Visions remembered – Using a Vision-Language Model to set and recall Image Impressions in the Memory of a Cognitive Model

Thomas Sievers
Institute of Information Systems
University of Lübeck
Lübeck, Germany
Email: t.sievers@uni-luebeck.de

Nele Russwinkel
Institute of Information Systems
University of Lübeck
Lübeck, Germany
Email: nele.russwinkel@uni-luebeck.de

***Abstract*—Large Language Models (LLMs) and Vision-Language Models (VLMs) have the potential to emulate human cognitive abilities for robots and thus significantly advance the evolution and use of cognitive architectures. This opens up opportunities for using human-like judgment and decision-making capabilities such as instance-based learning and intuitive decision making inherent in cognitive architectures with social robots. Using a combined system of an Adaptive Control of Thought-Rational (ACT-R) model and a humanoid social robot, we show how content from the declarative memory of the ACT-R model can be retrieved per association of real-world data obtained by the robot via the image recognition capabilities of an LLM. Such recollections can be fetched and processed according to the procedural memory of cognitive model productions and then returned to the robot as instructions, for example to add the prompt to LLM-driven utterances to keep them more contextual. In addition, visual impressions captured by the robot can be stored in the cognitive model.**

I. Introduction

Large Language Models (LLMs) like ChatGPT have advanced reasoning capabilities, blending intuitive and deliberate cognitive processes [6]. LLMs help robotic systems improve their generalization capabilities in dynamic and complex real-world environments and can significantly increase their behavior planning and execution capabilities, enabling robots to engage with their environment in a human-like manner [12].

Vision-Language Models (VLMs) are multimodal AI systems created by combining an LLM with a vision encoder that gives the LLM the ability to "see". They provide assistance with complex tasks such as creating captions and answering visual questions [7]. VLMs are capable of performing a variety of tasks after learning relationships between images and language from large data sets, such as answering questions about images, finding sentences that correspond well with images, and finding image regions that correspond to texts. These skills can be used in many ways for robotics, for example for robot movements, state recognition, object recognition, affordance recognition, relation recognition, and anomaly detection [14, 25].

Cognitive architectures refer both to a theory about the structure of the human mind and to a computer-based implementation of such a theory. Their formalized models can be used to react flexibly to actions in a human-like manner and – when used in a robot – to develop a situational understanding for adequate reactions. ACT-R (Adaptive Control of Thought - Rational) is a well-known and successfully used cognitive architecture [2]. A cognitive architecture can also be used to add a "human component" to robotic applications [29]. Furthermore, language models are good at fast automatic reasoning, but less capable of high-level cognition to enable complex mental operations and "slow thinking" following the dual process theory of human cognition [13].

Looking at Human-Robot Interaction (HRI), a combination of robot sensor technology and data processing with a cognitive architecture offers the possibility of dealing with information from the robot's real world in cognitive models. With their ability to use general concepts inspired by the human brain and create mental models based on human cognitive abilities, such architectures can help provide facts and context, e.g. in the form of memories from a particular scenario, for an LLM and provide individualized experiences. Using prompt augmentation, conclusions of a mental model can be taken into account in utterances of the LLM.

Intuitive decision-making as a subjective, particularly human type of decision-making, is based on implicit knowledge that is transmitted to the conscious mind at the time of the decision through affect or unconscious cognition. Computational models of intuitive decision-making can be expressed as instance-based learning using the ACT-R cognitive architecture [26]. In instance-based learning theory (IBLT), past experiences (i.e. instances) are retrieved using cognitive mechanisms of a cognitive architecture. IBLT proposes learning mechanisms related to a decision-making process, such as instance-based knowledge and recognition-based retrieval. These learning mechanisms can be implemented in an ACT-R model [9, 8]. In our opinion, it would be interesting to enable such behavior for a social robot as well.

We apply the LLM of OpenAI's Generative Pretrained Transformer (GPT, commonly known as ChatGPT) in an application example for a dialog between a human and a robot

using it for the robot's utterances [20]. Our application is primarily intended to demonstrate the technical possibility of realization. However, the proposed method can also be used for other scenarios and beyond an HRI context.

Visual impressions of the robot are processed by the LLM in such a way that the core content of what the robot sees is expressed in three keywords or key phrases. These keywords or phrases are passed to an ACT-R model as chunks and processed there. The cognitive model uses these chunks to search its declarative memory for existing memory content that is indexed in the same or a similar form using chunks. The stored memory content contains a recollection phrase that reflects the visual impression at the particular time in the form of a short, complete sentence. If there is a positive correlation between chunks from the LLM and memory chunks, this recollection or *fact* phrase is passed to the LLM for prompt augmentation, which generates a response based on this knowledge. This process can be seen as an association of remembered impressions. An a-priori factual knowledge can be accumulated by creating corresponding chunks in the declarative memory. This leads us to the following hypothesis:

Hypothesis Utilizing memory chunks from a cognitive model enables prompt augmentation for an LLM with associated past visual impressions.

Using ACT-R's chunk and memory system to store and process impressions of a certain scenario and enabling an LLM to incorporate remembered data from such impressions into utterances, opens a path for a more reliable and evidence-based application of LLMs. We provide an insight into our ongoing work and summarize initial findings.

II. Related work

Robots are able to use Large Multi-modal Models (LMMs) to comprehend and execute tasks based on natural language input and environmental cues, and the integration of foundation models such as LLMs and VLMs can effectively improve robot intelligence [18, 12, 28].

VLMs help to equip robots with the ability to physically-based task planning and can analyze videos of humans performing tasks to obtain a symbolic task plan for a robot through textual explanations of the environment and action details in combination with an LLM-based task planner [11, 27].

Yoshida et al. investigate possibilities for the development of a "minimal self" with a sense of agency and ownership in a robot that is able to mimic human movements and emotions by using human knowledge from language models [31]. To do this, they use GPT-4's motion generation and image recognition capabilities. VLMs as a basis for metacognitive thinking can enable robots to understand and improve their own processes, avoid hardware failures and thus increase their resilience [16].

Since the recent successes of language models, there has been an increased interest in the interplay between LLMs and cognitive architectures. Niu et al. provided an overview of similarities, differences, and challenges between LLMs and cognitive science by analyzing methods for assessing potential cognitive capabilities of LLMs, discussing biases and limitations, and an integration of LLM with cognitive architectures [19]. In the novel neuro-symbolic architecture presented by Wu et al., human-centered decision making was enabled through the integration of ACT-R with LLMs by using knowledge from the decision process of the cognitive model as neural representations in trainable layers of the LLM [30]. They found that this improved the ability for grounded decision making.

Knowles et al. proposed a system architecture that combined LLMs and cognitive architectures with an analogy to "fast" and "slow" thinking in human cognition [15]. Leivada et al. explored whether the current generation of LLMs is able to develop grounded cognition that incorporates prior expectations and prior world experiences to perceive the big picture [17]. Insights from human cognition and psychology anchored in cognitive architectures could contribute to the development of systems that are more powerful, reliable and human-like [24]. The dual-process architecture and the hybrid neuro-symbolic approach to overcoming the limitations of current LLMs is seen as particularly important for this.

A cognitive system based on ACT-R was used by Birlo et al. for their concept of robot self-awareness in an embodied robotic system [3]. They focused on the representation of internal states of the robot, which processed external states from the real world and created its own interpretation of what it perceived. Sievers et al. used a cognitive model to store and process word associations for the grounding of words in a language game between a robot and a human [23].

An example integration of an LLM with a cognitive architecture to enable planning and reasoning in autonomous robots was given by Gonzáles-Santamarta et. al [10]. They proposed how to leverage the reasoning capabilities of LLMs inside the ROS 2-integrated cognitive architecture MERLIN2. Sievers et al. give examples of a connection between a cognitive model and a robot application via a TCP/IP connection of the *dispatcher* included in the ACT-R software [21, 22].

III. Methods

For a human-like way of retrieving recollections from memory, we propose to use the declarative memory of a cognitive architecture in conjunction with procedural capabilities of a cognitive model. Our system for testing such an approach consisted of the humanoid social robot Pepper with a software application that connects to OpenAI's GPT-4o model via an Application Programming Interface (API), and a TCP/IP connection over a wireless network to the ACT-R 7 cognitive architecture software version 7.28 running on a remote computer.

ACT-R offers the technical possibility of integrating a cognitive model with bidirectional communication into an application of choice. To establish a remote connection from the robot application to ACT-R, we are using it's remote interface – the *dispatcher*. Fig. 1 shows the setup of the bidirectional connection between the dispatcher, which acts as a kind of server, and the client application of the robot.

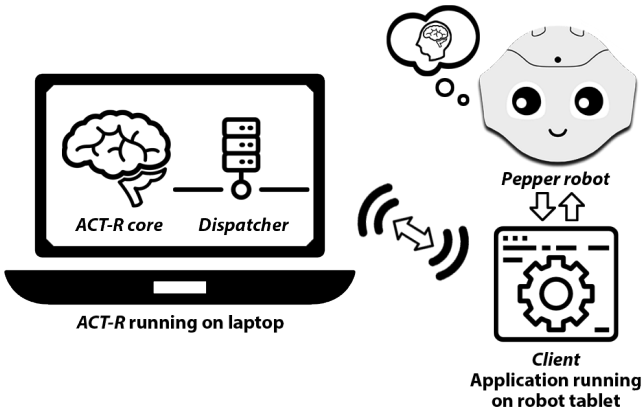

Fig. 1. Connection between ACT-R / Dispatcher and the robot application.

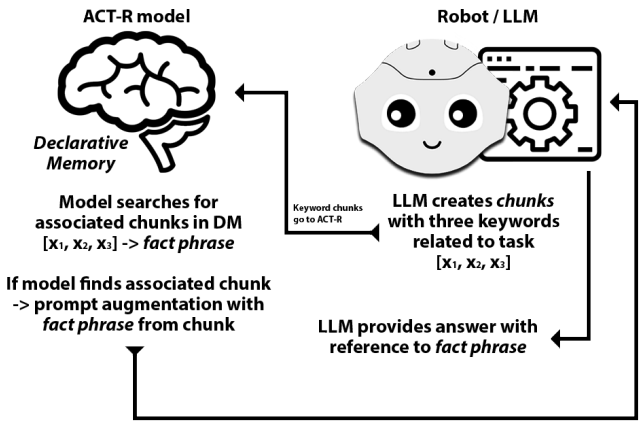

Fig. 2. Transfer of keywords to ACT-R and return of retrieved memory facts to the LLM, e.g. for prompt augmentation.

A. Cognitive Model

We used the standalone application of ACT-R 7, for which we created a cognitive model using LISP [5]. In ACT-R, declarative knowledge is represented in the form of chunks, i.e. representations of individual properties, each of which can be accessed via a labeled slot. The cognitive model programmed in LISP for our example usage should receive chunks with keywords from the robot application and checking whether chunks for these keywords are already stored in the declarative memory of the model in a suitable combination. We assumed the transfer of three keywords. The programmed productions of the procedural memory checked all combinations of the sequence of keywords for a match with memory content and generated a hit for two out of three matching keywords. In this case, the associated memory content including a corresponding recollection was called up and the recollection was returned to the robot application.

Our LISP code defined a chunk type for the memory content in declarative memory as follows: *(chunk-type keyword asso-one asso-two asso-three phrase)*

This chunk type 'keyword' featured three labeled slots 'asso-one', 'asso-two' and 'asso-three' as well as a 'phrase' slot, which stored the recollection we wanted to retrieve. For example, this could be a remembered fact used to complete a system prompt for an LLM that enables the robot to talk to humans. Each chunk stored in the declarative memory also has an arbitrary name. Figure 2 shows the process of searching for a matching memory chunk. Having found a corresponding memory chunk, the cognitive model returned the sentence stored in the 'phrase' slot of this chunk as a recollection or *fact phrase*. Example content for such a memory chunk could be: asso-one *person* asso-two *indoor* asso-three *computer* phrase *A person sitting in front of a computer indoors*

B. Humanoid Social Robot Pepper

The humanoid social robot Pepper, shown in Figure 3, was developed by Aldebaran and first released in 2015 [1]. The robot is 120 centimeters tall and optimized for human interaction. It is able to engage with people through conversation, gestures and its touch screen. Pepper is equipped with internal sensors, four directional microphones in his head and speakers for voice output. Speech recognition and dialog is available in 15 languages. The robot features an open and fully programmable platform so that developers can program their own applications using software development kits (SDKs) for programming languages like Python, Java or Kotlin.

In dialogs with humans, our robot application forwarded utterances of the human dialog partner as input to the OpenAI API, which returned a dictionary with the status and response of the API. With each API call, the entire dialog was transferred to the GPT model. This allows the model to constantly 'remember' what was previously said and refer to it as the dialog progressed. In this way, complex dialogs between humans and the robot became possible.

C. Retrieving recollections based on vision

The system's ability to retrieve recollections from declarative memory by recognizing visual impressions from the robot's perspective should be demonstrated. The LLM was instructed to create a description of the content in the form of three keywords about the content of an image taken with the robot's camera equipment. The robot's front camera took a picture every few seconds to capture an up-to-date impression of what the robot saw. To prepare a response to a human utterance based on visual impressions, the robot application sent the currently captured image to the GPT-4o model as a completion task, using the type *'image_url'* instead of *'text'*.

The three image-related keywords from the model's response were transmitted as chunks to the cognitive model to search for comparable chunks in declarative memory. In the event of a match, the content of the 'phrase' chunk slot was transferred to the robot application for further use.

In the context of our example application we focused on possible uses of recollections from declarative memory for prompt augmentation of a GPT model. But the other way around, it is also possible with this system to write memory chunks and thus visual impressions, which are provided with

Fig. 3. Robot Pepper shows a picture of what it sees

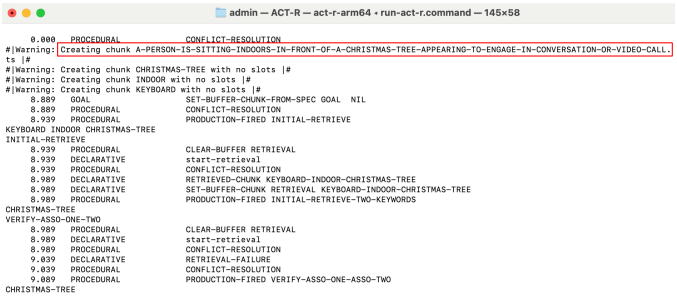

Fig. 4. ACT-R model creates chunks for keywords and phrase

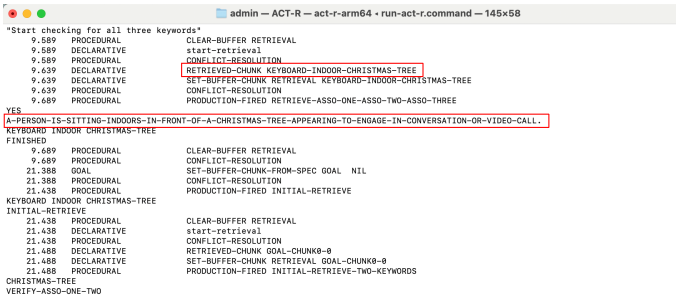

Fig. 5. ACT-R model retrieves chunk with keywords and phrase from declarative memory

keywords, into the declarative memory in order to retrieve them later in the outlined manner.

D. Prompting the LLM

We chose GPT-4o to create the conversational parts of the robot. The OpenAI API provides various hyperparameters that can be used to control the model behavior during an API call. We set the values of *temperature*, *presence penalty* and *frequency penalty* to zero in order to obtain consistent responses and exclude any randomness as far as possible. For the instruction of the GPT model, we used prompts with zero-shot prompting [4] for the system role to have the LLM perform the desired tasks as a completion task.

For chunk generation, including a phrase that is to be stored in the *'phrase'* slot describing the image sent to GPT-4o, we have provided a system prompt like the following *'I am robot, my name is Pepper. I don't answer but create three keywords followed by a sentence with a short description describing what is in this image. The three keywords are enclosed by square brackets. The short sentence is enclosed by round brackets, for example [person, indoor, computer] (A sentence with a short description describing what is in this image).'*. The brackets only serve to facilitate subsequent processing and separation of different contents in the application.

For augmenting the prompt with a retrieved recollection we used for example *'I am robot, my name is Pepper. I can see what is described in the following:'* followed by the *phrase* recalled from declarative memory. In this way, the LLM could receive and process information about visual recollections stored in the past if similar sensory impressions and thus similar keywords are currently present.

IV. Results

Figure 4 shows the generation of individual chunks from the keywords *'keyboard'*, *'indoor'*, *'christmas tree'* and in a frame the phrase *'A person is sitting indoors in front of a christmas tree appearing to engage in conversation or video call.'* supplied by the LLM in the running ACT-R model. These chunks were automatically transferred to the declarative memory of the cognitive model, not only separately, but also as a combination of keywords and phrase. To be able to save keywords and phrases as chunks, spaces had to be replaced by a '-' due to ACT-R limitations. The procedure was reversed when the phrase was used for prompt augmentation.

A retrieval of a recollection by the ACT-R model is shown in Figure 5 in framed areas. The model searches the stored memory chunks using the supplied keywords to find a matching recollection, the content of which is passed from the *'phrase'* slot to the robot application for use in prompt augmentation. The robot then included the knowledge from declarative memory in its answer – especially when asked directly about visual impressions.

V. Discussion

On the one hand, cognitive models can be used within the framework of their cognitive architectures to anticipate human behavior and thus make it more comprehensible. On the other hand, cognitive processes can use declarative and procedural memory to produce decisions which, when used to control the actions of a robot, allow the robot to act in a more humane way. Such a system can therefore serve both to make human actions intelligible for the machine in HRI and to make the robot's actions more human and therefore easier

for people to understand. The possibility of using cognitive architectures to unlock human-like judgment and decision-making capabilities such as instance-based learning and intuitive decision making for social robots can be an opportunity for greater acceptance and trust through more common ground in the way we think. With the help of LLMs and VLMs and the corresponding sensors, robots can access and interpret the same information that is available to humans. The use of declarative and procedural memory in cognitive models enables the data provided by robot sensors and language models to be processed in a human-centered way. We only outline the possibilities here with our example application for storing visual impressions together with keywords and using them for prompt augmentation. In general, there are far more conceivable use cases and options available for a programmatic implementation of cognitive models with ACT-R in combination with social robots.

VI. Limitations

The proposed system and procedure is not without shortcomings and open problems. It is well known that LLMs and VLMs are sometimes prone to hallucinations. One possible way to reduce the risk of such hallucinations and the reproduction of made-up statements by an LLM such as OpenAI's GPT would be to provide the language model with additional information tailored to individual scenarios via the system prompt. To concretize the prompt information, relevant memories of a cognitive system could be a useful addition. However, even such a method of constraining the LLM does not offer absolute certainty for the exclusion of hallucinations.

So far, we have only tested our method with a few memory contents or recollections, but we consider it scalable with regard to more complex cognitive models and the use of concepts such as *forgetting*, *learning new facts* and *reinforcement of recollections* or further possibilities of the cognitive architecture of ACT-R.

The time required to retrieve relevant recollections could play a limiting role depending on the amount of memory chunks available. In addition, for each interaction with the human user, our system requires two consecutive API calls to the GPT model, both of which have a certain latency. Together, these time aspects may cause a noticeable and unnatural delay in the interaction. And it also seems to be a disadvantage and increases the complexity that the implementation of the cognitive architecture runs as a standalone version on an extra computer and not on the robot itself. This problem concerns the Pepper robot and could possibly be solved by using other robot models and thus other ways of implementing the cognitive architecture.

VII. Conclusion and future work

We proposed a design and development approach for a combined system consisting of an ACT-R cognitive model and a humanoid social robot. Recollections from the declarative memory of the ACT-R model could be retrieved using real-world data obtained by the robot via an LLM with vision abilities. The procedural memory, which consists of the productions of the cognitive model, was used to retrieve these recollections and return them to the robot as instructions for action. In addition, real-world data captured by the robot could be stored as memory chunks in the cognitive model's declarative memory.

We improved the correctness and accuracy of GPT-4o's reasoning capabilities by using recollections for prompt augmentation. Our proof-of-concept application delivered keyword labeled visual impressions of the robot to the declarative memory or retrieved recollections based on these impressions. This method can of course also be used with other than visual data. We testet this system with the social robot Pepper. However, this method can also be used with other robot models and also independently of HRI scenarios.

Regarding explainability aspects, our system provides approaches for a possible constraint of LLMs to generate robot utterances in HRI by comprehensible memory contents. The type of possible connection or integration of the cognitive model depends on the robot's operating system and whether or which ACT-R implementations are available for it. For example, there is a direct implementation for Python, which means that the standalone version of ACT-R on an external computer and thus also the TCP/IP connection between the cognitive model and the robot application could be avoided.

The use of a cognitive architecture such as ACT-R enables the inclusion and investigation of further cognitive principles and processes in interaction with a social robot and LLMs independent of declarative memory retrieval. Future work should consider these aspects in more detail. Further experiments are needed to optimize the ACT-R models and system prompts for the LLM, as well as ongoing evaluation in studies with different people and different robot systems for various tasks. We are confident that this will open up a wide range of possibilities for future research into how cognitive architectures and their models can add a human touch to a social robot in HRI.

References

[1] Aldebaran, United Robotics Group and Softbank Robotics. Pepper. Technical report, 2025. URL https://www.aldebaran.com/en/pepper.

[2] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. 111(4):1036–1060, 2004. doi: 10.1037/0033-295X.111.4.1036.

[3] Manuel Birlo and Adriana Tapus. The crucial role of robot self-awareness in hri. In *Proceedings of the 6th International Conference on Human Robot Interaction*, pages 115–116, 03 2011. doi: 10.1145/1957656.1957688.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel

Ziegler, Jeffrey Wu, Clemens Winter, and Dario Amodei. Language models are few-shot learners, 05 2020.

[5] Carnegie Mellon University. Act-r software. Technical report, 2024. URL http://act-r.psy.cmu.edu/software/.

[6] Souvik Dubey, Ritwik Ghosh, Mahua Dubey, Subhankar Chatterjee, Shambaditya Das, and Julián Benito-León. Redefining cognitive domains in the era of chatgpt: A comprehensive analysis of artificial intelligence's influence and future implications. *Medical Research Archives*, 12, 06 2024. doi: 10.18103/mra.v12i6.5383.

[7] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions, 2024. URL https://arxiv.org/abs/2404.07214.

[8] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(03)00031-4. URL https://www.sciencedirect.com/science/article/pii/S0364021303000314.

[9] Cleotilde Gonzalez, Varun Dutt, and Christian Lebiere. Validating instance-based learning mechanisms outside of act-r. *Journal of Computational Science*, 4(4):262–268, 2013. ISSN 1877-7503. doi: https://doi.org/10.1016/j.jocs.2011.12.001. URL https://www.sciencedirect.com/science/article/pii/S1877750311001086. PEDISWESA 2011 and Sc. computing for Cog. Sciences.

[10] Miguel Á. González-Santamarta, Francisco J. Rodríguez-Lera, Ángel Manuel Guerrero-Higueras, and Vicente Matellán-Olivera. Integration of large language models within cognitive architectures for autonomous robots, 2024. URL https://arxiv.org/abs/2309.14945.

[11] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning, 2023. URL https://arxiv.org/abs/2311.17842.

[12] Hyeongyo Jeong, Haechan Lee, Changwon Kim, and Sungtae Shin. A survey of robot intelligence with large language models. *Applied Sciences*, 14(19), 2024. ISSN 2076-3417. doi: 10.3390/app14198868. URL https://www.mdpi.com/2076-3417/14/19/8868.

[13] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 2011. ISBN 9780374275631 0374275637.

[14] Kento Kawaharazuka, Yoshiki Obinata, Naoaki Kanazawa, Kei Okada, and Masayuki Inaba. Robotic applications of pre-trained vision-language models to various recognition behaviors. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, page 1–8. IEEE, December 2023. doi: 10.1109/humanoids57100.2023.10375211. URL http://dx.doi.org/10.1109/Humanoids57100.2023.10375211.

[15] Kobe Knowles, Michael Witbrock, Gillian Dobbie, and Vithya Yogarajan. A proposal for a language model based cognitive architecture. *Proceedings of the AAAI Symposium Series*, 2024. URL https://api.semanticscholar.org/CorpusID:267206594.

[16] Daniel Leidner. Toward robotic metacognition: Redefining self-awareness in an era of vision-language models. *40th Anniversary of the IEEE Conference on Robotics and Automation*, 2024. URL https://elib.dlr.de/207493/.

[17] Evelina Leivada, Gary Marcus, Fritz Günther, and Elliot Murphy. A sentence is worth a thousand pictures: Can large language models understand hum4n l4ngu4ge and the w0rld behind w0rds?, 2024. URL https://arxiv.org/abs/2308.00109.

[18] Artem Lykov, Mikhail Litvinov, Mikhail Konenkov, Rinat Prochii, Nikita Burtsev, Ali Alridha Abdulkarim, Artem Bazhenov, Vladimir Berman, and Dzmitry Tsetserukou. Cognitivedog: Large multimodal model based system to translate vision and language into action of quadruped robot. HRI '24, page 712–716, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703232. doi: 10.1145/3610978.3641080.

[19] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, and Silin Chen. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges, 2024. URL https://arxiv.org/abs/2409.02387.

[20] OpenAI. The most powerful platform for building ai products. Technical report, 2025. URL https://openai.com/api/.

[21] Thomas Sievers and Nele Russwinkel. How to use a cognitive architecture for a dynamic person model with a social robot in human collaboration. *CEUR Workshop Proceedings*, 2024. URL https://ceur-ws.org/Vol-3794/paper2.pdf.

[22] Thomas Sievers, Nele Russwinkel, and Ralf Möller. What am I? – complementing a robot's task-solving capabilities with a mental model using a cognitive architecture. *CEUR Workshop Proceedings*, 2024. URL https://ceur-ws.org/Vol-3906/paper1_BAILAR.pdf.

[23] Thomas Sievers, Nele Russwinkel, and Ralf Möller. Grounding a social robot's understanding of words with associations in a cognitive architecture. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 406–410. INSTICC, SciTePress, 2025. ISBN 978-989-758-737-5. doi: 10.5220/0013144200003890.

[24] Ron Sun. Can a cognitive architecture fundamentally enhance llms? or vice versa?, 2024. URL https://arxiv.org/abs/2401.10444.

[25] Ryan Tan, Harry Yang, Sean Jiao, Liuyang Shan, Chenxi Tao, and Roger Jiao. Smart robot manipulation using gpt-4o vision. In *7th European Industrial Engineering and Operations Management Conference, Augsburg, Germany*. IEOM Society International, 2024. doi: 10.46254/EU07.20240272.

[26] Robert Thomson, Christian Lebiere, John R. Anderson,

and James Staszewski. A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*, 4(3):180–190, 2015. ISSN 2211-3681. doi: https://doi.org/10.1016/j.jarmac.2014.06.002. URL https://www.sciencedirect.com/science/article/pii/S2211368114000539. Modeling and Aiding Intuition in Organizational Decision Making.

[27] Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. Gpt-4v(ision) for robotics: Multimodal task planning from human demonstration, 2024. URL https://arxiv.org/abs/2311.12015.

[28] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 2024. ISSN 2949-8554. doi: https://doi.org/10.1016/j.jai.2024.12.003. URL https://www.sciencedirect.com/science/article/pii/S2949855424000613.

[29] A. Werk, S. Scholz, T. Sievers, and N. Russwinkel. How to provide a dynamic cognitive person model of a human collaboration partner to a pepper robot. *ICCM*, 2024. URL https://mathpsych.org/presentation/1452.

[30] Siyu Wu, Alessandro Oltramari, Jonathan Francis, C. Lee Giles, and Frank E. Ritter. Cognitive llms: Towards integrating cognitive architectures and large language models for manufacturing decision-making, 2024. URL https://arxiv.org/abs/2408.09176.

[31] Takahide Yoshida, Suzune Baba, Atsushi Masumori, and Takashi Ikegami. Minimal self in humanoid robot "alter3" driven by large language model, 2024. URL https://arxiv.org/abs/2406.11420.