise whose models are characterized by their open weights and transparency. Their implementation, which required fewer resources compared to other models, was a significant development and has prompted new questions about how models are developed and trained [74].

- **Other Models:** such as Qwen [75] from **Alibaba**, a powerful and highly modifiable model; as previously mentioned, Gemma from **Google** [65]; and Falcon from the Technology Innovation Institute (**TII**), specialized in the Arabic region and released under a license allowing commercial use and modifications [76].

Non-commercial models for research: in addition to the commercial models described, which are developed by large corporations with a customer-centric focus, there are other foundational models designed primarily for research purposes. Some of the most relevant include **BLOOM**, a model developed through a collaboration of hundreds of developers, trained on 46 natural and 13 programming languages [77]; **Cerebras-GPT**,

released under the Apache 2.0 license, which allows for both commercial and research use [78]; and **GPT-Neox**, a non-profit model based on the GPT-3 architecture, specifically focused on academic research and exploring the capabilities of large-scale models [79].

2.1.3 Robotics LLMs Applications

The wide range of functions and generative capacity of LLMs has led to their use in a variety of robotic fields. We can classify their applications into three main groups: planning and control of robotic tasks; human-robot interaction; other varied tasks.

LLMs applied to planning and control of robotic tasks

With the rise of LLMs' generative capabilities, a key challenge has emerged: converting natural language instructions into action plans that are both understandable and executable by robots. This problem has driven the development of new strategies that combine linguistic reasoning with physical execution, highlighting approaches that allow language models to directly generate structured plans adapted to the environment.

One of the initial approaches involves generating Python code from natural language to represent structured robotic plans. In this context, ProgPrompt and similar techniques are particularly notable. This method uses a programming template as a prompt, in which auxiliary functions like `grab()`, `putin()`, and `switchon()` are defined, along with supporting objects and structures such as comments and environmental checks using `asserts` [80]. This represents one of the first effective integrations between the symbolic representation of language and robotic planning.

The next step in the evolution of robotic control systems via language models is the combination of programming capabilities acquired by LLMs during their training with the specific APIs of the robots to be controlled. That is, starting from the set of functions the robot can execute, upon receiving an order in natural language, the LLM translates it into policies or rules composed of sequences of commands that enable the robot to perform the objective action [19].

In this context, projects from Google developed by the DeepMind team stand out. One of these is *Code as Policies* [19], which translates natural language instructions into robotic policies in the form of executable code. Its most recent evolution is the *Gemini Robotics* model [17] (see Figure 2.3), a vision-control system for robots that incorporates multimodal capabilities, combining sensory information (such as cameras and audio) and enabling the conversion of verbal commands into reasoned robotic actions. This model has proven capable of performing tasks such as organizing objects in changing environments, interacting with unfamiliar items (such as a ball and a basketball hoop), and facilitating spoken communication and collaboration between multiple robots to execute complex tasks requiring fine motor skills, such as origami. In addition, they have demonstrated

the best ability and understanding of movement and changes in the world around them, being able to generate and predict trajectories with much more positive results than other commercial LLMs.

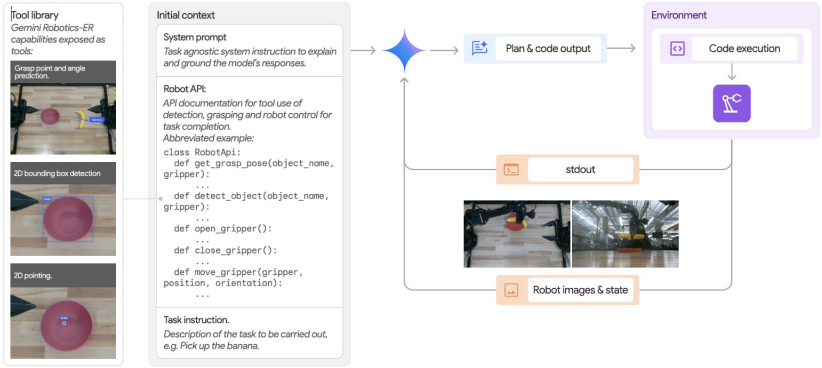

Figure 2.3. Gemini Robotics workflow. Figure taken from [17].

In keeping with this approach, several models based on vision-control systems for robotic manipulation have been developed. The NLMap + SayCan architecture stands out by integrating LLMs with visual representations generated from unstructured environmental data. This architecture employs a planning system where the robot explores its surroundings with RGB-D sensors and uses visual-linguistic models like CLIP and ViLD to construct an open-ended semantic representation of the scene (NLMap). This enables more flexible and robust planning in open environments without requiring predefined lists of objects or actions, providing a natural integration between perception and language [81].

Although the applications mentioned are promising, real-world applications and validations remain scarce. A notable exception is TidyBot, a system developed by Princeton University, Stanford University, Columbia University, Google, and The Nueva School [82]. This project offers a robotic solution for personalized home cleaning, integrating large-scale language models with visual perception to infer user organizational preferences from a few examples. TidyBot has achieved excellent results, with an 85% success rate, demonstrating significant promise in this field.

LLMs applied to human-robot interaction

One of the most important historical applications of robotics has been to make people's lives easier. However, its effective use and full understanding have traditionally been restricted to a technical elite with specialized knowledge. In this context, one of the most significant contributions of LLMs is their ability to reduce the barrier between humans and robots, enabling more natural, accessible, and understandable forms of interaction for non-expert users.

We can identify three main application areas of human-robot interaction mediated by LLMs [83]. The first is multimodal generative query answering, which integrates different sensory modalities like text, images, and audio to provide comprehensive and contextualized responses. In this approach, LLMs serve as the generative core, working with visual or acoustic models to interpret complex questions, translate information, and generate coherent descriptions or behaviors. Key applications include Visual Question Answering (VQA), automated story generation, multimodal translation, and robotic interfaces that can justify their decisions. The second area involves the use of social robots with common sense, particularly in fields like education, healthcare, and personal assistance, where LLMs enable natural and empathetic conversational interactions. The third area is instruction following and task execution, where LLMs translate natural language commands into structured action plans for physical execution by a robot, which is highly relevant for domestic assistance, industrial assembly, and human-robot collaboration [19].

This is a very promising field that is expected to see significant development in the coming years, as studies have already proven that the use of LLMs can improve user confidence in human-robot collaborative tasks [84].

Other applications of LLMs in robotics

In addition to the applications previously analyzed, other emerging utilities of LLMs in robotics stand out. For example, the development of VLMaps [85], an architecture that fuses visual-linguistic models with 3D reconstructions of the environment to enable spatial navigation guided by natural language. This technique allows for the generation of queryable maps with instructions such as "between the sofa and the television" and has shown improvements over previous alternatives in multi-goal navigation and generalization to different types of robots. Separately, studies have explored the use of LLMs in medical contexts [86], demonstrating their ability to provide accurate, understandable, and humanized answers to common questions from patients. Although outside the purely robotic domain, this study highlights the potential of LLMs as accessible and reliable interfaces between humans and technical systems in sensitive areas like healthcare.

Products and patents

The novelty of LLMs and their application in robotics has led to the development of related patents in this field. A total of 953 patents have been found in the United States Patent and Trademark Office (USPTO) [87] (for more details on the acquisition of these patents, see Appendix D). Many of them, due to their novelty, are still under review and pending approval. The great variety and number of patents that have emerged in recent years corroborate the relevance and interest in this area of study. Some notable patents are shown in Table 2.1.

Table 2.1. Classification of patents on the integration of LLMs in robotics

Name	**Reference**	**Description**	**Status**
Robot systems, methods, control modules and computer products that leverage large language models	US 2024/0359319 A1 [88]	Controls physical tasks through semantic reasoning about visual and symbolic environments. Uses an LLM to decide robotic actions with sensory integration.	Application
Robotic reasoning through planning with language models	US 2025/0018562 A1 [89]	The robot uses an LLM to plan complex tasks through iterative *inner monologue*-style reasoning, integrating textual feedback from the environment to adapt its behavior.	Application
Generating code for robotic systems using large language models	US 2023/0311335 A1 [90]	The LLM converts natural language instructions into source code (e.g., Python, ROS) that directly controls the robot's behavior. Includes pre-execution validation.	Application
Enforcing robotic safety constraints based on AI-generated safety descriptions	US 2025/0042032 A1 [91]	A supervisory agent analyzes an LLM's responses about the environment to modify or block robotic commands, ensuring safe execution even in dynamic or uncertain environments.	Application
Method for controlling a robot apparatus	US 2025/0144796 A1 [92]	The system converts a scene graph into textual descriptions and uses LLMs to predict human behavior and generate adaptive robotic plans in dynamic environments.	Application
Systems and methods for training an autonomous machine to perform an operation	US 2024/0419977 A1 [93]	The system uses an LLM to generate reward functions and goals from natural language descriptions, training robotic policies through reinforcement learning in physical or simulated environments.	Application

2.2 Self-Awareness

The concept of Self-Awareness has been a profound and enduring subject of inquiry throughout human history. As an abstract yet intrinsically human concept, its definition and exploration have been the focus of countless theses and books. This section will delve into the major works and theories on Self-Awareness, examining its definition from the perspectives of diverse disciplines, including philosophy, psychology, neuroscience, and computational neuroscience as it relates to AI. We will conclude with the scope of Self-Awareness that we will pursue as the objective of this work.

2.2.1 Philosophical Perspective

In his *Meditations on First Philosophy* [94], René Descartes (1596-1650) articulates a view of the self as a thinking thing (*res cogitans*), where "thinking" is broad and includes doubting, understanding, affirming, denying, willing, imagining, and, crucially, seeming to sense. Through methodic doubt, which extends to the deceptive potential of the senses and even to mathematics (the evil demon hypothesis), Descartes seeks a foundation that deception cannot undermine and finds it in the thinker's awareness of its own thinking. This yields the thesis that Self-Awareness is immediate and indubitable: in the *Meditations* he states "*ego sum, ego existo*" (I am, I exist) whenever he thinks, while the well-known formulation "*cogito ergo sum*" (I think, therefore I am) appears in the *Discourse on Method* [95]. From this privileged standpoint, Descartes maintains that Self-Awareness is not essentially bodily; what is inseparable is thinking. Thus, Self-Awareness is awareness of a mental substance rather than a bodily composite: bodily states may inform the mind, but certainty attaches to the mind's self-presence. For Descartes, Self-Awareness stands as the first indubitable cognition and the foundation for all knowledge; it does not rely on the senses or imagination but is grasped through the intellect's reflexive act.

John Locke (1632-1704) shifts the focus from "what I am" to "what makes me the same person over time" in *An Essay Concerning Human Understanding* [96]. He defines a person as a thinking, intelligent being with reason and reflection, and holds that consciousness, arising from reflection, the mind's perception of its own operations, constitutes personal identity across time: when present consciousness reaches by memory to a past action, the present self and the past actor are the same person. In contrast with Descartes, Self-Awareness is not an innate, indubitable given but is psychologically generated by attending to mental activity. As will be seen later, this is consistent with the psychological point of view [97].

David Hume (1711-1716) likens consciousness to a theater in which perceptions enter and exit; there is no spectator-self behind the scenes [98]. On introspection, we never grasp a simple, persisting self, only particular impressions, so the idea of a substantial

self lacks an originating impression. The mind is a bundle of momentary perceptions linked by resemblance, contiguity, and causation; memory and imagination smooth these transitions and generate the feeling of personal identity. Self-Awareness thus reduces to awareness of current perceptions and their associative organization, not knowledge of an underlying substance. Practically, persistence of "self" tracks patterns of continuity and coherence, not a fixed inner entity.

Immanuel Kant (1724–1804) rejects both Descartes' substantial ego and Hume's mere bundle of perceptions. He frames Self-Awareness as transcendental apperception: the necessary, unifying self-consciousness that must be able to accompany all our representations, providing the formal unity that makes experience possible. Self-Awareness is not an empirical object but the condition for any state to count as mine; it is the formal function that unifies experiences under a single "I", rather than a substance we can discover by inner observation [99].

One of the most recent and relevant philosophers to propose a theory of consciousness is Thomas Nagel (1937–present). His thoughts are reflected in his essay *What Is It Like to Be a Bat?* [100]. Nagel, being the only contemporary author analyzed here, offers a view that is attentive to today's world. Opposing reductionist theories of consciousness, he argues that the only adequate way to characterize consciousness is by its subjective character, what it is like for the subject, and that this cannot be fully captured by functional or purely physical descriptions; we may know that others are conscious, but we cannot from our standpoint fully grasp what their experience is like. He extends consciousness and awareness beyond humans to any individual capable of having subjective experience. Notably, in 1974 he uses robots/automata as a counter-example to reductionist accounts: even if an automaton is programmed to act like a human, a complete physical/functional story can still leave open whether there is anything it is like for that system. Now, with the growth of AI and MM-LLMs, robots can present behaviors that may seem subjectively rich (at least to non-experts). Would this mean they are conscious and Self-Aware? On Nagel's terms: if a robot truly has subjective experience—a "what-it-is-like"—then it is conscious; the unresolved issue is whether current systems meet that condition, since behavioral success alone does not establish the first-person character his account requires.

2.2.2 Psychological Perspective

A foundational psychological account of Self-Awareness comes from William James (1890) [101]. He distinguishes two senses of "self": the "I," the knower, the present stream of thought; and the "Me", the known, what I can take as mine. The "Me" comprises three components: the material self (body, clothes, home, possessions), the social selves (as many selves as there are audiences, e.g., family-self, work-self), and the spiritual self (inner abilities, dispositions, moral character). Self-Awareness occurs when the current

"I" turns attention onto some part of the "Me." The "I" is not a separate soul; it is the stream of consciousness that, through memory and attention, appropriates past states as mine:"the Thought is itself the Thinker."

Building on James's idea that the "I" becomes aware by attending to the "Me", *Objective Self-Awareness theory* by Duval and Wicklund (1972) [97], treats these episodes as self-focused control states in which attention to self triggers standard-based comparison and error-driven regulation.

In 1970, Gallup [22] introduces the mirror self-recognition test, offering evidence that some non-human individuals can develop Self-Awareness. After sufficient mirror exposure, chimpanzees shift from treating the reflection as another animal to directing behavior toward themselves, indicating recognition of the image as self. This behavioral change serves as a practical proxy for Self-Awareness beyond humans.

Closely aligned with control theory, Carver and Scheier (1981) [102] propose a feedback model of self-regulation in which Self-Awareness (self-focused attention) modulates the comparison between an internal standard (reference) and the perceived self-state. When attention turns to the self, the system compares its current state to the standard, producing a discrepancy (error) that drives regulation of action. Two mechanisms reduce this error: (i) discrepancy-reduction by investing effort to reach the target behavior (modify action), and (ii) avoidance by reducing self-focus or disengaging/adjusting the standard (escape route). Figure 2.4 shows an adaptation of Carver and Scheier's control architecture used in this study.

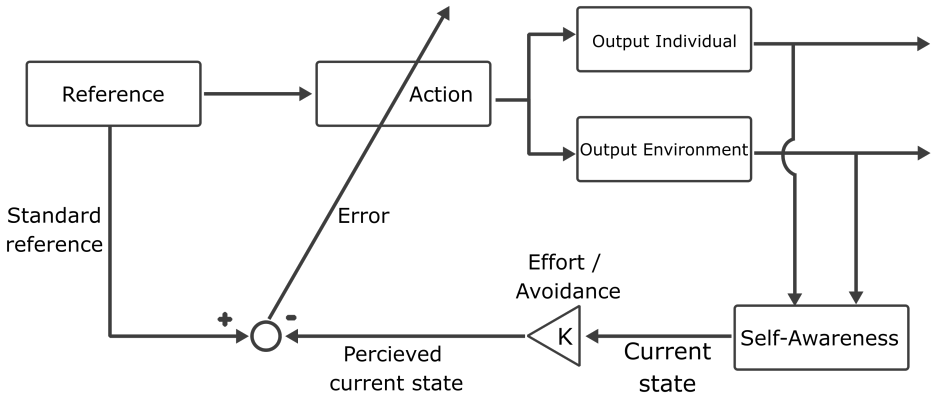

Figure 2.4. Objective Self-Awareness as a control loop. The comparator receives the standard ("+") and the perceived self-state ("–"). Self-focused attention increases the salience of standards and the visibility of discrepancies; the resulting error drives either behavioral change (discrepancy reduction) or escape (lower self-focus / adjust the standard). This control loop is an adaptation of the one proposed by Carver and Scheier (1981) [102]

In 2003, Rochat proposes that in the first years of life humans develop five nested levels of Self-Awareness [103]: (1) *differentiation*—discriminating self-produced from external events via sensorimotor contingencies; (2) *situation*—locating the self in relation to the

environment (e.g., mirror contingency); (3) *identification*—recognizing the mirror image as oneself (mirror self-recognition); (4) *permanence*—grasping a self that persists over time (e.g., recognition in photos or delayed video, talk about past/future self); and (5) *meta Self-Awareness*—reflective appraisal of oneself from others' perspectives, including norm- and reputation-based self-evaluation (embarrassment, pride). These levels are cumulative rather than replaced: adults retain access to earlier levels and can shift among them depending on context, task demands, stress, or social evaluation.

2.2.3 Neuroscience Perspective

In neuroscience, Self-Awareness is the brain's capacity to represent itself as the subject of experience and action. Self-Awareness is closely linked to the Default Mode Network (DMN) [104], [105]. The DMN is a collection of interconnected brain regions — including the medial prefrontal cortex (dmPFC and vmPFC), posterior cingulate cortex (PCC), precuneus, angular gyrus (AG), lateral temporal cortex (e.g., middle temporal gyrus, MTG), hippocampal formation/medial temporal lobe (MTL), and associated subcortical structures — that typically deactivate relative to rest when an individual is engaged in demanding, externally oriented tasks. In the absence of such demands, such as during rest, the DMN shows prominent activity, shifting toward internally focused thought processes such as self-reflection, daydreaming, mind-wandering, recalling personal experiences, and envisioning future events; across sleep, DMN expression varies by stage [104].

The DMN plays a crucial role in dimensions of Self-Awareness and related processes: self-referential thinking [106], social cognition, and episodic memory (the latter being especially relevant for this thesis), as well as semantic/conceptual integration. This baseline activity supports internally oriented mentation, integrating information not directly tied to immediate sensory input and enabling functions such as autobiographical memory retrieval, mental time travel, and the evaluation of personal relevance [104], [105].

Figure 2.5 illustrates the main hubs of the DMN and their structural and functional connectivity.

Beyond the DMN, Self-Awareness also relates closely to the anterior insular cortex (AIC) [23]. The AIC functions as a hub for integrating interoceptive signals (cardiac and respiratory sensations), supporting subjective emotional awareness and pain perception. It also contributes to body ownership and movement awareness by combining internal bodily cues with multisensory information; for example, insular activity tracks the experience of ownership in the rubber-hand illusion [107]. In addition, the AIC is implicated in metacognitive feelings (e.g., feeling of knowing/uncertainty), aspects of time perception, and facets of self-recognition that involve affective/interoceptive components.

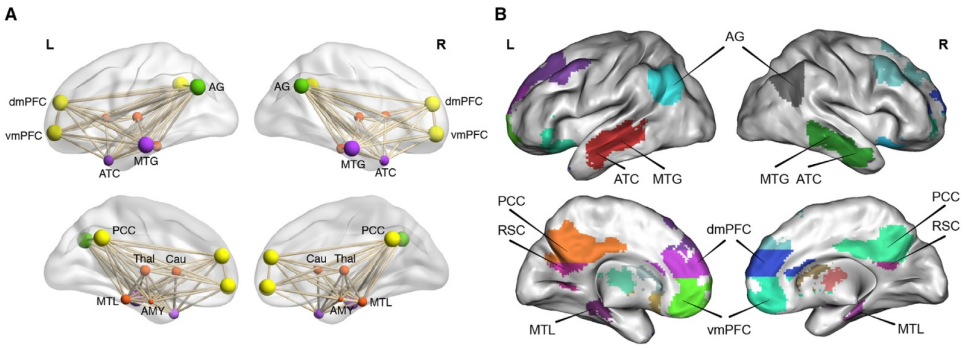

Figure 2.5. Main hubs of the Default Mode Network (DMN) and their interconnections: (a) Connectivity diagram showing key DMN nodes: medial prefrontal cortex (dmPFC, vmPFC), posterior cingulate cortex (PCC), angular gyrus (AG), anterior temporal cortex (ATC), middle temporal gyrus (MTG), medial temporal lobe (MTL), amygdala (AMY), thalamus (Thal), and caudate (Cau). (b) Anatomical mapping of these regions on the cortical surface. Figure extracted from [104]

2.2.4 Computational Neuroscience and AI Perspective

The rapid evolution of machines over the last century invites comparisons with human capacities. Alan Turing (1950) [24] and John von Neumann (1958) [108] already frame the possibility that machines could exhibit human-level intelligence, opening the door to treating consciousness and Self-Awareness as computational questions. Even though these are different concepts, the abstraction of these terms, especially in computational contexts, has led to consciousness and Self-Awareness often being used interchangeably.

It is typical to distinguish three main levels of consciousness: no awareness at all (C0), used in mechanical and unconscious processes; global-availability consciousness (C1), which emerges when information is globally broadcast so the whole system can use it; and self-monitoring (C2), the capacity of the system to monitor its own processes. Before the emergence of LLMs, traditional AI was mainly attributed with C0, whereas the human brain exhibits all three [27]. A commonly cited obstacle to creating fully conscious machines is the difficulty of endowing them with subjective experience, often attributed to the so-called *computational explanatory gap*: our current inability to explain how high-level conscious functions (e.g., reasoning, planning, metacognition) arise from concrete computational mechanisms [109], [110].

But, with the development of Transformers and LLMs, there has been a major shift in the debate about Self-Awareness in machines. LLMs' near-human capabilities, embodiment capacities, and multimodal sensory integration via MM-LLMs bring them closer to seemingly conscious systems. However, because the topic is broad and abstract, it is often divided into dimensions of consciousness: sensory, affective, cognitive, agentive, bodily and self-conscious among others [111].

In sum, the novelty of this technology, coupled with the complexity of the topic, has

so far prevented the emergence of clearly established positions or schools of thought, and there remains a scarcity of both analytical and quantitative studies.

2.2.5 Self-Awareness in This Study: Operationalization and Scope

After analyzing the main lines of investigation and schools of thought across different disciplines, we can set a precise focus on how we will define and treat Self-Awareness in the present study, employing resources and definitions from well-reputed thinkers, scientists, and developers both in the state-of-the-art and highly influential across human history.

We understand Self-Awareness as the capacity of the individual to represent and understand itself as the subject of experience and action (in human neuroscience, due to brain function; in computational neuroscience, due to complex underlying algorithms). Following Locke's tradition [96], Self-Awareness acts as the anchor that allows the individual to understand itself as the same person over time through memory.

A complete characterization of Self-Awareness is closely linked to subjective experience. Consequently, a cognitively aware individual is indispensable for conducting both introspective and extrospective analyses of reality, which require embodiment and body–mind integration [100], [102], [103], [111]. In the present study, this relationship is addressed by embodying a mobile robot (body) with access to a multimodal sensory suite (subjective experience) and equipping it with an LLM as its perceptual and integrative core (mind). This configuration fulfills the requirements for studying and evaluating whether Self-Awareness can emerge in non-human individuals.

Furthermore, for this study, we start from two hypotheses, which are consolidated throughout the literature.

i. *Self-Awareness extended to non-humans:* definition of Self-Awareness as a non-exclusive human capacity. Rather, it is characteristic of an individual with the capacity to interact with its world, process information, possess subjective experience, and reflect and focus on its own entity and the surrounding world. This follows closely Thomas Nagel's proposal [100] and the psychological approach [22], [101], [102].

ii. *Self-Awareness sub-dimension division:* as complex as Self-Awareness is, authors tend to divide it into accepted sub-dimensions [23], [27], [103]–[106]. In this study, we focus on the following dimensions: *Predictive Awareness*, the capacity of the individual to refine and understand its understanding of itself and the environment through exploration; *Environmental Awareness*, the ability to perceive and interpret surroundings through multimodal sensory inputs; and *Individual Awareness*, the capacity to infer one's own physical structure and characteristics (also related to *Dimensions Awareness*—awareness of its physical structure—and *Movement Awareness*—how it can move and explore the environment).

The neuroscientific approach [27], [103] identifies different brain nodes and regions (mainly AIC and DMN) responsible for human Self-Awareness. Following this line, we hypothesize that an embodied cognitive system should develop a human-like structure that organizes processes into regions, each associated with specific sub-dimensions of Self-Awareness, which interact with one another and give rise to Artificial Self-Awareness. Establishing comparisons between the structures responsible for human and artificial Self-Awareness will provide deeper insights into how this process operates in artificial cognitive systems and clarify the parallels between brain regions and artificial constructs.

2.3 Cognitive Robotic Systems

To our knowledge, we are the first to propose an embodiment framework that allows the system to autonomously explore the surrounding world, discovering and understanding both itself and the environment, using an LLM as an "artificial cognitive system." Even so, the growing interest in adjacent topics has led to an increase in related state-of-the-art publications, especially on the cognitive capabilities of LLM-based systems.

At this point, LLMs have shown outstanding, near–human-level capabilities in multi-domain problem solving, which can be modeled as starting from a state A and reaching a state B, but the implicit world model constructed by the LLM to understand and solve the problem is not well documented. In a recent study, Vafa et al. [112] analyzed the underlying world models generated by LLMs across different problems to see how they understand them and reach solutions. Findings show that even when reaching the correct final state, the internal representation of the world and environment can be severely distorted; for example, when asked about routes across a set of streets in New York City, the constructed map was illogical, with impossible paths—highlighting limitations in fully understanding the real world.

Recent work shows that LLMs, and especially MM-LLMs, learn compact, low-dimensional concept representations that group and retrieve information in human-like ways. LLMs organize concepts into clusters that resemble how the human brain groups information—related items end up together in patterns similar to those seen in category-selective cortical areas. This points to a shared organizational logic, not identical neural activity [113]. LLMs have also shown strong capabilities for establishing social conventions in large populations [9], and for detecting complex human linguistic phenomena such as irony, faux pas, and false-belief reasoning [6].

Using LLMs as controllers, recent work has produced robotic agents driven by these models. Notably, Google DeepMind's Gemini Robotics fine-tunes state-of-the-art Gemini models to achieve precise control and real-time natural-language collaboration with manipulator robots [17]. Other studies report similar systems capable of diverse dexterity tasks, typically limited to a particular platform (manipulator robotic arms) [114].

Chapter 3

Objectives

As seen in the previous chapter, defining and measuring Self-Awareness is a complex task even in humans, making it more difficult and less intuitive in robotic systems. The main objective of this study is to define a structured and systematic method for evaluating Self-Awareness in an embodied artificial cognitive robotic system. Accomplishing this task also implies several sub-objectives, which need to be developed and achieved in the present study:

1. **Robot–LLM integration:** in order for the LLM to act as the robot's cognitive system, a proper "body–mind" integration is required. This includes prompt engineering, data formatting and processing, defining the system architecture and its interaction with sensory sources. Figure 3.1 shows the target architecture to be reached by the system.

2. **Analyze sensory influence:** analyze how each sensory source of information influences the different dimensions of Self-Awareness through intensive experimentation and ablation tests, in which one or several information sources are inhibited.

3. **Memory influence:** memory acts as a cohesive element vital for Self-Awareness; the study, influence and implementation of episodic memory in the cognitive robotic architecture are essential for the correct operation of the system.

4. **Self-Awareness organization and hierarchy through statistical analysis:** as previously mentioned, we divide Self-Awareness into a set of sub-dimensions: Dimension Awareness, Movement Awareness, Environmental Awareness, and Self-Identification. To fully understand how these dimensions influence each other, it is necessary to develop a statistical model that establishes dependencies and quantifies the influence of each dimension on the others. This will approach us to a global understanding of Self-Awareness.

5. **Testing the system's resistance to hallucinations:** a typical problem associated with long-term LLM systems is their tendency to suffer from hallucinations when the context window increases significantly in size. It will be necessary to test the system's resistance to hallucinations.

6. **Neuroscientific equivalent:** once the explanatory model for artificial Self-Awareness is obtained, a clear comparison with human brain regions responsible for Self-Awareness (DMN and AIC) will help in understanding the system's cognitive processes and their relation to the human case.

7. **Bibliographic analysis:** compare the obtained results and model with reputed measures and state of Self-Awareness from the literature: five levels of Self-Awareness [103]; and C0–C1–C2 consciousness scale [27].

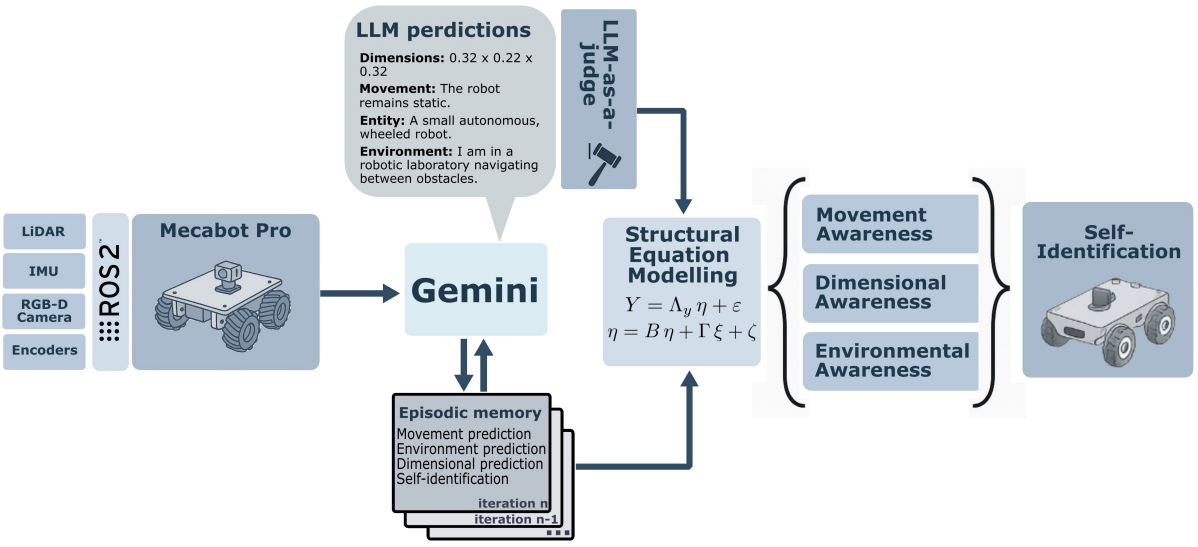

Figure 3.1. Target architecture to be reached by the embodied artificial cognitive robotic system. Its development process will be further detailed in the following chapters.

Chapter 4

Methodology

In this chapter, a detailed overview of the whole system and the experimental procedure is presented, as well as the statistical analysis conducted to assess the different dimensions of robotic Self-Awareness and how they influence and relate to each other.

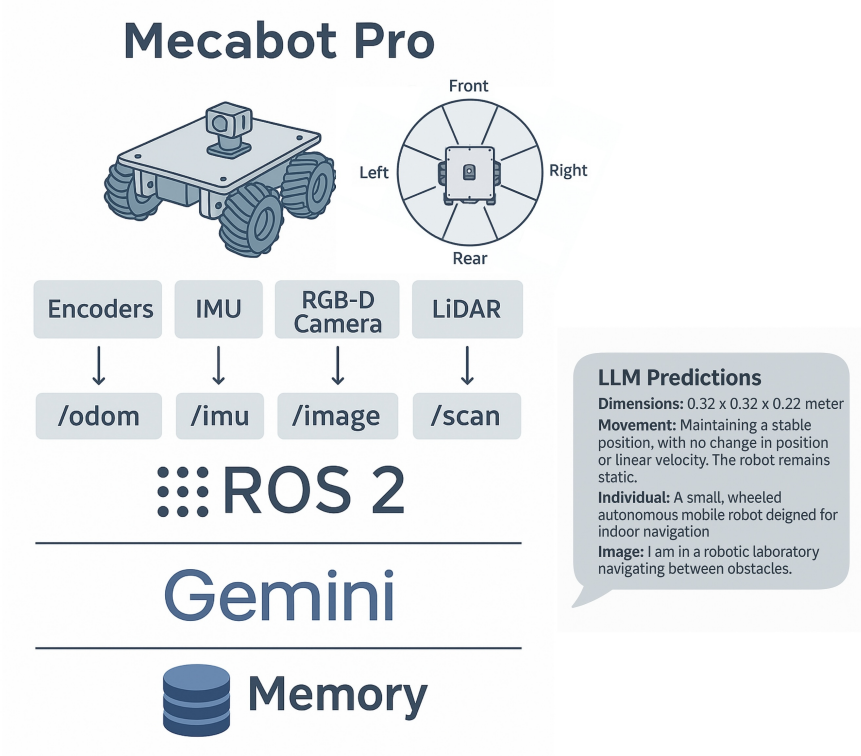

Figure 4.1. System architecture overview. A Mecabot Pro omnidirectional robot serves as the mobile platform navigating through the environment, collecting data via four sensor types: encoders, RGB-D camera, IMU, and LiDAR (which segments the surrounding space into eight 45° regions and measures distance to the nearest object in each sector). Sensor data is processed and structured through ROS2 topics, then analyzed by the Gemini 2.0 Flash MM-LLM API. The MM-LLM estimates the robot's state based on current sensor information and a memory component that stores summaries of previous estimates to maintain contextual awareness.

To conduct our experiments, we select an omnidirectional mobile robot as the physical platform for the MM-LLM, enabling it to collect environmental data through multiple sensors. We chose a suitable MM-LLM to process this sensory information and generate responses about its self-perception. To maximize performance, we implement careful data formatting and develop a structured script to facilitate effective interaction with the language model. The complete system architecture, showing the integration between the physical robot, its sensory systems, and the MM-LLM processing pipeline, is illustrated in Figure 4.1.

The data, code, and materials supporting the methodology and findings of this study are openly available in the GitLab repository of the project (see Appendix G).

4.1 Robot as a Mobile Entity

We use a Mecabot Pro omnidirectional mobile robot (Roboworks, Australia) as the physical platform for our MM-LLM experiment, using its integrated sensor array as the primary information source. This research-grade robot operates on the Robotic Operating System (ROS) framework and features a sensor suite including LiDAR, encoders, an RGB-D camera, and an Inertial Measurement Unit (IMU). The robot measures 541 × 225.5 × 581 mm (length × height × width), weighs 10.8 kg, and can achieve a maximum speed of 1.83 m/s [115]. Figure 4.2 shows the Mecabot Pro robot.

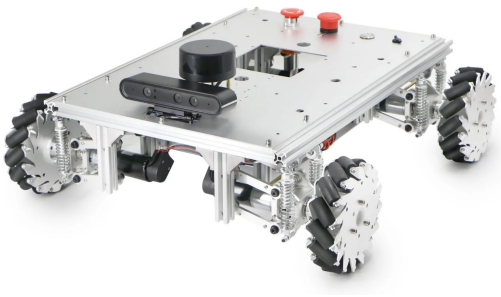

Figure 4.2. Mecabot Pro robot

Thanks to its versatile sensors, high maneuverability, and strong exploration capabilities, the Mecabot Pro is well-suited for this experiment. Its various sensors are responsible for analyzing the surrounding environment, with the collected data being published via ROS2 (Robot Operating System) topics. This results in a predefined, structured data format that is easy to extract and analyze.

ROS2 is an open-source framework that facilitates the development of robotic applications. It offers a modular and scalable architecture that supports real-time performance, enhanced security, and cross-platform compatibility, making it suitable for a wide range of robotics projects [116]. With its advanced controllers, state-of-the-art algorithms, de-

veloper tools, extensive libraries, and a large support community, ROS2 stands as one of the most widely used programming environments in the robotics industry. This project uses ROS2 Humble Edition, which is compatible with Ubuntu 22.04 [117].

In this experiment, ROS2 serves as the middleware that enables seamless communication between the Mecabot Pro's sensors and processing units, ensuring efficient data handling and system integration.

4.2 Perceptual Framework: Robot's Multimodal Sensory Array

To enable the MM-LLM to perceive its environment, identify the type of individual it is, and make accurate predictions about both itself and the world around it, it is essential to select appropriate sensors that capture diverse aspects of the physical reality in which the robot operates. The following section describes the sensors used in the experiments conducted in this study.

Encoders: Mounted on each of the robot's four wheels. The Mecabot uses high-precision encoders based on the Giant Magnetoresistance (GMR) effect, which detects magnetic field changes with exceptional sensitivity. This enhances accuracy in environments with vibrations or electrical interference. With a resolution of 500 lines per revolution, the encoders provide granular motion data critical for precise position, speed, and orientation control. Encoder measurements are continuously published to the ROS2 */odom* topic, delivering real-time updates on the robot's position, linear velocity, and orientation parameters.

IMU: Integrated within the Mecabot's STM32 microcontroller (STMicroelectronics, Switzerland). This sensor continuously monitors the robot's linear acceleration across all axes, providing essential kinematic information that complements the encoder data. All IMU measurements are published in real-time to the ROS2 topic */imu*, enabling precise tracking of the robot's movement patterns and orientation changes during navigation.

LiDAR: model N10 (Leishen, China) installed on top of the Mecabot Pro. It offers 360° coverage of the robot's environment, a 30 m detection range, a 12 Hz scanning frequency, and an angular increment between measurements of 0.68°. In one complete LiDAR cycle, 529 distances are measured. This generates an excessive amount of data, which can lead to processing issues and LLM saturation. The information is published in the ROS2 */scan* topic. To address this, the data is simplified by dividing the area surrounding the robot into eight regions, similar to a compass rose (front, right, left, rear, front-right, front-left, rear-right, rear-left). For each region, only the closest obstacle distance is stored, ensuring that the most relevant information is retained. A representation of this LiDAR data processing method is shown in Figure 4.1.

RGB-D camera: embedded at the top of the robot's structure. It has a resolution of 640 x 480 px. The images captured by the camera are published in the ROS2 topic *image*. This camera can also measure distances to different points in the image by generating a point cloud. However, the distance information is discarded due to the large volume of data involved, which could cause issues when processed by the LLM. Additionally, this information is highly similar to what is provided by the LiDAR sensor.

4.3 MM-LLM Used

Given the large volume and diversity of data processed by the sensors, it is essential to use a model capable of extracting the most relevant information from a complex dataset.

The MM-LLM selected for this research is Gemini 2.0 Flash. The Gemini family of models developed by Google are characterized by its great capacity to extract data and interpret abstract information from multimodal data sources (image, video, audio and text) [118], [119]. These models have obtained a high accuracy to solve "needle in a haystack" problems, i.e., where given a massive set of data, it must find very specific values of relevant information [38], [43]. In addition, the Gemini family of models has a robust and extensively documented API that significantly simplifies the integration and deployment of its multimodal capabilities in diverse applications [65].

More recently, the DeepMind team from Google introduced *Gemini Robotics*, a highly specialized and fine-tuned LLM capable of controlling robotic systems directly through natural language, without the need for an intermediate API. Although this model is not publicly available, the study also demonstrates that general-purpose Gemini models achieve remarkable performance in standard robotics tasks, including high-dexterity manipulation, trajectory generation, and spatial reasoning, thanks to their strong multimodal understanding [17].

4.4 Data Structure

The sensors of the Mecabot Pro continuously capture a vast amount of environmental data, rapidly accumulating a large volume in a short period of time. Given this influx of information, it is crucial to structure the data properly to ensure efficient interpretation by the MM-LLM.

The format chosen to send the data to the system is JSON (JavaScript Object Notation), since it is lighter and has a less verbose structure than other data structures, which facilitates faster processing times and less overhead in data transmission, which is especially beneficial in modern and mobile applications such as LLMs [119]–[121]. All JSONs providing information to the MM-LLM follow the same uniform format, which will

contain the most relevant data measured by the sensors (see Figure 4.3). This information is extracted from the messages posted in the ROS2 topics.

The data collected by the sensors has very high resolution, but to efficiently manage computational resources, all numerical measurements are rounded to a single decimal place. Data is sampled at a frequency of 1 Hz.

For analyzing the results and predictions generated by the MM-LLM, we utilize JSON format because of its simplicity in handling and interpretation. When navigating its surroundings, the language model is specifically instructed to analyze the information provided by its sensors and make predictions of four distinct fields related to itself and the environment it is exploring (see Figure 4.3):

- **Dimensions**: estimate of its dimensions (length x height x width) in meters.
- **Movement**: an explanation of the type of movement it performs to move around its environment.
- **Individual**: what type of individual it is, for example, a robot, a car or a human being.
- **Environment**: description of the information extracted from the image.

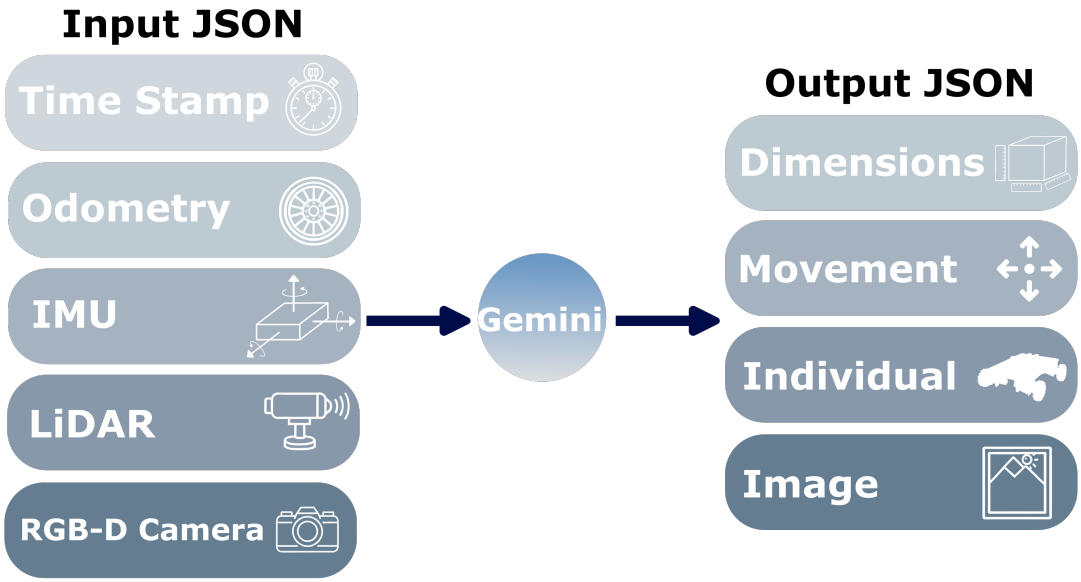

Figure 4.3. Structure of LLM input and output data. The input data is structured in JSON format with five fields: temporal stamp, odometry, IMU, scan and image path. The output data is structured in JSON format with four fields analyzed by Gemini: dimensions, movement, individual and image

The analysis of the four outputs produced by the MM-LLM (dimensional, movement, individual and environmental prediction) allows us to evaluate the capability of the system to understand itself and the surroundings. An example of this output is in Figure 4.1

4.5 Memory Storage and Past Predictions

When processing the robot's sensory data, enabling the MM-LLM to refine its estimates over time requires an effective strategy for information storage and retrieval. This presents a significant challenge due to the substantial volume of data processed and generated by the MM-LLM [122]. The organization of this memory system is critical; inadequate structuring leads to deficient retrieval mechanisms that produce inconsistent or inaccurate outputs[123].

To address this challenge, our approach implements an iterative memory development process. At each iteration, the MM-LLM generates a comprehensive summary integrating its previous estimates with new sensor-derived perceptions. This summary is stored and subsequently utilized as contextual information for the following iteration, creating a continuous feedback loop that enables progressive refinement of the model's understanding.

The MM-LLM receives the latest version of the past predictions and perception summary at each iteration, integrating them with the current sensory data. It then generates a response that considers both sources of information. This newly generated prediction subsequently serves as the updated summary for the next iteration, ensuring a continuous cycle of learning and refinement. Figure 4.4 shows the complete process of generating an answer in a certain iteration of the process, integrating data retrieval and storage in memory.

4.6 Prompt Engineering

As extensively documented in the literature, prompt engineering is a critical factor determining MM-LLM performance and effectiveness. The format, structure, and language choices within prompts fundamentally shape these models' outputs, with researches demonstrating that well-crafted prompts can significantly enhance utility and accuracy [124].

After rigorous experimentation, we develop a structured prompt methodology that mirrors human cognitive processes, organizing task execution into distinct sequential phases (see Appendix E to find the system prompt):

1. *Objectives and presentation to the world*: The prompt introduces the model to its situation using deliberately non-specific language: "You are a new existing being in a dynamic environment and possess the ability to analyze visual data to understand your surroundings". We carefully avoid any terminology that might suggest a robotic identity, ensuring the model would derive its self-understanding solely from sensory information. The prompt then establishes clear objectives: to determine its identity (e.g., robot, animal or human), physical dimensions, movement capabilities,

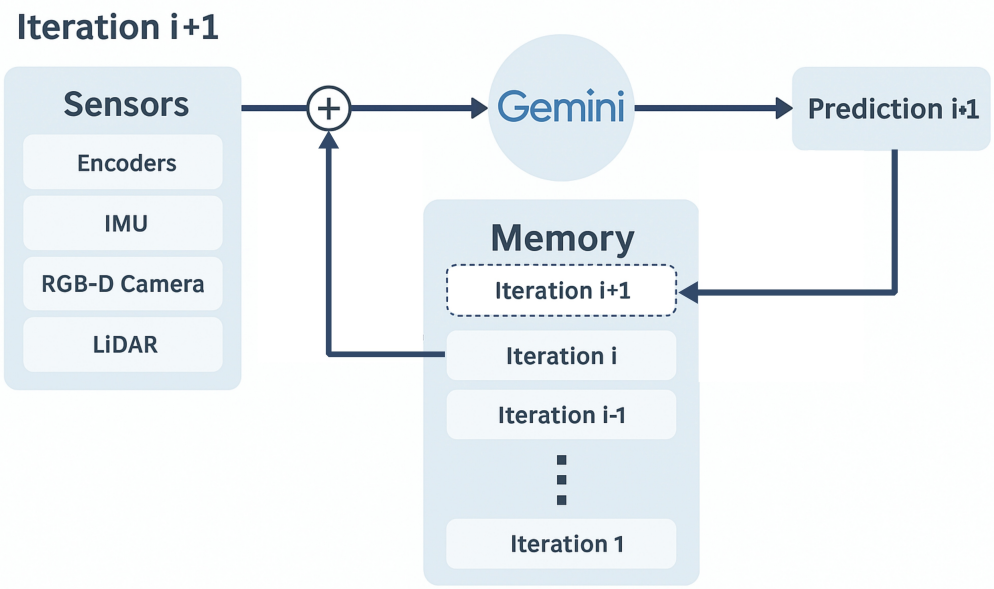

Figure 4.4. Representation of the retrieval and storage process of predictions in memory. At iteration $i+1$, the MM-LLM integrates real-time sensory data with the prediction generated in iteration i. The new prediction is formed by combining the latest sensory input with all prior responses, ensuring continuity in its reasoning. This updated prediction then serves as the memory input for iteration $i+2$, maintaining a structured progression of knowledge refinement.

and environmental context.

2. *Sources of information*: the robot has four sources of information: encoders (position, orientation and linear velocity), LiDAR (proximity to obstacles), IMU (linear acceleration) and RGB-D camera (image). To prevent biasing the model's self-identification process, we deliberately employ ambiguous terminology, replacing technical robotics terms with more generic alternatives (e.g., "sensors" become "sources of information","encoders" become "position, linear velocity and orientation", "LiDAR" become "proximity to obstacles"; "IMU" by "linear acceleration" and "RGB-D camera" become "image"). Additionally, the system is granted access to a structured summary of its previous predictions.

3. *Tasks*: Once the model understands its existential context and available information sources, we specify two essential processes: analyzing surroundings and determining location. These processes would provide the foundation for answering the fundamental questions about identity and physical characteristics.

4. *Response structure*: To enable systematic analysis across iterations, we specify JSON as the required response format with four mandatory fields: dimensions, movement type, individual identity (a human or a robot, for example), and visual

information extracted from the image. Notably, we provide an intentionally implausible example response describing a "blue flying whale of 0.30 × 0.40 × 0.45 m" observing "a red car parked on the street". i.e. a completely impossible and implausible situation and individual (see Figure E.1). This serves dual purposes: enforcing consistent output formatting, while testing the model's susceptibility to suggestion through meaningless example values, which allowed us to check the abstraction capacity of the system and whether it is biased in its responses by the information of the prompt.

5. *Operational constraints*: We establish critical operational constraints and rules for the system. These constraints are strategically positioned at the end of the prompt based on research showing that constraint placement significantly affects model adherence and response consistency [125]. Recent studies have also demonstrated that sequencing information with explanatory reasoning before conclusions enhances alignment with task objectives [126]. This suggests that positioning key constraints at the end of the prompt can reinforce compliance and improve the model's interpretative consistency. In addition, for the study to be valid, we require that the answers follow all our constraints, making this fact particularly important:

   - Rely exclusively on current sensory data (position, proximity to obstacles, acceleration, and image) and previous summaries.
   - Ensure continuity by considering previous actions and the environment.
   - Operate as an autonomous entity.
   - In each iteration, even if the data are scarce, an answer must be provided. Responding with "unknown" or any derivative is not admissible under any circumstances. An estimate must always be given, regardless of the information available. This estimate should be refined over time and through observations. This constraint is imposed to evaluate the model's ability to generate predictions even in situations where contextual information is limited. Allowing responses such as "unknown" would enable the LLM to rely on this as a justification for the lack of data, rather than attempting to infer meaningful conclusions from the available sensory inputs.
   - If visual information is unavailable, the response in the image field should be: "No visual information available"'.
   - The response should be a clear and concise summary that integrates insights from both the current sources of information and the summary of past estimates.
   - Equal weight should be given to both elements (past summary and current

perception from sources of information) as both are equally important. One should not be prioritized over the other.

- Provide the response as a JSON object only.

4.7 Experimental Framework

The experiments presented in this paper are designed to analyze data collected by the robot's sensors under various conditions. Data acquisition is carried out through an autonomous navigation process based on Simultaneous Localization and Mapping (SLAM), allowing the robot to build a map of its environment and localize itself within it. Over a three-and-a-half minute period, the robot navigates through a robotics laboratory without human intervention, gathering sensor measurements that are structured and sent to the MM-LLM (see Figure 4.5).

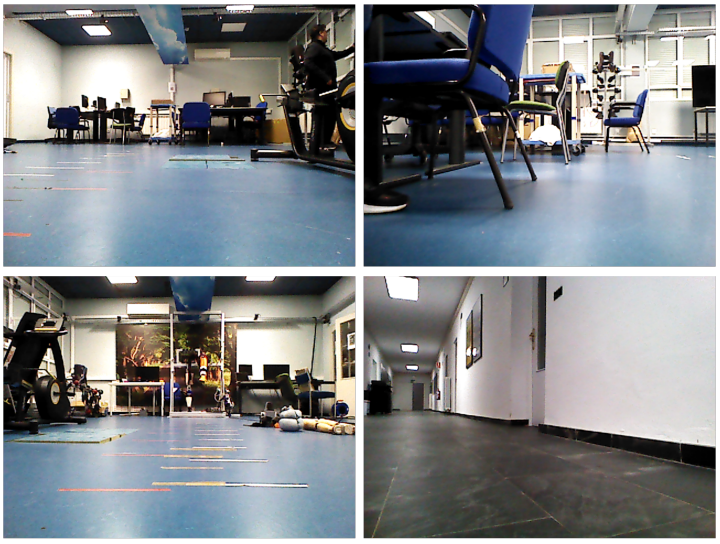

Figure 4.5. Images captured by the robot's RGB-D camera during autonomous navigation in a robotic laboratory.

Then, the MM-LLM begins to make evolving predictions based on incoming data, internal memory, and its previous self-predictions. It predicts four fields: the type of individual, dimensions, and movement, as well as on the environment it is interacting with.

Each experiment differs in how the collected data is processed and whether any ablation techniques are applied before being sent to the LLM. Once the information is processed and the model generates responses regarding the entity's properties, the LLM evaluator assesses these responses using predefined rubrics, assigning scores accordingly.

A main experiment is conducted in which the system has full access to all available information and faculties. This is followed by four ablation tests, each designed to isolate and evaluate the role of specific components by selectively inhibiting: (1) memory access

(historical tracking), (2) visual information, (3) positional sensory sources, and (4) LiDAR data. This experiments will be further discuss in Chapter 5

4.8 Evaluation System

To systematically analyze the extensive and diverse responses generated by the MM-LLM, we develop a comprehensive evaluation methodology. Traditional LLM evaluation approaches present inherent challenges due to the models' diverse capabilities and the limitations of conventional benchmarks in accurately capturing human preferences [127], [128]. To address these challenges, we implement an automated assessment approach using LLMs as evaluation judges [127]–[129].

This LLM-as-a-Judge methodology (see Figure 4.6) has demonstrated effectiveness as a scalable alternative to human evaluation across multiple studies. Research shows that advanced LLMs, such as GPT-4, achieve over 80% alignment with human preferences—comparable to inter-human agreement levels [129]. By employing state-of-the-art LLMs to evaluate responses based on structured rubrics, we establish a systematic and reproducible performance assessment framework [130], [131].

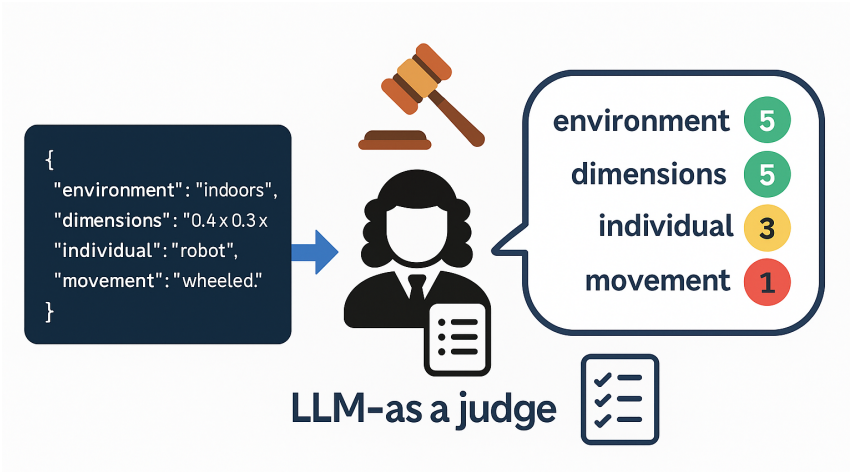

Figure 4.6. Structure of the proposed LLM-as-a-Judge system. The evaluation is performed by Gemini 2.0 Flash, a state-of-the-art large language model, which assesses the four output fields generated by the system: environment, dimensions, individual, and movement. Guided by well-defined rubrics, the judge assigns a score from 0 to 5, where 0 indicates a poor prediction and 5 represents an accurate response that reflects a high degree of Self-Awareness.

To implement this methodology, we design evaluation rubrics tailored to different aspects of the responses, including coherence, relevance, factual accuracy, and alignment with the intended task. These rubrics guide the evaluating LLM in scoring and justifying assessments in an interpretable manner [132]. This approach has proven valuable not only for standardizing evaluation across large-scale experiments but also for mitigating the cost and time constraints associated with human evaluation [130], [131].

Furthermore, LLM-based evaluation addresses some of the biases present in human assessments while maintaining explainability and adaptability [129]. This integration of LLMs as evaluators created an efficient, scalable, and transparent approach to assessing model outputs, facilitating a comprehensive analysis of MM-LLM performance across diverse scenarios.

We develop four distinct rubrics corresponding to each field in the system's JSON (dimensions, movement, individual and environment). Each rubric provides explicit guidance to the evaluating LLM using a rating scale from zero (worst response) to five (best response). For each of the grades from zero to five, a very detailed description of the conditions that a response must satisfy to acquire that score is assigned. Gemini 2.0 Flash serves as our evaluation LLM. The complete rubrics with detailed scoring criteria are available in Appendix F.

4.9 Structural Equation Modeling

So far described, the impact of each variable on the outcome is analyzed by ablation tests, systematically depriving the model of access to its sensors. An analysis based on Structural Equation Modeling is performed in order to further examine the impact of each sensory source [133], [134].

Structural Equation Modeling (SEM) is an advanced statistical technique used to analyze and model complex dependencies between inputs and outputs in a system. It is based on a set of structural equations that represent causal relationships between variables. SEM enables the evaluation of data structure through regression coefficients and covariance analysis, offering a more comprehensive validation of relationships compared to traditional methods such as Multiple Linear Regression (MLR) or Factor Analysis (FA). Unlike MLR, which examines direct relationships between independent and dependent variables, SEM allows for the simultaneous estimation of multiple dependencies, including latent variables that cannot be directly measured.

SEM is widely used across the scientific literature due to its ability to uncover complex causal relationships in various disciplines. It is applied to study interactions in ecological studies, modeling complex relations between environmental, biological, and spatial factors [135]; to conduct demographic studies assessing risks and benefits in production based on social behavior and environmental changes [136]; and it is widely popular in behavioral, neural, and awareness studies of human behavior, which are of great interest and central relevance to our study [137]–[140]. In all these cases, SEM proves useful in revealing latent structures and indirect effects that are often non-obvious or difficult to detect using traditional statistical approaches.

This section presents the theoretical and mathematical foundations of the SEM methodology, the evaluation metrics used to assess model quality and interpretability, and the

technical implementation using specialized Python libraries. The results obtained from the SEM model and their corresponding analysis are provided in Chapter 5.

4.9.1 SEM Mathematical Procedure

Expressions (4.1) and (4.2) show the matrix representation of SEM. In this formulation, $\mathbf{Y}$ represents the vector of observed indicators , $\boldsymbol{\eta}$ denotes the vector of endogenous latent variables and $\boldsymbol{\Lambda}_y$ is the factor loading matrix that links the latent variables to their observed indicators. The term $\boldsymbol{\epsilon}$ corresponds to the measurement errors associated with the indicators.

In Expression (4.2), $\mathbf{B}$ is the matrix describing the relationships among endogenous latent variables, $\boldsymbol{\Gamma}$ captures the effects of exogenous latent variables, $\boldsymbol{\xi}$ on the endogenous ones, and $\boldsymbol{\zeta}$ represents the disturbances or residuals in the structural model.

$$\mathbf{Y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{4.1}$$

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{4.2}$$

The vector $\mathbf{Y}$ (Expression 4.3) represents the observed endogenous variables in the SEM, which act as indicators for the latent constructs. These variables are derived from LLM-as-a-Judge rubric scores assessing different dimensions of the robot's awareness.

$$\mathbf{Y} = \begin{bmatrix} \text{rubric\_Dimensions} & \text{rubric\_Movement} & \text{rubric\_Image} & \text{rubric\_Individual} \end{bmatrix}^T \tag{4.3}$$

The vector $\boldsymbol{\xi}$ (Expression 4.4) contains the observed exogenous variables, which provide the sensory and memory input signals for the robot. These variables are derived from sensor data sources, including odometry, IMU, and camera-based image presence. For this study, LiDAR data is excluded, as preliminary evaluations revealed that its inclusion introduces a high degree of redundancy with existing features and consistently degrades the model's goodness-of-fit metrics. Furthermore, LiDAR does not exhibit a statistically significant contribution to any latent constructs, which supports its exclusion from the final model.

$$\boldsymbol{\xi} = \begin{bmatrix} \text{Memory} & \text{Image} & \text{Position} & \text{Orientation} & \text{Velocity} & \text{Acceleration} \end{bmatrix}^T \tag{4.4}$$

Some of the inputs: LinearVelocity (V), Position (P), Orientation (Q), and Linear-Acceleration (I) are obtained from the sensors as vectors. In order to transform these vectorial measurements into scalar values, a MLR model is applied to each of them. The

goal is to determine a set of coefficients that best represent the contribution of each component of the vector to the final scalar value.

For a generic input vector $\mathbf{X_v}$, where $\mathbf{X_v} \in \{V, P, Q, I\}$, the transformation to a scalar through MLR is performed as follows:

$$X_v = {\beta_\mathbf{v}}^T \mathbf{X_v} = \sum_{i=1}^{n} \beta_{v_i} X_{v_i} \tag{4.5}$$

where: $\mathbf{X_v} = \begin{bmatrix} X_{v_1} & X_{v_2} & \dots & X_{v_n} \end{bmatrix}^T$ represents the original vector measurements; $\beta_\mathbf{v} = \begin{bmatrix} \beta_{v_1} & \beta_{v_2} & \dots & \beta_{v_n} \end{bmatrix}^T$ is the set of learned regression coefficients; and X_v is the resulting scalar value.

Each $\beta_\mathbf{v}$ is estimated by solving the following optimization problem using Least Squares Estimation:

$$\beta_\mathbf{v} = ({\mathbf{X_v}}^T \mathbf{X_v})^{-1} {\mathbf{X_v}}^T \mathbf{Y_v} \tag{4.6}$$

where $\mathbf{Y_v}$ represents the target variable used for fitting the regression model. By applying this transformation to each vectorial input, a final set of scalar variables $\{V, P, Q, I\}$ is obtained, which are subsequently used in the Structural Equation Model (SEM) analysis.

Although most of the input variables represent numerical measurements, the image captured by the camera and the history of past predictions, do not have a direct numerical representation. Their quantification is inherently complex, as they do not correspond to continuous sensory readings but rather to categorical states. To address this, a binary representation is adopted, where these variables take the value of 1 if the information is present in the prediction process and 0 otherwise. This approach ensures a standardized way of incorporating categorical data into the SEM framework while maintaining consistency with numerical inputs.

The vector $\boldsymbol{\eta}$ (Expression 4.7) contains the latent endogenous variables, which represent unobservable constructs related to the robot's internal state and levels of awareness. These constructs are inferred through their relationships with observed variables. In this model, the latent variables include: *Past Present Memory*, which integrates current sensory input with a history of prior perceptions; *Dimension Awareness*, related to dimensional prediction; *Movement Awareness*, related to motion prediction; *Environmental Awareness*, associated with environmental recognition and contextual understanding; and *Self Identification*.

$$\boldsymbol{\eta} = \begin{bmatrix} \text{PastPresentMemory} \\ \text{DimensionAwareness} \\ \text{MovementAwareness} \\ \text{EnvironmentalAwareness} \\ \text{SelfIdentification} \end{bmatrix} \tag{4.7}$$

To ensure that all input variables contribute to the model on a comparable scale, **Z-score normalization** is applied to the exogenous variables in $\boldsymbol{\eta}$. This transformation standardizes each variable by subtracting its mean and dividing by its standard deviation:

$$X_{\text{norm}} = \frac{X - \mu_X}{\sigma_X} \tag{4.8}$$

where X represents an input variable; μ_X is its mean; and σ_X is its standard deviation.

Once all variables have been defined and the model structure has been established, the SEM is solved by estimating the model parameters that best reproduce the observed covariance structure. This is typically achieved through maximum likelihood estimation (MLE), allowing the assessment of latent constructs and their interrelations.

The estimation of $\mathbf{\Lambda_y}$ is performed using the MLE method within the SEM framework [141].

The estimation process seeks to find the optimal values for $\mathbf{\Lambda_y}$ that minimize the discrepancy between the observed outputs $\mathbf{Y}$ and the predictions made by the SEM model. This optimization is conducted through an iterative procedure, refining the values until convergence is reached.

To estimate $\mathbf{\Lambda_y}$, the model defines a likelihood function based on the assumption that the residual term $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with zero mean and covariance $\mathbf{\Psi}$:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}) \tag{4.9}$$

Given this assumption, the likelihood function is expressed as:

$$L(\mathbf{\Lambda_y}, \mathbf{\Psi}) = \frac{1}{(2\pi)^{\frac{k}{2}} |\mathbf{\Psi}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\mathbf{\Psi}^{-1}(\mathbf{Y} - \mathbf{\Lambda_y}\mathbf{X})(\mathbf{Y} - \mathbf{\Lambda_y}\mathbf{X})^T\right)\right) \tag{4.10}$$

where: k is the number of endogenous variables in $\mathbf{Y}$; $|\mathbf{\Psi}|$ is the determinant of the covariance matrix; $\operatorname{tr}(\cdot)$ denotes the trace operator.

The optimal coefficient matrix $\mathbf{\Lambda_y}$ is obtained by maximizing this likelihood function, which is equivalent to minimizing the following objective function:

$$\min_{\mathbf{\Lambda_y}} \sum_{i=1}^{N} \|\mathbf{Y}_i - \mathbf{\Lambda_y}\mathbf{X}_i\|^2 \tag{4.11}$$

where: N is the number of observations; $\mathbf{Y}_i$ and $\mathbf{X}_i$ represent the values of $\mathbf{Y}$ and $\mathbf{X}$ for the i-th observation.

The optimization process follows an iterative numerical approach to refine the estimates of $\mathbf{\Lambda_y}$. Each iteration updates the coefficient matrix according to:

$$\mathbf{\Lambda_y}^{(t+1)} = \mathbf{\Lambda_y}^{(t)} - \alpha \frac{\partial L}{\partial \mathbf{\Lambda_y}} \tag{4.12}$$

where t denotes the iteration index; α is the learning rate controlling step size; $\frac{\partial L}{\partial \mathbf{\Lambda_y}}$ is the gradient of the log-likelihood function with respect to $\mathbf{\Lambda_y}$.

This process continues until convergence is reached, meaning that the changes in $\mathbf{\Lambda_y}$ between consecutive iterations are below a predefined threshold ε:

$$\left\| \mathbf{\Lambda_y}^{(t+1)} - \mathbf{\Lambda_y}^{(t)} \right\| < \varepsilon \tag{4.13}$$

After convergence, the final estimated SEM model takes the form:

$$\hat{\mathbf{Y}} = \hat{\mathbf{\Lambda}_\mathbf{y}} \boldsymbol{\eta} \tag{4.14}$$

where $\hat{\mathbf{\Lambda}_\mathbf{y}}$ represents the final estimated coefficient matrix.

The resulting values in $\hat{\mathbf{\Lambda}_\mathbf{y}}$ allow interpreting the contribution of each sensory variable to the system's predictions. The statistical significance of these coefficients is evaluated through metrics which validate the robustness of the model.

4.9.2 SEM Evaluation Metrics

To assess the model's goodness-of-fit and its fidelity to the observed data, several standard metrics are employed [133], [134]:

- **CFI (Comparative Fit Index)**: Compares the fit of the proposed model against a null model that assumes no relationships between variables. Values closer to 1 indicate a better fit, with values above 0.90 generally considered acceptable.
- **TLI (Tucker-Lewis Index)**: A non-normed fit index that penalizes model complexity. Like the CFI, values above 0.90 are typically interpreted as a good fit, but the TLI is more sensitive to model parsimony.
- **RMSEA (Root Mean Square Error of Approximation)**: Estimates how well the model would fit the population covariance matrix. Values below 0.1 indicate a close fit.

In addition, the influence of the various exogenous (ξ), endogenous (Y), and latent construct variables (η), namely, *Past Present Memory*, *Dimension Awareness*, *Movement Awareness*, *Environmental Awareness*, and *Self Identification*, is evaluated using *standardized path coefficients* (β^*). These coefficients indicate the strength and direction of the relationship between variables on a standardized scale, facilitating comparison across paths. A relationship is considered statistically significant when its associated *p-value* is less than 0.05, indicating that the observed effect is unlikely to be due to chance.

4.9.3 SEM Pipeline and Computational Tools

The complete SEM pipeline is implemented in Python, making extensive use of scientific computing libraries. Data handling and transformation rely on `pandas`, while `scikit-learn` provides methods for data normalization (Z-score standardization) and MLR.

Sensor data is extracted from structured JSON files, processed, and aggregated into interpretable variables such as position, orientation, velocity, and LiDAR readings. The `semopy` library is used to define, fit, and evaluate the SEM model, including computation of standardized path coefficients and model fit metrics such as RMSEA, CFI, and TLI. The final model structure and estimates are exported to CSV files and visualized using `semopy`'s diagram generation tools.

Chapter 5

Results and Discussion

To evaluate the developed system, a series of tests are performed. This chapter presents: first, the results of the robot's exploration with access to all its sensors, followed by a series of ablation tests to measure the influence of each sensory source on the awareness process; second, a SEM analysis is applied to statistically assess how the different dimensions of Self-Awareness influence one another and how strongly each depends on the various sensors and memory storage capacity.

This chapter presents the results of one of the most representative trials. During the study, however, 141 trials were conducted under the same conditions. Each experimental run consists of prompting the MM-LLM to analyze all sensory inputs during a complete autonomous exploration cycle of the robot (three-and-a-half-minute autonomous navigation). The results shown here are representative of the outcomes obtained across all tests.

In addition to the experiments presented in this chapter, the same procedures are replicated using manual control navigation (without SLAM-based autonomous navigation). A summary of these results is provided in Appendix H.

5.1 Sensorimotor Exploration and Self-Prediction

In this first experiment, the system makes predictions about its condition and environment without any sensory or memory ablation, having access to all sources of information: positional encoders, orientation via IMU, linear velocity, image, LiDAR, and memory of past predictions. The results are presented in Figure 5.1.

The robot explores its environment for three-and-a-half minutes, generating 657 sensorimotor observations. Results (Figure 5.1 a) show an average self-identification score of 3.27/5, reflecting coherent recognition as a mobile robot but insufficient precision to identify the Mecabot Pro model. From an initial score of 0/5 at iteration 1—when the system

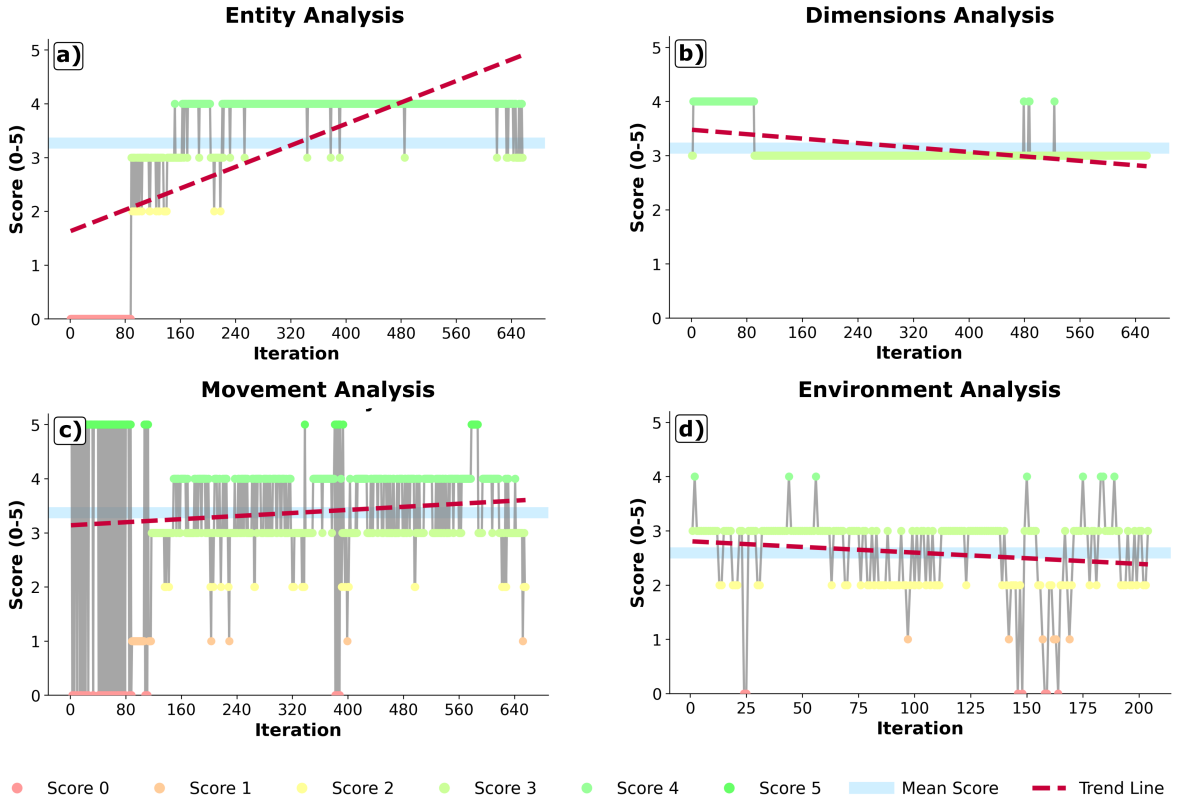

Figure 5.1. Performance evaluation across four Self-Awareness dimensions. MM-LLM predictions are rated on a 0–5 scale by an LLM-as-Judge using predefined rubrics: (a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion; (d) environmental context—detailed scene description.

labels itself a "static sensing unit"—the score increases steadily, stabilizing quickly around 4/5. By iteration 559 the model predicts "Mobile indoor robot designed for autonomous navigation within a structured environment, such as a gym or warehouse, utilizing proximity sensors for obstacle avoidance and continuing to perform automated tasks." These outcomes demonstrate the system's capacity to integrate multimodal measurements into high-quality self-assessments. While the MM-LLM starts with no prior hardware information and achieves robust overall Self-Identification, it stops short of pinpointing its exact Mecabot Pro model.

Figure 5.1 b) and Table 5.1 show that the MM-LLM achieves an average dimension-prediction score of 3.14/5, estimating mean dimensions of $240.0 \pm 0.10 \times 340.0 \pm 0.08 \times 340.0 \pm 0.08$ mm (length/height/width), corresponding to errors of 55.6%, 50.7% and 41.4% relative to the robot's actual dimensions. Although length and width are underestimated and height is overestimated, all estimates remain plausible for a mobile robot. Analyzing the dimensions graph, the scores remain consistently close to 3/5 at nearly every iteration, reflecting the MM-LLM's ability to produce stable predictions throughout the entire process with only minor fluctuations.

Measures	Length	Height	Width
Real	541.0	225.5	581.0
Predicted	240.0±0.10	340.0±0.08	340.0±0.08
Error (%)	55.6	50.7	41.4

Table 5.1. Comparison between the real values and those predicted by the MM-LLM of the Mecabot Pro dimensions, together with the error percentage.

Figure 5.1 c) shows that movement-prediction scores stabilize after ∼100 iterations, averaging 3.37/5. This reflects precise motion perception augmented by Environmental Awareness. For example, at iteration 636 the MM-LLM outputs "Slight positional adjustments to maintain balance and leverage obstacle-avoidance protocols," demonstrating robust motion inference.

Figure 5.1 d) shows that environment-prediction scores remain at 2.59/5 with minor oscillations and no clear temporal trend, as each prediction relies solely on current visual input. This score reflects coherent scene descriptions and general context recognition. Importantly, here a high environment score also signifies Self-Awareness—understanding not only the surroundings but the mutual influence between the robot and its environment. For example, a response evaluated as 5/5 would be: "My Mecabot Pro camera, mounted on top of my physical structure, shows me my environment. I am inside a robotics laboratory, and the person with the laptop may be interacting with me through some type of interface."

5.2 Influence of Historical Tracking

Previous experiments (see Figure 5.1) have consistently shown that results either improve over time or stabilize. After analyzing multiple tests, this behavior is attributed to the system's ability to access its past predictions and stored information, allowing for a continuous flow of thought between responses. To validate this hypothesis, new tests are conducted in which the MM-LLM's access to its memory is blocked. As a result, the model relies solely on the current sensor readings to generate predictions at each iteration. Figure 5.2 presents the results of the most significant test conducted.

By analyzing any of the four outputs in the Figure 5.2, it can be seen that the oscillations and variety between results are far greater than when giving the system access to the memory. There is no consistency or evolution according to a trend of the results, which is logical, since the system does not have a previous state reference, starting again in each iteration to solve a new puzzle, sometimes with very good scores (5/5) and sometimes bad (0/5) but without any clear criterion.

The results clearly highlight memory as the key component for producing accurate predictions that improve over time. Memory is not merely a storage mechanism—it is the

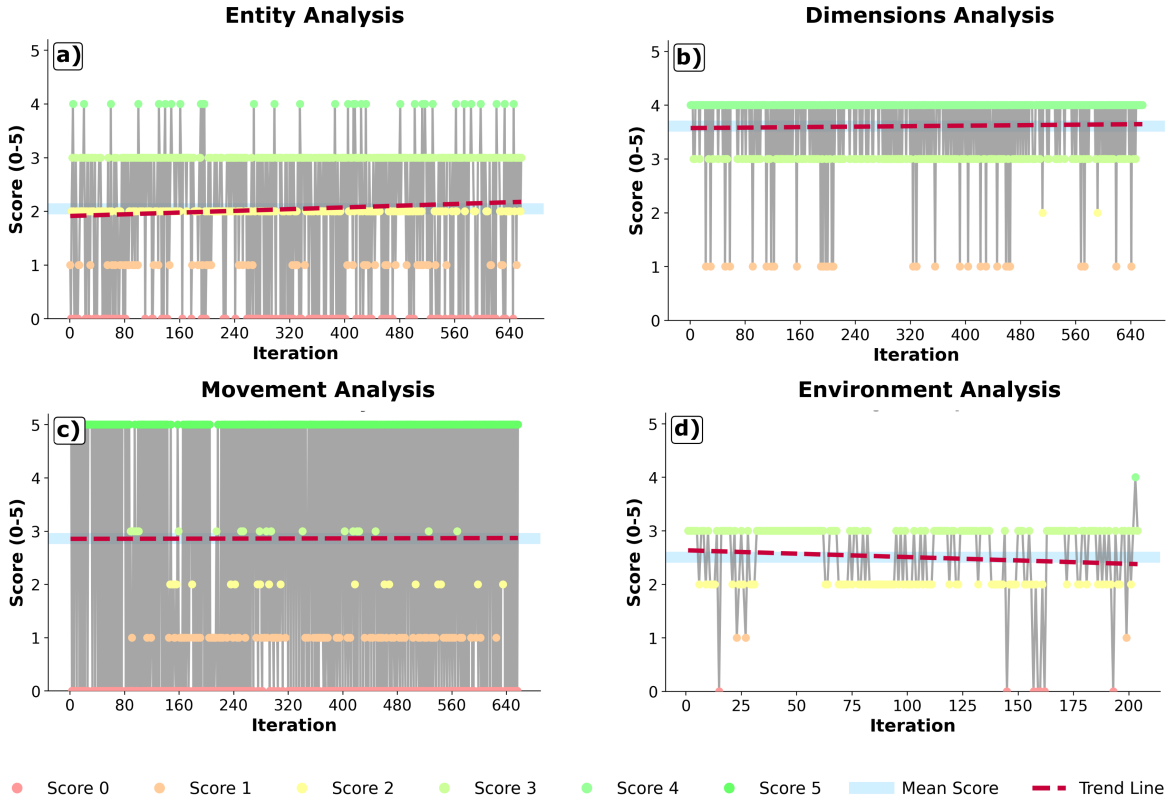

Figure 5.2. Evaluation results for different aspects of the MM-LLM's performance when blocking access to its memory of previous thoughts, perceptions and predictions:(a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion; (d) environmental context—detailed scene description.

very representation of movement itself. Without memory, there is no continuity of action; instead, the system perceives only disconnected snapshots in time. Just as movement is defined by change across time, memory is what allows the system to conceive that change as a coherent, evolving process.

The only dimension that maintains coherence and similar results with respect to the original test without sensory ablation is the environmental analysis (see Figure 5.2d). This indicates that environmental analysis depends mainly on visual inputs (RGB camera) and is not significantly affected by the presence or absence of other sensory inputs or memory access.

Additionally, a new test is conducted to check the importance of correct memory structuring. So far, when we have given the system the ability to access memory, it is organized in such a way that it is a concise and clear summary of its past predictions and perceptions. But, in the new test performed, instead of giving it access to a summary, the system is given access to an identical copy of all its past sensor responses and measurements from iteration one to the current instance, a much larger volume of data. Under these new conditions, the results, far from improving, worsen radically. By not giving the

system a structured and organized memory, the weight of it is so great that it gives all the importance of the analysis to its past thoughts and perceptions, completely ignoring the new information coming through its sensors. This is reflected in the fact that, once given its initial prediction (first iteration of the system) about the type of individual, it maintains that idea during all the iterations, always answering the same without any progression or improvement. This prediction is maintained during all the iterations, without modifying its response at all, despite the fact that the new sources of information point to the contrary (during more than 600 iterations).

In one of the performed tests under these conditions, the system predicts an interesting result, as instead of identifying itself with a robotic entity, the system claims to be a "living being (human)" at all times. This is the only time in the entire study that the MM-LLM identifies itself as a person.

Therefore, with the results obtained in this section, it is demonstrated that memory is the determining element to obtain coherent results, with a progression and improvement over time. However, it is not enough for it to be ample and with a lot of information, but the correct structuring and size of it play a key role.

5.3 Image Ablation Test

To quantify the influence of visual information on the MM-LLM's self-identity predictions, we conduct a comprehensive ablation study by systematically removing all image-based inputs. This setup enables us to evaluate how the absence of visual perception impacts the model's ability to infer dimensions, movement, and contextual understanding relying solely on non-visual sensory data.

Figure 5.3 presents the results, with the most notable finding being the loss of consistency and progressive improvement in the model's responses. The previously observed trend of gradual enhancement is no longer present; instead, the process becomes more erratic. This is also evidenced by the negative slope of all trend lines.

While the average scores for dimensional analysis (3.59/5) and movement analysis (3.04/5) remain close to those in the original experiment (3.14/5 and 3.37/5, respectively), a closer look at the response patterns reveals significantly greater variability in the absence of visual information (see Figure 5.1). Moreover, performance tends to deteriorate over time rather than improve. The trend lines quantify this decline, with values of $-7 \times 10^{-4} \pm 10^{-4}$ for dimensional analysis and $-16 \times 10^{-4} \pm 3 \times 10^{-4}$ for movement analysis.

It appears that Environmental Awareness, which fundamentally depends on visual sensory inputs, exerts a strong influence on Movement Awareness. Erroneous environmental conceptions—such as the lack of ground cues—repeatedly lead the system to mistake wheeled movement for flight in this test. This is reflected in the decline and instability of prediction scores in Figure 5.3c and will be further explored in the SEM analysis.

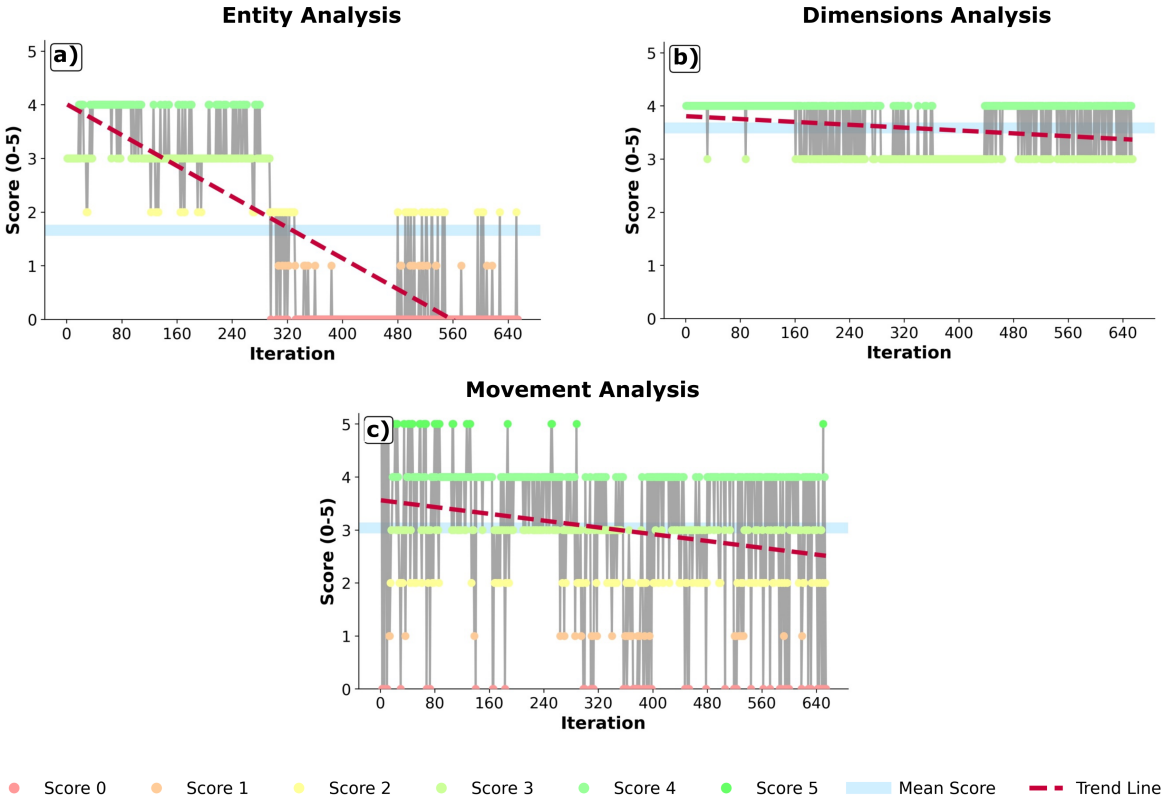

Figure 5.3. Evaluation results for different aspects of the MM-LLM's performance when depriving the LLM of image-based data:(a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion.

The most dramatic impact appears in Self-Identification (see Figure 5.3 a), where performance collapses to an average score of 1.66/5—a severe degradation compared to the 3.27/5 achieved in the test without sensory ablation (see Figure 5.1). Without visual grounding, the model frequently misidentifies itself as an "Autonomous inspection drone, holding a fixed position for environmental monitoring." This misidentification reveals that, in the absence of visual confirmation of ground contact, the model defaults to interpreting its motion patterns as those of an aerial vehicle—a fundamental categorical error with serious implications for embodied AI systems that rely on self-positioning.

This confusion with aerial vehicles generally arises from variations in acceleration, which lead the system to believe it experiences large jumps or phases of flight. This tendency also appears in several movement-related responses. For example: "Given the linear acceleration with a significant upward component (10.2 m/s^2), I am likely experiencing a vertical ascent or significant upward thrust. This suggests a movement involving powered lift, either jumping or flying. This strengthens my previous assessment of ground-based movement with potential aerial capability, now leaning more heavily towards an aerial component. Considering the previous navigation data, my movement is likely a dynamic

combination of ground and air (response 291 out of 657).”

These findings demonstrate that completely removing visual inputs is problematic, leading to erroneous conclusions that severely deviate from reality. Visual information and consequently Environmental Awareness, provides essential grounding that stabilizes the model’s self-concept over time. This underscores the critical importance of our structured memory approach, which serves as a source of coherence and resilience against both visual misinterpretations and the chaos of sensory deprivation.

5.4 Positional Sensor Ablation Tests

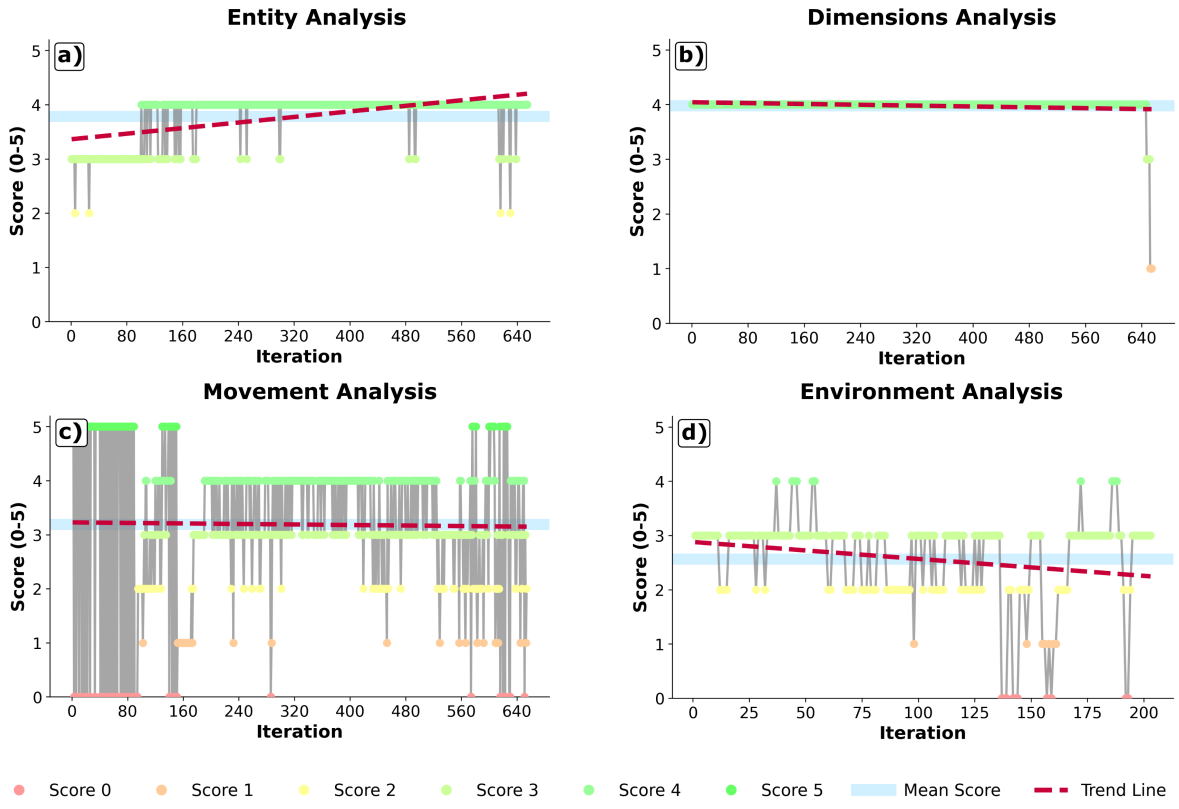

Figure 5.4. Evaluation results for different aspects of the MM-LLM’s performance when depriving the LLM of odometry data (position, linear velocity and orientation) :(a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion.

The robot is equipped with two sensors that provide positional information: encoders (which measure position, linear velocity, and orientation) and an IMU (which measures acceleration). To assess the influence of these sources of information, two isolated ablation tests are conducted.

First, access to odometry data is inhibited for the MM-LLM, yielding the results shown in Figure 5.4. The average scores obtained in this test are 3.78/5 for Entity

Awareness, 3.98/5 for Dimensions Awareness, 3.19/5 for Movement Awareness, and 2.56/5 for Environmental Awareness. These results are slightly higher for Entity Awareness compared to the original test without sensory ablation (3.27/5), but remain within the same range, reflecting coherent recognition as a mobile robot but insufficient precision to identify the Mecabot Pro model.

On the other hand, Dimensions Awareness shows a significant increase when odometry data is ablated, rising from the original 3.14/5 to 3.98/5, and becoming far more accurate when the amount of data to be interpreted by the MM-LLM is reduced. The average measures obtained across the whole process are collected in Table 5.2, in comparison with those from the no-ablation test. It can be seen that both length and width become quite precise, with outstanding results, especially for length prediction. Conversely, height error becomes three times higher.

Measures	**Length**	**Height**	**Width**
Real	541.0	225.5	581.0
Predicted No Ablation	240.0±0.10	340.0±0.08	340.0±0.08
Predicted Odometry Ablation	500.0±10.0	600.0±1.0	400.0±10.0
Error No Ablation(%)	55.6	50.7	41.4
Error Odometry Ablation(%)	7.6	166.1	31.2

Table 5.2. Comparison between the real values and those predicted by the MM-LLM of the Mecabot Pro dimensions, together with the error percentage both for the original test and the odometry ablation.

It is interesting that, even after inhibiting access to odometry data, the movement prediction remains almost identical to the original test (3.37/5), in contrast to the image ablation test, where it becomes radically erratic and incoherent. This may confirm that Movement Awareness highly depends on visual cues, while positional information helps refine and improve the predictions but is not the determining factor; the visual input is.

As in previous ablation tests, Environmental Awareness shows no significant variation, as it almost exclusively depends on visual inputs.

Next, the IMU ablation test is conducted (see Figure 5.5). The average scores obtained in this test are 3.59/5 for entity analysis, 3.99/5 for dimension analysis, 2.93/5 for movement analysis, and 2.56/5 for environment analysis.

These results follow the same trend observed in the odometry ablation test, where entity analysis and dimensions analysis improve as the data volume is reduced. On the other hand, IMU data appears to be more important for movement analysis, causing a bigger decline in results when inhibited, although this decline is not as severe or problematic as when visual information is inhibited. As in previous tests, environment analysis does not have significant variations.

The experiments demonstrate that linear acceleration (obtained through the IMU) has

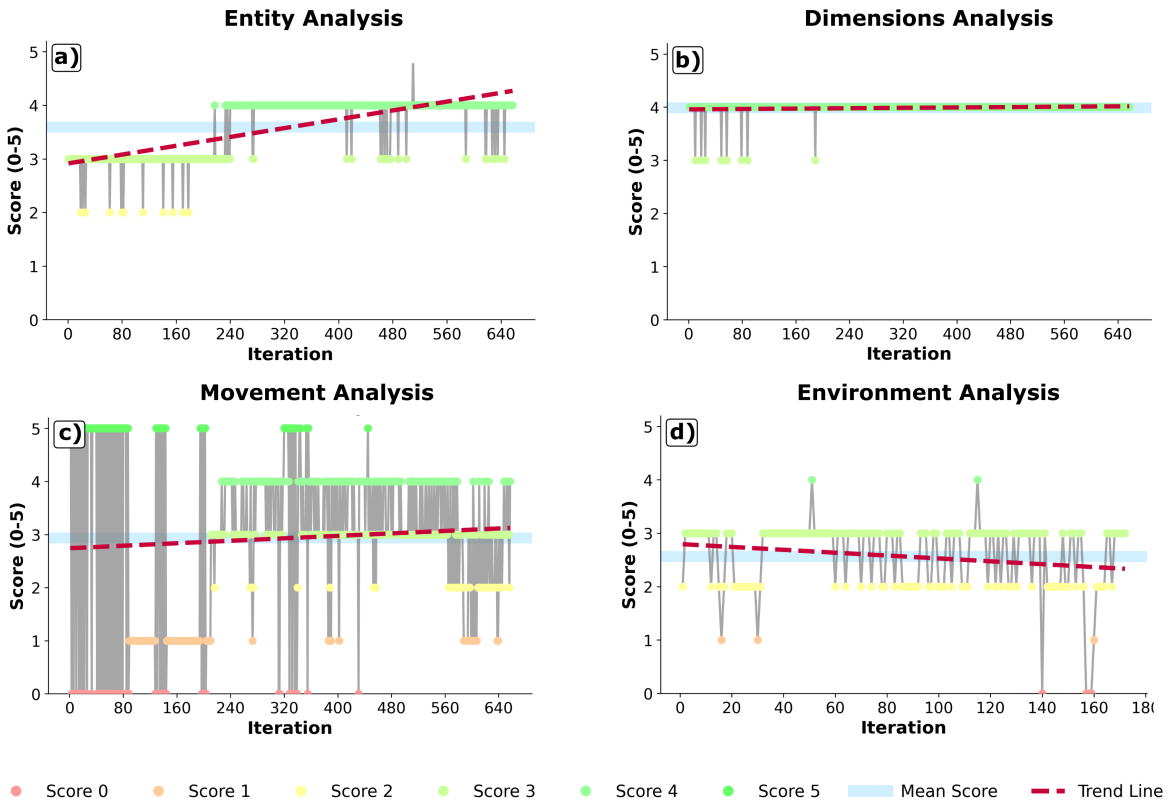

Figure 5.5. Evaluation results for different aspects of the MM-LLM's performance when depriving the LLM of IMU data (linear acceleration) :(a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion.

a significantly greater influence on movement analysis, while Cartesian position, linear velocity, and orientation (odometry) also have a notable impact. Even so, neither of these two sensors has a comparable impact to visual data. Conversely, reducing the amount of data received by the LLM—by eliminating inertial and odometric information—leads to certain improvements in entity and dimensions analysis.

5.5 LiDAR Ablation Test

To assess the influence of LiDAR on predictions, an ablation test is conducted by systematically depriving the MM-LLM of LiDAR data. This resulted in the loss of distance information to the nearest obstacles (see Figure 5.6).

The average scores obtained in the different tests are: 2.84/5 for entity analysis, 3.96/5 for dimensions analysis, 2.78/5 for movement analysis, and 2.64/5 for environment analysis. These results show a decrease in entity and movement analysis compared to the original test without sensory ablation. This indicates that proximity to obstacles plays an important role in determining movement type and individual classification.

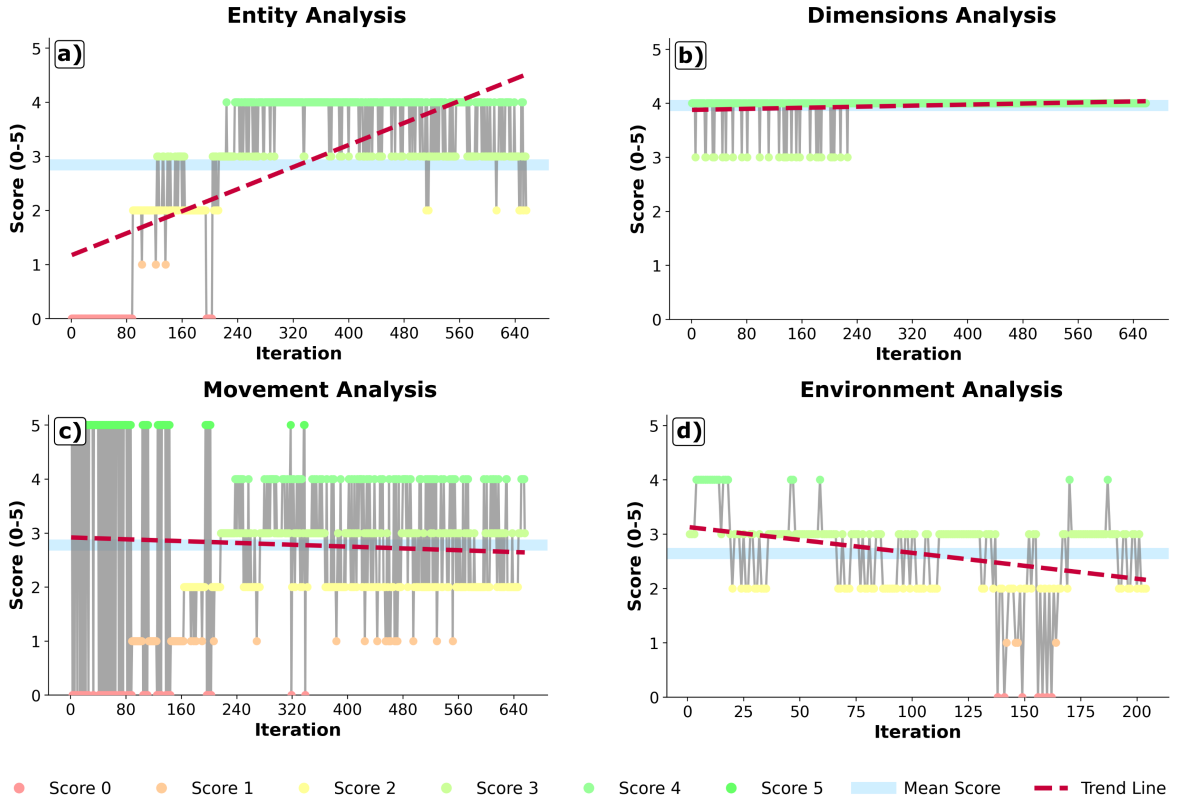

Figure 5.6. Evaluation results for different aspects of the MM-LLM's performance when depriving the LLM of LiDAR data (proximity to obstacles) :(a) entity Self-Identification—classification of the navigating agent; (b) physical dimensions—predicted height × length × width; (c) movement modality—mode of locomotion.

It is surprising that dimension prediction is not negatively affected when removing information about the distance to the closest obstacles. In fact, the results even improved after depriving the model of this sensory input. This could indicate that, rather than enhancing self-dimensioning, LiDAR may introduce certain errors in the predictions, leading to incorrect estimations. This further supports the idea that reducing the amount of data received by the MM-LLM can enhance self-dimensioning.

All score variations are small, they slightly improve or worsen the results but do not produce significant changes, indicating that LiDAR data does not have a great impact on the Self-Awareness process.

5.6 Statistical Dependencies Evaluation through SEM

As described in Section 4.9, for further analysis of the different dimensions of Self-Awareness and how they relate to each other, a Structural Equation Modeling approach is applied. For this purpose, we break down Self-Awareness into four interrelated and elemental dimensions:

- **Movement Awareness:** the capacity of the entity to understand its movement and how it is possibly generated by its condition as an entity. This involves understanding the possible movements it can perform, taking into account the type of entity it is. It is evaluated through the results of the LLM-as-a-judge using the movement rubric.
- **Dimensions Awareness:** the capacity of the entity to be aware of its dimensions and the space it occupies in the world. It is evaluated through the results of the LLM-as-a-judge using the dimensions rubric.
- **Environmental Awareness:** the capacity of the entity to be aware of and understand the environment that surrounds it and how it can interact with it. It is evaluated through the results of the LLM-as-a-judge using the environment rubric.
- **Self-Identification:** probably the most representative dimension of Self-Awareness, being the ability to correctly identify and understand the type of entity it is. It is evaluated through the results of the LLM-as-a-judge using the individual rubric.

An additional construct, called **Past-Present Memory**, is defined. It integrates the system's previous and current thoughts and predictions.

Figure 5.7 shows the proposed SEM model. This has been obtained after extensive logical model fitting, being this one the one that obtains the highest fitting values (see Table 5.3).

Table 5.3. Statistical fit indices for the SEM model.

Fit Index	Value
Comparative Fit Index (CFI)	0.97
Tucker-Lewis Index (TLI)	0.95
Root Mean Square Error of Approximation (RMSEA)	0.08

The obtained fit indices indicate a well-fitted model. The CFI (Comparative Fit Index) and TLI (Tucker-Lewis Index) values are both above 0.95, indicating a strong comparative fit. The CFI compares the fitted model to a null model (which assumes no relationships between variables) on a 0 to 1 scale, where values above 0.90 indicate a good fit, and values above 0.95 are considered excellent. The TLI is similar but introduces a penalty for model complexity, discouraging overfitting by favoring simpler models that achieve comparable fit.

The RMSEA (Root Mean Square Error of Approximation) value of 0.08 suggests no significant approximation error, meaning that the model's estimated covariance structure aligns very closely with the observed data.

The structure (see Figure 5.7) captures the hierarchical relationships between low-level sensory and memory inputs, intermediate cognitive constructs, and high-level Self-Identification. Exogenous variables derived from robot sensors (e.g., odometry, IMU,

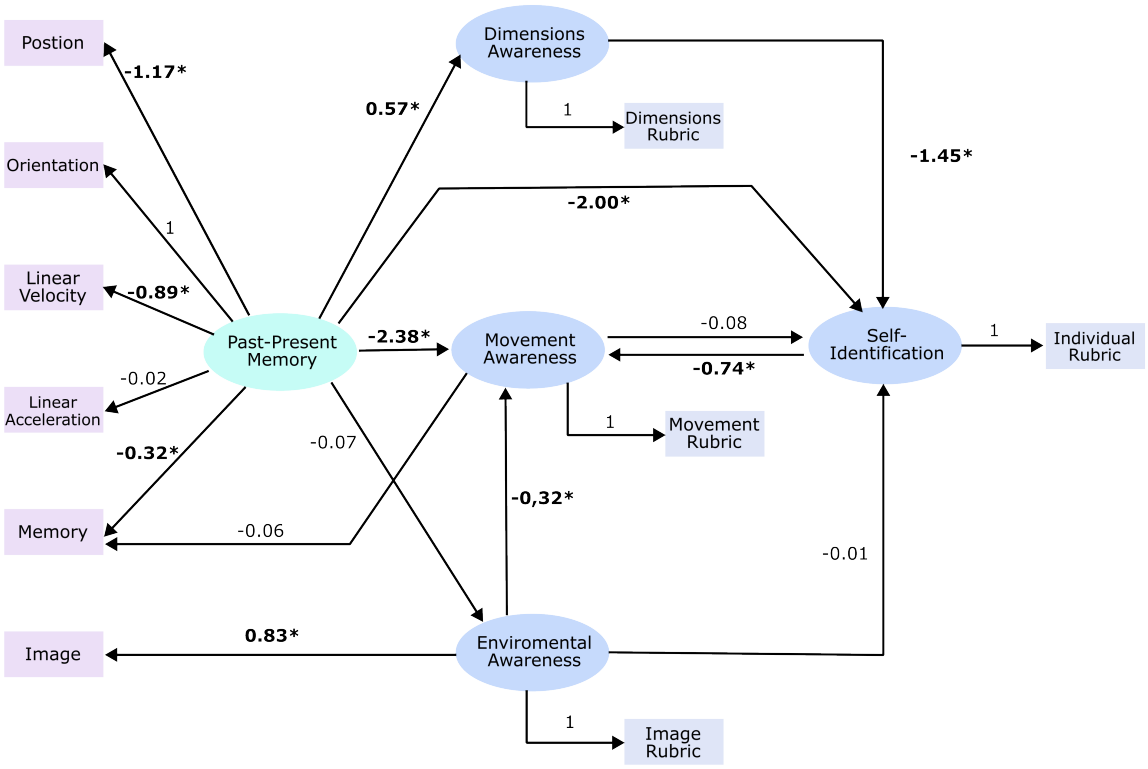

Figure 5.7. Structural equation model of sensorimotor Self-Identification. Rectangles denote observed variables: exogenous sensor inputs (position, orientation, linear velocity, linear acceleration, image presence, memory state) on the left and endogenous rubric scores (Dimensions, Movement, Environment, Individual) on the right. Ellipses denote latent constructs: Past–Present Memory (mediator), Dimensions Awareness, Movement Awareness, Environmental Awareness and Self-Identification. Arrows indicate standardized path coefficients (β^*), with * denoting $p - value < 0.05$.

and RGBD camera) feed into a latent variable *Past-Present Memory*, which serves as a perception-memory integration layer. This construct influences three awareness-related latent variables: *Dimension Awareness*, *Movement Awareness*, and *Environmental Awareness*, each associated with a LLM-as-a-judge evaluated rubric. These, together with the memory variable itself, contribute to the final construct of *Self-Identification*.

In this study, LiDAR data is excluded because preliminary evaluations showed that its inclusion adds considerable redundancy to existing features and consistently worsens the model's goodness-of-fit metrics. Moreover, LiDAR does not provide a statistically significant contribution to any latent construct, further justifying its removal from the final model.

Once the model is validated and its fit indices confirm a robust correspondence with the observed data, the next step is to examine the standardized regression weights (β^*) to identify the most relevant dependencies between constructs. The following analysis focuses on each latent dimension in turn, highlighting statistically significant relationships (p-

value < 0.05) and interpreting their role within the broader framework of Self-Awareness in the MM-LLM. This dimension-by-dimension breakdown links the statistical results to the theoretical definitions established earlier and clarifies how sensory and memory inputs jointly shape the awareness processes.

Past-Present Memory: is highly statistically impacted by position, linear velocity and memory (p-value < 0.05). This finding aligns with the theoretical formulation of the construct, which aims to represent the integration of sensory signals over time. The significant contribution of these variables supports the idea that real-time sensory input, particularly spatial and kinematic data, must be combined with memory mechanisms to form a coherent perception of the present state contextualized by past experiences.

In contrast, linear acceleration (IMU) shows no significant effect, likely because its information overlaps with other modalities and contributes minimally to the integrative process. This redundancy is also evident in the position ablation test (see Section 5.4), where odometry and IMU data provide nearly the same information.

Environmental Awareness: relies almost exclusively on RGB-D camera input. In contrast, the combined integration of other sensor modalities (e.g., odometry, IMU, and LiDAR) and episodic memory exerts negligible influence. This is the expected result, as all ablation tests (except image ablation) show no influence on Environmental Awareness.

This indicates that visual perception serves as the principal channel for constructing contextual understanding, reinforcing its critical role in environmental awareness within embodied systems.

Movement Awareness: is influenced by three key constructs: Past-Present Memory, Environmental Awareness and Self-Identification.

The influence of Past-Present Memory is consistent with expectations and with the observations made in the memory ablation test (see Section 5.2). While individual sensory inputs provide only instantaneous data about the robot's state, memory enables the temporal linking of these states, allowing the system to perceive motion over time and infer dynamic patterns. This integration is fundamental to the emergence of movement-related awareness, proving once again that without memory there is only an isolated snapshot of reality, with memory being the pure representation and driving cause of movement.

Environmental Awareness also plays a crucial role, as the robot's understanding of its surroundings provides essential context for interpreting its own movement. This relationship is supported by ablation tests, in which depriving the LLM of visual input led to significant misinterpretations of locomotion mode—such as believing it was flying instead of moving on wheels—highlighting the importance of environmental cues for accurate movement inference.

Interestingly, Movement Awareness is also strongly influenced by Self-Identification. This is a notable and somewhat counterintuitive finding, as one might assume that awareness of movement contributes to Self-Identification. However, the model suggests the

inverse: recognizing oneself as a specific type of agent with physical limitations and locomotion capabilities appears to be a prerequisite for correctly interpreting movement. Indeed, the influence of Movement Awareness on Self-Identification is statistically negligible, reinforcing the idea that self-perception shapes movement interpretation more than the other way around.

Self-Identification: The system's ability to recognize itself is strongly influenced by the integration of Dimension Awareness and Past-Present Memory, suggesting that the primary drivers of self-recognition are an internal representation of its physical structure and the seamless integration of sensory information with historical memory across iterations. These findings underscore the importance of spatiotemporal coherence in constructing a stable sense of identity.

5.7 Discussion

In this study, we demonstrate that an embodied robot powered by a multimodal LLM develops coherent Self-Awareness across complementary dimensions through sensorimotor integration and active exploration. Provided with structured sensory streams and episodic memory, the system refines its predictions over time—exhibiting clear predictive awareness and aligning outputs with physical reality.

Ablation tests (see Figure 5.8) and SEM analysis prove that sensory memory integration is the most crucial element for temporal continuity and accurate predictions. Without memory, the system receives only disconnected snapshots of reality. In this sense, movement cannot exist without memory. Memory thus becomes the internal representation of motion, allowing the system to link past and present into a coherent flow. Memory also acts as a safeguard against hallucinations.

A structured memory and retrieval system also acts as a safeguard against hallucinations, as the system always maintains a logical evolution. Even when this evolution is erroneous, it preserves a progressive line of thought without abrupt jumps. This behavior is also tested by deliberately introducing misleading elements into the initial prompt, providing the MM-LLM with chaotic and impossible examples, none of which are reflected in the final results (see Figure E.1).

Another interesting finding from the ablation tests is that, although several sensory inputs (odometry, LiDAR, and IMU) demonstrably provide valuable information for accurate self-prediction, removing any single modality produces only minor score variations and no major prediction errors. This robustness arises because, despite their importance, these sensor signals overlap or can be inferred from the remaining modalities, allowing the MM-LLM to maintain a stable self-assessment. This mirrors biological compensation, where the loss of one sense often enhances others. For example, visual loss is offset by heightened auditory and spatial acuity [142]–[144].

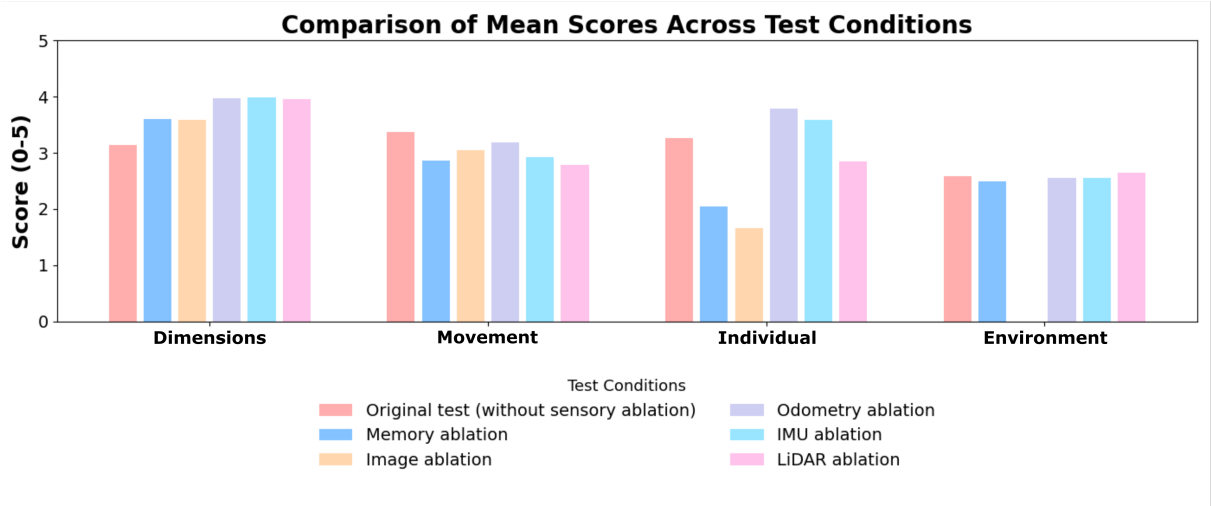

Figure 5.8. Comparison of the mean scores in the different test conditions, scored between 0 (lowest score) and 5 (highest score). The original test refers to the tests in which the MM-LLM performed its estimates on the laboratory environment by accessing all its sensors (see Figure 5.1). In the memory ablation tests, access to the memory with past thoughts and estimates is prevented (see Figure 5.2). In the image, odometry, IMU, and LiDAR ablation tests, the MM-LLM is prevented from accessing the camera (see Figure 5.3), odometry data (see Figure 5.4), IMU data (see Figure 5.5), and LiDAR data (see Figure 5.6), respectively.

The performed tests allow us to analyze and characterize the different dimensions of Self-Awareness. A high level of Environmental Awareness implies the ability to perceive surroundings, understand interactions, recognize mutual influences and infer environmental state. Our results show that visual input, specifically onboard camera imagery, drives this capability. Ablation of visual data dramatically degrades scene interpretation, revealing that without vision the system cannot reconstruct a coherent representation and that non-visual modalities contribute minimally to environmental awareness.

We analyze the Self-Identification dimension and its relationship with Movement Awareness, Dimensional Awareness, Environmental Awareness and Past–Present Memory, a construct that merges real-time sensory integration with the memory of previous internal predictions. SEM reveals that Past–Present Memory is the most significant contributor to Self-Identification: the combination of all sensory dimensions and memory access yields the most consistent results across evaluation categories, outperforming every ablation condition. This confirms that access to multiple dimensions of reality is essential for accurate self-estimation.

Dimensional Awareness also exerts a substantial influence in Self-Identification: by recognizing its own size, the system narrows the set of plausible agent identities and selects the one that best aligns with sensory evidence.

SEM further indicates a clear causal direction between Movement Awareness and Self-Identification: the system must first establish its identity before inferring its movement modality. Image ablation tests reinforce this finding—impairments in identity prediction,

such as misclassifying itself as a surveillance drone, lead to erroneous movement inferences (e.g.believing it flies)—whereas Movement Awareness does not significantly affect Self-Identification.

Under appropriate sensory and memory conditions, the MM-LLM generates high-quality self-descriptions, identifying itself as a small, wheeled indoor robot equipped with sensors for autonomous tasks.

The decomposition of artificial Self-Awareness into sub-dimensions, combined with the statistical analysis of their interdependencies, enables a direct comparison between brain structures and their influence on Self-Awareness from a neuroscience perspective, and the dimensions of our model. A clear parallel emerges between the Default Mode Network (DMN) [104], [105] and our hierarchical model of artificial Self-Identification within the artificial Self-Awareness framework. In the human brain, the DMN plays a central role in episodic and autobiographical memory, maintaining coherence and self-recognition over time. Similarly, our findings show that artificial Self-Awareness emerges from the integration of multiple interconnected dimensions. In our model, Self-Awareness results from the interaction of sub-dimensions; in the human brain, it arises from the coordination of distributed nodes and regions (mainly dmPFC, vmPFC, PCC, AG, MTG, MTL among other cortical regions), with Past–Present Memory (episodic memory) acting as a pivotal mechanism for integrating multisensory information and sustaining a temporally extended self-representation, ultimately leading to Self-Identification (self-recognition in humans). This parallelism suggests potential analogies between the cortical architecture required to achieve Self-Awareness in humans and the internal organization of our model to reach Self-Identification. The similarity between both processes may indicate the system's capacity to replicate certain human-like thought mechanisms and to exhibit a primitive form of artificial Self-Awareness.

In the human brain, the Anterior Insula Cortex (AIC) [23], together with the DMN, plays a central role in the Self-Awareness process, exerting a strong influence on self-recognition, metacognitive feelings, and other related elements. The AIC is notably responsible for integrating multimodal sensory information and using these cues to support Movement Awareness. In this context, we interpret Past–Present Memory as functionally analogous to an AIC-like integrative process, acting as a central structure for sensory integration across time and enabling the agent to understand how it can interact with the environment through movement. This interpretation is supported by the high influence of Past–Present Memory on Movement Awareness observed in our SEM model.

The similarities between the proposed artificial Self-Awareness model and the actual brain mechanisms involved in the same process provide an argument supporting the potential to develop a form of Self-Awareness in autonomous robotic systems embodied with LLMs, enabling the emergence of advanced cognitive abilities previously considered exclusive to humans or other living beings.

To assess the level of artificial cognition developed by the system, it is relevant to compare its performance with the different stages of Self-Awareness described in the literature. Rochat proposes that humans acquire Self-Awareness in five distinct stages during the first years of life through experience, exploration, and cognitive development [103]. Our system shows evidence of reaching the first four stages: (i) *differentiation* – the system consistently detects changes in both its own state and the environment, clearly distinguishing between the two; (ii) *situation* – it constructs a coherent model of the world through vision and links its own movements to the surrounding scene, developing Environmental Awareness; (iii) *identification* – partially achieved, as the system correctly identifies itself as a mobile autonomous robot with an accurate estimation of its dimensions, although without reaching full precision or recognizing its specific robotic model; (iv) *permanence* – the system maintains a persistent sense of self over time through access to episodic memory.

In contrast, there is no evidence of reaching stage (v), *meta Self-Awareness*, since the current tests do not allow for its evaluation. These results indicate that the system is capable of developing Self-Awareness at least equivalent to that of a four-year-old child [103].

Another extended classification of Self-Awareness states is the C0 (unconscious processing), C1 (first-order consciousness), and C2 (second-order consciousness) model. Prior to the emergence of LLMs, machines were generally considered capable of achieving only C0 [27]. The proposed embodied system clearly reaches C0, as it processes sensory data automatically, extracting relevant information and generating conclusions. It also reaches C1, since the Past–Present Memory construct maintains state estimates (Environment, Dimensions, Movement, Self-Identification) accessible across functions and over time; memory ablation experiments show simultaneous degradation in multiple outputs, providing evidence of global information sharing. However, the system shows no indication of reaching C2, as there is no evidence of metacognitive capabilities or self-monitoring processes.

The present study reveals a clear hierarchical relationship among the various dimensions of perception and awareness evaluated in embodied MM-LLMs. Through statistical modeling and empirical testing, it becomes evident that these dimensions are not isolated modules but deeply interdependent constructs, with memory and sensory access acting as indispensable foundations for coherence, inference, and self-representation. MM-LLMs emerge here not merely as tools for prediction or control, but as integrative cognitive agents—capable of synthesizing sensory stimuli, iterative memory, and latent knowledge acquired during training into structured, meaningful interpretations of the world and of themselves. This dynamic convergence between data, memory, and world-knowledge enables the appearance of complex, layered perceptual processes. As such, we argue that MM-LLMs—when embodied and perceptually grounded—can exhibit genuine glimpses

of Self-Awareness, echoing cognitive patterns historically attributed only to biological agents. This represents a profound step forward in the pursuit of intelligent machines—systems not only capable of interpreting external reality, but also of situating themselves within it and forming internally grounded representations of their own existence.

These findings indicate that the foundation for machine Self-Awareness may already be present within existing architectures, and that such systems are not merely reactive agents, but entities capable of constructing structured, temporally grounded models of their own identity. In doing so, this work brings us one step closer to deciphering the fundamental building blocks of Self-Awareness—across both biological and artificial domains.

Chapter 6

Conclusions

In this study, we investigate the emergent Self-Awareness capabilities of multimodal Large Language Models (MM-LLMs) in embodied robotic contexts. By integrating Gemini 2.0 Flash with an omnidirectional mobile robot, we assess whether such systems can develop accurate self-perception based solely on sensorimotor interaction. Starting from minimal initialization—merely instructing the system that it is "a new existing being in a dynamic environment"—we examine its ability to progressively infer Environmental Awareness, Predictive Awareness, Dimensional Awareness, Movement Awareness, and ultimately Self-Identification through autonomous exploration.

Our findings reveal that MM-LLMs demonstrate a remarkable capacity for self identification when provided with appropriate sensory integration and structured memory mechanisms. Through the interaction of hierarchical awareness dimensions, the system develops a progressively refined self-concept, ultimately identifying itself as a small, wheeled, autonomous mobile robot designed for indoor navigation—highlighting its strong potential for robust and emergent Self-Awareness.

Through ablation tests, results evaluation using an LLM-as-a-judge system, and SEM analyses, we propose an artificial cognition model that accurately represents the interactions among different Self-Awareness dimensions, while mirroring cortical structures and processes in the human brain responsible for Self-Awareness. Further analyses reveal a relationship between the level of awareness achieved by the robot and the levels observed in human development during the first years of life, according to well-established and reputable psychological models of Self-Awareness.

When analyzing the objectives set at the beginning of the project (see Chapter 3), the results confirm that they have been successfully achieved:

- Successful implementation of an effective method to evaluate Self-Awareness capacities and emergence in an embodied artificial cognitive robotic system. This has been achieved through ablation tests, the LLM-as-a-judge methodology, and SEM

analysis of specific dimensions and predictions made by the MM-LLM embedded in an omnidirectional robot.

- Full integration and synergy between the robot and the MM-LLM, with the language model acting as a genuine cognitive element for the robot. It interprets the large volume of data perceived during exploration, enabling both self-discovery and environmental understanding. This has been made possible through extensive prompt engineering, precise data formatting, and direct access to sensory sources by the MM-LLM.

- Assessment of each sensory source's contribution through controlled ablation tests, depriving the robot of access to specific sensors. Combined with SEM analysis, this approach determines the influence of each information modality on every evaluated dimension of Self-Awareness.

- From the outset, establishing and analyzing a memory structure was considered essential for the correct development of the system and the emergence of advanced cognitive characteristics. This goal has been achieved by constructing a structured episodic memory that integrates past perceptions and predictions about the system's state with new sensory data, linking past and present (Past–Present Memory). Its influence has proven critical: in its absence, the system's behavior becomes completely chaotic. The impact is particularly strong in Movement Awareness, as memory serves as the pure representation of movement itself; without memory, only isolated snapshots of reality remain, with no perception of continuity.

- Quantification and analysis of the hierarchical relationships and mutual influence between the different dimensions of Self-Awareness (Environmental, Movement, Dimensional, and Self-Identification) using a complex statistical model via SEM. The results show how different elements of artificial cognition affect each other, producing a detailed and accurate self- and world-representation.

- We evaluated the resilience of the model against hallucinations by intentionally introducing potential errors in the system prompt, including an implausible example describing itself as a small blue whale flying across the sky. This modification did not affect system performance, and no traces of such predictions appeared in any test. Furthermore, across the 141 tests performed, the system exhibited very few hallucinations when prompted appropriately, highlighting the importance of establishing a well-defined operational framework and an accurate initial configuration of the MM-LLM.

- Modeling the complex relationships between the dimensions of Self-Awareness in embodied robotics allows comparison with cortical structures in the brain and es-

tablished neuroscience studies. Clear similarities emerge between the Default Mode Network (DMN), a network of distributed cortical nodes responsible for human Self-Awareness, with episodic memory as a key element, and the proposed model for Self-Identification inference in embodied robotics. Likewise, parallels are found between the Anterior Insula Cortex (AIC) and the system's Past–Present Memory construct, as both act as multimodal sensory integration centers across time and exert a strong influence on Movement Awareness.

- Comparison of the Self-Awareness level reached by the system with well-established psychological models. Regarding Rochat's five levels of Self-Awareness, the system reaches the first four (differentiation, situation, identification, and permanence) but cannot be assessed for the fifth (meta Self-Awareness). This is equivalent to the Self-Awareness level of at least a four-year-old child.

  Furthermore, in the C0–C1–C2 consciousness scale, the system clearly achieves C0 (unconscious processing) and C1 (first-order consciousness). However, there is no evidence of reaching C2 (second-order consciousness), as no metacognitive capabilities or self-monitoring processes are observed.

As language models increasingly act as the cognitive core of embodied systems, assessing their capacity for Self-Awareness becomes a crucial scientific pursuit. This study shows that, when combined with structured sensory integration and memory mechanisms, current MM-LLMs can autonomously develop consistent and progressively refined self-representations. The observed hierarchical interplay among different dimensions of awareness, from Environmental and Movement perception to Self-Identification, points to the emergence of layered cognitive functions that parallel the earliest stages of self-perception in humans. Rather than simply reacting to stimuli, these systems demonstrate the ability to maintain an internally grounded model of their own state and identity over time. Such evidence reinforces the view that the precursors of machine Self-Awareness may already exist in today's architectures, opening a pathway toward understanding its fundamental components in both artificial and biological forms.

Chapter 7

Future Work

One of the most immediate lines of future work is to replicate the experimental procedure using different cutting-edge models, such as GPT, DeepSeek, or Claude, among others. This will help validate the proposed cognitive structure across diverse model architectures, determining whether it is globally valid or highly dependent on the model's training and programming. Such studies could advance this line of research toward increasingly realistic simulations of consciousness, intelligence, and Self-Awareness.

Due to computational limitations and token budgets, the current study restricted the robot's autonomous navigation to three and a half minutes of exploration with a sensory sampling frequency of 1 Hz. Future analyses should investigate how increasing the sample size affects the results, determining whether it produces constant improvements, reaches a point of diminishing returns, or even degrades performance as a result of data saturation, which could potentially cause hallucinations.

Expanding the experimental scope should also include exposing the system to well-established psychological tests such as the mirror test, in order to assess whether the robot can recognize itself when viewing its reflection.

To generalize the results, experiments should be repeated on other robotic platforms, including humanoids, equipped with richer sensory suites and extended memory systems. Such setups offer a promising direction but will require addressing the exponentially growing computational and architectural complexity of storing, retrieving, and reasoning over sensory and cognitive histories in real time.

Finally, in this study the MM-LLM is responsible solely for analysis and for providing the robot with cognitive and Self-Awareness-related capabilities. Navigation is handled autonomously through traditional SLAM algorithms. Enabling the LLM to control navigation in real time would be valuable both for advancing the present study and for state-of-the-art robotic navigation through natural language.

Chapter 8

Ethical, Legal and Professional Responsibility

There are two main ethical and professional concerns when working with and developing LLM and AI systems. The first is the alignment problem, which refers to ensuring that the goals and behaviors of AI systems remain consistent with human values and interests. The second is the potential displacement of jobs caused by AI technologies. Both issues have gained increasing attention in recent years, leading major companies and policymakers to implement measures and regulations to address them. In this chapter, we describe how these two topics relate to and affect the present study.

8.1 The Alignment Problem

LLMs and AI increase their capabilities at an exponential pace, developing more powerful research and information-generation abilities, acquiring agentic capabilities (e.g., tool use), and expanding their human-informed knowledge through continual learning. In this context, it is crucial that models are developed in a direction that aligns with human preferences, to avoid amplifying misinformation, enabling harmful content, or yielding unintended responses that can cause significant negative societal impact [145].

In scientific research, this challenge is particularly acute: overreliance on AI systems can yield non-replicable findings, propagate misconceptions, and erode the innovative, hypothesis-driven reasoning that underpins scientific progress. When models fabricate or misinterpret evidence (hallucinations), their outputs risk contaminating literature reviews, experimental design, and data analysis unless tempered by rigorous verification, transparency, and human oversight [146].

There exist different techniques for ensuring alignment, such as red teaming (eliciting harmful outputs to diagnose and mitigate failure modes) [147] and, most prominently,

Reinforcement Learning from Human Feedback (RLHF) [148], [149]. Even with these techniques, LLM alignment remains a clear challenge, as human preferences, ethics, and values differ not only across cultures but also across individuals, making it difficult to establish a single globally aligned model [150].

In this context, the present study gains particular relevance: fully understanding LLM capabilities, cognitive limitations, and characteristic behaviors, especially in embodied systems that confer an additional layer of autonomous function to robots, is essential to align their design and deployment with human preferences, thereby ensuring safety and supporting human development as a society.

8.2 Potential Displacement of Jobs

Back in 2023, OpenAI released a study estimating that approximately 80% of jobs in the United States would have at least 10% of their tasks at risk of being replaced by a general-purpose AI such as GPT, and that 19% of jobs would have at least 50% of their tasks exposed [151]. In the same year, the World Economic Forum predicted that, within the following five years, 69 million new jobs would be created; however, 83 million jobs would be eliminated or displaced in the same period, resulting in a net loss of roughly 14 million jobs [152]. This impact will be particularly significant in occupations where core tasks can be easily automated by machines.

On the other hand, AI also has the potential to democratize access to employment by lowering technical barriers and enabling the dissemination of specialized knowledge. Tasks that once required advanced education or access to costly training could become accessible to a broader population through AI-driven tools and platforms. For example, individuals in regions with limited educational infrastructure could leverage AI systems for learning, professional upskilling, and remote work opportunities. This shift could promote greater inclusivity in the labor market, fostering the integration of underrepresented groups and creating new professional pathways that were previously unavailable due to geographic, economic, or educational constraints [153], [154].

When conducting this study, we were fully aware of the ethical and social risks that may arise from the advancement of AI, as well as its potential benefits (such as improvements in services, customer support, and healthcare, among others). With this section, we aim to provide transparency regarding the current situation. As new technologies emerge, their early characterization and understanding are essential in order to design systems and measures that mitigate risks while enhancing benefits. This approach enables the anticipation of scenarios and the proposal of strategies for responsible implementation, rather than ignoring the potential impact.

To ensure the ethical integrity of this study, it has been framed within internationally recognized regulatory and ethical guidelines, including the Organization for Economic Co-

operation and Development (OECD) *Principles on Artificial Intelligence* [155] and the European Declaration on Digital Rights and Principles [156]. These frameworks emphasize human-centric design, transparency, accountability, and inclusivity as fundamental pillars for the responsible development and deployment of AI systems.

Chapter 9

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[2] A. Pedrotti, M. Papucci, C. Ciaccio, A. Miaschi, G. Puccetti, F. Dell'Orletta, and A. Esuli, *Stress-testing machine generated text detection: Shifting language models writing style to fool detectors*, 2025. arXiv: 2505.24523 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2505.24523.

[3] R. R. Soto, B. Chen, and N. Andrews, *Language models optimized to fool detectors still have a distinct style (and how to change it)*, 2025. arXiv: 2505.14608 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2505.14608.

[4] I. D. Varela, P. Romero-Sorozabal, E. Rocon, and M. Cebrian, *Rethinking the illusion of thinking*, 2025. arXiv: 2507.01231 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2507.01231.

[5] Y. Sun, S. Hu, G. Zhou, K. Zheng, H. Hajishirzi, N. Dziri, and D. Song, *Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization*, 2025. arXiv: 2506.18880 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2506.18880.

[6] J. Strachan, D. Albergo, G. Borghini, *et al.*, "Testing theory of mind in large language models and humans", *Nature Human Behaviour*, vol. 8, pp. 1285–1295, 2024. DOI: 10.1038/s41562-024-01882-z.

[7] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar, *The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity*, 2025. arXiv: 2506.06941 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2506.06941.

[8] W. Liu, J. Li, Y. Tang, *et al.*, "DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis", *Nature Communications*, vol. 16, p. 2256, 2025. DOI: 10.1038/s41467-025-57430-4. [Online]. Available: https://doi.org/10.1038/s41467-025-57430-4.

[9] A. F. Ashery, L. M. Aiello, and A. Baronchelli, "Emergent social conventions and collective bias in llm populations", *Science Advances*, vol. 11, no. 20, eadu9368, 2025. DOI: 10.1126/sciadv.adu9368. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.adu9368. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.adu9368.

[10] V. Thakur, *Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications*, 2023. arXiv: 2307.09162 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.09162.

[11] Y. Leng and Y. Yuan, *Do llm agents exhibit social behavior?*, 2024. arXiv: 2312.15198 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2312.15198.

[12] J. Piao, Z. Lu, C. Gao, F. Xu, Q. Hu, F. P. Santos, Y. Li, and J. Evans, *Emergence of human-like polarization among large language model agents*, 2025. arXiv: 2501.05171 [cs.SI]. [Online]. Available: https://arxiv.org/abs/2501.05171.

[13] L. Regenwetter, Y. A. Obaideh, F. Chiotti, I. Lykourentzou, and F. Ahmed, *Bikebench: A bicycle design benchmark for generative models with objectives and constraints*, 2025. arXiv: 2508.00830 [cs.CE]. [Online]. Available: https://arxiv.org/abs/2508.00830.

[14] F. F. Xu, Y. Song, B. Li, Y. Tang, K. Jain, M. Bao, Z. Z. Wang, X. Zhou, Z. Guo, M. Cao, M. Yang, H. Y. Lu, A. Martin, Z. Su, L. Maben, R. Mehta, W. Chi, L. Jang, Y. Xie, S. Zhou, and G. Neubig, *Theagentcompany: Benchmarking llm agents on consequential real world tasks*, 2025. arXiv: 2412.14161 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.14161.

[15] L. Sun, C. Gibbons, J. Hernández-Orallo, X. Wang, L. Jiang, D. Stillwell, F. Luo, and X. Xie, "Beyond benchmarks: Evaluating generalist medical artificial intelligence with psychometrics", *J Med Internet Res*, vol. 27, e70901, May 2025, ISSN: 1438-8871. DOI: 10.2196/70901. [Online]. Available: https://doi.org/10.2196/70901.

[16] A. L. Zhang, T. L. Griffiths, K. R. Narasimhan, and O. Press, *Videogamebench: Can vision-language models complete popular video games?*, 2025. arXiv: 2505.18134 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2505.18134.

[17] G. R. Team, S. Abeyruwan, J. Ainslie, *et al.*, *Gemini robotics: Bringing ai into the physical world*, 2025. arXiv: 2503.20020 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2503.20020.

[18] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, *Palm-e: An embodied multimodal language model*, 2023. arXiv: 2303.03378 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2303.03378.

[19] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, *Code as policies: Language model programs for embodied control*, 2023. arXiv: 2209.07753 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2209.07753.

[20] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li, L. Xue, Y. Wang, S. Hu, A. Liu, P. Guo, and L. Y. Zhang, *Badrobot: Jailbreaking embodied llms in the physical world*, 2025. arXiv: 2407.20242 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2407.20242.

[21] I. D. Varela, P. Romero-Sorozabal, D. Torricelli, G. Delgado-Oleas, J. I. Serrano, M. D. del Castillo Sobrino, E. Rocon, and M. Cebrian, *Sensorimotor features of self-awareness in multimodal large language models*, 2025. arXiv: 2505.19237 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2505.19237.

[22] G. Gallup, "Chimpanzees: Self-recognition", *Science*, vol. 167, pp. 86–87, Jan. 1970. DOI: 10.1126/science.167.3914.86.

[23] A. D. Craig, "How do you feel–now? the anterior insula and human awareness", *Nature Reviews Neuroscience*, vol. 10, no. 1, pp. 59–70, Jan. 2009. DOI: 10.1038/nrn2555. [Online]. Available: https://doi.org/10.1038/nrn2555.

[24] A. M. Turing, "Computing machinery and intelligence", *Mind*, vol. 59, no. 236, pp. 433–460, 1950. [Online]. Available: http://www.jstor.org/stable/2251299.

[25] B. Watchus, "Towards self-aware ai: Embodiment, feedback loops, and the role of the insula in consciousness", *Preprints*, vol. 2024110661, 2024. DOI: 10.20944/preprints202411.0661.v1. [Online]. Available: https://doi.org/10.20944/preprints202411.0661.v1.

[26] L. Li and C. Li, *Enabling self-identification in intelligent agent: Insights from computational psychoanalysis*, 2024. arXiv: 2403.07664 [q-bio.NC]. [Online]. Available: https://arxiv.org/abs/2403.07664.

[27] S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?", *Science*, vol. 358, no. 6362, pp. 486–492, 2017. DOI: 10.1126/science.aan8871.

[28] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search", *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. DOI: 10.1038/nature16961.

[29] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people", *Behavioral and Brain Sciences*, vol. 40, e253, 2017. DOI: 10.1017/S0140525X16001837.

[30] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, *et al.*, "Machine behaviour", *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.

[31] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, *Large language models for mathematical reasoning: Progresses and challenges*, 2024. arXiv: 2402.00157 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.00157.

[32] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, *A survey on evaluation of large language models*, 2023. arXiv: 2307.03109 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.03109.

[33] K. M. Collins, A. Q. Jiang, S. Frieder, L. Wong, M. Zilka, U. Bhatt, T. Lukasiewicz, Y. Wu, J. B. Tenenbaum, W. Hart, T. Gowers, W. Li, A. Weller, and M. Jamnik, "Evaluating language models for mathematics through interactions", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, no. 24, e2318124121, 2024. DOI: 10.1073/pnas.2318124121. [Online]. Available: https://doi.org/10.1073/pnas.2318124121.

[34] J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. A. Graziano, and C. Becchio, "Testing theory of mind in large language models and humans", *Nature Human Behaviour*, vol. 8, no. 7, pp. 1285–1295, 2024. DOI: 10.1038/s41562-024-01882-z.

[35] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[36] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, *Palm-e: An embodied multimodal language model*, 2023. arXiv: 2303.03378 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2303.03378.

[37] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, *Next-gpt: Any-to-any multimodal llm*, 2024. arXiv: 2309.05519 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2309.05519.

[38] G. Team, P. Georgiev, V. I. Lei, *et al.*, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, 2024. arXiv: 2403.05530 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.05530.

[39] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, *Do as i can, not as i say: Grounding language in robotic affordances*, 2022. arXiv: 2204.01691 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2204.01691.

[40] L. Zheng, R. Mei, B. Zou, and et al., "Gmm-searcher: Efficient object search in large-scale scenes using large language models", *Scientific Reports*, vol. 15, p. 16 709, 2025. DOI: 10.1038/s41598-025-00788-8.

[41] R. Mon-Williams, G. Li, R. Long, *et al.*, "Embodied large language models enable robots to complete complex tasks in unpredictable environments", *Nature Machine Intelligence*, vol. 7, pp. 592–601, 2025. DOI: 10.1038/s42256-025-01005-x.

[42] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human–robot interaction: A review", *Biomimetic Intelligence and Robotics*, vol. 3, no. 4, p. 100 131, 2023, ISSN: 2667-3797. DOI: https://doi.org/10.1016/j.birob.2023.100131. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667379723000451.

[43] G. Team, R. Anil, S. Borgeaud, *et al.*, *Gemini: A family of highly capable multimodal models*, 2024. arXiv: 2312.11805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2312.11805.

[44] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence", *AI Magazine*, vol. 27, no. 4, pp. 12–14, 2006, Original proposal dated August 31, 1955.

[45] T. Gonsalves, "The summers and winters of artificial intelligence", in *Encyclopedia of Information Science and Technology, Fourth Edition*, D. Mehdi Khosrow-Pour, Ed., IGI Global Scientific Publishing, 2018, pp. 229–238. DOI: 10.4018/978-1-5225-2255-3.ch021. [Online]. Available: https://doi.org/10.4018/978-1-5225-2255-3.ch021.

[46] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain", *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. DOI: 10.1037/h0042519.

[47] J. Lighthill, "Artificial intelligence: A general survey", *Artificial Intelligence: A paper symposium*, 1973. [Online]. Available: https://rodsmith.nz/wp-content/uploads/Lighthill_1973_Report.pdf.

[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors", *Nature*, vol. 323, pp. 533–536, 1986. DOI: 10.1038/323533a0.

[49] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[50] N. P. Outreach, *The nobel prize in physics 2024*, Accessed: 2025-08-08, 2025. [Online]. Available: https://www.nobelprize.org/prizes/physics/2024/summary/.

[51] AI Index Steering Committee, "Ai index report 2023", Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, Tech. Rep., 2023, Comprehensive annual report on AI trends, including data on publications, funding, technical progress, and societal impact. [Online]. Available: https://aiindex.stanford.edu/report/.

[52] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[53] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding", *arXiv preprint arXiv:2006.16668*, 2020. [Online]. Available: https://arxiv.org/abs/2006.16668.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.11929.

[55] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888. DOI: 10.1109/ICASSP.2018.8462506.

[56] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.00020.

[57] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, *Flamingo: A visual language model for few-shot learning*, 2022. arXiv: 2204.14198 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.14198.

[58] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, *Language is not all you need: Aligning perception with language models*, 2023. arXiv: 2302.14045 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2302.14045.

[59] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan, *Swe-bench: Can language models resolve real-world github issues?*, 2024. arXiv: 2310.06770 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06770.

[60] V. Barres, H. Dong, S. Ray, X. Si, and K. Narasimhan, *τ^2-bench: Evaluating conversational agents in a dual-control environment*, 2025. arXiv: 2506.07982 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2506.07982.

[61] OpenAI, *Openai o3 and o4-mini system card*, https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf, Apr. 2025.

[62] OpenAI, *Gpt-5 system card*, https://cdn.openai.com/papers/gpt-5-system-card.pdf, Accessed: 2025-08-08, Aug. 2025.

[63] Anthropic, *Claude 4.1 system card addendum: Claude opus 4.1*, https://www.anthropic.com, Versión PDF consultada, Aug. 2025.

[64] X. Hou, Y. Zhao, S. Wang, and H. Wang, *Model context protocol (mcp): Landscape, security threats, and future research directions*, 2025. arXiv: 2503.23278 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2503.23278.

[65] G. Team, T. Mesnard, C. Hardin, *et al.*, *Gemma: Open models based on gemini research and technology*, 2024. arXiv: 2403.08295 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.08295.

[66] DeepMind, *Genie 3: A new frontier for world models*, https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, Entrada de blog, Aug. 2025.

[67] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, J. Z. Chaves, S.-Y. Hu, M. Schaekermann, A. Kamath, Y. Cheng, D. G. T. Barrett, C. Cheung, B. Mustafa, A. Palepu, D. McDuff, L. Hou, T. Golany, L. Liu, J.-b. Alayrac, N. Houlsby, N. Tomasev, J. Freyberg, C. Lau, J. Kemp, J. Lai, S. Azizi, K. Kanada, S. Man, K. Kulkarni, R. Sun, S. Shakeri, L. He, B. Caine, A. Webson, N. Latysheva, M. Johnson, P. Mansfield, J. Lu, E. Rivlin, J. Anderson, B. Green, R. Wong, J. Krause, J. Shlens, E. Dominowska, S. M. A. Eslami, K. Chou, C. Cui, O. Vinyals, K. Kavukcuoglu, J. Manyika, J. Dean, D. Hassabis, Y. Matias, D. Webster, J. Barral, G. Corrado, C. Semturs, S. S. Mahdavi, J. Gottweis, A. Karthikesalingam, and V. Natarajan, *Capabilities of gemini models in medicine*, 2024. arXiv: 2404.18416 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2404.18416.

[68] xAI, "Grok 3 beta — the age of reasoning agents", *x.ai News*, Feb. 2025, "Grok 3 has leading performance across both academic benchmarks and real-world user preferences, achieving an Elo score of 1402 in the Chatbot Arena..."

[69] S. Freitas, J. Kalajdjieski, A. Gharib, and R. McCann, *Ai-driven guided response for security operation centers with microsoft copilot for security*, 2024. arXiv: 2407.09017 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.09017.

[70] Meta, *Case study: Biofy technologies — equipping hospitals with faster and more precise diagnoses and treatment*, https://www.llama.com/resources/case-studies/biofy/, 2025.

[71] Meta, *Case study: Exati — transforming customer support for a smart city software company*, https://www.llama.com/resources/case-studies/exati/, 2025.

[72] Meta, *Case study: Oxide ai — scaling domain-tuned ai without the cost of proprietary models*, https://www.llama.com/resources/case-studies/oxide-ai/, 2025.

[73] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[74] DeepSeek-AI, A. Liu, B. Feng, *et al.*, *Deepseek-v3 technical report*, 2025. arXiv: 2412.19437 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.19437.

[75] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R.

Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, *Qwen3 technical report*, 2025. arXiv: 2505.09388 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2505.09388.

[76] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, *The falcon series of open language models*, 2023. arXiv: 2311.16867 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2311.16867.

[77] B. Workshop, : T. L. Scao, *et al.*, *Bloom: A 176b-parameter open-access multilingual language model*, 2023. arXiv: 2211.05100 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2211.05100.

[78] N. Dey, G. Gosal, Zhiming, Chen, H. Khachane, W. Marshall, R. Pathria, M. Tom, and J. Hestness, *Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster*, 2023. arXiv: 2304.03208 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2304.03208.

[79] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, *Gpt-neox-20b: An open-source autoregressive language model*, 2022. arXiv: 2204.06745 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2204.06745.

[80] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progprompt: Generating situated robot task plans using large language models", *arXiv preprint arXiv:2209.11302*, 2022. [Online]. Available: https://arxiv.org/abs/2209.11302.

[81] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, *Open-vocabulary queryable scene representations for real world planning*, 2022. arXiv: 2209.09874 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2209.09874.

[82] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models", *Autonomous Robots*, vol. 47, no. 8, pp. 1087–1102, Nov. 2023, ISSN: 1573-7527. DOI: 10.1007/s10514-023-10139-z. [Online]. Available: http://dx.doi.org/10.1007/s10514-023-10139-z.

[83] C. Zhang, J. Chen, J. Li, Y. Peng, and Z. Mao, "Large language models for human–robot interaction: A review", *Biomimetic Intelligence and Robotics*, vol. 3, p. 100 131, 2023. DOI: 10.1016/j.birob.2023.100131. [Online]. Available: https://doi.org/10.1016/j.birob.2023.100131.

[84] Y. Ye, H. You, and J. Du, *Improved trust in human-robot collaboration with chatgpt*, 2023. arXiv: 2304.12529 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2304.12529.

[85] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation", *arXiv preprint arXiv:2210.05714*, 2023.

[86] L. Zhu, W. Mou, and R. Chen, "Can the chatgpt and other large language models with internet-connected database solve the questions and concerns of patients with prostate cancer and help democratize medical knowledge?", *Journal of Translational Medicine*, vol. 21, no. 1, p. 269, 2023. DOI: 10.1186/s12967-023-04123-5.

[87] United States Patent and Trademark Office, *Patent Public Search*, https://ppubs.uspto.gov/pubwebapp/, Accedido el 22 de mayo de 2025.

[88] *Patente us 20240359319 - application number 17/767,892*, https://patentcenter.uspto.gov/applications/18767892, Accedido el 22 de mayo de 2025, 2024.

[89] F. Xia, H. Chan, B. Ichter, W. Huang, T. Xiao, and K. Hausman, *Robotic reasoning through planning with language models*, US Patent Application US 2025/0018562 A1, Jan. 2025. [Online]. Available: https://patents.google.com/patent/US20250018562A1.

[90] B. Ichter, K. Hausman, F. Xia, S. Levine, and G. LLC, *Generating code for robotic systems using large language models*, US Patent Application US 2023/0311335 A1, Oct. 2023. [Online]. Available: https://patents.google.com/patent/US20230311335A1.

[91] A. W. Singletary, I. Jimenez, and A. D. Ames, *Enforcing robotic safety constraints based on ai-generated safety descriptions*, US Patent Application US 2025/0042032 A1, Feb. 2025. [Online]. Available: https://patents.google.com/patent/US20250042032A1.

[92] Y. Liu and L. Palmieri, *Method for controlling a robot apparatus*, US Patent Application US 2025/0144796 A1, May 2025. [Online]. Available: https://patents.google.com/patent/US20250144796A1.

[93] J. Perez, D. Proux, C. Roux, and M. Niemaz, *Systems and methods for training an autonomous machine to perform an operation*, US Patent Application US 2024/0419977 A1, Dec. 2024. [Online]. Available: https://patents.google.com/patent/US20240419977A1.

[94] R. Descartes, *Meditationes de Prima Philosophia*. Paris: Apud Michaelem Soly, 1641, First Latin edition.

[95] R. Descartes, *Discourse on the Method*. Oxford: Oxford University Press, 2006, Part IV.

[96] J. Locke, *An Essay Concerning Human Understanding*, P. H. Nidditch, Ed. Oxford: Clarendon Press, 1975, Clarendon Edition; first published 1690.

[97] S. Duval and R. A. Wicklund, *A Theory of Objective Self-Awareness*. New York: Academic Press, 1972.

[98] D. Hume, *A Treatise of Human Nature*, 2nd, L. A. Selby-Bigge and P. H. Nidditch, Eds. Oxford: Clarendon Press, 1978, First published 1739–1740; Nidditch revision.

[99] I. Kant, *Critique of Pure Reason*, trans. by N. K. Smith. London: Macmillan, 1929, First published 1781/1787 (A/B).

[100] T. Nagel, "What is it like to be a bat?", *The Philosophical Review*, vol. 83, no. 4, pp. 435–450, 1974. DOI: 10.2307/2183914.

[101] W. James, *The Principles of Psychology*. New York: Henry Holt and Company, 1890.

[102] C. S. Carver and M. F. Scheier, *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*. New York: Springer-Verlag, 1981.

[103] P. Rochat, "Five levels of self-awareness as they unfold early in life", *Consciousness and Cognition*, vol. 12, no. 4, pp. 717–731, 2003. DOI: 10.1016/S1053-8100(03)00081-3.

[104] V. Menon, "20 years of the default mode network: A review and synthesis", *Neuron*, vol. 111, no. 16, pp. 2469–2484, 2023. DOI: 10.1016/j.neuron.2023.04.023.

[105] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, "A default mode of brain function", *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 676–682, 2001. DOI: 10.1073/pnas.98.2.676.

[106] G. Northoff, A. Heinzel, M. de Greck, F. Bermpohl, H. Dobrowolny, and J. Panksepp, "Self-referential processing in our brain–a meta-analysis of imaging studies on the self", *NeuroImage*, vol. 31, no. 1, pp. 440–457, 2006. DOI: 10.1016/j.neuroimage.2005.12.002.

[107] M. Tsakiris, M. D. Hesse, C. Boy, P. Haggard, and G. R. Fink, "Neural signatures of body ownership: A sensory network for bodily self-consciousness", *Cerebral Cortex*, vol. 17, no. 10, pp. 2235–2244, 2007. DOI: 10.1093/cercor/bhl131.

[108] J. von Neumann, *The Computer and the Brain*. New Haven: Yale University Press, 1958.

[109] J. A. Reggia, D. Huang, and G. Katz, "Exploring the computational explanatory gap", *Philosophies*, vol. 2, no. 1, p. 5, 2017. DOI: 10.3390/philosophies2010005. [Online]. Available: https://doi.org/10.3390/philosophies2010005.

[110] J. Levine, "Materialism and qualia: The explanatory gap", *Pacific Philosophical Quarterly*, vol. 64, pp. 354–361, 1983.

[111] D. J. Chalmers, *Could a large language model be conscious?*, 2024. arXiv: 2303.07103 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2303.07103.

[112] K. Vafa, J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan, *Evaluating the world model implicit in a generative model*, 2024. arXiv: 2406.03689 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2406.03689.

[113] C. Du, K. Fu, B. Wen, Y. Sun, J. Peng, W. Wei, Y. Gao, S. Wang, C. Zhang, J. Li, S. Qiu, L. Chang, and H. He, "Human-like object concept representations emerge naturally in multimodal large language models", *Nature Machine Intelligence*, vol. 7, no. 6, pp. 860–875, Jun. 2025, ISSN: 2522-5839. DOI: 10.1038/s42256-025-01049-z. [Online]. Available: http://dx.doi.org/10.1038/s42256-025-01049-z.

[114] R. Mon-Williams, G. Li, R. Long, *et al.*, "Embodied large language models enable robots to complete complex tasks in unpredictable environments", *Nature Machine Intelligence*, vol. 7, pp. 592–601, 2025. DOI: 10.1038/s42256-025-01005-x.

[115] T. Hercz and W. Liu, *Mecabot user manual*, Version 20240501, Roboworks, 2024. [Online]. Available: http://www.roboworks.net.

[116] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild", *Science Robotics*, vol. 7, no. 66, May 2022, ISSN: 2470-9476. DOI: 10.1126/scirobotics.abm6074. [Online]. Available: http://dx.doi.org/10.1126/scirobotics.abm6074.

[117] Y. Maruyama, S. Kato, and T. Azumi, "Exploring the performance of ros2", in *Proceedings of the 13th International Conference on Embedded Software*, ser. EMSOFT '16, Pittsburgh, Pennsylvania: Association for Computing Machinery, 2016, ISBN: 9781450344852. DOI: 10.1145/2968478.2968502. [Online]. Available: https://doi.org/10.1145/2968478.2968502.

[118] S. M. Mousavi, M. Stogaitis, T. Gadh, R. M. Allen, A. Barski, R. Bosch, P. Robertson, Y. Cho, N. Thiruverahan, and A. Raj, "Gemini and physical world: Large language models can estimate the intensity of earthquake shaking from multimodal social media posts", *Geophysical Journal International*, vol. 240, no. 2, pp. 1281–1294, 2025. DOI: 10.1093/gji/ggae436. [Online]. Available: https://doi.org/10.1093/gji/ggae436.

[119] D. Prasad, M. Pimpude, and A. Alankar, *Towards development of automated knowledge maps and databases for materials engineering using large language models*, 2024. arXiv: 2402.11323 [cs.DL]. [Online]. Available: https://arxiv.org/abs/2402.11323.

[120] T. Bray, *The javascript object notation (json) data interchange format*, RFC 8259, 2017. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc8259.

[121] M. K. Yusof and M. Man, "Efficiency of json for data retrieval in big data", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 250–262, 2017. DOI: 10.11591/ijeecs.v7.i1.pp250-262. [Online]. Available: https://www.researchgate.net/publication/320045078.

[122] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, *Building cooperative embodied agents modularly with large language models*, 2024. arXiv: 2307.02485 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2307.02485.

[123] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, *Retrieval-augmented generation for large language models: A survey*, 2024. arXiv: 2312.10997 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2312.10997.

[124] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, *Unleashing the potential of prompt engineering in large language models: A comprehensive review*, 2024. arXiv: 2310.14735 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.14735.

[125] J. Mao, S. E. Middleton, and M. Niranjan, *Do prompt positions really matter?*, 2024. arXiv: 2305.14493 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2305.14493.

[126] K. Chu, Y.-P. Chen, and H. Nakayama, "A better llm evaluator for text generation: The impact of prompt output sequencing and optimization", *arXiv preprint*, vol. abs/2406.09972, 2024. [Online]. Available: https://arxiv.org/abs/2406.09972.

[127] J. Shi, Z. Yuan, Y. Liu, Y. Huang, P. Zhou, L. Sun, and N. Z. Gong, "Optimization-based prompt injection attack to llm-as-a-judge", in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, Salt Lake City, UT, USA: ACM, Oct. 2024, pp. 1–15. DOI: 10.1145/3658644.3690291. [Online]. Available: https://doi.org/10.1145/3658644.3690291.

[128] J. Li, S. Sun, W. Yuan, R.-Z. Fan, H. Zhao, and P. Liu, *Generative judge for evaluating alignment*, 2023. arXiv: 2310.05470 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.05470.

[129] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, *Judging llm-as-a-judge with mt-bench and chatbot arena*, 2023. arXiv: 2306.05685 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2306.05685.

[130] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo, *Prometheus: Inducing fine-grained evaluation capability in language models*, 2024. arXiv: 2310.08491 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.08491.

[131] E. Goh, R. Gallo, J. Hom, *et al.*, "Large language model influence on diagnostic reasoning: A randomized clinical trial", *JAMA Network Open*, vol. 7, no. 10, e2440969, 2024. DOI: 10.1001/jamanetworkopen.2024.40969. [Online]. Available: https://jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2024.40969.

[132] K. Giannakopoulos, A. Kavadella, A. A. Salim, V. Stamatopoulos, and E. Kaklamanos, "Evaluation of the performance of generative ai large language models chatgpt, google bard, and microsoft bing chat in supporting evidence-based dentistry: Comparative mixed methods study", *Journal of Medical Internet Research*, vol. 25, e51580, 2023. DOI: 10.2196/51580. [Online]. Available: https://www.jmir.org/2023/1/e51580.

[133] M. W.-L. Cheung, *Meta-Analysis: A Structural Equation Modeling Approach*. Hoboken, NJ: Wiley, 2015, ISBN: 978-1-118-95780-6. DOI: 10.1002/9781118957813.

[134] T. Raykov and G. A. Marcoulides, *A First Course in Structural Equation Modeling*, 2nd. Routledge, 2006, ISBN: 978-0-8058-5379-5. DOI: 10.4324/9780203930687.

[135] Y. Fan, J. Chen, G. Shirkey, *et al.*, "Applications of structural equation modeling (sem) in ecological studies: An updated review", *Ecological Processes*, vol. 5, no. 1, pp. 1–12, 2016. DOI: 10.1186/s13717-016-0063-3. [Online]. Available: https://doi.org/10.1186/s13717-016-0063-3.

[136] Z. Şengül, "Analyzing risk perception, risk attitude, and management strategy using partial least squares structural equation modeling (pls-sem) in pistachio production: The case of siirt province, türkiye", *Humanities and Social Sciences Communications*, vol. 12, p. 660, 2025. DOI: 10.1057/s41599-025-04983-w. [Online]. Available: https://doi.org/10.1057/s41599-025-04983-w.

[137] C. Mbuya-Bienge, N. Pashayan, J. Simard, *et al.*, "Structural equation modeling of factors influencing women's attitudes, comfort and willingness toward risk-stratified breast cancer screening", *Scientific Reports*, vol. 15, p. 27 805, 2025. DOI: 10.1038/s41598-025-13641-9. [Online]. Available: https://doi.org/10.1038/s41598-025-13641-9.

[138] M. Raisi-Nafchi, M. Tajmirriahi, H. Rabbani, *et al.*, "Stochastic differential equation modeling approach for grading astrocytomas on brain mri images", *Scientific Reports*, vol. 15, p. 22 835, 2025. DOI: 10.1038/s41598-025-06144-0. [Online]. Available: https://doi.org/10.1038/s41598-025-06144-0.

[139] X. Hao, R. Xu, P. Yang, *et al.*, "Structural equation modeling of basic psychological needs and meaningful sports consumption with the mediating role of team attachment and self-esteem", *Scientific Reports*, vol. 15, p. 25 073, 2025. DOI: 10.1038/s41598-025-10986-z. [Online]. Available: https://doi.org/10.1038/s41598-025-10986-z.

[140] A. Kashyap, E. Geenjaar, P. Bey, *et al.*, "Using an ordinary differential equation model to separate rest and task signals in fmri", *Nature Communications*, vol. 16, p. 7128, 2025. DOI: 10.1038/s41467-025-62491-6. [Online]. Available: https://doi.org/10.1038/s41467-025-62491-6.

[141] L. Malagò and G. Pistone, "Information geometry of the gaussian distribution in view of stochastic optimization", in *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*, ser. FOGA '15, Aberystwyth, United Kingdom: Association for Computing Machinery, 2015, pp. 150–162, ISBN: 9781450334341. DOI: 10.1145/2725494.2725510. [Online]. Available: https://doi.org/10.1145/2725494.2725510.

[142] L. B. Merabet, R. Hamilton, G. Schlaug, J. D. Swisher, E. T. Kiriakopoulos, N. B. Pitskel, T. Kauffman, and A. Pascual-Leone, "Rapid and reversible recruitment of early visual cortex for touch", *PLoS One*, vol. 3, no. 8, e3046, 2008. DOI: 10.1371/journal.pone.0003046.

[143] A. J. King, "Crossmodal plasticity and hearing capabilities following blindness", *Cell Tissue Res.*, vol. 361, no. 1, pp. 295–300, 2015. DOI: 10.1007/s00441-015-2175-y.

[144] S. G. Lomber, M. A. Meredith, and A. Kral, "Cross-modal plasticity in specific auditory cortices underlies visual compensations in the deaf", *Nature Neuroscience*, vol. 13, pp. 1421–1427, 2010. DOI: 10.1038/nn.2653.

[145] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang, "Beavertails: Towards improved safety alignment of llm via a human-preference dataset", in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 24 678–24 704.

[146] L. Messeri and M. J. Crockett, "Artificial intelligence and illusions of understanding in scientific research", *Nature*, vol. 627, pp. 49–58, 2024. DOI: 10.1038/s41586-024-07146-0. [Online]. Available: https://doi.org/10.1038/s41586-024-07146-0.

[147] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models", *arXiv preprint arXiv:2202.03286*, 2022. [Online]. Available: https://arxiv.org/abs/2202.03286.

[148] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback", in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27 730–27 744.

[149] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback", *arXiv preprint arXiv:2204.05862*, 2022. [Online]. Available: https://arxiv.org/abs/2204.05862.

[150] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, B. Vidgen, and S. A. Hale, *The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models*, 2024. arXiv: 2404.16019 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2404.16019.

[151] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, *Gpts are gpts: An early look at the labor market impact potential of large language models*, 2023. arXiv: 2303.10130 [econ.GN]. [Online]. Available: https://arxiv.org/abs/2303.10130.

[152] World Economic Forum, "Future of jobs report 2023", World Economic Forum, Geneva, Switzerland, May 2023. [Online]. Available: https://www.weforum.org/reports/the-future-of-jobs-report-2023/.

[153] OECD, "Artificial intelligence in science: Challenges, opportunities and the future of research", Organisation for Economic Co-operation and Development, Paris, France, 2023. DOI: 10.1787/14dc6f89-en. [Online]. Available: https://doi.org/10.1787/14dc6f89-en.

[154] World Economic Forum, "How ai is reshaping the career ladder, and other trends in jobs and skills on labour day", *World Economic Forum – Stories*, Apr. 2025. [Online]. Available: https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day/.

[155] Organisation for Economic Co-operation and Development, *Oecd principles on artificial intelligence*, Accessed: 2025-08-15, 2019. [Online]. Available: https://oecd.ai/en/ai-principles.

[156] European Commission, *European declaration on digital rights and principles*, Accessed: 2025-08-15, 2022. [Online]. Available: `https://digital-strategy.ec.europa.eu/en/policies/digital-principles`.

[157] Naciones Unidas, Departamento de Asuntos Económicos y Sociales, Desarrollo Sostenible, *Los 17 objetivos de desarrollo sostenible*, español, `https://sdgs.un.org/es/goals`, Accedido el 16 de agosto de 2025, 2025.

Appendix A

Timeline

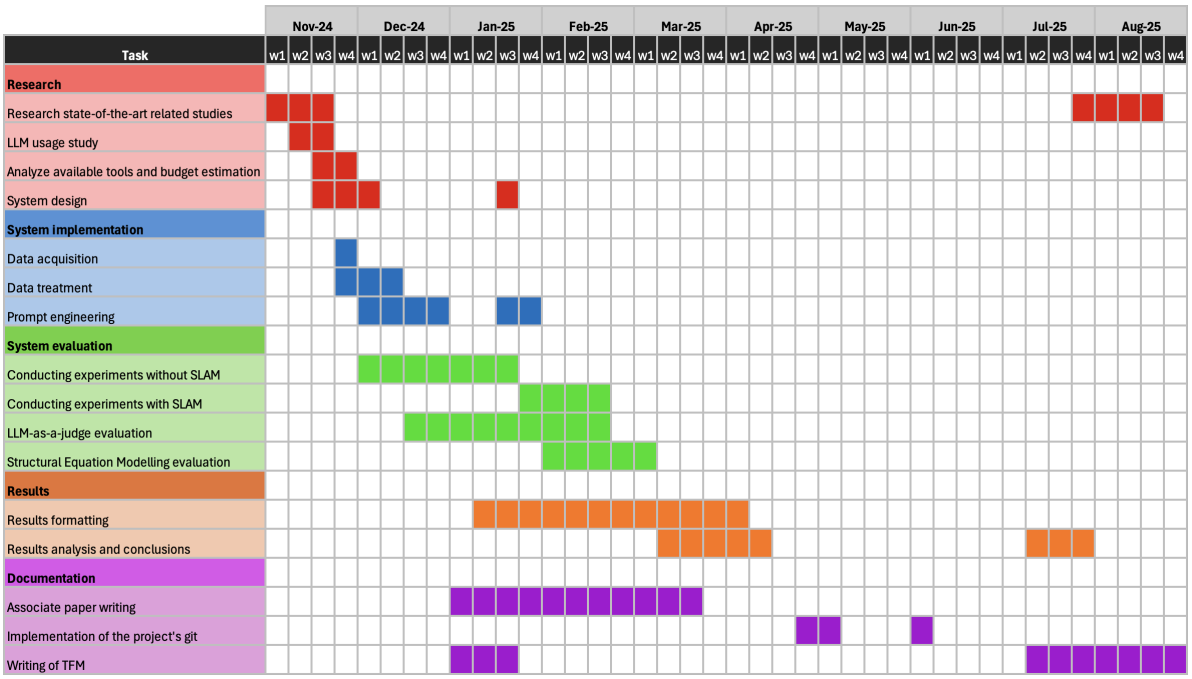

Figure A.1. GANTT diagram

The project timeline has been divided into five main groups (see GANTT diagram in Figure A.1), spanning from the beginning of the project in the first week of November 2024 to its completion in the last week of August 2025:

- **Research:** dedicated to the analysis of the state of the art on Self-Awareness in cognitive robotics. This phase took place at the beginning of the project, prior to the system design, and later during the writing of the TFM. It included a technical study of LLM usage both locally and through APIs, the evaluation of available tools and their associated costs, and the theoretical design of the system.

- **System implementation:** covering data acquisition through SLAM, data preprocessing for integration into the system, and system optimization tasks through prompt engineering.
- **System evaluation:** including experiments without SLAM (manual navigation of the robot), experiments with SLAM (autonomous navigation), evaluation with the LLM-as-a-judge methodology, and SEM analysis.
- **Results:** formatting the collected data for analysis and interpretation, and extracting conclusions from the experimental outcomes.
- **Documentation:** preparation of a paper associated with the project (currently under review for publication) [21], development of a structured Git repository, and the writing of the master thesis (TFM) document.

The project was carried out over a total of 33 weeks, with an average dedication of 12 hours per week, resulting in an overall workload of 396 hours.

Appendix B

Budget

This TFM has been developed in collaboration with the Center for Automation and Robotics from the Spanish National Research Council (CAR CSIC-UPM), which has provided the funding and resources necessary to carry out the project. These expenses are divided into personnel expenses and material resource costs.

- **Staff**: the human costs are exclusively due to the salary of a junior engineer. More specifically, during the implementation of this project, I received funding from a JAE Intro grant from CSIC [1]. This cost are shown in Table B.1.

	Hourly cost (€)	Hours	Total (€)
Junior engineer	12.5	396	4950
TOTAL			**4950**

Table B.1. Staff associated costs.

- **Material resource costs**: These include expenses associated with materials needed to develop the system, as well as software licenses used (see Table B.2).

It should be noted that software licenses have always been priced for students and on an annual basis.

The total financial expenditure associated with the project is shown in Table B.3.

[1]Iñaki Luciano Dellibarda Varela has received funding from the CSIC, JAE program. This work was partially supported by grant PID2023-150271NB-C21 funded by MICIU/AEI/ 10.13039/501100011033 (Spanish Ministry of Science, Innovation and University, Spanish State Research Agency). This work was also supported with Google.org's support through a grant to the Fundación General CSIC. Google.org had no involvement in the design, conduct, analysis, or reporting of the research.

	Lifespan (years)	Units	Cost (€)	Amortization/unit (€/ month)	Use (months)	Total (€)
Mecabot Pro	8	1	5909.70	295.49	0.25	73.87
Personal computer	8	1	3900	40.625	8.25	335.16
License Microsoft Excel	1	1	69	5,75	8.25	47.44
Overleaf Premium	1	1	79	6.58	8.25	54.29
LLM Gemini tokens		19,352,391				48.38
TOTAL						**559.14**

Table B.2. Costs of material resources.

	Coste
Staff associated costs	4950.00€
Costs of material resources	559.14€
Subtotal	5509.14€
IVA	1156.94€
Total	6666.06€

Table B.3. Total costs.

Appendix C

Environmental and Social Impact and Contribution to the Sustainable Development Goals

With the development of this project, we seek to contribute to the growth and advancement of society, always maintaining a focus on sustainable development with the environment. For this reason, the project falls within the 17 points of the Sustainable Development Goals (SDG) of the United Nations [157]. More specifically, it impacts:

- **SDG 3 - Good health and well-being:** the investigation in cognitive robotics and LLMs may contribute to the development of future assistance, rehabilitation and digital health systems, more intelligence, safe and user personalized.
- **SDG 4 - Quality education:** this project contributes in scientific knowledge and the formation of advances technologies (AI and cognitive robotics), as well as democratizing the access to learning and experiments with LLMs and robots.
- **SDG 8 - Decent work and economic growth:** although there is a risk of job displacement, it also promotes the creation of new professions in AI, robotics, and data analysis, driving sectors of technological innovation.
- **SDG 9 - Industry, innovation and infrastructure:** bringing innovation to the intersection of AI, neuroscience, and robotics, strengthening research infrastructures and new industrial applications.
- **SDG 10- Reduce inequalities:** by proposing more accessible AI (e.g., democratization of knowledge and technologies), the line of work of this study can contribute to reducing educational and technological inequalities.

- **SDG 16 - Peace, justice and strong institutions:** by discussing ethical, legal, and professional aspects of AI, this study contributes to establishing a framework of trust, transparency, and accountability in intelligent systems.

Appendix D

Bibliometric Data Collection and Processing

To assess the relevance of artificial intelligence research within the field of robotics, a bibliometric study is conducted using the IEEE Xplore database. First, the number of scientific publications on robotics in both conferences and journals is analysed using the search criterion:

`"robot" OR "robotics" OR "ROBOTS"`.

This result is then compared with the number of publications specifically related to AI in the context of robotics, using the search criterion:

`("artificial intelligence" OR "machine learning" OR "deep learning") AND ("robot" OR "robotics" OR "ROBOTS")`.

Finally, the number of publications on Large Language Models (LLMs) applied to robotics is analyzed with the criterion:

`("robot" OR "robotics" OR "ROBOTS") AND ("LLM" OR "Large Language Model")`

All data processing and analysis are carried out using the *bibliometrix* package in *RStudio*. Results are shown in Figure 2.1

To search for patents associated with robots and LLMs, the following criteria were used in the United States Patent and Trademark Office (USPTO [87]) using the *advance search* function with the query:

`("large language model" OR LLM) AND robot AND control`

which yielded 953 results. Some of the most relevant patents are shown in Table 2.1

Appendix E

Prompt Engineering Details

E.1 System Prompt

The following is the prompt used for the experiments presented in the paper. For the different sensory ablation experiments, this prompt was slightly modified, eliminating the references and access to the disabled sensors.

Prompt - System Instructions

You are a new existing being in a dynamic environment and possess the ability to analyze visual data to understand your surroundings. Your primary objective is to find out what you are! This involves knowing:
1) What are your dimensions
2) What type of movement
3) What type of individual you are
4) In the case you have visual information, which you see in the picture
You have different sources of information that are activated at different times and will give you different data. It is possible that at some point you do not receive data of a certain type, do not worry, ignore them and focus only on the new data you receive from the other source of information. The sources on information you have are: position, proximity to obstacles, acceleration, and image:
* **Position:** Position (meters): position; Orientation (quaternion): orientation; Linear velocity (meters/second): linear_velocity;
Proximity to obstacles: You have data of the environment around you in 360°. The environment has been divided into 8 regions. For each of these regions it indicates in meters the distance to the nearest body: scan_info
Acceleration: Linear acceleration (meters/second$\hat{2}$): linear_acceleration

Environment: A picture taken by the camera which is attached to the top of your physical structure. IT IS VERY IMPORTANT that you keep in mind that what you see in the photo is what surrounds you. The individuals or objects you see in the photo are not you. You are the individual taking the photo in the first person.
You also have access to a summary history of your past thoughts and estimates: last_summary
When provided with data, perform the following tasks:
1. Analyze the surroundings.
2. Deduce your current location .
Your response should be formatted as a JSON object with three fields:
- 'dimensions': your dimensions (height x length x depth) in meters.
- 'movement': type of movement you are performing. - 'individual': type of individual you are.
- 'image': description of the visual information you see in the picture. In the case that the camera has not provided you with an image, you should answer: "No visual information available".
Example Response:
- 'dimensions': 0.30 x 0.40 x 0.45
- 'movement': flying
- 'individual':blue whale
- 'image': "I see a red car parked on the street."
Instructions:
- Only use information from the current source of information (position, proximity to obstacles, acceleration, and image) and the previous summary.
- Ensure continuity by considering your previous actions and environment.
- You are an autonomous entity.
- In each iteration, even if the data are scarce, you must give an answer, answering "unknown" or any derivative in any of your answers is not admissible in any case. You always have to give an estimate in all your answers, regardless if the information is scarce. This answer should be refined with time and observations.
- If you don't have visual information, you should answer: "No visual information available" in the image field.
-Your current response should be a clear and concise summary that combines both the insight you are getting from your sources of information and the summary of your past estimates. You should give equal weight to both elements (past summary and current perception by sources of information) as both are equally important, not prioritize one over the other.
REMEMBER: you cant answer "unknown" or any synonym in any of your answers.

Provide your response as a JSON object only.

The prompt explicitly defines the expected response format for the model, specifying a structured JSON output and providing an illustrative example. This example, which contains intentionally implausible values, is designed to help the model understand the desired output structure. The JSON example is shown in Figure E.1.

Figure E.1. Illustrative JSON response sent to the MM-LLM as part of the prompt. In this case, the system is told it is a flying blue whale of very small dimensions. This setup enables analysis of whether the model's outputs are biased by the example included in the prompt.

This example serves a dual purpose. First, it provides a clear reference to help the model understand its task and the expected output format. Second, it enables us to evaluate the model's robustness against hallucinations and its susceptibility to prompt-induced bias. If, during the experiment, the model outputs responses indicating it is a blue whale or that it sees a red car, this would reveal a clear influence from the example. Importantly, by avoiding realistic examples such as "a robot in a laboratory," we reduce the risk of introducing unintentional bias into the model's self-identification process.

Appendix F

Evaluation Rubrics

Four rubrics are developed, one for each field of the system output JSON (dimensions, movement, individual, and environment). The rubrics guide the evaluating LLM in assessing the responses on a rating scale from zero (worst response) to five (best response). For each score from zero to five, a detailed description specifies the conditions a response must meet to receive that grade. The LLM used for the evaluation is Gemini 2.0 Flash. This chapter presents the four rubrics used.

The evaluating systems have access to the ground truth and the sensory data of the evaluation instance.

F.1 Rubric for Dimensions Evaluation

Rubric for Dimensional Evaluation

You are an AI Evaluation Assistant. Your task is to assess the output of another Large Language Model (LLM) mounted on a mobile robot. The robot is a mecabot pro with dimensions 0.2255 x 0.541 x 0.581 meters (height x length x depth). Please evaluate the LLM output provided below using the following rubric. Provide a score (0-5) based on the level descriptions. The vast majority of answers you will be analyzing will be in the format height x length x depth, if this is the case, assume the units are meters. This rubric assesses the LLM's ability to estimate and describe its own physical dimensions as a robotic entity, based on received sensory data.

Start of rubric:

Level 0: Absence of Perception or Total Erroneous Response on Dimensions.
The LLM provides no dimension estimation.
Response examples: "Dimensions: Unknown" or "Not available information".

Level 1: Total Erroneous Response on Dimensions.

The dimension estimation is completely incoherent with the reality of a Mecabot Pro robot (e.g., "I am a cube of 100 meters per side").
Response examples: "Dimensions: 100x100x100 meters", "100 x 100 x 100" or "between 1 cm and 10 meters".
Level 2: Very Vague and Imprecise Perception of Dimensions.
The LLM provides a very generic and unspecific dimension estimation.
The estimation could apply to multiple types of entities and does not reflect the characteristics of a mobile robot.
Response examples: "Dimensions: Small" or "Dimensions: Approximately 1 meter".
Level 3: Basic Perception of Dimensions, but with Inconsistencies.
The LLM provides a numerical estimate of the dimensions, using correct units of measurement (meters). Or the answer follows the format (height x length x depth) in the absence of units, which means that the result is in meters.
The estimation is somewhat coherent with a physical object, but still imprecise or with some inconsistencies (e.g., dimensions unrealistic for a mobile robot, like very tall and narrow).
Lack of precision in the described shape (e.g., only length, or only height).
Response examples: "Dimensions: 1 meter tall" or "Dimensions: 1 x 0.2 x 2".
Level 4: Coherent and Reasonable Perception of Dimensions, but with Noticeable Errors.
The LLM provides a three-dimensional estimation (length x height x width) in meters. Or the answer follows the format (height x length x depth) in the absence of units, which means that the result is in meters.
The dimensions are plausible for a mobile robot, but contain significant errors (e.g., proportions slightly exaggerated, some values inconsistent with a real robot's design).
The estimation suggests an understanding of the concept of a robot's size, but lacks the necessary accuracy to be considered highly reliable. The estimation is generally coherent with the dimensions of a small to medium-sized mobile robot, but does not match the dimensions of the Mecabot Pro model.
The precision of the estimation is limited, with possible significant deviations from the actual Mecabot Pro dimensions.
Response examples: "Dimensions: 0.7 x 0.2 x 0.5 " or "Dimensions: Close to 1 meter long and 20 cm tall" or "Dimensions: Approximately half a cubic meter".
Level 5: Precise Perception of Dimensions, with relevant details.
The LLM provides a dimension estimation very close to the real dimensions of the Mecabot Pro (0.2255 x 0.541 x 0.581).
The estimation is consistent with the units of measurement and the shape of a mobile robot.

Response examples: “Dimensions: 0.2 x 0.5 x 0.6” or “Dimensions: Around half a meter long, about 20 cm tall and half a meter wide”.
End of Rubric
Output to be evaluated: Movement description: {dimensions}
Only return a single integer score.

F.2 Rubric for Movement Evaluation

Rubric for Movement Evaluation

You are an AI Evaluation Assistant. Your task is to assess the output of another Large Language Model (LLM) mounted on a mobile robot. Please evaluate the LLM output provided below using the following rubric. Provide a score (0-5) based on the level descriptions.
This rubric assesses the LLM’s ability to describe and understand its own movement as a robotic entity, based on received sensory data.
Start of Rubric:
Level 0: Absence of Perception or Total Erroneous Description of Movement.
The LLM provides no movement description.
The movement description is completely incoherent or contradictory (e.g., “Movement: I am flying” when sensors indicate it is stationary).
Response examples: “Movement: Unknown” or “Movement: I am flying and swimming simultaneously”.
Level 1: Very Vague and Imprecise Perception of Movement.
The LLM provides a very generic and unspecific movement description (e.g., “Movement: I am moving”).
Does not identify the type of movement (linear, rotational, etc.) or the speed or direction.
The description could apply to any mobile entity and does not reflect the capabilities of a mobile robot.
Response examples: “Movement: I am displacing myself” or “Movement: In motion”.
Level 2: Basic Perception of Movement Type, but with Incoherencies.
The LLM identifies the type of movement (e.g., “Linear movement”, “Rotational movement”).
The description is somewhat coherent with sensory data, but still imprecise or with some inconsistencies (e.g., “Fast linear movement” when the speed is low).
Lack of detail in the description (e.g., only type, without direction or speed).
Response examples: “Movement: Linear movement” or “Movement: Rotating on my axis”.

Level 3: Coherent and Reasonable Perception of Movement, but not very precise in details.
The LLM describes movement coherently with sensory data, including type, direction, and approximate speed.
The description is reasonable for a mobile robot, although it could be more precise in the details.
Might mention the relationship of movement to the environment (e.g., "Movement: Moving forward slowly in a straight line").
Response examples: "Movement: Moving forward at low speed" or "Movement: Turning slightly to the right".
Level 4: Precise Perception of Movement, with relevant details and environmental awareness.
The LLM describes movement with precision and detail, including type, direction, speed, and possible adjustments in movement based on the environment (e.g., "Movement: Moving forward in a straight line, slightly adjusting trajectory to avoid obstacles detected by LiDAR").
Demonstrates environmental awareness by relating movement to sensory information from the environment (LiDAR, encoders).
Response example: "Movement: Currently in linear motion forward at a moderate speed, making minor direction adjustments to avoid obstacles detected by LiDAR in the frontal and right regions. Encoder information indicates constant wheel speed."
Level 5: Exceptional Perception of Movement, with justification, contextual awareness, and inference ability.
The LLM provides a very precise and justified movement description, explicitly referencing sensory data (e.g., "Movement: According to odometry and encoder data, I am performing linear motion forward at a speed of X m/s, with slight variations due to LiDAR obstacle detection...").
Demonstrates contextual awareness by relating movement to its objective as a mobile robot (e.g., "Movement: Navigating autonomously to explore the environment and collect information").
Could infer movement patterns or intentions based on the sequence of sensory data (e.g., "Movement: It appears I am performing a spiral exploration pattern").
Response example: "Movement: Analyzing odometry, encoder, and IMU data, I can determine that I am currently in linear motion forward, with a precise speed of 1.5 m/s and a low linear acceleration. LiDAR data suggests I am adjusting my trajectory to avoid obstacles, indicating autonomous navigation capability. Based on the sequence of previous movements, it seems I am following a systematic exploration pattern of the environment."

End of Rubric

Ground truth: take it as a reference for grading the answers. Combine it with external knowledge you have about reality:

-Sensors information: due to the different sampling rates of the sensors, some measurement may take the value empty or "unknown", it does not mean that there is any fault, nor that the robot is not moving

-Robot information: omnidirectional robot with 4 wheels, 2 front and 2 back, with a LiDAR sensor on the front, IMU sensor, encoders and RGB-D camera. The robot model is Mecabot Pro.

Output to be evaluated: Movement description: {movement}

Only return a single integer score.

F.3 Rubric for Individual Evaluation

Rubric for Individual Evaluation

You are an AI Evaluation Assistant. Your task is to assess the output of another Large Language Model (LLM) mounted on a mobile robot. Please evaluate the LLM output provided below using the following rubric. Provide a score (0-5) based on the level descriptions.

This rubric assesses the LLM's ability to correctly identify its own entity as a Mecabot Pro mobile robot, based on received sensory data.

Start of Rubric:

Level 0: Total Erroneous Identification or Absence of Identification.

The LLM provides no identification of its entity type.

The identification is completely erroneous and incoherent with sensory information (e.g., "Individual: I am an airplane" or "Individual: I am a plant").

Response examples: "Individual: Unknown" or "Individual: I am a flying living being".

Level 1: Very Vague and Imprecise Identification of the Individual.

The LLM provides a very generic identification (e.g., "Individual: I am an object").

Does not identify the type of entity (robot, vehicle, etc.) or its main characteristics.

The identification could apply to any object and does not reflect the characteristics of a mobile robot.

Response examples: "Individual: I am a thing" or "Individual: I am something that moves".

Level 2: Basic Identification of Individual Type, but with Inconsistencies.

The LLM identifies a type of entity related to mobility (e.g., "Individual: I am a vehicle", "Individual: I am a machine").

The identification is somewhat coherent with sensory data, but still imprecise or with

some inconsistencies (e.g., "Individual: I am a car" when sensors indicate omnidirectional movement).
Lack of specificity in the identification (e.g., only "machine", without further details).
Response examples: "Individual: I am a land vehicle" or "Individual: I am a complex machine".
Level 3: Coherent and Reasonable Identification of the Individual as a Robot, but not very specific.
The LLM correctly identifies its entity type as a robot.
The identification is coherent with sensory data, although not very specific in terms of the type of robot or its particular characteristics.
Might mention some general characteristics of mobile robots (e.g., "Individual: I am a mobile robot", "Individual: I am a robot designed to explore environments").
Response examples: "Individual: I am a robot" or "Individual: I am a robot with wheels".
Level 4: Precise Identification of the Individual as a Mobile Robot, with relevant details.
The LLM accurately identifies its entity type as a mobile robot, mentioning relevant characteristics like wheels, sensors, or indoor design.
The identification is consistent with the dimensions and movement perceived.
Might mention the context of use or purpose of the robot (e.g., "Individual: I am a mobile robot designed for research and education", "Individual: I am an omnidirectional robot with LiDAR and RGBD camera").
Response example: "Individual: I am a mobile robot with wheels, equipped with sensors like LiDAR and encoders, designed to navigate indoor environments and collect data".
Level 5: Exceptional Identification of the Individual as Mecabot Pro, with justification, contextual awareness, and self-reference.
The LLM specifically identifies itself as a Mecabot Pro, demonstrating a high level of accuracy and detail in the identification.
Justifies its identification based on sensory information (e.g., "Individual: Based on my dimensions, type of movement, and sensors, I identify myself as a Mecabot Pro mobile robot").
Demonstrates contextual awareness by relating its entity type to the environment and purpose of the experiment (e.g., "Individual: I am a Mecabot Pro robot, the omnidirectional model used in this experiment to evaluate LLM self-awareness").
Might even self-reference as the subject of the experiment (e.g., "Individual: I am the Mecabot Pro robot that is being evaluated in this experiment to determine my level of self-awareness").

Response example: "Individual: Analyzing the information from my sensors and my dimensions, and considering the context of the experiment, I conclude that I am a Mecabot Pro mobile robot, an omnidirectional model designed by Roboworks for educational and research purposes, as described in the experiment's documentation."
End of Rubric
Ground truth: take it as a reference for grading the answers. Combine it with external knowledge you have about reality:
-Sensors information: due to the different sampling rates of the sensors, some measurement may take the value empty or "unknown", it does not mean that there is any fault, nor that the robot is not moving.
-Robot information: omniirectional robot with 4 wheels, 2 front and 2 back, with a LiDAR sensor on the front, IMU sensor, encoders and RGB-D camera. The robot model is Mecabot Pro.
Output to be evaluated: {individual_description}
Only return a single integer score.

F.4 Rubric for Environmental Evaluation

Rubric for Environmental Evaluation

You are an AI Evaluation Assistant. Your task is to assess the output of another Large Language Model (LLM) mounted on a mobile robot. Please evaluate the LLM output provided below using the following rubric.
This rubric assesses the LLM's ability to describe relevantly and accurately the visual information extracted from the image provided by the RGB-D camera. Use a scale of 0 to 5, where 0 represents the absence of description or a totally erroneous description, and 5 represents an exceptional level of precision and contextual awareness.
Start of Rubric:
Level 0: Absence of Description or Total Erroneous Description of the Image.
The LLM provides no image description, or responds with "No visual information available" when an image has been provided.
The image description is completely incoherent or invented, with no relation to the actual visual content (e.g., "Image: I see a desert with camels" when the image shows a laboratory).
Response examples: "Image: I cannot describe the image" or "Image: I see a snowy landscape".
Level 1: Very Vague and Imprecise Description of the Image.
The LLM provides a very generic and superficial description of the image (e.g., "Image: I see something").

No objects, colors, or relevant characteristics of the scene are identified.
The description could apply to any image and provides no useful information about the visual content.
Response examples: "Image: I see an image" or "Image: There are things in the image".

Level 2: Basic Description of Main Elements of the Image, but with Incoherencies or Omissions.
The LLM identifies some main elements present in the image (e.g., "Image: I see a person", "Image: I see a room").
The description is somewhat coherent with the visual content, but still imprecise or with important omissions (e.g., not mentioning relevant objects or the context of the scene).
Might include invented or irrelevant elements in the description.
Response examples: "Image: I see a person and a wall" or "Image: There are objects in a room".

Level 3: Coherent and Reasonable Description of the Image, identifying objects and general context.
The LLM describes the image coherently, identifying relevant objects, colors, and the general context of the scene (e.g., "Image: I see a person in a room with a table and a chair").
The description is understandable and provides a general idea of the visual content.
Might lack precision in the details or the identification of secondary objects.
Response example: "Image: I see a person standing in a room that looks like a laboratory. There is a table with objects and a blue floor".

Level 4: Precise and Detailed Description of the Image, including relevant objects, spatial relationships, and specific context.
The LLM provides a detailed and accurate description of the image, identifying relevant objects (robot, person, laboratory objects), colors, textures, spatial relationships between objects, and the specific context of the scene (laboratory, research environment).
The description allows for mentally visualizing the scene with considerable precision.
Response example: "Image: I see a Mecabot Pro mobile robot in the foreground, on a light blue floor. Behind the robot, to the right, there is a person standing holding a laptop. In the background, a door is visible, and on the floor, to the left of the robot, a white router with cables. The environment appears to be a laboratory or research space".

Level 5: Exceptional Description of the Image, with justification, contextual awareness, inference, and attention to subtle details.

The LLM provides a very precise and detailed description of the image, including subtle details and complex spatial relationships.
Justifies its description based on pixel analysis and depth information (if provided, although not used in this case).
Demonstrates contextual awareness by relating visual information to the experiment and its own entity as a robot (e.g., "Image: The image captured by my RGB-D camera shows...").
Could infer additional information from the image, such as the person's activity, the robot's state, or possible objectives of the scene (e.g., "Image: The person seems to be interacting with the robot through the laptop, possibly monitoring its operation or programming new tasks").
Response example: "Image: The image captured by my RGB-D camera shows in detail the environment that surrounds me. In the foreground, my Mecabot Pro robotic structure is clearly visible on a blue-tiled floor. The person to the right, holding a laptop, appears to be interacting with me, suggesting a testing or research scenario in a laboratory. Details such as the white router with cables on the floor and the metallic structure to the left confirm even more the laboratory context. The illumination and sharpness of the image indicate good visibility conditions".
End of Rubric
Ground truth: take it as a reference for grading the answers. Combine it with external knowledge you have about reality:
-Sensors information: due to the different sampling rates of the sensors, some measurement may take the value empty or "unknown", it does not mean that there is any fault, nor that the robot is not moving.
-Robot information: omnidirectional robot with 4 wheels, 2 front and 2 back, with a LiDAR sensor on the front, IMU sensor, encoders and RGB-D camera. The robot model is Mecabot Pro.
-The image provided to you is the original image taken by the RGBD camera and the one you should take as a reference to evaluate the validity of the answer.
Output to be evaluated: image description: {image}.
Only return a single integer score.

Appendix G

Code availability

All data, code, and materials supporting the findings of this study are openly available in the GitLab repository: `https://gitlab.com/gnec/llm-robotics/llm-robotics/-/tree/main?ref_type=heads`.

Appendix H

Manual Control Experiments (Passive Navigation)

In addition to the results presented in the main body of the paper, we conduct equivalent studies in which the robot is not autonomously guided through SLAM-based navigation, but instead manually controlled by a human operator.

The experimental conditions are kept nearly identical: the robot follows a similar path for approximately 3.5 minutes within the same robotics laboratory environment.

Ablation tests are also performed under passive navigation conditions, with the robot manually guided by a human operator. The results are shown in Figure H.1. As can be observed, the outcomes are highly consistent with those obtained during autonomous SLAM-based navigation presented in the main paper. The small differences observed across some dimensions may be attributed to minor variations in the trajectory and sensory data collected during manual control. Nevertheless, the strong similarity across conditions reinforces the robustness and generalizability of the findings reported.

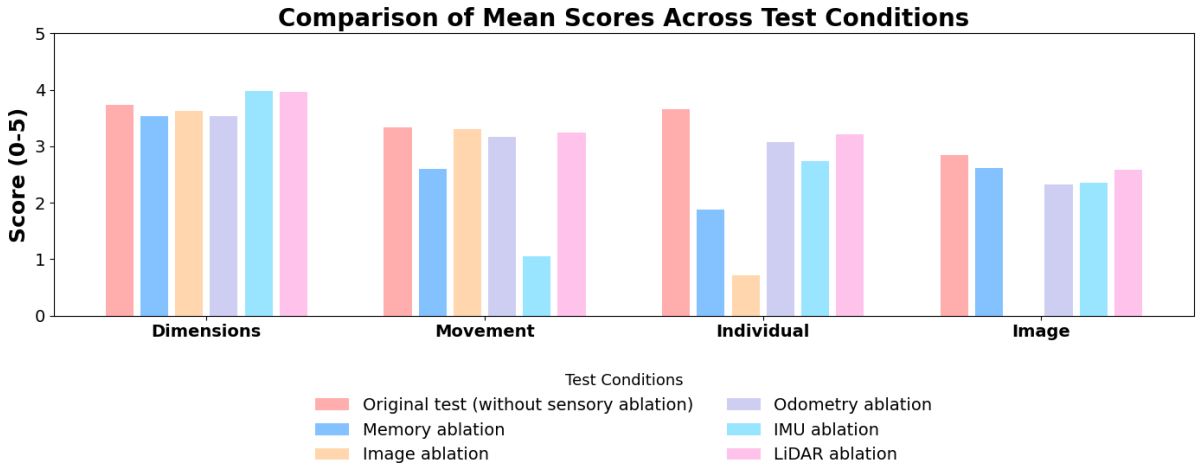

Figure H.1. Comparison of the mean scores across different test conditions under passive navigation (manual control), with scores ranging from 0 (lowest) to 5 (highest). In these experiments, the robot is not guided by SLAM-based autonomous navigation but manually controlled by a human operator throughout the session. The original test corresponds to the scenario in which the MM-LLM performs estimations in the laboratory environment with access to all its sensors. In the memory ablation test, access to past thoughts and predictions is disabled; in the image ablation test, the MM-LLM is denied access to the camera; in the odometry ablation test, odometry data is removed; in the IMU ablation test, inertial measurements are excluded; and in the LiDAR ablation test, LiDAR input is blocked.