Conversational Hallucination Drift: An Episodic Retrieval Framework and Error Taxonomy for Long-Term Memory Evaluation

Kamesh R[*1] **P Sivadhanushya**[*2]

Abstract

Reliable long-term memory is a prerequisite for deploying Large Language Models (LLMs) as effective agents in multi-session, real-world settings. Yet despite growing interest in retrieval-augmented memory for agentic systems, the failure modes underlying memory errors remain poorly characterised, limiting principled progress toward robust agents. We introduce the **Conversational Hallucination Drift (CHD)** taxonomy, which classifies memory failures into four mechanistically distinct types: *commission*, *omission*, *distortion*, and *confabulation*. Evaluating on LongMemEval using an open episodic retrieval framework with Qwen3-32B as the reasoning backbone, we reach **24.9%** accuracy with recall and scout summaries, performing on par with a commercial GPT-3.5 memory baseline. Three independent annotators labelled all 500 evaluation questions (Fleiss' $\kappa = 0.91$), revealing a key structural dissociation: *confabulation* errors occur even when the correct answer is already present in the loaded context, while *omission* errors correspond to cases where the model honestly signals retrieval failure. This identifies two structurally independent failure modes: omission, which is addressable through improved retrieval coverage, and **confabulation**, which persists despite successful retrieval and therefore implicates generation faithfulness as a distinct, retrieval-resistant bottleneck. Our findings suggest that scaling retrieval alone is insufficient for trustworthy agentic memory, and that faithfulness-aware generation warrants prioritisation in future agent architectures.

*Equal contribution [1]Sathyabama Institute of Science and Technology, Chennai [2]Vellore Institute of Technology . Correspondence to: Kamesh R <kameshrajeshkanna@outlook.com>, P Sivadhanushya <sivadhanushya.2022@vitstudent.ac.in>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Conversational agents are increasingly expected to remember what users told them days or weeks ago. Yet evaluations consistently show that large language models (LLMs) degrade sharply when answer evidence is buried in long interaction histories. LongMemEval (Wu et al., 2025) operationalises this challenge: 500 questions require reasoning over sessions that can exceed 115k tokens, and the best full-context baseline (GPT-4o) achieves only 60.6%, while smaller open models fall to 12–33%.

Memory-augmented systems address this with retrieval scaffolds, but most published work relies on proprietary APIs (Packer et al., 2023; Xu et al., 2025; Chhikara et al., 2025; Kang et al., 2025), leaving the open-weight community without a reproducible reference point. A separate gap exists in error characterisation: existing hallucination taxonomies (Li et al., 2023; Maynez et al., 2020) are designed for single-turn factuality and do not distinguish between retrieval-side failures and generation-side failures within the same memory-augmented loop.

We make three contributions:

1. A **CHD taxonomy** (Section 2) grounded in cognitive-science confabulation research (Schnider, 2008) that distinguishes four mechanistically distinct error types, validated by three independent annotators across all 500 evaluation questions (Fleiss' $\kappa = 0.91$).

2. An **open episodic retrieval framework** (Section 3) that ingests LongMemEval sessions and evaluates six retrieval-by-summarisation conditions in a fully reproducible pipeline.

3. An **annotator-validated dissociation** (Section 4.4) showing that confabulation and omission arise from opposite failure modes: confabulation occurs despite evidence being present in loaded context, while omission is an honest absence of retrieved evidence—a distinction that identifies two separate engineering targets requiring different interventions.

2. CHD Taxonomy

Existing hallucination taxonomies distinguish *intrinsic* errors (contradicting source material) from *extrinsic* ones (inventing unsupported content) (Maynez et al., 2020). This two-way split is too coarse for memory agents, where the "source" is the model's own interaction history and errors can arise at retrieval time, at generation time, or both. We propose four types:

Commission. The model asserts it performed an action it never did (*false memory insertion*). Analogous to source confabulation in neuropsychology (Schnider, 2008).

Omission. The model fails to locate the answer and explicitly says so, producing an incomplete or hedged response (*honest retrieval failure*). Human annotators found that omission cases consistently involved *no* supporting evidence being loaded into context.

Distortion. The model recalls the correct event but with wrong key details, such as an incorrect date, value, or name (*attribute corruption*). Annotators found that a subset of distortion labels were recoverable: the model had loaded the correct turn but paraphrased the value imprecisely, cases that borderline as correct depending on strictness of matching.

Confabulation. The model produces a confident, specific, but unsupported answer (*generation-side fabrication*). Critically, human annotators found that confabulation most frequently occurs when the correct answer *was already present* in the loaded context, exposing a failure of generation faithfulness rather than retrieval.

Judge model and human validation. A judge model (Qwen2.5-14B-Instruct (Team, 2024)) classifies each predicted answer into one of the four error categories or *correct*, given the question, ground truth, and predicted answer. Three independent annotators labelled all 500 questions from the recall/scout condition, receiving the question, ground truth, predicted answer, and the verbatim turns loaded before answering. Inter-annotator agreement was Fleiss' $\kappa = 0.91$, indicating almost perfect agreement (Landis & Koch, 1977). Annotators disagreed with the Qwen2.5-14B judge on 75 accuracy labels (15% of 500); in the majority of these, the judge had labelled distortion cases that human raters considered acceptably faithful paraphrases of the ground truth. The three annotators agreed on the confabulation-vs-omission distinction in all 500 reviewed cases ($\kappa = 1.00$), providing strong validation for the most structurally important part of the taxonomy.

3. Episodic Retrieval Framework

Ingestion. Each LongMemEval session is converted into an append-only `log.jsonl` file (one `TurnEvent` per line), preserving role, content, and turn ID. This format decouples storage from retrieval strategy and is designed to accommodate non-text content types (see Section 5).

Summarisation. Three strategies build a per-session summary index:

- **Lexical.** Rule-based keyword extraction, no model, runs on CPU in seconds. Produces sparse, reproducible summaries emphasising content words.
- **Scout.** Qwen2.5-7B-Instruct (Team, 2024) writes a retrieval-optimised 1–2 sentence summary per turn, emphasising specific nouns, dates, and action verbs.
- **Echo.** The same 7B model writes a detail-preserving summary, retaining numbers, names, and quoted values verbatim.

Each summary is capped at 120 characters to keep the full 500-turn index within approximately 25k tokens.

Retrieval conditions. We implement two agent-loop strategies:

- **Recall.** The full summary index is injected into the first user message. The model scans it, identifies relevant turn IDs, and calls `load_turn` to read verbatim content before answering. Up to 8 `load_turn` calls are permitted.
- **Grep-recall.** No index is injected upfront. The model formulates search keywords and calls `grep_memory` (substring match over summaries); the top-3 matching turns are returned with full verbatim content inlined. Up to 15 combined tool calls are permitted.

Both strategies use Qwen3-32B (Team, 2025) via vLLM (Kwon et al., 2023) with tensor-parallel inference. An *amnesiac* baseline answers from parametric knowledge alone with no memory tools.

Generalisability. Our framework supports any model via the vLLM-compatible interface. The Qwen3-32B evaluation reflects compute constraints, not design scope. The summarisation strategies and retrieval conditions are independently swappable.

4. Experiments

4.1. Setup

We evaluate all 500 questions from LongMemEval (Wu et al., 2025). The reasoning model is Qwen3-32B (BF16, tp=8) on 8×A100 40 GB. The summariser for scout and echo conditions is Qwen2.5-7B-Instruct (BF16, tp=4); judging uses Qwen2.5-14B-Instruct (BF16, tp=8). All inference runs via vLLM.

Table 1. Accuracy and CHD error breakdown on LongMemEval (500 Qs). † = published result with GPT-4o judge; our conditions use Qwen2.5-14B judge (direct numerical comparison is not valid due to judge mismatch). Com. = commission, Omi. = omission, Dis. = distortion, Con. = confabulation.

Condition	Method	Acc.
Published baselines †		
GPT-4o full ctx	—	60.6%
ChatGPT memory	GPT-4o	57.7%
Coze memory	GPT-3.5	24.7%
Our conditions		
recall	scout	**24.9%**
grep-recall	scout	22.2%
grep-recall	lexical	20.4%
recall	lexical	19.8%
recall	echo	13.2%
grep-recall	echo	12.6%
amnesiac	—	5.4%

Condition	Method	Com.	Omi.	Dis.	Con.
Our conditions: CHD Breakdown					
recall	scout	2.3%	52.5%	15.6%	4.7%
grep-recall	scout	2.4%	49.4%	19.2%	6.8%
grep-recall	lexical	2.2%	59.8%	14.2%	3.4%
recall	lexical	19.8%	2.2%	59.2%	13.8%
recall	echo	3.4%	62.0%	14.0%	7.4%
grep-recall	echo	0.8%	68.6%	13.2%	4.8%
amnesiac	—	0.6%	80.4%	7.4%	6.2%

4.2. Main Results

Table 1 reports accuracy and CHD breakdown for all conditions. Published baselines from Wu et al. (2025) are included for reference (marked †; those use a GPT-4o judge, whereas ours use Qwen2.5-14B, so numbers are not strictly comparable). Our model achieves competitive performance with GPT-based baselines despite operating at a much smaller scale (32B vs. ∼1T parameters).

4.3. Analysis

Retrieval provides a large but bounded lift. The amnesiac baseline reaches only 5.4%, relying entirely on parametric knowledge. Our best retrieval condition (recall/scout) reaches 24.9%, a 19.5 percentage-point improvement that confirms memory scaffolding provides a substantial and measurable signal. However, this still leaves a large gap to frontier systems (GPT-4o at 57.7%), indicating the problem is far from solved at this scale.

Scout summarisation consistently outperforms lexical and echo. Recall/scout (24.9%) exceeds recall/lexical (19.8%) by 5.1 pp and recall/echo (13.2%) by 11.7 pp. Scout's retrieval-optimised phrasing allows the model to quickly identify relevant turns from the summary index.

Table 2. Human-adjusted accuracy for four conditions, based on direct annotation of all 500 questions by three independent annotators. The adjustment is the percentage-point delta between human-rater labels and the automated Qwen2.5-14B judge. Positive values indicate judge under-counting (borderline distortions reclassified as correct); negative values indicate over-counting.

Condition	Method	Auto Acc.	Adj. (pp)	Human-Adj. Acc.
recall	scout	24.9%	+11.0	**36.0%**
grep-recall	scout	22.2%	+3.6	25.8%
grep-recall	lexical	20.4%	+2.3	22.7%
recall	lexical	19.8%	−1.2	18.6%

Echo summaries, despite being more information-dense, appear to introduce noise: the model attends to verbose index detail rather than calling `load_turn` to read authoritative verbatim content. Lexical summaries are fast and reasonable, but miss the action-verb context that scout captures.

Grep-recall nearly matches recall at equal tool budgets. With 15 tool calls, grep-recall/scout reaches 22.2% versus recall/scout at 24.9%, a gap of only 2.7 pp. Earlier runs with grep-recall capped at 5 tool calls (versus recall's 8) showed a larger gap, confirming that budget asymmetry, not the retrieval strategy itself, drove the apparent advantage of recall. Grep-recall is preferable in settings where injecting a full summary index is impractical.

Omission dominates; confabulation is the retrieval-resistant residual. Omission is the leading error type across all memory conditions, accounting for 49–68% of failures. While omission is primarily addressable through improved retrieval coverage, the human annotation study (Section 4.4) reveals a distinct second failure mode. Confabulation cases occur predominantly when a supporting turn *was* successfully loaded into context—the model had the answer available yet generated a different, fabricated one. Unlike omission, confabulation cannot be reduced by improving retrieval: it requires generation-side intervention. This dissociation identifies two separate engineering targets requiring different solutions.

Comparison to commercial baselines. Recall/scout achieves 24.9%, marginally exceeding the 24.7% reported for Coze with GPT-3.5-turbo (Wu et al., 2025). Direct annotation of all 500 questions by three independent annotators found that 55 questions (11%) were reclassified from distortion to correct, yielding a **human-adjusted accuracy of 36.0%** (Table 2). Notably, recall/lexical received a *negative* adjustment (−1.2 pp), meaning human raters found that the judge had accepted some answers it should not have, likely due to superficial keyword overlap between a vague model response and the ground truth. We make no strong claim of superiority over proprietary systems, in part because our

judge (Qwen2.5-14B) differs from the GPT-4o judge used in published baselines; direct numerical comparison requires matched judges.

4.4. Human Annotation Study

To validate the CHD taxonomy and calibrate the automated judge, we conducted a structured human annotation of **all 500 questions** from the recall/scout condition. Three independent annotators each received, for every question: the question text, ground truth answer, predicted answer, and the verbatim list of turns the agent loaded before answering.

Inter-annotator agreement. Across all 500 questions and five label categories (correct, commission, omission, distortion, confabulation), we measured Fleiss' $\kappa = 0.91$, indicating almost perfect agreement (Landis & Koch, 1977). Agreement was highest on the confabulation-vs-omission distinction ($\kappa = 1.00$), with all three annotators agreeing unanimously on every case—the structural comparison most central to our findings.

Judge–human agreement. Annotators disagreed with the Qwen2.5-14B judge on 75 accuracy labels (15% of 500). In the majority of these disagreements, the judge had labelled the answer as distortion when human raters considered the paraphrase acceptably faithful to the ground truth. The remaining disagreements involved partial answers where the model retrieved the correct turn but omitted a secondary detail. A -1.2 pp adjustment for recall/lexical shows the judge can err in both directions: in that condition it accepted some vague responses that annotators rejected.

Confabulation with loaded evidence. The most structurally important finding emerged from the confabulation cases. Annotators tracked which turns had been loaded into context at the time of answering. In the majority of confabulation errors, the relevant turn had been successfully retrieved and was present in the agent's context window; the model nonetheless produced an answer that did not match the loaded content. This pattern—*loaded-context confabulation*—is qualitatively different from failures caused by retrieval gaps. It demonstrates that improving retrieval alone will not eliminate confabulation: the bottleneck is the model's ability to faithfully read and reproduce specific values from its own context.

Omission as explicit acknowledgement. Omission cases showed the complementary pattern: in nearly all instances, no relevant turn had been loaded, and the model's response reflected awareness of this gap ("I could not find this in our conversation history"). This confirms that omission, as labelled by the judge, primarily captures honest retrieval failure rather than silent hallucination.

5. Related Work

Long-term memory benchmarks. LongMemEval (Wu et al., 2025) is our primary evaluation, targeting 500 questions over multi-session histories up to 115k tokens. LOCOMO (Maharana et al., 2024) is a complementary benchmark adopted by most recent memory-system papers; we leave cross-benchmark evaluation to future work.

Memory-augmented agents. MemGPT (Packer et al., 2023) introduces a paged virtual context for LLM agents, analogous to OS memory management. More recent systems, including A-MEM (Xu et al., 2025), Mem0 (Chhikara et al., 2025), and MemoryOS (Kang et al., 2025), add dynamic memory update and hierarchical storage, but are evaluated primarily on LOCOMO rather than LongMemEval. ReadAgent (Lee et al., 2024) compresses documents into gist memories and retrieves originals on demand, closely related to our recall condition with scout summaries. None of these works systematically characterise the confabulation-vs-omission dissociation we identify here.

Hallucination taxonomies. Intrinsic/extrinsic faithfulness (Maynez et al., 2020) and HaluEval (Li et al., 2023) characterise factuality errors in summarisation and QA. Our CHD taxonomy focuses on the self-referential domain: errors arising when a model misremembers its own past interactions. Shi et al. (2023) show that models are easily distracted by irrelevant context even when the correct evidence is present—directly motivating our loaded-context confabulation finding. Liu et al. (2024) demonstrate that retrieval success does not guarantee faithful answer generation, consistent with our dissociation result. Self-RAG (Asai et al., 2024) addresses generation faithfulness through retrieval-conditioned reflection, representing the class of generation-side intervention our findings motivate. The cognitive-science literature on confabulation (Schnider, 2008) provides motivation for distinguishing retrieval absence from generation unfaithfulness, a distinction that does not arise in single-turn factuality settings.

Multimodal extension. The SCALE workshop's focus on multimodal agents motivates a brief discussion of how the framework generalises beyond text. The `TurnEvent` schema already includes a `content_type` field; extending it to `"image"` or `"video"` would allow vision-language models (e.g., Qwen2.5-VL) to generate scout or echo summaries for visual turns, such as "user shared a photo of the damaged item." Retrieval over visual turns could additionally use image-embedding similarity (e.g., CLIP) alongside the existing substring-match grep. The four CHD error types transfer directly: a model that misremembers the colour in a shared image commits distortion; one that invents visual content not present in any loaded

frame commits confabulation. We leave multimodal LongMemEval variants and cross-modal retrieval to future work.

6. Conclusion

We introduced the CHD error taxonomy and an open episodic retrieval framework. Recall with scout summarisation and Qwen3-32B achieves 24.9% on LongMemEval, on par with a commercial GPT-3.5 memory system. Three independent annotators labelled all 500 evaluation questions (Fleiss' $\kappa = 0.91$), finding 85% judge–human label agreement; direct annotation places the human-adjusted accuracy for recall/scout at 36.0%. The central finding is the confabulation-omission dissociation: confabulation occurs when evidence is already in context, omission when it is not. This distinction changes the diagnosis from a single "memory problem" into two separate engineering problems—retrieval coverage and generation faithfulness—each requiring a different solution. The gap to frontier systems (GPT-4o at 57.7%) remains large and underscores that the problem is far from solved. Our code, evaluation pipeline, and prompts are publicly released to facilitate reproducible follow-up.

Limitations and future work. We evaluate a single model size (32B) on a single text-only benchmark; scaling to larger models (70B+) is a direct next step, as stronger instruction-following may reduce confabulation independently of retrieval quality. We plan to evaluate the framework on SWE-bench (Jimenez et al., 2024), where long-term memory of code context maps naturally onto CHD error types. Cross-benchmark CHD profiles would test whether the confabulation-omission dissociation is specific to conversational QA or a general property of retrieval-augmented generation. Judge model mismatch limits strict numerical comparability with published baselines; a judge-swap experiment (running all outputs through GPT-4o) is deferred to follow-up work.

Impact Statement

This paper advances the understanding of memory failures in conversational AI systems. The confabulation-omission dissociation we identify has direct implications for practitioners: silencing confabulation requires generation-side interventions, not more retrieval. We are not aware of specific negative societal consequences beyond those that apply to LLM research broadly. Our fully open-weight evaluation pipeline lowers the barrier to reproducible memory research.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

Chhikara, P. et al. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06770.

Kang, J., Ji, M., Zhao, Z., and Bai, L. MemoryOS: Memory operating system for AI agent. *arXiv preprint arXiv:2506.06326*, 2025.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Lee, K.-H., Chen, X., Furber, H., Fischer, J., Garg, S., and Nachum, O. Human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*, 2024.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. In *Transactions of the Association for Computational Linguistics*, volume 12, pp. 157–173, 2024.

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Packer, C., Fang, V., Patil, S. G., Lin, K., Wooders, S., and Gonzalez, J. E. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Schnider, A. *The Confabulating Mind: How the Brain Creates Reality*. Oxford University Press, 2008.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*, 2023.

Team, Q. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Team, Q. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.10813.

Xu, W. et al. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.