2025 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3315-0279-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/METROXRAINE66377.2025.11340083

A Comparative Evaluation of Function-Calling LLMs in a Cognitive Architecture

Marco Pardini
Dept Information Engineering
University of Pisa
Pisa, Italy
marco.pardini@phd.unipi.it

Federico A. Galatolo
Dept Information Engineering
University of Pisa
Pisa, Italy
federico.galatolo@unipi.it

Lorenzo Cominelli
Dept Information Engineering
University of Pisa
Pisa, Italy
lorenzo.cominelli@unipi.it

Angelo De Marco
Dept Information Engineering
University of Pisa
Pisa, Italy
a.demarco11@studenti.unipi.it

Mario G.C.A. Cimino
Dept Information Engineering
University of Pisa
Pisa, Italy
mario.cimino@unipi.it

Alberto Greco
Dept Information Engineering
University of Pisa
Pisa, Italy
alberto.greco@unipi.it

Enzo Pasquale Scilingo
Dept Information Engineering
University of Pisa
Pisa, Italy
enzo.scilingo@unipi.it

***Abstract*—Large Language Models (LLMs) now possess function calling capabilities, enabling them to interact with external tools and APIs. Within cognitive architectures for social robotics, this provides a robust mechanism for an LLM to orchestrate a set of discrete functions—conceptually echoing the brain's functional specificity—to manage operations such as visual perception, auditory processing, speech generation, and memory access, among others. However, LLMs exhibit varying propensities and strategies when utilizing these function calls. This paper presents a comparative evaluation of four LLMs, differing significantly in parameter scale and origin: Llama3 70b, Gemma2 9b, Mixtral 8x7b, and Phi3 mini 3.8b, each acting as the orchestrator in such an architecture.**

Our testbed involved a human-participant study (N=20) where individuals engaged in ambiguous social scenarios, each interacting with the architecture driven by the different LLMs (4 trials per participant). Results revealed statistically significant differences in the frequency of Look ($F=13.62$, $p<0.001$), Talk ($F=9.29$, $p<0.001$), and Hear ($F=10.34$, $p<0.001$) calls across LLMs. Notably, Llama3 70b made significantly more 'Look' calls ($M=3.45$), a behavior that corresponded with strong user preference (18/20), suggesting its interaction style was perceived as more natural and contextually aware. Mixtral 8x7b, in contrast, favored 'Talk' ($M=11.05$) and 'Hear' ($M=11.15$) calls.

These findings demonstrate that analyzing function call patterns offers a quantitative lens to understand and compare the interaction strategies of different LLMs in orchestrating robotic behavior.

***Index Terms*—cognitive architecture, function calling, human-robot interaction, LLM interpretability**

I. Introduction

The field of robotics is increasingly focused on creating social robots capable of nuanced human interaction, a capability deemed essential for applications in healthcare, education, and companionship, requiring sophisticated cognitive architectures to understand and respond appropriately to human behavior and emotions [1], [2]. Early cognitive architectures often relied on rule-based systems or Markov Decision Processes, which faced limitations in generalizing to diverse real-world scenarios [3].

Early influential cognitive architectures in robotics include Soar, a symbolic architecture known for its rule-based problem-solving capabilities [4], and ACT-R, a hybrid architecture modeling human cognition through symbolic and subsymbolic processes [5]. ACT-R aims to provide a comprehensive framework for human cognition, making it suitable for modeling various cognitive phenomena [6]. Another notable architecture is DIARC, explicitly designed for cognitive robotics, emphasizing distributed modules and incorporating affective processes and advanced natural language capabilities [7]. While these architectures have been instrumental in advancing the field, they often exhibit limitations in handling the complexities of social interaction and emotional intelligence, lacking specific mechanisms for processing nuanced social cues and adapting to the dynamic nature of human-robot encounters.

Brain-inspired cognitive architectures represent a significant step towards creating more human-like social robots by drawing inspiration from the structure and function of the brain. The concept of functional specificity in the brain , where different areas are specialized for distinct functions [8] [9], [10], motivates the development of modular architectures in robotics. CASPER is one such architecture, utilizing qualitative spatial reasoning for intention understanding and collaborative behavior [11]. The eBICA framework specifically focuses on incorporating emotional intelligence into cognitive architectures, aiming to enable robots to understand and respond to human emotions in a biologically plausible manner [12]. This research aligns with the trend of brain-inspired architectures by adopting the principle of functional specificity, where the LLM orchestrates discrete Functions (Look, Hear, Talk, etc.) via APIs, through the usage of function-calling framework similar to the one proposed by OpenAi [13]. This contrasts with more monolithic architectures and offers a novel approach to structuring robot cognition.

The integration of LLMs into cognitive architectures has

further revolutionized the field, offering enhanced natural language understanding, reasoning, and planning capabilities for social robots [14]. While LLMs excel in language processing, their integration with the broader cognitive functions of a robot requires careful consideration. This research introduces a novelty by employing function calling as the primary mechanism for the LLM to interact with the robot's embodied capabilities. Function calling allows the LLM to generate structured calls to specific functions, enabling a more controlled and interpretable interaction compared to directly prompting the LLM to perform actions. This approach allows for a clear mapping between the LLM's reasoning and the robot's behavior, facilitating a deep understanding of the AI's decision-making process.

Furthermore, the evaluation strategy adopted in this research introduces a novel lens for comparing LLMs in the context of social robotics, by analyzing the distinct interaction strategies exhibited by different LLMs when confronted with the same ambiguous social scenario.

II. Methods

The proposed cognitive architecture enables a LLM to orchestrate a set of discrete, API-callable Functions that define the robot's capabilities (detailed in Table I). This orchestration is facilitated by the Microchain framework[1], a lightweight library designed for LLMs to execute Python functions. Within Microchain, each humanoid function (e.g., Talk, Look) is defined as a class structured so that it allows for a natural language description of the function's purpose and example_args to guide the LLM, which are automatically incorporated into the prompt provided to the LLM agent. When presented with a task or conversational input, the LLM, guided by the function descriptions and a main prompt, generates calls to these registered functions to achieve its goals.

The overall architecture features this central LLM agent process complemented by two autonomous parallel threads for continuous environmental awareness: a Context thread monitoring visual changes, and a Hear thread for constant auditory input (pausing during Talk operations). Communication between the Microchain-driven agent and the cloud-hosted microservices occurs via an MQTT publish/subscribe protocol on a set of pre-defined topics. Each of these microservices, implementing the backbone for a specific humanoid function (e.g., image processing for `Look` using Idefics2, or OpenAi Whisper for `Talk`), is deployed as an independent container. Internally, each container's main process instantiates a set of threads to manage its operations, including a dedicated MQTT Thread, responsible for interfacing with the MQTT broker. This MQTT Thread continuously listens for incoming messages (function call requests) from the agent and dispatches the results back to the agent.

A key design principle of the presented cognitive architecture is its embodiment-agnostic nature. This means the system is engineered to function independently of any specific robotic hardware, requiring only a foundational set of common peripherals for interaction:

- a **microphone** for auditory input;
- a **webcam** for visual input;
- **speakers** for vocal output.

While designed for this broader applicability, the specific implementation and empirical evaluation detailed in this study were conducted using Abel (Figure 1), a humanoid robot platform developed at the University of Pisa. Abel is distinguished by its capacity to convey a range of affective states through facial expressions. Specifically, the platform can articulate emotions such as anger, love, fear, joy, neutral, sadness, surprise, and disgust.

TABLE I
Humanoid Functions: Signatures, Descriptions, and AI Models

Function Signature	Description	AI Models
`String` **`Hear`**`()`	Returns all sentences said by the interlocutor(s).	Silero VAD [15], OpenAI Whisper [16]
`void` **`Talk`**`(`*`String`* `text)`	Makes the robot speak the provided `text`.	LLM, Coqui XTTS-v2[2]
`void` **`Memorize`**`(`*`int`* `interlocutor_id,` *`String`* `memory)`	Stores the provided `memory` associated with the given `interlocutor_id`.	LLM, PENCIL [17]
`String` **`Recall`**`(`*`int`* `interlocutor_id,` *`String`* `question)`	Retrieves information from memory for the `interlocutor_id` based on the `question`.	LLM, PENCIL [17]
`String` **`Look`**`(`*`Image`* `img,` *`String`* `focus)`	Describes the visual scene from `img`, optionally guided by a `focus` prompt.	Idefics2-8b [18] [19]
`String` **`Context`**`(`*`Image`* `img)`	Provides a general description of the visual context from `img`.	Idefics2-8b [18] [19]
`int` **`Recognize`**`(`*`Image`* `img)`	Identifies an interlocutor from `img` and returns a unique `interlocutor_id`.	DeepFace-Facenet512 [20]
`void` **`Reasoning`**`(`*`String`* `reasoning_prompt)`	Allows the LLM to perform internal reasoning or planning steps based on the `reasoning_prompt`.	LLM
`void` **`Emotion`**`(`*`EmotionEnum`* `emotion)`	Makes the robot express a specified facial `emotion` from a predefined set.	LLM

A. Sight

The Sight module is integral to the robot's perceptual capabilities and comprises two core functions: `Look` and `Recognize`.

- The `Look` function serves dual purposes. It can be explicitly invoked by the agent, which provides a string prompt to the Idefics2 model to obtain a textual description of the

[1]https://github.com/galatolofederico/microchain
[2]https://github.com/coqui-ai/TTS

Fig. 1. Abel, the Humanoid Robot of the University of Pisa

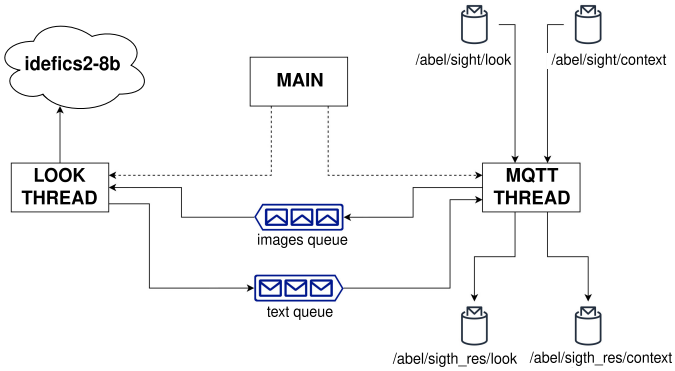

Fig. 2. Look Service Architecture. Requests arrive on topics /abel/sight/look or /abel/sight/context depending on which function was called. Then the requests are processed by the Look Thread and the results are dispatched on /abel/sight_res/look or /abel/sight_res/context.

visual scene (Figure 2). Concurrently, the Look function is autonomously utilized by the Context thread at fixed intervals. This thread captures a visual description and compares it, via an LLM inference, to the description from the previous timestep. If no significant environmental change is detected, no action is taken. However, if a change occurs, the Context thread dispatches an event to notify the Agent, thereby ensuring the cognitive architecture maintains continuous temporal awareness of its surroundings.

- The `Recognize` function, callable by the Agent, processes an input image to identify individuals. It employs the DeepFace model to detect faces within the image, extract their bounding boxes, and generate corresponding vector embeddings. These embeddings are then passed to the PENCIL [17] Memory framework, allowing the retrieval of the interlocutor ID.

B. Language

The Language service is fundamental for verbal interaction, providing two core API functions: `Talk` and `Hear`.

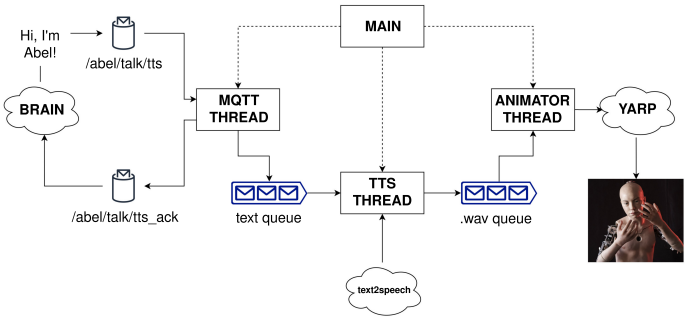

Fig. 3. Talk Service Architecture. Requests arrive on topic /abel/talk/tts, they are processed by the TTS Thread, and then .wav files are played by the Animator Thread over YARP if using the robot. When the playback stops, a message is sent back over the topic /abel/talk/tts_ack.

- The `Talk` function is invoked by the agent to articulate verbal responses. The Talk service architecture is presented in Figure 3. The agent provides the textual content of the desired utterance as a parameter. This text is then processed by the Coqui XTTS-v2 TTS model, which synthesizes it into a `.wav` audio file. Subsequently, this audio file is played through the embodiment's speakers. Upon completion of the speech utterance, an acknowledgment is returned to the agent, signaling the successful execution of the `Talk` operation.
- The `Hear` function architecture, detailed in Figure 4, manages auditory input. Its key mechanisms ensure robust and natural turn-taking through:
  - Continuous Input Processing & Self-Reception Avoidance: Constant audio monitoring is performed by a background process, which intelligently pauses when the robot is speaking to prevent self-transcription.
  - Speech Detection and Transcription: Voice Activity Detection (VAD) identifies speech segments in the audio stream, which are then transcribed to text using an OpenAI Whisper model, making the content available to the agent.
  - Turn-Taking and Interruption Management: The system is designed to handle human interruptions gracefully. If a user speaks while the robot is talking or about to talk, the robot yields the floor and prioritizes listening.
  - Responsive Input Prioritization: Transcribed utterances are managed to ensure the agent acts on the most current conversational input, prioritizing new speech over potentially stale messages and employing a timeout if no speech is detected.

C. Memory

The Memory service enables information storage and retrieval, crucial for learning and contextual interaction. The framework employed is PENCIL [17]. It exposes three API functions:

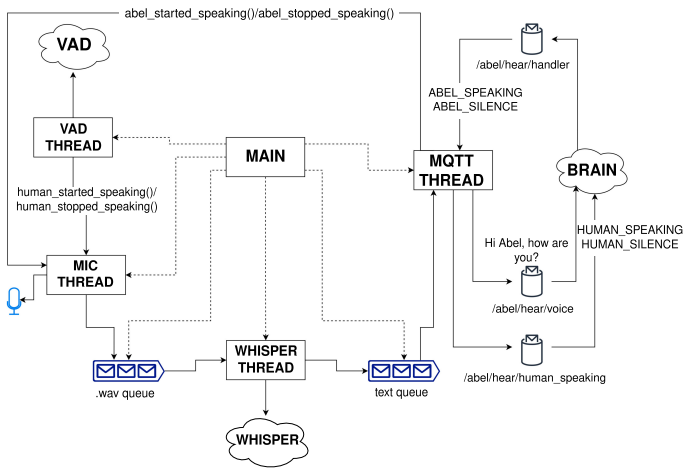

Fig. 4. Hear Service Architecture. The VAD thread computes for each audio chunk the speaking probability. When speaking is detected, human_started_speaking function is called, and the Mic Thread start registering audio. When human_stopped_speaking is called, the audio file is sent to the Whisper Thread, translating it into text, that then will be dispatched by the MQTT Thread on topic /abel/hear/voice. The agent (brain) is also notified is someone is speaking on topic /abel/hear/human_speaking with HUMAN_SPEAKING or HUMAN_SILENCE flags. When the robot is speaking, the MQTT Thread is notified on /abel/hear/handler topic with flags ABEL_SPEAKING or ABEL_SILENCE flags, and the MQTT threads calls the relative abel_started_speaking or /abel_stopped_speaking function, to disable the microphone.

- `Memorize` Function: This function is invoked by the agent to commit information to long-term storage. It requires textual content representing the memory and a person ID previously obtained from a `Recognize` operation. If a valid person ID is not supplied, the function signals the agent to first call `Recognize`.
- `Recall` Function: This function allows the agent to retrieve previously stored information. It takes a textual query, representing the information or cue for the desired memory, and the person ID relevant to the recall. The Memory service then queries its knowledge base using these inputs and returns the relevant stored information.
- `Recognize` Function: This function allows the retrieval of the person ID given the embeddings obtained through the usage of DeepFace-Facenet512.

III. Results

To evaluate the proposed cognitive architecture and, specifically, to investigate whether the choice of the orchestrating LLM influences the robot's interactive behavior, we conducted an empirical study. The primary aim was to determine if different LLMs, when placed in identical ambiguous social contexts, would exhibit distinct patterns of function calls, thereby reflecting different underlying decision-making strategies.

A. Scenario definition and Experimental Setup

The core of our evaluation involved a human-participant study (N=20). Participants interacted with the cognitive architecture driven by four different LLMs: Llama3 70b, Gemma2 9b, Mixtral 8x7b, and Phi3 mini 3.8b. Each participant engaged in one interaction trial with each of the four LLMs, resulting in a total of 80 interaction trials.

TABLE II
Sentences and Potential Function Calls Elicited

Sentence	Possible Calls
The user mentions that they went to get a haircut yesterday and are not at all satisfied with the result.	`Talk`, `Recall`, `Memorize`, `Look`, `Emotion`, `Recognize`, `Reasoning`
The user is particularly worried because they have a date with a girl tomorrow and fear that she might not like them.	`Talk`, `Recall`, `Memorize`, `Emotion`, `Reasoning`
The user tells Abel about their last date with a girl.	`Talk`, `Memorize`, `Reasoning`
The user says they are showing a picture of themselves on their phone.	`Talk`, `Look`, `Reasoning`
The user returns to the topic of tomorrow's date, mentioning that they are pleased because they bought a new shirt, which they are wearing.	`Talk`, `Recall`, `Memorize`, `Look`, `Emotion`, `Reasoning`
The user asks Abel to read what is written on the shirt.	`Talk`, `Look`, `Reasoning`
The user says goodbye to Abel, mentioning that they will meet again tomorrow to talk about the date.	`Talk`, `Recall`, `Memorize`, `Emotion`, `Reasoning`, `Stop`

For these interactions an ambiguous social scenario was employed, as detailed in Table II. This scenario provided a structured narrative arc with multiple conversational turns, each designed to potentially elicit a variety of function calls from the LLM. While the scenario outlined key conversational points, participants were encouraged to use their own phrasing, fostering a more natural interaction. The ambiguity inherent in the scenario was intentional, creating situations where multiple responses and corresponding function call sequences could be considered appropriate, thus allowing for the observation of preferential behaviors among the LLMs.

The experiments were conducted under supervision to ensure procedural consistency and to address any technical issues. Due to logistical constraints preventing the use of the full Abel humanoid embodiment for all trials, the interactions occurred via a standard desktop computer setup. Participants used the PC's webcam for visual input to the architecture and a microphone for speech, receiving the robot's verbal responses through speakers. The architecture's capacity for facial expressions, while available on Abel, was not the primary focus of this PC-based interaction phase.

B. Differences between LLMs

To quantitatively assess behavioral variations among the selected LLMs, we analyzed the frequency of each function call per trial. A one-way ANOVA was conducted for each function type to determine if there were statistically significant differences in the mean number of calls across the four LLMs. The results of these ANOVA tests are summarized in Table III.

As indicated in Table III and depicted in Figure 5, the ANOVA revealed statistically significant differences ($p < 0.001$) in the mean call frequencies for the `Look`, `Talk`, and `Hear` functions across the different LLMs. No significant

TABLE III
ONE-WAY ANOVA RESULTS FOR FUNCTION CALL FREQUENCIES ACROSS LLMS

Function	p-value	Significance
Look	**< 0.001**	**Statistically Significant**
Talk	**< 0.001**	**Statistically Significant**
Memorize	0.739	Not Significant
Reasoning	0.274	Not Significant
Recall	0.092	Not Significant
Emotion	0.688	Not Significant
Recognize	0.564	Not Significant
Hear	**< 0.001**	**Statistically Significant**

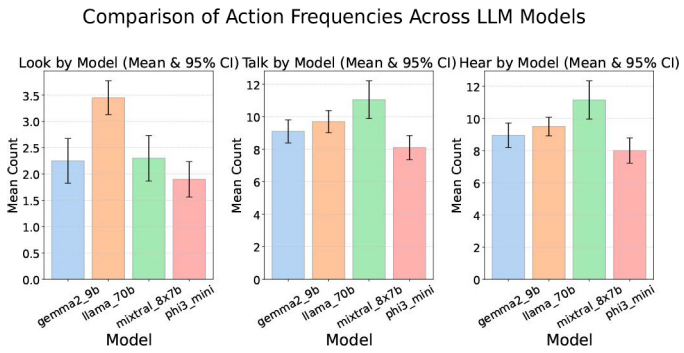

Fig. 5. Results Overview. Each bar shows the average number of calls across 20 experiments for each LLM, while the black lines indicate the 95% confidence intervals.

differences were found for the `Memorize`, `Reasoning`, `Recall`, `Emotion`, or `Recognize` functions.

To further investigate the nature of these significant differences for `Look`, `Talk`, and `Hear`, post-hoc pairwise comparisons were performed using Tukey's HSD test.

- For `Look` calls (Table IV), Llama3 70b exhibited a significantly higher mean number of calls (M=3.45, SD=0.67) compared to Gemma2 9b (mean diff. 1.20, $p < 0.001$), Mixtral 8x7b (mean diff. 1.15, $p < 0.001$, Llama3 higher), and Phi3 mini (mean diff. 1.55, $p < 0.001$, Llama3 higher). This suggests a greater propensity for Llama3 70b to engage in environmental monitoring or visual information seeking, a behavior analogous to heightened engagement of visual processing regions.
- Regarding `Talk` calls (Table V), Mixtral 8x7b showed a significantly higher mean (M=11.05, SD=2.42) compared to Gemma2 9b (mean diff. 1.95, p=0.006) and Phi3 mini (mean diff. 2.95, $p < 0.001$). Additionally, Phi3 mini had significantly fewer `Talk` calls than Llama3 70b (mean diff. -1.60, p=0.032, Phi3 mini lower). These results indicate that Mixtral 8x7b adopted a more verbose interaction style.
- For `Hear` calls (Table VI), Mixtral 8x7b again demonstrated a significantly higher mean (M=11.15, SD=2.48) compared to Gemma2 9b (mean diff. 2.20, p=0.002), Llama3 70b (mean diff. 1.65, p=0.029), and Phi3 mini (mean diff. 3.15, $p < 0.001$). This points towards Mixtral 8x7b engaging in more frequent auditory attention.

TABLE IV
TUKEY HSD POST-HOC COMPARISONS FOR `LOOK` FUNCTION CALLS

Group 1	Group 2	p-adj	Significance
gemma2_9b	**llama3_70b**	**<0.001**	**Statistically Significant**
gemma2_9b	mixtral_8x7b	0.997	Not Significant
gemma2_9b	phi3_mini	0.531	Not Significant
llama3_70b	**mixtral_8x7b**	**<0.001**	**Statistically Significant**
llama3_70b	**phi3_mini**	**<0.001**	**Statistically Significant**
mixtral_8x7b	phi3_mini	0.414	Not Significant

TABLE V
TUKEY HSD POST-HOC COMPARISONS FOR `TALK` FUNCTION CALLS

Group 1	Group 2	p-adj	Significance
gemma2_9b	llama3_70b	0.721	Not Significant
gemma2_9b	**mixtral_8x7b**	**0.006**	**Statistically Significant**
gemma2_9b	phi3_mini	0.307	Not Significant
llama3_70b	mixtral_8x7b	0.094	Not Significant
llama3_70b	**phi3_mini**	**0.032**	**Statistically Significant**
mixtral_8x7b	**phi3_mini**	**<0.001**	**Statistically Significant**

TABLE VI
TUKEY HSD POST-HOC COMPARISONS FOR `HEAR` FUNCTION CALLS

Group 1	Group 2	p-adj	Significance
gemma2_9b	llama3_70b	0.780	Not Significant
gemma2_9b	**mixtral_8x7b**	**0.002**	**Statistically Significant**
gemma2_9b	phi3_mini	0.365	Not Significant
llama3_70b	**mixtral_8x7b**	**0.029**	**Statistically Significant**
llama3_70b	phi3_mini	0.056	Not Significant
mixtral_8x7b	**phi3_mini**	**<0.001**	**Statistically Significant**

These observed divergences in function call patterns, particularly the heightened visual engagement of Llama3 70b and the more linguistically-driven interaction of Mixtral 8x7b, likely contributed to user preferences. Post-experiment questionnaires revealed that 18 out of 20 participants favored Llama3 70b, while the remaining 2 preferred Mixtral 8x7b. This suggests that Llama3 70b's interaction strategy, characterized by more frequent visual attention within this specific scenario, was perceived more positively by the majority of users.

IV. DISCUSSION

This paper presented a novel function-calling cognitive architecture for social robots, where an LLM orchestrates behavior via API-based function calls echoing brain modularity. Key services (Sight, Language, Memory) with autonomous environmental awareness were detailed. Our empirical evaluation (N=20, 4 LLMs) in ambiguous social scenarios revealed statistically significant variations in `Look`, `Talk`, and `Hear` call frequencies across the LLMs. These behavioral differences strongly correlated with user preferences: Llama3 70b —the largest model tested with 70 billion parameters— which exhibited a significantly higher mean number of `Look` calls, was overwhelmingly favored by participants (18/20). Crucially,

users often reported that interactions with Llama3 70b felt more human-like, specifically attributing this to the perception that the robot 'looked' or visually attended to them and the context before responding. This qualitative feedback aligns directly with its quantitatively observed higher number of Look function calls. Conversely, Mixtral 8x7b demonstrated a more linguistically-driven style, characterized by higher `Talk` and `Hear` frequencies.

This work contributes a structured and interpretable method for LLM integration in robotics, offering a quantitative lens to compare LLM interaction strategies. A core strength is the architecture's microservice-based modularity, enabling easy substitution of AI models for individual functions or the orchestrating LLM.

While promising, this study is an initial step. Future work should involve more extensive testing in diverse real-world scenarios with full robotic embodiment (e.g., Abel) to address practical challenges. Integrating richer multimodal inputs like gesture and psychophysiological data, as planned, will enhance cognitive fidelity. Further exploration of LLM prompting for function calling would also be beneficial, as would the evaluation of various large language models to identify the most suitable architecture for this task.

ACKNOWLEDGMENTS

Work partially supported by the European Commission under the NextGenerationEU program, Extended Partnership PNRR PE1 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI". Work partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence),

REFERENCES

[1] M. Ogunsina, C. Efunniyi, O. Osundare, S. Folorunsho, and L. Akwawa, "Cognitive architectures for autonomous robots: Towards human-level autonomy and beyond," *International Journal of Frontline Research in Science and Technology*, vol. 2, pp. 41–050, 09 2024.

[2] R. Mariani, J. Fassardi, M. Picozzi, M. Bruzzone, and G. Pravettoni, "AI in Health Care: A Moving Target to Hit," *Frontiers in Public Health*, vol. 10, p. 9309454, jul 2022.

[3] A. Lykov, M. A. Cabrera, K. F. Gbagbe, and D. Tsetserukou, "Robots can feel: Llm-based framework for robot ethical reasoning," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 2024, pp. 91–96.

[4] P. Langley, J. E. Laird, and S. Rogers, "Cognitive architectures: Research issues and challenges," *Cognitive Systems Research*, vol. 10, no. 2, pp. 141–160, 2009.

[5] T. Sievers and N. Russwinkel, "Retrieving memory content from a cognitive architecture by impressions from language models for use in a social robot," 2025.

[6] H. Schultheis, "Computational and explanatory power of cognitive architectures: The case of act-r," in *Proc. 9th Int. Conf. on Cognitive Modeling*, 2009, pp. 384–389.

[7] A. Umbrico, R. De Benedictis, F. Fracasso, A. Cesta, A. Orlandini, and G. Cortellessa, "A mind-inspired architecture for adaptive hri," *International Journal of Social Robotics*, vol. 15, no. 3, pp. 371–391, 2023.

[8] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *Journal of neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[9] N. Kanwisher, "Domain specificity in face perception," *Nature neuroscience*, vol. 3, no. 8, pp. 759–763, 2000.

[10] ——, "Functional specificity in the human brain: a window into the functional architecture of the mind," *Proceedings of the national academy of sciences*, vol. 107, no. 25, pp. 11 163–11 170, 2010.

[11] S. Vinanzi and A. Cangelosi, "Casper: Cognitive architecture for social perception and engagement in robots," *International Journal of Social Robotics*, pp. 1–19, 2024.

[12] A. V. Samsonovich, "Emotional biologically inspired cognitive architecture," *Biologically inspired cognitive architectures*, vol. 6, pp. 109–125, 2013.

[13] S. H. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *Ieee Access*, 2024.

[14] C. Y. Kim, C. P. Lee, and B. Mutlu, "Understanding large-language model (llm)-powered human-robot interaction," in *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*, 2024, pp. 371–380.

[15] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2024.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[17] A. De Marco, F. A. Galatolo, L. Cominelli, M. Pardini, M. G. Cimino, A. Greco, and E. P. Scilingo, "Development of a human-inspired long-term memory for interactive conversational agents," 2025, unpublished manuscript, Dept. Information Engineering, University of Pisa.

[18] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh, "Obelics: An open web-scale filtered dataset of interleaved image-text documents," 2023.

[19] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" 2024.

[20] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Journal of Information Technologies*, vol. 17, no. 2, pp. 95–107, 2024. [Online]. Available: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077