

Systematic Diagnosis and Benchmarking of Memory Systems in Autonomous AI Research Agents: A Low-Resource Framework

Nur Arifin Akbar
Università degli Studi di Palermo
Dipartimento Matematica e
Informatica
Palermo, Italy
nurarifin.akbar@unipa.it

Rahool Dembani
Singularlogic
Athens, Greece
rdembani@singularlogic.eu

Gregorius Airlangga
Atma Jaya Catholic University of
Indonesia
Jakarta, Indonesia
gregorius.airlangga@atmajaya.ac.id

Ripto Mukti Wibowo
Atma Jaya Catholic University of
Indonesia
Jakarta, Indonesia
ripto.wibowo@atmajaya.ac.id

Biagio Lenzitti
Università degli Studi di Palermo
Dipartimento Matematica e
Informatica
Palermo, Italy
biagio.lenzitti@unipa.it

Domenico Tegolo
Università degli Studi di Palermo
Dipartimento Matematica e
Informatica
Palermo, Italy
domenico.tegolo@unipa.it

Abstract

Autonomous AI research agents can execute long-running scientific workflows, revise code, inspect results, and reuse prior observations. Their reliability depends on memory architectures that retrieve relevant evidence without overwhelming the agent or discarding useful context. This hands-on tutorial presents a low-resource framework for diagnosing, benchmarking, and improving memory systems in autonomous research agents. Participants learn taxonomies of agent memory, trace-based diagnostics for retrieval and utilisation failures, LLM-as-judge evaluation protocols, and standardised benchmarking across heterogeneous memory providers. Exercises use real autonomous-research traces and lightweight open-source tooling that runs on a standard laptop with modest API credits.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; • **Information systems** → **Information retrieval**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

memory systems; autonomous agents; retrieval-augmented generation; benchmarking; LLM evaluation; low-resource AI

ACM Reference Format:

Nur Arifin Akbar, Rahool Dembani, Gregorius Airlangga, Ripto Mukti Wibowo, Biagio Lenzitti, and Domenico Tegolo. 2026. Systematic Diagnosis and Benchmarking of Memory Systems in Autonomous AI Research Agents: A Low-Resource Framework. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3770855.3816477>



This work is licensed under a Creative Commons Attribution 4.0 International License. KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
<https://doi.org/10.1145/3770855.3816477>

1 Tutorial Summary

Autonomous agents increasingly combine reasoning, tool use, retrieval, and long-horizon adaptation [7–9]. Yet memory remains a practical bottleneck: retrieved context may be irrelevant, useful evidence may be lost in long contexts [5], and agents may ignore information that was correctly retrieved.

This tutorial introduces a low-resource framework for diagnosing and benchmarking agent memory systems. It covers retrieval-augmented generation foundations [4], agent memory architectures such as MemGPT [6], trace-based separation of retrieval and utilisation failures, and LLM-as-judge evaluation protocols [10]. The material builds on recent work on RAG-driven memory, weather-energy pattern learning, and neurocognitive-inspired hybrid memory architectures [1–3].

2 Author Biographies

2.1 Nur Arifin Akbar

Nur Arifin Akbar is a researcher at the University of Palermo specialising in AI systems, autonomous agents, retrieval-augmented generation, and memory architectures for long-horizon research workflows.

2.2 Rahool Dembani

Rahool Dembani is a researcher and AI engineer at Singularlogic, Greece. His work focuses on production deployment of LLM systems, hybrid memory architectures, enterprise AI, scalability, reliability, and cost-effective autonomous systems.

2.3 Gregorius Airlangga

Gregorius Airlangga is a researcher and faculty member at Atma Jaya Catholic University of Indonesia. His interests include retrieval-augmented generation, natural language processing, AI evaluation, reproducibility, and resource-efficient AI technologies.

2.4 Ripto Mukti Wibowo

Ripto Mukti Wibowo is a researcher at King Abdulaziz University. His work focuses on autonomous AI systems, machine learning, computational intelligence, memory optimisation, resource-constrained deployment, and systematic AI evaluation.

2.5 Biagio Lenzitti

Biagio Lenzitti is a researcher in the Dipartimento Matematica e Informatica at the Università degli Studi di Palermo. His work spans health informatics, machine learning, cyber risk assessment, distributed data sharing, tailored health information systems, data security, and medical informatics.

2.6 Domenico Tegolo

Domenico Tegolo is an Associate Professor in the Dipartimento Matematica e Informatica at the Università degli Studi di Palermo. His research spans computational vision, mathematical modelling, biomedical image analysis, three-dimensional feature detection, neural networks, artificial intelligence, and image-analysis architectures.

Acknowledgments

This project has received funding from the European Union's Horizon research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101073381.

References

- [1] Nur Arifin Akbar, Rahool Dembani, Biagio Lenzitti, and Domenico Tegolo. 2025. RAG-Driven Memory Architectures in Conversational LLMs—A Literature Review With Insights Into Emerging Agriculture Data Sharing. *IEEE Access* (2025). <https://doi.org/10.1109/ACCESS.2025.3589241>
- [2] Nur Arifin Akbar, Shahin Rezvanian, and Rahool Dembani. 2025. Memory-Augmented LLMs for Sustainable Urban Energy Management via Weather-Energy Pattern Learning. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25)*. 589–592. <https://doi.org/10.1145/3714394.3750593>
- [3] Nur Arifin Akbar, Rahool Dembani, Clare Sullivan, Biagio Lenzitti, and Domenico Tegolo. 2025. Neurocognitive-Inspired Memory Architectures for Agricultural Knowledge Systems: Performance Analysis of Hybrid Approaches. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Vienna, Austria, 668–671. <https://doi.org/10.1109/SMC58881.2025.11342728>
- [4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. 9459–9474.
- [5] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
- [6] Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560* (2023).
- [7] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Article 2, 1–22.
- [8] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*.
- [9] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.
- [10] 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*.