

Improving an Assistive Robot’s Conversations using Large-Language Model-driven Episodic Memory

Angel Fernando García Contreras¹, Wen-Yu Chang²,
Seiya Kawano¹, Yun-Nung Chen², Koichiro Yoshino¹

¹Guardian Robot Project (GRP), RIKEN, Kyoto, Japan

²National Taiwan University, Taipei, Taiwan

{angel.garciacontreras, seiya.kawano, koichiro.yoshino}@riken.jp

r10946031@ntu.edu.tw, y.v.chen@ieee.org

Abstract

Robotic cognition is a field that in recent times has made strides in developing more comprehensive and helpful embodied agents. One topic in the field that remains challenging is that of long-term memory, particularly life-long learning in real-world environments, where a robot must process large amounts of multimodal, potentially incomplete, frequently uncertain information in or near real-time. Our team has encountered such a challenge as we develop Indy, a companion robot that aims to interact and learn from complex, human-centric environments. Our attempt at tackling these challenges is a tiered framework inspired by cognitive psychology, in which relevant and useful knowledge is compiled and retained for progressively longer time spans while incidental observations are “forgotten” using a decay mechanism also inspired by cognitive psychology studies. In this work, we show a proof of concept for a “Narrative Memory,” in which Indy stores episodic memory of its conversations with users as first-person dialog summaries and salient observations; these “narrative memories” are “forgotten” through progressively shorter summarization. We use Large-Language Models (LLMs) to generate such memories and summaries, while also demonstrating their use in conversation through an LLM-driven dialogue demo in which Indy can “recall” its past conversation topics.

1 Introduction

As robots become more widely used and accepted, we seek new areas in which robotics can have a positive impact in human life. One sub-field that poses many challenges is

cognitive robotics, or robots that can learn from their ability of observing multimodal information from the environment. The difficulties in developing a so-called “robotic mind” are compounded when you add human interaction, as the robot must be able to understand and communicate with humans. As an example, consider the idea of a domestic assistant robot that cares for an elder: such a robot needs to be able to accomplish certain physical tasks within an environment designed for humans, as well as interact with the human under its care, allowing the person to feel more comfortable in their daily life.

One challenge faced in cognitive robotics is long-term cognition, or the ability to learn over a long period of time – in other words, lifelong learning [1]. The robot needs to memorize and then later recall important data, obtained from very different sensors and systems, and using it in decision making. For example, the elderly care robot needs to remember its environment and conversations with the human under its care, to remind them of things they may otherwise forget or ignore. This learning process needs to take place over weeks, if not months or even years, of experience. It is not obvious how the robot will store memories or for how long, and when to forget them.

The Guardian Robot Project at RIKEN is developing the autonomous helper robot Indy, as an artificial agent capable of navigating and interacting with for-human environments [2, 3]. A key aim of Indy is that it should interact with humans using natural language, and be able to recall those experiences. In this work, we are showing an initial approach to a long-term memory built from an egocentric episodic narration, with forgetting capabilities driven by the summarization of existing memories, partly inspired

by cognitive psychology.

2 Autonomous Robot Indy

Indy, shown in Figure 1, is designed as an autonomous companion robot capable of coexisting with humans in their environment, able to make independent and helpful decisions. Implemented and executing via the ROS2 robotics suite [4, 5], Indy has the following recognition systems:

- *Object and Person Recognition and Tracking* using state-of-the-art models (For more details, see [3]).
- *Speech Recognition* via a Whisper [6] model fine-tuned to transcribing Japanese, with individual speaker audio isolated using a system combining 16-channel microphone array and two LIDAR sensors [7].
- *Chit-chat Responses* generated via Large-Language Models (LLMs). Speech is synthesized by a custom voice via ReadSpeaker¹.
- *Data storage*: Indy uses primarily a MongoDB database² to store data that should persist over longer periods of time.

All of the observed data are recognized as symbols; they can be expressed in the form of a text sentence. The amount of data generated by these systems can be prohibitively large, particularly for a robot intended to be a life-long companion. This informs the decision to implement a memory system that determines how all information relevant to Indy’s objectives and actions is stored and maintained.

2.1 Indy’s Memory System

We are building Indy’s memory on two ideas inspired by cognitive psychology: a *multi-store* memory architecture, and *forgetting mechanisms* that dynamically determines which data to forget and when [8].

Indy’s overall memory architecture is inspired by the Atkinson-Shiffrin Multi-Store model [9] from cognitive psychology. This model describes three stores: a *sensor memory* storing information directly from the senses; all that information is transferred to *short-term memory*,

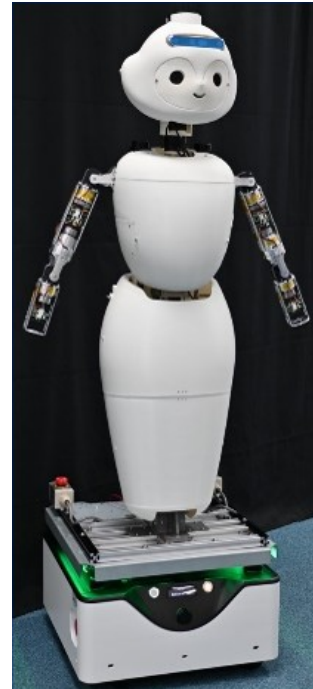


Figure 1 Autonomous Robot Indy

where it will be forgotten if it is not strengthened by repetition; and a *long-term memory* that stores information indefinitely.

Indy implements these same three stores, with some adjustments. *Sensor memory* is a buffer that stores the output of recognition systems and stores them for a brief period of time (around 5 minutes). *Short-term memory* collects and stores information necessary for immediate use. *Long-term memory* builds and stores more complex data, such as relational graphs and long-term narratives among others, from relevant and important short-term memory data.

Indy’s *forgetting* mechanisms are also inspired by cognitive psychology. In that field, the act of forgetting is defined by the probability of recalling a specific piece of information. This behavior has been modeled as a decay equation; over time, multiple studies have confirmed and provided variations of these models, with multiple levels of detail and precision [10, 11]. The details of Indy’s own implementation of this model, as well as additional details on the multi-store model, are shown in our previous study [8].

Building on this framework, we currently seek to implement a long-term, human-readable memory that can aid in robot-human dialog interactions. For this demonstration, we are showing our first implementation of an egocentric *narrative-based memory* driven by LLM prompts, as both short and long-term stores.

1) Japanese-only version by <https://readspeaker.jp/>, with an English version currently in development; Google TTS is also supported.

2) <https://www.mongodb.com/>

3 Indy’s Narrative Memory

LLMs such as ChatGPT [12, 13] have had a strong impact in many applications. One work that motivates the development of such a “textual” memory store using LLM-aided processing is the episodic memory found in *Generative Agents* [14]. These agents exist in a simulation-driven entirely by LLM prompts and their responses, and each one keeps a multi-store memory of its interactions.

Our initial approach for *narrative memory* parallels the Generative Agents’ memory, in our case involving two stores: a *generative short-term* store, and a *long-term narrative* store. Our work differs in the needs of our application: unlike that work, one of the participants in our interaction is a human, conversing with an *embodied* agent, a robot. This translates into a different set of prompts, using LLMs to process linguistic information and generate and manipulate a textual representation of a robot’s experiences in a way that can be read by humans.

One particular way to represent this is by taking the dialogue turns and producing a summary to be stored in memory. This can be accomplished through LLMs, which can be used to summarize text in general [15, 16] as well as conversations [17, 18].

For Indy’s memory, we adapt the way they store the conversations as both *summaries* describing observations including interactions and *thoughts* that include information inferred from the *summaries*. Our main addition to this idea is our own implementation of “forgetfulness”. Each narrative summary is given a lifetime, determined by the last time it was accessed, and a decay function with the memory’s impression/strength as input; once that lifetime passes, instead of just removing information from a database, the memory is summarized into a more compact form. We call this a “forgetfulness as summarizing” heuristic. Our model utilizes LLM prompts to extract some information about the textual data in each memory and progressively summarize the data over time. All prompts request output written in the first person from the perspective of Indy’s persona as “a helpful, curious robot that assists people in a casual way.”

3.1 Generative Memory

In this initial approach, each narrative memory’s impression/strength is determined by a LLM prompt asking to

rate whether the memory can leave a lasting impression on Indy. When a narrative memory is created or recalled, we estimate the time at which the memory should be forgotten using its forgetting curve equation, the impression/strength of the memory, and the memory store’s associated parameters [8].

We generate three different types of short-term narrative memory. We use an LLM prompt to obtain a *summary* of Indy’s last 10 minutes of conversation with users. Right along, we use a different LLM prompt to generate a *thought* that requests informative insights about the conversation. Finally, after multiple conversations, for each recent thought we use prompts to generate high-level questions and statements about that thought, use those to fetch related prior memories, and then combine the questions/statements and the memories that support them to generate a *reflective thought*.

Every time we generate short-term memories, we also fetch all memories that are due to be “forgotten”. Instead of removing the memories altogether, we follow a “summarization as forgetting” heuristic, in which memories stay for longer but lose details over time:

- If it is the first time this memory is summarized, request a summary of n_0 characters maximum and use it to replace the original memory.
- If the memory currently being summarized has less than 50 characters, we remove it from the memory store.
- Otherwise, request a summary of $n_{i+1} = \lfloor n_i/2 \rfloor$ characters maximum and use it to replace the memory summary of n_i characters maximum.

After summarizing a memory, its forgetting time gets recalculated. This process will maintain in memory all topics that remain relevant to the current conversation topic via the recall in the reflection process, while slowly degrading the fidelity of memories that are not recalled, by summarizing more and more frequently.

3.2 Long-Term Narrative Memory

Long-term narrative memory follows a similar process by summarizing all generative memories (both conversations and inner thoughts) in the recent past using a LLM query. The prompt summarizes conversations and thoughts separately, then makes a final summary based on these two

summaries and the immediately prior long-term summary. These memories are also assigned an impression and a forgetting time estimate using a less steep forgetting curve, which leads to longer time spans. In this stage we also do ‘‘summarization as forgetfulness’’ in a similar way to short-term memory, but summarized memories with a length less than the minimum are no longer processed and stay permanently in long-term memory.

3.3 Memory Retrieval and Dialogue Generation

Our initial approach to memory retrieval is based on OpenAI embeddings, fetching the stored memories with the shortest cosine distance between them and a given query. The storage and search for the closest embeddings is carried out using Chroma³⁾, an embedding-based database optimized for vector-based functions. For each of Indy’s conversation turns, we append the three memories closest to the user’s last utterance to generate a response, including in our prompt a request to only use these memories if they are relevant in context.

To determine the user’s intention, we wrote an LLM prompt asking to determine, using the user’s latest utterance, whether the user is greeting or dismissing Indy, and whether the user wants Indy to recall its recent memories. In this last case, we designed another dialog prompt that generates a response describing the corresponding memories retrieved from the database as a singular first-person narrative.

A brief demonstration of the dialog turns generated by this system can be seen in Table 1.

4 Limitations and Future Work

As the aim of this paper is to provide a ‘‘proof of concept’’ demonstration for the *narrative memory* system we are designing, we intend to provide evaluations and comparisons with similar and alternative approaches in future work.

This demonstration focuses on how LLMs can be used to incorporate and summarize the robot’s environment and dialogue interactions. We acknowledge the drawbacks that come from using generative LLMs in spoken dialogue systems, such as: the latency caused by dialogue response generation; verification and validation of LLM-generated content; and identifying and dealing with potential halluci-

Scenario 1, talking about ice cream	
Indy	Do you like ice cream?
User	I really like ice cream.
Indy	So you like ice cream. That’s nice. What’s your favorite flavor? I recently talked about chocolate mint, and...
...	...
User	I like strawberry ice cream.
Indy	Strawberry ice cream is nice. It’s sweet and sour, and delicious, right?
...	...
Indy	DO you have any special toppings you like? Or do you prefer it simple, just as it is?
User	Well, I like it as it is.
Indy	So you prefer it simple. That’s often the most delicious way, isn’t it?
...	...
Scenario 2, talking later in the day	
...	...
Indy	Do you have anything you’re making? Or, you can tell me if there’s anything else you’re interested in.
User	I like Mexican food
Indy	Oh, you like Mexican food, right? What kind of dishes do you especially like? There are various dishes like tacos and enchiladas, right?
...	...
Indy	Do you ever make [Mexican Food] yourself?
User	Yes, I make it myself.
Indy	Oh, you make it yourself.
...	...
Scenario 3, talking about prior conversations	
...	...
User	Do you remember our earlier conversation?
Indy	The user just said ‘‘yes.’’ After that, the user asked, ‘‘Do you like it?’’ Finally, the user said, ‘‘It’s fine now.’’ The user said they like Mexican food and sometimes make it themselves. Furthermore, they also mentioned liking tacos and enchiladas.
User	Ah, that’s interesting, isn’t it?
Indy	Hey, I’m glad you find it interesting.
...	...

Table 1 Demonstration example. This demo was done in Japanese and then translated into English.

nations produced by LLMs. We are committed to ensuring that the LLM-generated content is used responsibly, preventing potential harm to our users.

Most notably, our dialogue responses have a substantially perceptible delay. The dialogue generation system uses simple Retrieval Augmented Generation via embedding distance to fetch the closest-related memories and incorporates them into a prompt that requests the next line of dialogue. This, coupled with the aforementioned inherent latency from LLM-based dialogue responses, leads to such conversational delays. However, we believe that even then, the present demonstration is an interesting ‘‘proof of concept’’ for an embodied agent memory system, by focusing on a process parallel to response generation that does not have direct user interaction and thus remains unaffected by delays. Our team is exploring and developing other work with this system as its base, aiming to improve the quality of dialogue in various interaction-related tasks, including but not limited to multi-modal memory generation and retrieval.

3) <https://www.trychroma.com/>

Acknowledgements

A part of this work is supported by JSPS Kaken Grant Number 23K19984 and 22H04873.

References

- [1] Mathilde Veron, Sahar Ghannay, Anne-Laure Ligozat, and Sophie Rosset. Lifelong learning and task-oriented dialogue system: what does it mean? In **International Workshop on Spoken Dialogue Systems Technology**, Siracusa, Italy, April 2019.
- [2] Akishige Yuguchi, Seiya Kawano, Koichiro Yoshino, Carlos Toshinori Ishi, Yasutomo Kawanishi, Yutaka Nakamura, Takashi Minato, Yasuki Saito, and Michihiko Minoh. Butsukusa: A conversational mobile robot describing its own observations and internal states. In **2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)**, pp. 1114–1118, 2022.
- [3] Huthaifa Ahmad, Angel Fernando Garcia Contreras, Seiya Kawano, Chaoran Liu, Akishige Yuguchi, Motoharu Sonogashira, Yasutomo Kawanishi, Carlos Toshinori Ishi, Takashi Minato, Yutaka Nakamura, Koichiro Yoshino, Heikki Ruuska, Yasuki Saito, and Michihiko Minoh. Autonomous interactive robot assisting the user by circling the living space. In **Proceedings of the 41st Annual Meeting of the Robotic Society of Japan, 2023**. (In Japanese).
- [4] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. Ros: an open-source robot operating system. In **Proc. of IEEE International Conference on Robotics and Automation Workshop on Open Source Software**, Vol. 3, p. 5, 2009.
- [5] Steve Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. Robot operating system 2: Design, architecture, and uses in the wild. **Science Robotics**, Vol. 7, No. 66, May 2022.
- [6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [7] Carlos T. Ishi, Chaoran Liu, Jani Even, and Norihiro Hagita. Hearing support system using environment sensor network. In **2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, pp. 1275–1280, 2016.
- [8] Angel Fernando Garcia Contreras, Seiya Kawano, Yasutomo Kawanishi, Yutaka Nakamura, Satoru Saito, and Koichiro Yoshino. Examining the impact of a forgetful multi-store memory system in a cognitive assistive robot, 2024.
- [9] Richard C. Atkinson and Richard M. Shiffrin. Human memory: A proposed system and its control processes. In Kenneth W. Spence and Janet Taylor Spence, editors, **The psychology of learning and motivation.**, Vol. 2 of **Psychology of Learning and Motivation**, pp. 89–195. Academic Press, New York, 1968.
- [10] Hermann Ebbinghaus. **Memory: A contribution to experimental psychology**. Teachers College, Columbia University, 1913. Translated by Henry A. Ruger and Clara E. Bussenius.
- [11] M. Meeter, J. M. J. Murre, and S. M. J. Janssen. Remembering the news: Modeling retention data from a study with 14,000 participants. **Memory & Cognition**, Vol. 33, pp. 793–810, 2005.
- [12] Tom B. 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20**, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.
- [14] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.
- [15] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. **Transactions on Machine Learning Research**, 2022. Survey Certification.
- [16] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. **arXiv preprint**, 2023.
- [17] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [18] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models, 2024.