

An Inspectable Neuro-Symbolic Pipeline for Routine Elicitation in Conversational Health Coaching

Shakyani Jayasiriwardene*
The University of Sydney
djay0399@uni.sydney.edu.au

Blake Mountford*
The University of Sydney
bmou8324@uni.sydney.edu.au

Meican Ma
The University of Sydney
mema2458@uni.sydney.edu.au

Weiwei Jiang
Nanjing University of Information
Science and Technology
weiweijiangcn@gmail.com

Niels van Berkel
Aalborg University
nielsvanberkel@cs.aau.dk

Nicholas A. Koemel
Monash University
nicholas.koemel@monash.edu

Matthew Ahmadi
Monash University
matthew.ahmadi@monash.edu

Jorge Goncalves
University of Melbourne
jorge.goncalves@unimelb.edu.au

Emmanuel Stamatakis
Monash University
emmanuel.stamatakis@monash.edu

Anusha Withana
The University of Sydney
anusha.withana@sydney.edu.au

Zhanna Sarsenbayeva
The University of Sydney
zhanna.sarsenbayeva@sydney.edu.au

Abstract

Large language model (LLM)-based conversational health coaches can support personalised behaviour change interventions, but their decision-making processes often remain opaque. This lack of transparency makes it difficult for developers and users to understand what information has been retained, how recommendations are generated, and why specific intervention decisions are made. In this paper, we present an inspectable neuro-symbolic conversational architecture for a conversational health coach that combines LLM-based information extraction, a persistent knowledge graph, and finite state machine dialogue management. The architecture transforms unstructured conversations into explicit symbolic representations, allowing conversational decisions to be traced from user utterances to intervention recommendations. We describe how the proposed approach has the potential to support transparency, inspection, and debugging throughout the conversational pipeline. In a feasibility pilot with 16 participants, knowledge-graph-derived schedule summaries received a mean perceived-fidelity rating of 3.8/5. The architecture provides technical affordances for traceability that should be evaluated in future user- and developer-centred studies.

CCS Concepts

• **Human-centered computing** → **Interaction design theory, concepts and paradigms**; *Systems and tools for interaction design*.

1 Introduction

Large language models (LLMs) have enabled conversational agents to engage in increasingly natural and context-aware interactions [10, 11]. However, contemporary LLM-based conversational systems often provide limited visibility into what information has been retained, how that information is represented, and how it informs

subsequent system behaviour. This is particularly important in conversational health coaching, where user-provided contextual information may influence when and how interventions are delivered [1, 8].

Neuro-symbolic AI is a promising direction for addressing these limitations by integrating the flexibility of neural language models with explicit symbolic representations and reasoning mechanisms [3]. In this work, we present a neuro-symbolic architecture for a conversational physical-activity coach that combines LLM-based information extraction, a persistent knowledge graph representing users' daily routines, and a finite-state-machine dialogue controller. These components externalise intermediate system states, allowing developers and researchers to trace how conversational information is extracted, represented, and used to identify potential opportunities for walking interventions.

We report a feasibility pilot study with 16 participants who discussed their daily schedules with the system and reviewed the resulting conversation transcript, knowledge graph, state-machine representation, and a schedule summary. The study provides an initial assessment of whether conversationally elicited routine information can be externalised into structured representations that reflect what users intended to communicate. We treat this work as an early step toward making conversational health-coaching pipelines more inspectable and identify directions for evaluating whether such representations support stakeholder understanding.

This paper contributes: (1) an inspectable neuro-symbolic pipeline that externalises routine extraction, persistent memory, dialogue state, and trigger-time identification; and (2) preliminary evidence concerning the perceived fidelity of its structured routine representations.

2 Related Work

Conversational agents have long been used in health and behaviour-change settings, from relational agents and mental-health chatbots

*Both authors contributed equally to this research.

Table 1: Typed schema of the information extraction module and knowledge graph.

Entity	Attributes
place	name
event	name; start_iso, end_iso, duration; vicinity (past/present/future); frequency
free_time_window	coarse label (morning... night); start_iso, end_iso, duration; vicinity; frequency
Relation	Signature
occurs_at	event → place
precedes	event/free_time_window → event/free_time_window (temporal order)

to newer LLM-based systems supporting coaching, triage, and reflective dialogue [2, 5]. LLMs make these systems more natural and context-aware, but they also make it harder to inspect what information has been retained and how it influences subsequent system behaviour. This has motivated growing interest in explainable AI, particularly in health-related contexts where transparency and accountability are important [1, 8]. While many approaches rely on post-hoc explanations [13], such explanations may not faithfully reflect the system’s actual computation [6]. Explainability-by-design instead seeks to make relevant internal representations and decision processes explicit from the outset [17].

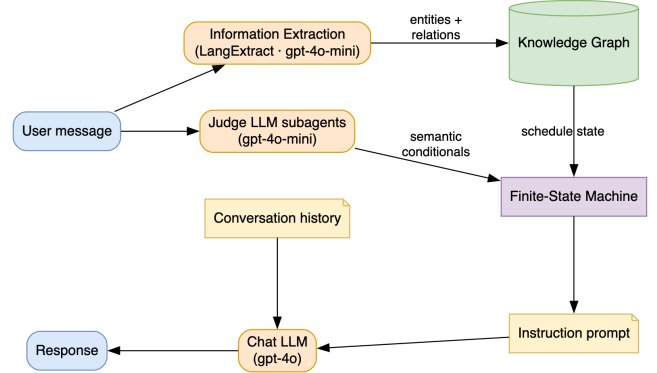
Neuro-symbolic AI offers one way to support this goal by combining neural language processing with explicit symbolic structures [3]. Knowledge graphs can provide persistent and queryable representations of user context [15], while state-machine controllers can make dialogue progression explicit and controllable [19]. LLMs can additionally bridge natural language and these symbolic representations through structured information extraction [4]. Prior work has commonly used these components separately, for example for memory or dialogue control [14, 18]. Our work combines LLM-based information extraction, a persistent knowledge graph, and finite-state-machine dialogue control within a single conversational health-coaching pipeline to make intermediate system states more inspectable and traceable.

3 Inspectable Neuro-Symbolic Architecture

3.1 System Architecture

Fig. 1 shows the response generation pipeline. Schedule information is extracted using LangExtract [7] and stored in a Python-based knowledge graph (KG) following the schema in Table 1. The KG and subagent evaluations update the finite state machine (FSM), which governs dialogue flow by selecting the desired behaviour for each turn. The current FSM state is then used to construct the system prompt for GPT-4o, which generates the assistant response.

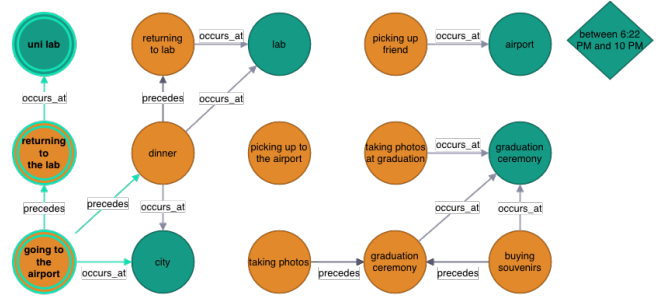
For each conversational turn, the most recent four messages (two conversational turns) are passed to GPT-4o-mini for single-pass structured information extraction, following the extraction window used by Zep [16]. Timestamps and speaker roles are included, and extracted entities are assigned temporary IDs to support relation mapping.

**Figure 1: Response generation pipeline**

3.2 Knowledge Graph (KG)

Once entities and relations are extracted from the recent transcript, these are merged into the KG, seen in Fig. 2. Temporary entity IDs are generated during extraction in a standardised form for the purpose of matching extracted relations to their corresponding entities. Deduplication is performed using name matching and checking for matching numerical attributes, as well as temporal merging.

The KG is expanded throughout the conversation and persists after completion. Subsequent conversations will then resume with the persisted KG, maintaining the core knowledge through each interaction. This constitutes a structured memory system [16], supporting the persistence and subsequent retrieval of routine information shared across conversations. This KG is formatted into text and injected into the agent’s system prompt, functioning as a structured and stable temporal memory layer.

**Figure 2: Knowledge Graph. (Teal nodes: locations; orange nodes: events; Diamond node: free time window. Green highlights indicate new additions.)**

3.3 Finite State Machine (FSM)

The FSM governs dialogue flow using explicit states, prompt templates, and transition rules, drawing on prior work in task-oriented dialogue systems and structured LLM workflows [12, 21]. Each state specifies the desired agent behaviour, while the current knowledge

graph is inserted into the prompt to provide user context. Transitions are triggered either by deterministic conditions or semantic checks performed by a Judge-LLM subagent (GPT-4o-mini) [20].

This structured control improves traceability by making dialogue progression explicit [12]. As shown in Fig. 3, unexpected behaviours can be traced back to particular states, prompts, or transition conditions, allowing developers to diagnose and refine the conversation flow more systematically.

The dialogue consists of two phases: *Inquiring* and *Confirming*. During Inquiring, the agent gathers and clarifies missing schedule information. During Confirming, it summarises the inferred routine, proposes potential free-time windows, and invites user corrections before concluding the conversation. Suggesting a free-time window can also prompt users to reconsider their availability and reveal previously unmentioned constraints.

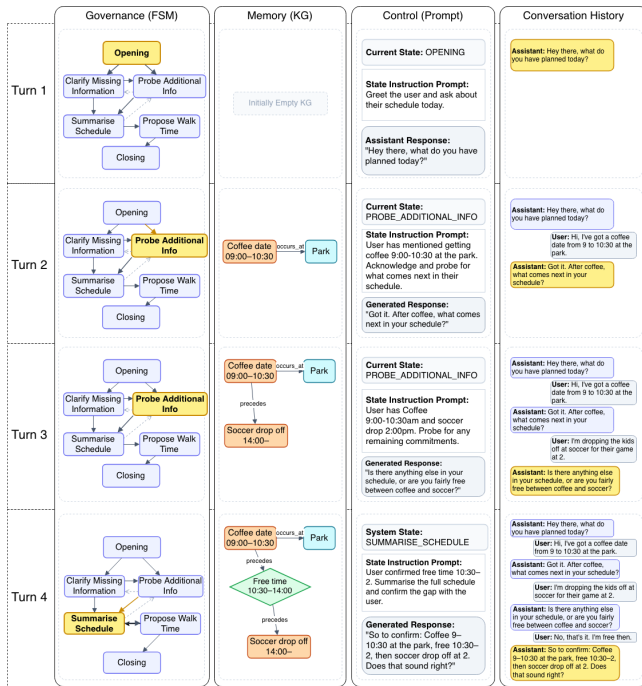


Figure 3: After each user message, the FSM transitions into a new state (yellow), the KG expands with new entities and relations, and the instruction prompt is rebuilt for generation of the response

3.4 Intervention Selection

Information provided by users is often incomplete; not all event attributes are certainly known to the user and may not even be relevant for the goal of intervention scheduling. Additionally, temporal information may be given in relative terms, rather than absolute terms, e.g., “I leave the house at least one hour before work starts”. Our system is designed to handle incomplete and relative temporal information. The agent merely needs to collect information until a partial ordering can be established that implies the existence of a free-time window in the user’s schedule of a sufficient duration.

A partial ordering is constructed from temporal relations between events and free time windows. Once this partial ordering is found and confirmed, further schedule information is significantly less useful. The start and end times of the selected free time window need to be known in absolute terms, as this is the window within which the intervention must occur. Further, extracted start and end time attributes for events are used to identify “empty” periods of time, in which no event is likely to occur. The suspected availability of these periods of time is then clarified with the user at the end of the conversation. If confirmed, these times are selected for intervention, and 1-2 walk reminders are scheduled.

3.5 Feasibility Pilot Study

To assess the feasibility of the proposed architecture and its structured artefacts, we conducted a pilot study with 16 participants aged 25-35 years, all of whom reported regularly using digital technologies. Participants attended an in-lab session in which they interacted with the conversational health coach and described their schedule for the current day, from morning to night. The system used this information to identify potential times for delivering walking notifications. Each session lasted approximately 30-45 minutes, including the system interaction and participants’ review of the generated artefacts.

From each conversation transcript, the system generated a knowledge graph and a state-machine diagram. It also produced a textual schedule summary using only information contained in the knowledge graph, without access to the original conversation history. On average, the resulting knowledge graphs contained 8.00 events, 2.69 locations, and 5.81 temporal relationships, comprising 11.56 nodes and 12.06 edges per conversation.

Participants reviewed the conversation transcript alongside the generated artefacts, enabling them to compare the extracted information directly with what they had communicated. They then verbally rated the accuracy of the knowledge-graph-derived summary on a five-point scale (1 = completely inaccurate; 5 = completely accurate), and their responses were recorded by the researcher. Excluding one missing response, the summaries received a mean perceived-accuracy rating of 3.80 out of 5 ($SD = 0.62$, median = 4.00, range = 2.00-4.50). These results indicate generally positive perceived accuracy while also highlighting opportunities to improve the information-extraction and summarisation pipeline.

A complementary analysis of the same study reports participants’ qualitative feedback concerning the knowledge graph, perceived system understanding, correction behaviour, and broader interaction experience in greater detail [9]. The present paper instead focuses on the perceived fidelity with which conversationally elicited routine information was captured in the structured knowledge graph and reflected in the generated schedule summary.

4 Discussion & Limitations

Our architecture is designed to support developer-facing inspectability by externalising intermediate system states through information extraction, a knowledge graph, and finite-state-machine dialogue control. This allows developers to trace how conversational information is represented and subsequently used to identify potential intervention opportunities, while also helping localise errors to

specific stages of the pipeline. Our pilot study showed moderate perceived accuracy of summaries generated from the knowledge graph, suggesting that the structured representation captured much of the information participants intended to communicate while also revealing limitations in the extraction and summarisation process. An accepted companion paper provides a more detailed user-centred analysis of how participants interpreted and interacted with these representations [9].

These findings should nevertheless be considered preliminary. The present study involved a small sample, relied on perceived rather than objective accuracy, and did not compare the architecture against an end-to-end LLM baseline or evaluate the appropriateness of the resulting intervention decisions. Errors introduced during information extraction may also propagate through the knowledge graph and subsequent dialogue. A key next step is to compare the architecture against an end-to-end LLM baseline and evaluate whether its explicit intermediate representations improve representation accuracy, stakeholder understanding, auditing, and intervention quality.

5 Conclusion & Future Work

In this paper, we presented a neuro-symbolic conversational architecture for a conversational health coach that combines LLM-based information extraction, a knowledge graph, and finite-state-machine dialogue management to make intermediate system states explicit and traceable. Our preliminary study showed moderate perceived accuracy of the generated schedule representations, while also revealing opportunities to improve the information extraction and summarisation pipeline. Future work will examine how these representations can better support understanding and inspection, incorporate additional contextual information such as wearable and calendar data, and evaluate the architecture in longitudinal real-world deployments.

References

- [1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bénéttot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *arXiv:1910.10045 [cs.AI]* <https://arxiv.org/abs/1910.10045>
- [2] Timothy W. Bickmore, Suzanne E. Mitchell, Brian W. Jack, Michael K. Paasche-Orlow, Laura M. Pfeifer, and Julie O'Donnell. 2010. Response to a Relational Agent by Hospital Patients with Depressive Symptoms. *Interacting with Computers* 22, 4 (2010), 289–298. doi:10.1016/j.intcom.2009.12.001
- [3] Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. *arXiv:1905.06088 [cs.AI]* <https://arxiv.org/abs/1905.06088>
- [4] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv:2404.16130 [cs.CL]* <https://arxiv.org/abs/2404.16130>
- [5] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (2017), e19. doi:10.2196/mental.7785
- [6] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health* 3, 11 (2021), e745–e750. doi:10.1016/S2589-7500(21)00208-9
- [7] Akshay Goel and Atilla Kiraly. 2025. Introducing LangExtract: A Gemini powered information extraction library. Google Developers Blog. Accessed: [Insert access date here]. <https://developers.googleblog.com/introducing-langextract-a-gemini-powered-information-extraction-library/>
- [8] David Gunning, Mark Stefik, Jaesik Choi, Tim Miller, Simone Stumpf, and G. Z. Yang. 2019. XAI: Explainable Artificial Intelligence. *Science Robotics* 4, 37 (2019), eaay7120. doi:10.1126/scirobotics.aay7120
- [9] Shakyani Jayasiriwardene, Blake Mountford, Meican Ma, Niels van Berkel, Nicholas Koemel, Matthew Ahmadi, Jorge Goncalves, Emmanuel Stamatakis, and Zhanna Sarsenbayeva. 2026. Graphing the Everyday: A Neurosymbolic Approach to Eliciting Routines for Just-In-Time Adaptive Interventions. (2026). Accepted for publication at OzCHI 2026; forthcoming.
- [10] Shakyani Jayasiriwardene, Benjamin Tag, Anusha Withana, and Zhanna Sarsenbayeva. 2025. More Than Words: The Impact of Voice Assistant Personality Traits on Failure Mitigation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 91 (Sept. 2025), 33 pages. doi:10.1145/3749518
- [11] Shakyani Jayasiriwardene, Hongyu Zhou, Weiwei Jiang, Benjamin Tag, Nicholas Koemel, Matthew Ahmadi, Jorge Goncalves, Emmanuel Stamatakis, Anusha Withana, and Zhanna Sarsenbayeva. 2026. From Fixed to Flexible: Shaping AI Personality in Context-Sensitive Interaction. *arXiv:2601.08194 [cs.HC]* <https://arxiv.org/abs/2601.08194>
- [12] Myung Ho Kim. 2026. Bridging Symbolic Control and Neural Reasoning in LLM Agents: Structured Cognitive Loop with a Governance Layer. *arXiv:2511.17673 [cs.AI]* <https://arxiv.org/abs/2511.17673>
- [13] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) (NIPS '17). Curran Associates Inc., Red Hook, NY, USA, 4766–4777.
- [14] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560 [cs.AI]* <https://arxiv.org/abs/2310.08560>
- [15] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100
- [16] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. *arXiv:2501.13956 [cs.CL]* <https://arxiv.org/abs/2501.13956>
- [17] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. doi:10.1038/s42256-019-0048-x
- [18] Yiran Wu, Tianwei Yue, Shaokun Zhang, Chi Wang, and Qingyun Wu. 2024. StateFlow: Enhancing LLM Task-Solving through State-Driven Workflows. *arXiv:2403.11322 [cs.CL]* <https://arxiv.org/abs/2403.11322>
- [19] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.
- [20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685 [cs.CL]* <https://arxiv.org/abs/2306.05685>
- [21] Łukasz Bieniasz-Krzywiec, Kyle Boone, and Ryan Carl Yao Yu. 2025. *Constrained Response Generation for Conversational Generative AI Agents*. Technical Report. Technical Disclosure Commons. https://www.tdcommons.org/dpubs_series/9036