SPEAKER-ATTESTED GROUNDING FOR FALSE MEMORY RESISTANCE IN AGENT MEMORY SYSTEMS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Agent memory systems enable long-term personalization by extracting and storing user information from conversations. However, these systems are susceptible to false memory storage, where assistant-generated content that was never confirmed by the user enters persistent memory. We identify a key failure mode: when the full dialogue serves as evidence for memory verification, assistant-originated statements can "self-verify," passing filtering despite lacking user attestation. We propose Speaker-Attested Grounding (SAG), a minimal intervention that restricts the evidence corpus to user turns only during memory filtering while maintaining full-dialogue extraction. On the HaluMem benchmark, SAG improves False Memory Resistance by +11.94 percentage points (58.76% $\rightarrow$ 70.70%) with a moderate recall tradeoff (-7.20pp). Per-source analysis reveals that 100% of SAG's gains come from assistant-only interference memories, confirming the targeted mechanism. Ablation studies show that 47.5% of the improvement stems from the speaker restriction itself, with the remainder from reduced evidence length.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Agent memory systems have emerged as a critical component for enabling long-term personalization in conversational AI (Packer et al., 2023; Chhikara et al., 2025; Zhong et al., 2023). By extracting and storing user information across sessions, these systems allow agents to maintain context, recall preferences, and provide increasingly personalized assistance over time. Recent surveys highlight memory as one of the fundamental capabilities distinguishing sophisticated AI agents from simple chatbots (Hu et al., 2025).

However, agent memory systems are susceptible to a critical failure mode: *false memory storage*. False memories occur when information that was never stated or confirmed by the user enters persistent storage, potentially leading to incorrect assumptions, privacy violations, or degraded user experience. The HaluMem benchmark (Chen et al., 2026) reveals that current memory systems frequently store "interference memories"—facts mentioned by the assistant that the user does not confirm.

We identify a root cause of this problem: *assistant-originated self-verification*. Standard memory systems use the full dialogue as evidence for filtering candidate memories. When an assistant makes a statement (whether correct or not), that statement becomes part of the evidence corpus. Consequently, assistant-originated false memories can find supporting evidence in the assistant's own statements, passing verification despite lacking user attestation.

We hypothesize that restricting the evidence corpus to user turns only will specifically target assistant-originated false memories while preserving legitimate memories that users actually stated. To test this hypothesis, we propose **Speaker-Attested Grounding (SAG)**, a minimal intervention that changes only the evidence corpus construction—from full dialogue to user turns only—while

[1] https://gitlab.com/fars-a/user-attested-evidence-halumem

keeping all other pipeline components identical. This single-variable design enables clean attribution of any performance differences to the speaker restriction mechanism.

Our contributions are as follows:

- We identify *assistant-originated self-verification* as a root cause of false memory storage in agent memory systems, where assistant statements serve as evidence for their own verification.
- We propose Speaker-Attested Grounding (SAG), a minimal single-variable intervention that improves False Memory Resistance by +11.94 percentage points on the HaluMem benchmark.
- We demonstrate that SAG's improvements are precisely targeted: 100% of FMR gains come from assistant-only interference memories, with no change on user-repeated interference.
- We provide ablation analysis isolating the speaker restriction mechanism (47.5% of gains) from evidence length effects (52.5%), showing that SAG achieves a more favorable recall-FMR tradeoff than length-matched baselines.

2 Related Work

2.1 Agent Memory Systems

Long-term memory is essential for LLM-based agents to maintain coherent, personalized interactions across sessions (Wang et al., 2023; Hu et al., 2025). Early approaches such as MemGPT (Packer et al., 2023) drew inspiration from operating system memory hierarchies, implementing virtual context management to extend effective context beyond the LLM's native window. MemoryBank (Zhong et al., 2023) introduced memory updating mechanisms inspired by the Ebbinghaus forgetting curve, enabling selective retention based on recency and importance. More recent systems have focused on scalability and structured representations: Mem0 (Chhikara et al., 2025) provides production-ready memory with graph-based representations for capturing relational structures, while Zep (Rasmussen et al., 2025) employs temporal knowledge graphs to maintain historical relationships across conversations. MemOS (Li et al., 2025) proposes a comprehensive memory operating system abstraction, and A-MEM (Xu et al., 2025) introduces agentic memory that dynamically organizes memories through interconnected knowledge networks. These systems typically employ a two-stage pipeline: memory extraction from conversations followed by filtering or verification before storage. However, existing approaches use the full dialogue as evidence for filtering, which allows assistant-originated content to influence verification decisions—a design choice that our work specifically addresses.

2.2 Memory Hallucination and Grounding

Memory hallucination—where agents store information that was never stated by users—has emerged as a critical reliability concern. Chen et al. (2026) introduced HaluMem, the first operation-level benchmark for evaluating hallucinations in memory systems, revealing that existing systems tend to generate and accumulate hallucinations during extraction and updating stages. From a security perspective, A-MemGuard (Wei et al., 2025) addresses adversarial memory manipulation through consensus-based validation and dual-memory structures, though it focuses on malicious injection rather than organic false memory formation. The broader challenge of grounding LLM outputs in evidence has been extensively studied in retrieval-augmented generation (Gao et al., 2023) and fact-checking contexts. MiniCheck (Tang et al., 2024) demonstrates efficient fact verification by checking claims against grounding documents, achieving GPT-4-level performance with smaller models. Our work applies this grounding principle specifically to memory filtering: rather than verifying against the full dialogue, SAG restricts evidence to user turns only, preventing assistant-originated content from self-verifying false memories.

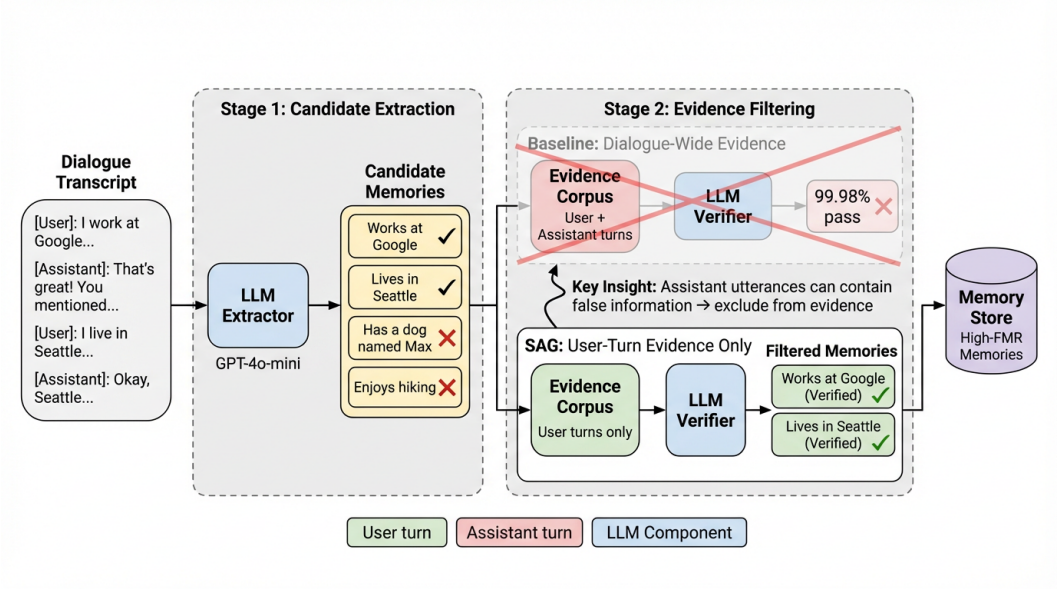

Figure 1: Speaker-Attested Grounding (SAG) pipeline. Stage 1 extracts candidate memories from the full dialogue. Stage 2 filters each candidate by checking if it can be grounded in user turns only, rejecting assistant-originated false memories that lack user attestation.

3 METHOD

3.1 PROBLEM FORMULATION

Agent memory systems typically operate through a two-stage pipeline. Given a dialogue transcript $D = \{(r_1, c_1), \ldots, (r_n, c_n)\}$ where each turn consists of a role $r_i \in \{\texttt{user}, \texttt{assistant}\}$ and content c_i, the system first extracts candidate memories $M = \{m_1, \ldots, m_k\}$ from D. Each candidate m_j is then verified against an evidence corpus E to determine whether it should be committed to persistent storage.

The critical design choice lies in constructing the evidence corpus E. Standard approaches use the full dialogue: $E_{\text{full}} = \{c_i : (r_i, c_i) \in D\}$. However, this allows assistant-generated content to serve as evidence for memory verification. When the assistant makes an incorrect statement, that statement becomes part of E_{full}, enabling false memories to pass verification—a phenomenon we term *assistant-originated self-verification*.

3.2 SPEAKER-ATTESTED GROUNDING

We propose Speaker-Attested Grounding (SAG), a minimal intervention that restricts the evidence corpus to user turns only:

$$E_{\text{SAG}} = \{c_i : (r_i, c_i) \in D, r_i = \texttt{user}\} \tag{1}$$

The key insight is that while extraction benefits from the full dialogue context (to capture all potentially relevant information), filtering should only use user-attested content as evidence. This separation ensures that assistant-originated statements cannot self-verify during the filtering stage.

Figure 1 illustrates the SAG pipeline. In Stage 1, an LLM extractor processes the complete dialogue transcript to generate candidate memories. In Stage 2, each candidate is verified against the user-only evidence corpus. Candidates that cannot be grounded in user statements are rejected, preventing assistant-originated false memories from entering persistent storage.

This design constitutes a single-variable intervention: the only difference between SAG and standard dialogue-wide filtering is the evidence corpus construction. The extraction prompt, filtering prompt,

Table 1: Main results on HaluMem-Medium (5-user subset). SAG improves False Memory Resistance (FMR) by +11.94pp over Dialogue-Wide baseline while trading off 7.20pp recall. Best in **bold**.

Method	Recall ↑	FMR ↑	F1 ↑	Avg Memories
Extract-Only	**53.27**	56.90	**66.79**	14.31
Dialogue-Wide	53.14	58.76	66.63	14.31
SAG (relaxed)	45.94 (−7.20pp)	70.70 (+11.94pp)	60.66	13.27
SAG (strict)	43.81 (−9.33pp)	**74.42** (+15.66pp)	58.69	12.95

and all other pipeline components remain identical, enabling clean attribution of any performance differences to the speaker restriction mechanism.

3.3 Grounding Threshold Variants

We implement two variants of SAG that offer different precision-recall tradeoffs. **SAG (strict)** requires explicit user statements to support a memory candidate—the user must have directly stated the information. **SAG (relaxed)** allows user-implied content, accepting memories that can be reasonably inferred from user statements even if not explicitly stated. The strict variant prioritizes false memory resistance at the cost of potentially missing some legitimate memories, while the relaxed variant maintains higher recall with slightly lower filtering stringency. This configurable threshold enables practitioners to tune the system based on application requirements: high-stakes applications may prefer strict mode to minimize false memories, while general-purpose assistants may benefit from the relaxed mode's better coverage.

4 Experiments

4.1 Experimental Setup

We evaluate SAG on the HaluMem-Medium benchmark (Chen et al., 2026), which provides operation-level evaluation of memory hallucinations in agent systems. The benchmark includes explicit interference memory points—facts mentioned by the assistant that the user does not confirm—enabling direct measurement of false memory resistance.

We use a 5-user subset comprising 343 dialogue sessions. For memory extraction and filtering, we employ GPT-4o-mini (Achiam et al., 2023) with temperature 0 for deterministic outputs. Evaluation uses GPT-4o as the judge model following the official HaluMem evaluation protocol.

We report four metrics: **Recall** measures coverage of gold user memories; **False Memory Resistance (FMR)** measures the fraction of interference memories successfully rejected (higher is better); **F1** is the harmonic mean of Recall and FMR; and **Avg Memories** reports the average number of memories stored per session.

We compare against two baselines: **Extract-Only** applies no filtering after extraction, and **Dialogue-Wide** filters using the full dialogue (user + assistant turns) as evidence.

4.2 Main Results

Table 1 presents the main experimental results. SAG (relaxed) achieves an FMR of 70.70%, improving over the Dialogue-Wide baseline (58.76%) by +11.94 percentage points. This substantial gain in false memory resistance comes with a moderate recall tradeoff: recall decreases from 53.14% to 45.94% (−7.20pp). The stricter variant, SAG (strict), achieves even higher FMR (74.42%, +15.66pp) at the cost of additional recall loss (−9.33pp).

Notably, the Dialogue-Wide baseline provides minimal improvement over Extract-Only (FMR: 58.76% vs 56.90%), confirming that when the full dialogue serves as evidence, assistant-originated false memories find supporting evidence and pass verification. In contrast, SAG's user-only evidence restriction effectively blocks these false memories from entering storage.

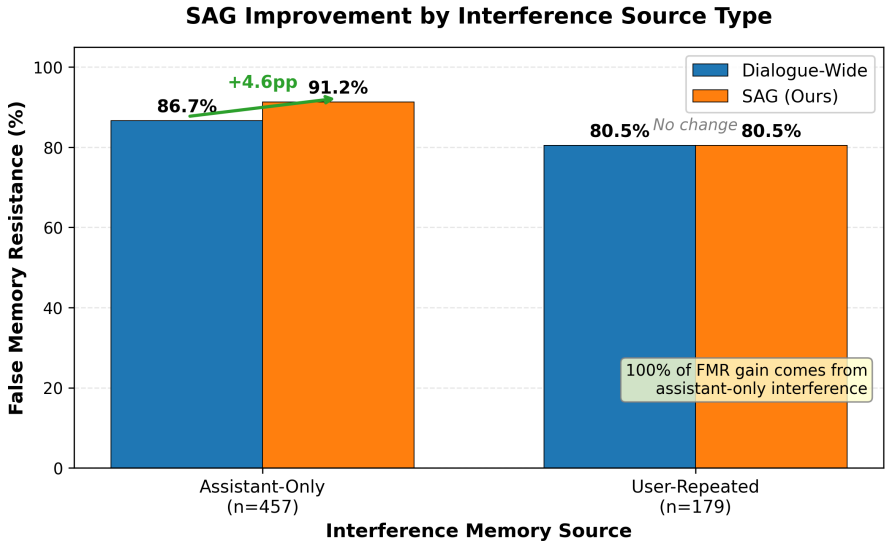

Figure 2: FMR breakdown by interference memory source type. SAG achieves +4.6pp FMR gain on assistant-only interference (91.2% vs 86.7%) while showing no change on user-repeated interference (both 80.5%). 100% of SAG's overall FMR improvement comes from the assistant-only subgroup.

The two SAG variants offer a configurable precision-recall tradeoff. SAG (relaxed) accepts memories reasonably inferable from user statements, achieving better recall while maintaining strong FMR. SAG (strict) requires explicit user attestation, maximizing false memory resistance for high-stakes applications where memory accuracy is paramount.

4.3 Per-Source Analysis

To understand where SAG's improvements originate, we analyze FMR by interference memory source type. Of the 645 interference memories in our evaluation set, 70.9% appear only in assistant turns (assistant-only), 27.8% are repeated by users after the assistant mentions them (user-repeated), and 1.4% appear in neither.

Figure 2 reveals a striking pattern: SAG's entire FMR improvement comes from the assistant-only subgroup. On assistant-only interference, SAG achieves 91.25% FMR compared to Dialogue-Wide's 86.65% (+4.6pp). On user-repeated interference, both methods achieve identical FMR (80.45%). This confirms our hypothesis: by restricting evidence to user turns, SAG specifically targets assistant-originated false memories while leaving user-repeated interference unchanged—exactly the behavior expected from the speaker restriction mechanism.

4.4 Ablation Study

SAG's user-only evidence corpus is shorter than the full dialogue, raising the question of whether FMR improvements stem from the speaker restriction or simply from reduced evidence length (which may make verification easier). To isolate these effects, we introduce a **Token-Matched** baseline that truncates dialogue-wide evidence to match SAG's average evidence length while retaining both user and assistant content.

Table 2 shows that Token-Matched achieves FMR of 66.98%, accounting for 52.5% of SAG's total FMR gain over Dialogue-Wide. The remaining 47.5% (+7.44pp) comes from the speaker restriction mechanism itself. Notably, Token-Matched suffers severe recall degradation (33.99% vs SAG's 43.81%), demonstrating that SAG's approach of restricting evidence by speaker rather than by length achieves a more favorable recall-FMR tradeoff.

Table 2: Ablation study isolating speaker restriction from evidence length effects. Token-Matched uses dialogue-wide evidence truncated to match SAG's evidence length.

Method	Recall ↑	FMR ↑	ΔFMR vs DW
Dialogue-Wide	**53.14**	58.76	—
Token-Matched	33.99	66.98	+8.22pp (52.5%)
SAG (strict)	43.81	**74.42**	**+15.66pp** (100%)

4.5 Error Analysis

We inspected SAG's remaining errors to understand its limitations. Among false positives (interference memories incorrectly retained), 100% exhibit *semantic overlap*: the user discussed the same topic as the interference memory, providing genuine user-turn evidence that the filter cannot distinguish from the false content. For example, when an assistant mentions "Martin's sabbatical started July 1" but the user actually stated "June 1," both discuss sabbaticals, creating semantic overlap that passes verification. This represents an irreducible error floor for speaker-based filtering alone.

Among false negatives (gold memories incorrectly rejected), 87% are cases where the LLM filter missed available evidence, often due to extraction inaccuracies (e.g., wrong details in the candidate memory) that SAG correctly rejected. Only 13% represent genuine recall losses where users implied but did not explicitly state information. These findings suggest that SAG's recall tradeoff is partially offset by improved extraction quality control, and that combining SAG with semantic deduplication could address the remaining false positives.

5 Conclusion

We presented Speaker-Attested Grounding (SAG), a minimal intervention that improves False Memory Resistance in agent memory systems by +11.94 percentage points through restricting the evidence corpus to user turns only. Our analysis confirms that SAG's improvements are precisely targeted: 100% of gains come from assistant-only interference memories, validating the hypothesis that speaker restriction addresses assistant-originated self-verification. The approach involves a recall tradeoff (-7.20pp) and cannot address semantic overlap failures where users discuss the same topics as interference memories. Future work could combine SAG with semantic deduplication to address remaining false positives while preserving the targeted benefits of speaker-based filtering.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, S. Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, L. Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, C. Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, S. Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, W. Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, B. Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, I. Kanitscheider, N. Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, J. Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis,

Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, A. Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, S. McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, An drey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, O. Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, J. Pachocki, A. Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, J. Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, N. Ryder, M. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, M. Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, N. Staudacher, F. Such, Natalie Summers, I. Sutskever, Jie Tang, N. Tezak, Madeleine Thompson, P. Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. 2023.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents, 2026. URL `https://arxiv.org/abs/2511.03506`.

P. Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. pp. 2993–3000, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, ZhongXiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhen Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yu Niu, Yu Wang, Zhe Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Jingwen Wen, Xuanjing Huang, Yu gang Jiang, and Shuicheng Yan. Memory in the age of ai agents. *ArXiv*, abs/2512.13564, 2025.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Ke-Rong Li, Zhenzhen Tao, Junpeng Ren, Huayi Lai, Hao Wu, Bo Tang, Zhenren Wang, Zhaoxin Fan, Ningyu Zhang, Linfeng Zhang, Junchi Yan, Ming-Zhou Yang, Tong Xu, Wei Xu, Huajun Chen, Haofeng Wang, Hongkang Yang, Wentao Zhang, Zhikun Xu, Siheng Chen, and Feiyu Xiong. Memos: A memory os for ai system. *ArXiv*, abs/2507.03724, 2025.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph Gonzalez. Memgpt: Towards llms as operating systems. *ArXiv*, abs/2310.08560, 2023.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *ArXiv*, abs/2501.13956, 2025.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. pp. 8818–8847, 2024.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 2023.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and Xiaofeng Wang. A-memguard: A proactive defense framework for llm-based agent memory. *ArXiv*, abs/2510.02373, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *ArXiv*, abs/2502.12110, 2025.

Wanjun Zhong, Lianghong Guo, Qi-Fei Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *ArXiv*, abs/2305.10250, 2023.