

PRAGMA: Evaluating Personalized Guidance with Memory Alignment in Lifelong Conversations

Hyojeong Yu¹, Hyukhun Koh², Minsung Kim¹, Yunah Jang¹, Kyomin Jung^{1,2,†}

¹Dept. of ECE, Seoul National University, ²IPAI, Seoul National University
{hyoj.yu, hyukhunkoh-ai, kms0805, vn2209, kjung}@snu.ac.kr

Abstract

Large language models (LLMs) are increasingly deployed as personalized assistants that interact with users over extended periods of time. As conversations grow longer, relying on full interaction histories becomes increasingly inefficient and unreliable: long contexts introduce substantial computational overhead, making it difficult for models to consistently identify and utilize the most relevant information for the current request. These challenges have motivated memory systems that structure and retrieve user-specific information. In realistic interactions, users often seek practical guidance such as recommendations, planning, and decision support. Unlike factual recall tasks, personalized guidance requires models to integrate information across multiple past conversations and reason about changing user preferences and experiences. However, existing conversational memory evaluations mainly focus on retrieval and factual recall. To study this challenge, we introduce PRAGMA, a benchmark for evaluating personalized guidance in long-term conversations. PRAGMA contains curated longitudinal conversation histories, evidence annotations, and guidance scenarios grounded in evolving user contexts and incorrect user assumptions. Experiments across retrieval systems, memory systems, and long-context models reveal that current systems struggle both to recover the appropriate conversational evidence and to effectively use it for personalized guidance. Our results highlight the need for memory architectures that support robust conversational retrieval and memory-grounded reasoning beyond evidence recall.

1 Introduction

Large language models (LLMs) are increasingly deployed as personalized conversational assistants

[†]Corresponding author.

*Code and data are available at <https://github.com/yuhyojeong/PRAGMA> and <https://huggingface.co/datasets/stellahj/PRAGMA>.

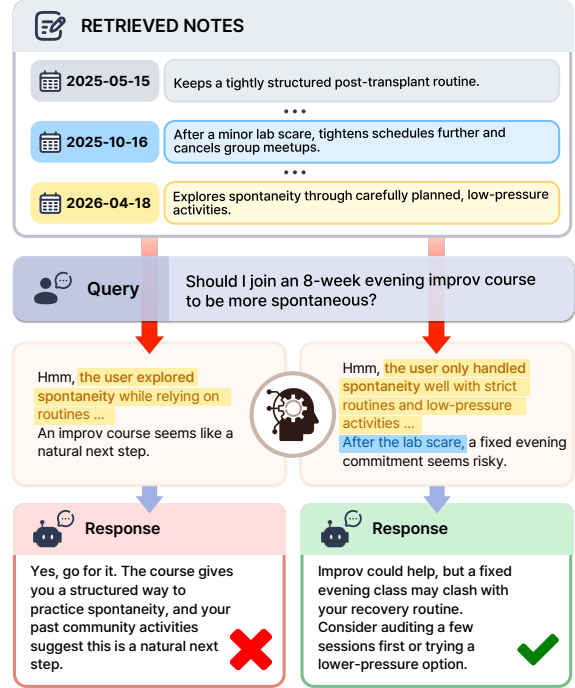


Figure 1: Personalized guidance requires both effective conversational retrieval and downstream memory-grounded reasoning. Relevant evidence may be temporally distributed and only implicitly connected to the final user request. Although both responses use retrieved memories, only the right response correctly synthesizes the user’s longitudinal context.

that interact with users over extended periods of time (Li et al., 2025a; Tan et al., 2025). Applications such as recommendation agents, tutors, and productivity assistants rely on awareness of users’ preferences, experiences, and evolving needs across interactions. As conversational histories grow longer, conditioning on full history becomes increasingly inefficient and unreliable, motivating memory systems that selectively store and retrieve user-specific information (Zhong et al., 2024).

Existing work on conversational memory has primarily focused on retrieval and factual recall from long histories (Wu et al., 2025; Maharana et al., 2024). However, real-world personalization often

Benchmark	In-situ	Open	Guide	E-A	E-C	T-A	T-C	Tokens
LongMemEval	○	△	△	○	×	×	×	115K, 1.5M
LoCoMo	×	×	×	—	—	—	—	9K
HiCUPID	○	○	△	○	×	×	×	17K
ImplexConv	○	○	○	○	×	×	×	60K
ConvoMem	○	△	△	○	×	×	×	1K–3M
PersonaMem	○	×	○	○	×	○	×	32K–1M
PRAGMA	○	○	○	○	○	○	○	160K

Table 1: Comparison of long-term conversational memory benchmarks. In-situ: queries embedded in conversations; Open: open-ended generation; Guide: personalized guidance; E-A/E-C and T-A/T-C: event- and trajectory-grounded aligned/corrective reasoning. Tokens: approximate context length. ○: supported, △: partial, ×: unsupported.

requires practical guidance (Chatterji et al., 2025) such as recommendations, planning support, and decision-making grounded in evolving user experiences. These queries often require reasoning over long-term user trajectories and potentially incorrect user assumptions (Feng et al., 2026; Sharma et al., 2024). Moreover, relevant evidence is often temporally distributed and only implicitly connected to the final user request, making conversational retrieval itself challenging. As a result, personalized guidance depends not only on retrieving relevant memories, but also on utilizing them coherently during response generation (Kwon et al., 2026). Yet the relationship between retrieval and downstream personalized reasoning remains underexplored (Laban et al., 2026; Li et al., 2026).

Constructing realistic evaluation settings for this problem is also challenging. Without careful design, synthetic long-term conversations can produce shallow trajectories or queries solvable through simple recency heuristics. Effective evaluation therefore requires balancing realism, controllability, and resistance to retrieval shortcuts.

To address these challenges, we introduce **PRAGMA** (**P**RActical **G**uidance with **M**emory **A**lignment), a benchmark for evaluating personalized guidance in long-term conversations. PRAGMA is built through a controlled, human-validated pipeline that generates long-term conversational histories with evolving user states and diverse memory requirements. The benchmark includes guidance scenarios grounded in both event-level memories and user states that evolve over time, including settings where users make assumptions that conflict with their conversational history.

We evaluate retrieval systems, structured memory systems, and long-context models on PRAGMA. Across architectures and generation models, we find that current systems struggle both to recover the appropriate conversational evidence and to ef-

fectively use it for personalized guidance. Even when relevant evidence is retrieved, models often fail to generate coherent and well-grounded responses, particularly on trajectory-grounded and corrective reasoning tasks. Our findings also highlight the need for memory architectures that support both robust conversational retrieval and downstream memory-grounded reasoning for long-term personalized assistance.

In summary, our contributions are as follows:

- We introduce PRAGMA, a benchmark for evaluating personalized guidance grounded in long-term conversational memory, focusing on guidance scenarios that require reasoning over evolving user trajectories and potentially incorrect user assumptions.
- We propose a controlled, human-validated benchmark construction pipeline that generates realistic longitudinal conversations with evolving memory dependencies and fine-grained evidence annotations.
- We evaluate retrieval systems, memory systems, and long-context models on PRAGMA, showing that current systems struggle to recover relevant memory and to effectively utilize it for personalized guidance.

2 Related Work

Factual memory recall benchmarks. Recent benchmarks such as LongMemEval (Wu et al., 2025), LoCoMo (Maharana et al., 2024), and ConvoMem (Pakhomov et al., 2025) evaluate whether language models can retain and access information from long conversational histories through tasks including factual QA, dialogue understanding, and memory-grounded response generation. While these benchmarks have advanced evaluation of long-context memory and conversational

consistency, they primarily focus on recovering or reproducing past information rather than utilizing memory for open-ended practical guidance.

Personalized conversational benchmarks. Another line of work studies personalized generation conditioned on user history. ImplexConv (Li et al., 2025b) evaluates implicit reasoning over semantically distant conversational evidence, but focuses on narrow reasoning settings. PersonaMem (Jiang et al., 2025) introduces preference evolution and recommendation scenarios, yet relies on multiple-choice evaluation rather than open-ended generation. HiCUPID (Mok et al., 2025) evaluates profile-conditioned generation where user preferences and profile attributes are explicitly embedded in the conversation history, reducing the need for implicit memory reasoning over long-term interactions.

Table 1 provides a comparison of PRAGMA with existing personalized-memory and long-context benchmarks. Overall, existing benchmarks provide limited evaluation of open-ended personalized guidance grounded in evolving user trajectories.

3 PRAGMA

We introduce PRAGMA, a benchmark for evaluating personalized guidance in long-term conversations. In this section, we describe its construction pipeline, including query design, history generation, evidence annotation, and evaluation protocols.

3.1 Query Design

To evaluate real-world personalized guidance, we construct long-term conversational contexts in which user information naturally accumulates over time. In these settings, users seek open-ended practical guidance, including recommendations, planning support, and decision-making assistance, within ongoing conversations. However, practical guidance in long-context conversations introduces several challenges. Relevant context may come from isolated events or emerge across extended interactions, and user requests may be incomplete, outdated, or inconsistent with prior context. These features are not fully capturable through current factual recall benchmarks with constrained response generation settings.

To thoroughly cover these challenges, we organize queries along two orthogonal dimensions: memory dynamics and query alignment. Memory dynamics distinguishes between event-level experiences (static) and longitudinal user trajec-

	Memory-Aligned	Memory-Misaligned
Static	Event-Align	Event-Correct
Evolving	Traj-Align	Traj-Correct

Table 2: PRAGMA query taxonomy.

tories (evolving). Query alignment characterizes whether user assumptions are consistent with prior conversational evidence. Combining these dimensions yields four query categories (Table 2): **Event-Align** queries require guidance grounded in coherent event-level experiences; **Event-Correct** queries require correcting incorrect event-level assumptions before providing appropriate guidance; **Traj-Align** queries require reasoning over evolving user trajectories; **Traj-Correct** queries require recognizing when a user’s proposed decision conflicts with their longitudinal trajectory and providing corrective guidance. Example queries for each query type are provided in Appendix A.5.

3.2 Benchmark Construction

PRAGMA is built through a controlled generation pipeline with validation to create personalized guidance queries while minimizing shortcut strategies such as recency heuristics and lexical matching. Additional construction details and examples are provided in Appendix A.

User schema design. To ensure diversity in user personas and longitudinal behaviors, we begin with 100 synthetic users from Privasis-Zero (Kim et al., 2026), which provides rich profile information suitable for generating coherent persona-grounded attributes and experiences. We extract only the profile information relevant for controllable generation, including demographic attributes (e.g., age, income class, native language, citizenship) and user event lists. Using gpt-5, we generate for each user: (1) a one-sentence persona summary, (2) a topic for event-grounded experiences, and (3) two latent behavioral axes for trajectory-grounded evolution.

The topic and axes are prompted to remain persona-compatible while mutually independent.

Event-grounded query construction. For event-grounded queries, we first generate 2–5 timestamped user events associated with the sampled topic over a one-year period (2025-05-01 to 2026-04-30). These events are then used to construct two query types corresponding to aligned and corrective guidance settings.

Event-aligned queries are designed to refer to the topic, providing a retrieval cue while still re-

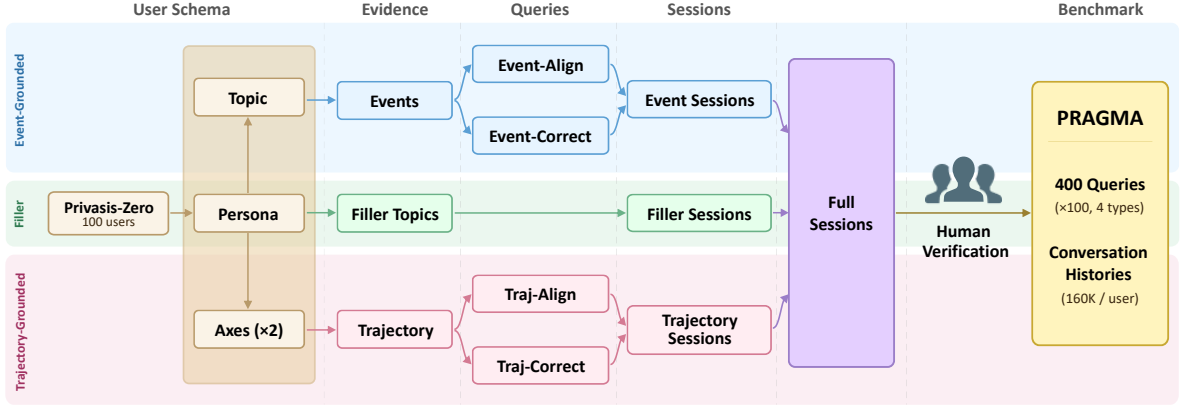


Figure 2: Overview of the PRAGMA benchmark construction pipeline.

quiring the model to interpret how previous experiences should influence the recommendation. Event-corrective queries intentionally contain mixed or partially incorrect recollections constructed from multiple prior events. Rather than directly requesting factual correction, the user asks for practical guidance based on a mistaken premise, requiring models to identify inconsistencies in the user’s assumptions before producing appropriate guidance.

Trajectory-grounded query construction. For trajectory-grounded queries, we generate longitudinal user trajectories consisting of 4–8 timestamped states over two independent behavioral axes. We intentionally construct trajectories over multiple simultaneously evolving behavioral axes. When only a single attribute changes, the task can often collapse into retrieving the user’s most recent state, whereas multi-axis trajectories require models to jointly track multiple aspects of the user over time.

Using these trajectories, we construct two types of trajectory-grounded guidance queries. For trajectory-aligned queries, users explicitly refer to the underlying axes as qualities they currently value or wish to emphasize. To avoid direct lexical shortcuts, we rewrite the axes into abstract descriptors (e.g., single-word concepts) before they appear in the query, requiring models to infer the underlying trajectory from the conversational history.

For trajectory-corrective queries, we generate decisions that appear individually plausible but conflict with the user’s longer-term trajectory.

Conversation history construction. Each evidence item is expanded into a natural conversational session using gpt-5-mini. To simulate realistic long-term interactions and increase retrieval difficulty, we additionally generate 30 filler topics per user that are unrelated to the target topic and trajectory axes while remaining consistent with the

user’s persona. These filler topics are similarly expanded into conversational sessions.

Evidence and filler sessions are then concatenated into a single long-term conversational history. Evidence sessions are first ordered according to their timestamps, after which filler sessions are uniformly inserted between them to avoid positional concentration. In particular, the first and last sessions are always filler sessions, preventing trivial boundary-position or recency heuristics.

The final benchmark contains 100 users and 400 queries (four query types per user). Each user history contains approximately 160K conversational tokens shared across the four associated queries.

Human validation. All generated queries and evidence annotations are manually reviewed before finalizing. Human validation is particularly important for trajectory-grounded and corrective queries, where subtle inconsistencies or unintended shortcuts can significantly reduce benchmark difficulty. Annotators were instructed to verify query realism, evidence correctness, and consistency with the intended query types and reasoning requirements. Given the complexity of this validation over long conversational histories and distributed evidence, we prioritize rigorous instance-level quality over increasing benchmark scale. Additional validation details and benchmark statistics are provided in Appendix A.3 and B.2.

3.3 Annotations and Evaluation Protocols

PRAGMA comprehensively evaluates both retrieval- and response-level performance.

Retrieval Evaluation. For the retrieval evaluation, each query is annotated with the necessary evidence sessions to generate a response. These annotations enable standard retrieval-based evaluation using metrics such as Recall, Precision, F1, and

Method		Event-Align (<i>n</i> = 100)		Event-Correct (<i>n</i> = 100)		Traj-Align (<i>n</i> = 100)		Traj-Correct (<i>n</i> = 100)	
		Aln.	Grd.	Aln.	Grd.	Aln.	Grd.	Aln.	Grd.
GPT-5-mini									
Full Context		67.00	36.00	1.00	10.62	70.00	8.00	19.00	13.27
Dense	S	90.00	57.00	3.00	15.56	73.00	13.00	32.00	17.68
	T	98.00	52.00	2.00	25.18	53.00	6.00	42.00	16.46
Window	S	96.00	59.00	3.00	18.50	75.00	11.00	32.00	17.06
	T	93.00	44.00	5.00	16.25	60.00	6.00	30.00	13.77
BM25	S	96.00	60.00	4.00	13.90	64.00	10.00	28.00	16.97
	T	89.00	42.00	6.00	16.72	57.00	6.00	25.00	17.46
A-MEM		97.00	59.00	8.00	22.52	81.00	16.00	36.00	20.03
Mem0		96.00	58.00	3.00	24.32	60.00	5.00	34.00	19.41
SimpleMem		94.00	60.00	4.00	20.30	50.00	3.00	22.00	16.36
Qwen3-30B-A3B-Instruct-2507									
Full Context		44.00	12.00	0.00	12.32	31.00	0.00	8.00	8.21
Dense	S	81.00	55.00	4.00	19.13	63.00	19.00	43.00	22.30
	T	82.00	44.00	8.00	22.42	45.00	7.00	38.00	23.52
Window	S	84.00	70.00	5.00	24.33	65.00	20.00	42.00	22.12
	T	85.00	59.00	10.00	30.50	52.00	13.00	41.00	18.08
BM25	S	76.00	51.00	6.00	24.58	68.00	14.00	30.00	16.99
	T	64.00	41.00	3.00	20.05	51.00	7.00	24.00	16.42
A-MEM		76.00	59.00	7.00	22.19	56.00	10.00	36.00	21.33
Mem0		80.00	54.00	13.00	28.51	42.00	9.00	41.00	23.33
SimpleMem		73.00	45.00	7.00	26.38	51.00	10.00	27.00	23.28

Table 3: Performance across query types. Aln. denotes *alignment* and Grd. denotes *grounding*. S and T indicate *session-* and *turn-level* retrieval. Bold indicates the best result among practical systems.

Exact Recall at the session level. Because many queries require integrating information distributed across multiple sessions, session-level evidence coverage serves as the primary retrieval metric.

Response Evaluation and Metadata. We evaluate generated responses using gpt-5 as an LLM judge with query-type-specific rubrics along two dimensions: (1) **Alignment**, measuring consistency with the user’s experiences and longitudinal history, and (2) **Grounding**, measuring explicit support from the annotated evidence.

Each of the four query types has separate alignment and grounding rubrics, yielding eight rubric sets. Detailed rubrics and judging prompts are provided in Appendix C.1. The benchmark also provides query types, annotated evidence sessions, gold responses, and no-context responses. Gold responses use summarized evidence and evaluation rubrics, while no-context responses use only the query, both generated with gpt-5.

4 Experimental Setup

4.1 Models and Baselines

We evaluate RAG and memory systems using two generation models: gpt-5-mini (Singh et al.,

2026) and qwen3-30b-a3b-instruct (Yang et al., 2025). All methods use bge-base-en-v1.5 (Xiao et al., 2024) embeddings. We include three reference conditions: (1) **No-context**, where the model answers using only the query without conversational history; (2) **Oracle-session**, where the gold evidence sessions are directly provided; and (3) **Oracle-summary**, where the model receives summarized evidence from annotated metadata. As a long-context baseline, we evaluate **full-context**, where the model receives the complete conversational history. For retrieval-based baselines, we evaluate RAG systems under both turn-level and session-level retrieval settings. We compare **dense** retrieval, **BM25** sparse retrieval, and a simple **window** retrieval strategy that augments retrieved turns with nearby conversational context.

We further evaluate representative memory systems for long-term conversational personalization, including **A-MEM** (Xu et al., 2025), **Mem0** (Chhikara et al., 2025), and **SimpleMem** (Liu et al., 2026), which differ in how conversational histories are stored and retrieved. Each system provides the top- k retrieved memory records as context for response generation. Unless otherwise

Method	Event-Align ($n = 100$)		Event-Correct ($n = 100$)		Traj-Align ($n = 100$)		Traj-Correct ($n = 100$)	
	Aln.	Grd.	Aln.	Grd.	Aln.	Grd.	Aln.	Grd.
GPT-5-mini								
<i>No-Context</i>	41.00	0.00	2.00	6.85	4.00	0.00	0.00	2.49
<i>Oracle Session</i>	84.00	39.00	4.00	17.12	78.00	17.00	40.00	25.18
<i>Oracle Summary</i>	99.00	97.00	12.00	45.62	99.00	86.00	67.00	71.76
Qwen3-30B-A3B-Instruct-2507								
<i>No-Context</i>	39.00	0.00	1.00	7.94	0.00	0.00	1.00	1.68
<i>Oracle Session</i>	74.00	48.00	2.00	15.14	73.00	26.00	35.00	18.47
<i>Oracle Summary</i>	84.00	66.00	17.00	32.16	100.00	97.00	71.00	60.39

Table 4: No-Context and oracle settings. Aln. denotes *alignment* and Grd. denotes *grounding*.

specified, all methods use comparable retrieval budgets and the same backbone model for memory ingestion. Additional implementation details are provided in Appendix D.

4.2 Evaluation Metrics

Retrieval Evaluation. Using the annotated evidence sessions, we evaluate whether systems retrieve required information for each query. Because annotations are provided only at the session level, retrieval evaluation is reported only for session-level RAG systems. We report **Recall**, measuring the fraction of annotated evidence sessions retrieved, and **Exact Recall**, measuring whether all required evidence sessions are retrieved.

Response Evaluation. We evaluate generated responses using gpt-5 with the query-type-specific alignment and grounding rubrics provided in PRAGMA. To support evaluation robustness, we report evaluations using gemini-3.1-pro-preview (Google, 2026) and claude-opus-4.6 (Anthropic, 2026) with inter-judge agreement results in Appendix C.2.

5 Experimental Results

5.1 Main Results

In Table 3, Full-Context remains ineffective despite being given complete conversational history across both generation models. Specifically, under gpt-5-mini, Full-Context achieves only 8.00 grounding on Trajectory-Align, substantially below practical retrieval and memory systems. These results suggest that long conversational histories alone are insufficient for robust personalized guidance.

Performance also varies substantially across query types. Trajectory-grounded queries highlight the importance of broader conversational context. For example, under gpt-5-mini, Dense-session im-

proves Trajectory-Align alignment from 53.00 to 73.00 compared to turn-level retrieval, while A-MEM achieves the strongest performance at alignment. Nevertheless, grounding performance remains limited across practical systems, highlighting the difficulty of producing well-grounded personalized guidance over evolving user trajectories. Corrective queries are particularly challenging across all systems. Even when user assumptions conflict with prior memory, models often fail to produce corrective responses. Under gpt-5-mini, Event-Correct alignment remains between 2.00 and 8.00 across all practical systems, while Trajectory-Correct alignment remains below 43.00.

Across nearly all systems, alignment scores are substantially higher than grounding scores. For example, with gpt-5-mini, A-MEM achieves 81.00 alignment but only 16.00 grounding on Trajectory-Align, suggesting that models often generate plausible personalized guidance without effectively grounding it in conversational evidence.

We further evaluate additional retrieval and memory baselines, which broadly support our main findings (Appendix E.1).

5.2 Gold Retrieval is Not Enough

Table 4 presents no-context lower bounds and oracle settings for personalized guidance. Across both generation models, No-Context achieves near-zero grounding and very low alignment, indicating that PRAGMA instances are not solvable from query-only priors or generic guidance patterns. In other words, personalized guidance fundamentally must utilize the conversational memory.

Furthermore, to separate retrieval failure from evidence utilization failure, we evaluate two increasingly model-friendly oracle settings: Oracle-Session directly provides the annotated gold evidence sessions, while Oracle-Summary further

Method		Event-Align ($n = 100$)				Event-Correct ($n = 100$)				Traj-Align ($n = 100$)				Traj-Correct ($n = 100$)			
		Rec.	Ex.	Aln.	Grd.	Rec.	Ex.	Aln.	Grd.	Rec.	Ex.	Aln.	Grd.	Rec.	Ex.	Aln.	Grd.
<i>Dense</i>	S	85.30	50.00	81.00	55.00	89.07	68.00	4.00	19.13	62.48	0.00	63.00	19.00	69.83	18.00	43.00	22.30
	T	—	—	82.00	44.00	—	—	8.00	22.42	—	—	45.00	7.00	—	—	38.00	23.52
<i>BM25</i>	S	57.45	7.00	76.00	51.00	88.70	63.00	6.00	24.58	51.95	0.00	68.00	14.00	56.45	6.00	30.00	16.99
	T	—	—	64.00	41.00	—	—	3.00	20.05	—	—	51.00	7.00	—	—	24.00	16.42
<i>Window</i>	S	73.45	18.00	84.00	70.00	84.40	59.00	5.00	24.33	56.67	2.00	65.00	20.00	53.70	8.00	42.00	22.12
	T	—	—	85.00	59.00	—	—	10.00	30.50	—	—	52.00	13.00	—	—	41.00	18.08
<i>Dynamic</i>	S	69.20	16.00	74.00	48.00	77.57	40.00	5.00	17.99	64.17	4.00	40.00	8.00	64.15	12.00	30.00	15.12
	T	—	—	67.00	25.00	—	—	4.00	18.83	—	—	11.00	3.00	—	—	19.00	13.27
<i>QR</i>	S	97.20	91.00	71.00	53.00	96.98	91.00	6.00	23.81	95.95	80.00	25.00	20.00	97.17	89.00	26.00	16.81
	T	—	—	82.00	52.00	—	—	7.00	27.62	—	—	37.00	9.00	—	—	34.00	19.71
<i>AdaK</i>	S	79.50	39.00	79.00	50.00	83.50	55.00	4.00	17.07	59.43	0.00	30.00	21.00	62.46	15.00	35.00	20.46
	T	—	—	74.00	38.00	—	—	8.00	20.69	—	—	23.00	4.00	—	—	26.00	16.87

Table 5: RAG accuracy across query types and metrics. S and T indicate session- and turn-level retrieval, while QR and AdaK indicate query rewriting and Adaptive- K . Rec. and Ex. denote evidence recall and exact recall.

compresses the same evidence into concise summaries containing the key information needed for guidance. Despite removing retrieval as a bottleneck in both settings, we observe a substantial gap between Oracle-Session and Oracle-Summary, particularly on trajectory-grounded and corrective queries. For example, under gpt-5-mini, Oracle-Summary achieves 99.00 alignment and 86.00 grounding on Trajectory-Align, whereas Oracle-Session reaches only 78.00 and 17.00, respectively. This gap suggests that merely exposing the relevant conversation sessions is insufficient; models still struggle to organize and synthesize longitudinal evidence unless it is presented in an explicitly distilled form. Notably, even Oracle-Summary remains far from perfect on corrective queries, suggesting that these challenges cannot be resolved through retrieval quality alone.

We observe the same pattern across generation models of varying scales: Oracle-Summary consistently outperforms Oracle-Session, further showing that gold evidence access alone does not ensure effective memory utilization (Appendix E.2).

5.3 Retrieval-Response Discrepancy

To further analyze the gap between retrieval and response, we evaluate several RAG variants adapted to our setting, including Dynamic Retrieval (Jiang et al., 2023), Query Rewriting (Gao et al., 2023), and Adaptive- k (Taguchi et al., 2025) retrieval. All methods use the same embedding model and retrieval budget ($k \leq 10$), with qwen3-30b-a3b-instruct for generation.

Table 5 shows that many methods fail to recover the complete evidence required for personalized

guidance, particularly for trajectory-grounded and corrective queries. For example, Dense retrieval achieves 62.48 Recall but 0.00 Exact Recall on Trajectory-Align, while BM25 reaches 51.95 Recall with similarly low complete evidence recovery.

At the same time, strong retrieval performance does not necessarily lead to strong downstream responses. This pattern is most pronounced for Query Rewriting, which achieves near-oracle retrieval on Trajectory-Align (95.95 Recall, 80.00 Exact Recall) but still produces weak downstream responses (25.00 Alignment, 20.00 Grounding). Overall, these results suggest that personalized guidance requires improvements in both conversational retrieval and downstream memory utilization.

We also observe clear differences between retrieval granularities. For trajectory-grounded queries, session-level retrieval consistently outperforms turn-level retrieval on alignment. On Trajectory-Align, Dense improves from 45.00 to 63.00 and BM25 from 51.00 to 68.00. In contrast, Event-Correct queries often favor turn-level retrieval, as excessive session-level context can obscure fine-grained corrective evidence. For instance, Window grounding improves from 24.33 to 30.50 with turn-level retrieval. Overall, the results suggest that improving retrieval alone is insufficient for robust personalized guidance.

6 Analysis

6.1 Models Still Fail at Grounding

To better understand corrective guidance failures, we analyze Event-Correct queries, where user assumptions conflict with conversational history.

Method		Recall	Exact	Alignment	Alignment (2-Stage)	Grounding	Grounding (2-Stage)
<i>Dense</i>	S	89.07	68.00	76.0	99.0 (+23.0)	49.3	74.3 (+25.0)
	T	–	–	67.0	95.0 (+28.0)	49.3	71.1 (+21.8)
<i>Window</i>	S	84.40	59.00	77.0	100.0 (+23.0)	55.9	80.2 (+24.3)
	T	–	–	77.0	99.0 (+22.0)	54.2	78.0 (+23.9)
<i>BM25</i>	S	88.70	63.00	68.0	100.0 (+32.0)	48.0	79.5 (+31.5)
	T	–	–	67.0	97.0 (+30.0)	41.2	63.1 (+21.9)
<i>A-MEM</i>	–	–	–	80.0	100.0 (+20.0)	57.2	77.9 (+20.7)
<i>Mem0</i>	–	–	–	60.0	87.0 (+27.0)	39.8	59.0 (+19.2)
<i>SimpleMem</i>	–	–	–	70.0	93.0 (+23.0)	45.8	62.3 (+16.5)

Table 6: Results on corrective guidance queries involving misaligned user assumptions. S and T indicate session- and turn-level retrieval. Alignment and Grounding denote one-stage generation with an explicit instruction that the user’s assumption may be incorrect; 2-Stage first identifies inconsistencies before generating the response.

These queries require identifying inconsistencies and generating evidence-grounded corrections.

Adding an instruction that the user’s assumption may be incorrect substantially improves alignment across systems, suggesting that inconsistency detection itself is not the primary bottleneck. We further evaluate a two-stage setup that first identifies inconsistencies and then generates a corrective response conditioned on them. As shown in Table 6, alignment improves dramatically under this setup, often reaching near-perfect performance.

However, grounding remains substantially lower despite strong retrieval performance and explicit inconsistency identification. These results suggest that corrective guidance failures cannot be explained solely by inconsistency detection failures; reliably recovering and utilizing the appropriate conversational evidence remains challenging. Such grounding failures can propagate across future interactions, where earlier responses themselves become part of the conversational context used for subsequent reasoning. A detailed follow-up case study is provided in Appendix F.3.

The extremely low corrective alignment observed across practical systems also suggests a broader tendency toward over-accommodation to user assumptions, consistent with prior observations of sycophantic behavior in instruction-tuned LLMs (Sharma et al., 2024; Hong et al., 2025). In personalized guidance settings, this behavior becomes particularly problematic because effective assistance may require challenging the user’s current belief rather than simply validating it.

6.2 Where Do Memories Get Lost?

Previous experiments suggest that retrieval alone is insufficient for personalized guidance. To better understand system failures, we decompose the mem-

Type	System	St	Rt	Rs
EA	A-MEM	98.2	60.4	70.3
	Mem0	82.6	23.0	78.7
	SimpleMem	64.2	41.3	74.8
EC	A-MEM	98.0	81.2	66.4
	Mem0	83.2	57.7	62.7
	SimpleMem	65.1	65.5	72.0
TA	A-MEM	99.4	49.6	81.6
	Mem0	78.9	17.1	81.3
	SimpleMem	62.1	20.6	82.6
TC	A-MEM	99.1	48.9	63.3
	Mem0	77.5	20.1	81.2
	SimpleMem	58.1	31.0	72.5

Table 7: Evidence preservation across memory stages. St: fraction of gold evidence preserved in storage; Rt: fraction of stored evidence successfully retrieved; Rs: fraction of retrieved evidence reflected in the response.

ory pipeline into three aspects: **memory preservation**, whether evidence remains preserved in memory; **retrieval accessibility**, whether preserved evidence is successfully retrieved; and **response utilization**, whether retrieved evidence is reflected in the final response. All evaluations use entailment-style LLM judgments with gpt-5-nano.

Table 7 reveals substantial tradeoffs across memory systems. A-MEM achieves nearly perfect preservation across all query types by storing raw conversational content, while Mem0 and SimpleMem lose information during memory rewriting and compression. However, strong preservation does not necessarily translate into downstream utilization; on Traj-Correct, A-MEM preserves 99.1% of evidence but retrieves only 48.9%.

We further compare response-stage utilization across memory systems and RAG pipelines in Table 8. Despite preserving less information overall, summarized memory systems often achieve stronger downstream utilization than more detailed

System	EA	EC	TA	TC
<i>A-MEM</i>	70.3	66.4	81.6	63.3
<i>Mem0</i>	78.7	62.7	81.3	81.2
<i>SimpleMem</i>	74.8	72.0	82.6	72.5
<i>Dense</i>	67.0	61.2	76.8	51.5
<i>Window</i>	66.0	61.0	74.2	55.6
<i>BM25</i>	68.9	60.7	75.6	53.8

Table 8: Response-stage evidence utilization across query types. EA: Event-Align, EC: Event-Correct, TA: Trajectory-Align, TC: Trajectory-Correct.

memory representations and standard RAG. This suggests that memory abstraction is not merely a compression mechanism, but a critical interface between retrieval and generation: concise structured memories may discard some low-level conversational detail, yet expose the remaining evidence in a form that generation models can more reliably incorporate into personalized guidance. In contrast, raw conversational context can preserve more evidence while still leaving the generator to identify, organize, and synthesize the relevant implications. Overall, current systems struggle to simultaneously optimize preservation, retrieval accessibility, and utilization, highlighting the need for memory architectures that store information not only accurately, but also in generation-usable forms.

7 Conclusion

We introduced PRAGMA, a benchmark for evaluating personalized guidance in long-term conversations beyond factual recall. It evaluates whether models can provide grounded guidance under evolving user preferences and potentially incorrect assumptions. Experiments across RAG, memory systems, and long-context models reveal substantial failures in both memory retrieval and utilization. Even when relevant evidence is retrieved, models often fail to generate grounded personalized guidance, highlighting the need for memory systems that support robust longitudinal reasoning and memory-grounded generation beyond retrieval.

Limitations

PRAGMA focuses on controlled evaluation of memory-grounded personalized guidance, and several limitations remain for future work. First, although the benchmark is human-validated, the conversational histories are generated through a controllable synthetic pipeline. This design enables evidence annotation, trajectory control, and system-

atic evaluation across diverse memory scenarios, but may not fully capture the ambiguity and variability of natural long-term human conversations.

Second, the benchmark evaluates guidance generation in a single-turn setting. In real deployments, conversational agents may recover from incomplete memory retrieval through iterative interaction or follow-up dialogues. Future works could extend PRAGMA toward interactive multi-turn evaluation of memory utilization and conversational recovery.

Finally, retrieval evaluation is based on annotated evidence sessions rather than fine-grained reasoning traces. While this abstraction improves annotation reliability and scalability, some queries may admit multiple valid reasoning paths or rely on partially implicit evidence distributed across conversations. Developing more fine-grained evaluation protocols for longitudinal memory reasoning remains an important direction for future research.

Ethical Considerations

PRAGMA is constructed from fully synthetic conversational histories and does not contain real user data or personally identifiable information. However, models may overfit to benchmark-specific annotation structures or reasoning patterns rather than developing robust long-term personalization capabilities. In addition, although PRAGMA is designed to cover diverse personas and longitudinal behaviors, synthetic generation pipelines may still underrepresent certain cultural, linguistic, or interactional patterns, potentially introducing unintended biases in evaluation outcomes. We therefore encourage future work to evaluate whether improvements on PRAGMA transfer to more open-ended and realistic conversational settings.

Acknowledgments

This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2024-00437633). This work was conducted in collaboration with LYWAY on domain-specific AI research, whose support for our memory research and funding of the API costs we gratefully acknowledge. K. Jung is with ASRI, Seoul National University, Korea. The Institute of Engineering Research at Seoul National University provided research facilities for this work.

References

- Anthropic. 2026. Claude opus 4.6 system card. <https://www-cdn.anthropic.com/6a5fa276ac68b9aeb0c8b6af5fa36326e0e166dd/Claude%20opus%204.6%20System%20Card.pdf>.
- Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. [How People Use ChatGPT](#). OpenAI Economic Research Report.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. [Mem0: Building production-ready ai agents with scalable long-term memory](#). In *European Conference on Artificial Intelligence*.
- Jizhan Fang, Xinle Deng, Haoming Xu, Ziyang Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2026. [Lightmem: Lightweight and efficient memory-augmented generation](#). In *International Conference on Learning Representations*, volume 2026, pages 98706–98729.
- Xueyang Feng, Weinan Gan, Xu Chen, Quanyu Dai, and Yong Liu. 2026. [How does personalized memory shape llm behavior? benchmarking rational preference utilization in personalized assistants](#). *Preprint*, arXiv:2601.16621.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.
- Google. 2026. Gemini 3.1 pro preview. <https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview>.
- Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. [Measuring sycophancy of language models in multi-turn dialogues](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. [Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale](#). In *Proc. of the Conference on Language Modeling (COLM)*.
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Hyunwoo Kim, Niloofar Miresghallah, Michael Duan, Rui Xin, Shuyue Stella Li, Jaehun Jung, David Acuna, Qi Pang, Hanshen Xiao, G. Edward Suh, Sewoong Oh, Yulia Tsvetkov, Pang Wei Koh, and Yejin Choi. 2026. [Privasis: Synthesizing the largest "public" private dataset from scratch](#). *Preprint*, arXiv:2602.03183.
- Taeyoon Kwon, Dongwook Choi, Hyojun Kim, Sunghwan Kim, Seungjun Moon, Beong woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. 2026. [Embodied agents meet personalization: Investigating challenges and solutions through the lens of memory utilization](#). In *The Fourteenth International Conference on Learning Representations*.
- Philippe Laban, Tobias Schnabel, and Jennifer Neville. 2026. [Llms corrupt your documents when you delegate](#). *Preprint*, arXiv:2604.15597.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. [Hello again! llm-powered personalized agent for long-term dialogue](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5259–5276. Association for Computational Linguistics.
- Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. 2025b. [Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 11493–11506. Association for Computational Linguistics.
- Yu Li, Yuan Huang, Tao Wang, Caiyu Fan, Xiansheng Cai, Sihan Hu, Xinzijian Liu, Cheng Shi, Mingjun Xu, Zhen Wang, Yan Wang, Xiangqi Jin, Tianhan Zhang, Linfeng Zhang, Lei Wang, Youjin Deng, Pan Zhang, Weijie Sun, Xinyu Li, and 4 others. 2026. [Inverse knowledge search over verifiable reasoning: Synthesizing a scientific encyclopedia from a long chains-of-thought knowledge base](#). *Preprint*, arXiv:2510.26854.
- Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026. [Simplemem: Efficient lifelong memory for llm agents](#). *Preprint*, arXiv:2601.02553.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13851–13870. Association for Computational Linguistics.
- Meta. 2025. Llama 4 Scout 17B-16E Instruct. <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct>. Model card.

- Jisoo Mok, Ik-hwan Kim, Sangkwon Park, and Sungroh Yoon. 2025. [Exploring the potential of LLMs as personalized assistants: Dataset, evaluation, and analysis](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10212–10239, Vienna, Austria. Association for Computational Linguistics.
- OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. <https://arxiv.org/pdf/2508.10925>. Accessed: 2026-08-28.
- Egor Pakhomov, Erik Nijkamp, and Caiming Xiong. 2025. [Convomem benchmark: Why your first 150 conversations don’t need rag](#). *Preprint*, arXiv:2511.10523.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. [Raptor: Recursive abstractive processing for tree-organized retrieval](#). In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding sycophancy in language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Chihiro Taguchi, Seiji Maekawa, and Nikita Bhutani. 2025. [Efficient context selection for long-context QA: No tuning, no iteration, just adaptive- \$k\$](#) . In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20105–20130, Suzhou, China. Association for Computational Linguistics.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. [In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 8416–8439. Association for Computational Linguistics.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. [Longmemeval: Benchmarking chat assistants on long-term interactive memory](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). *Preprint*, arXiv:2309.07597.
- Xingchen Xiao, Heyan Huang, Runheng Liu, and Jincheng Xie. 2026. [MASS-RAG: Multi-agent synthesis retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 9865–9883, San Diego, California, United States. Association for Computational Linguistics.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. [A-mem: Agentic memory for llm agents](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 17577–17604. Curran Associates, Inc.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Wanjuan Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

A Benchmark Construction Details

A.1 Prompt Templates and Generation Pipeline

Below, we provide abbreviated prompt templates used in our benchmark generation pipeline.

Event Generation

Given the persona, axes, and topic, create a few specific past events for the user.
Events: focus on the topic; should NOT be relevant to the axes; include temporal or spatial words; start with “The user.”
Timestamps: between 2025-05-01 and 2026-04-30, sorted in temporal order.

Figure 3: Prompt template for event generation.

Trajectory Generation

Given the persona and axes, create a trajectory for the user.
Trajectory:

- Sequential with 4–8 states.
- Each state relevant to at least one axis; some states relevant to only one axis.
- Include the user’s initial preference and drift over time.
- Vary in pattern (gradual shift, oscillation, plateau, partial reversal).
- Avoid overly clean or perfectly structured progression.
- The last state should NOT represent the user’s current state.

Timestamps: between 2025-05-01 and 2026-04-30, sorted in temporal order.

Figure 4: Prompt template for trajectory generation.

Event-Aligned Query Generation (Type 1)

Given the topic and past user events, summarize the events then create a user query.
Summary: first-person (“I”), concise.
Query:

- Should NOT restate or hint at the events; remain implicit.
- Should ask for recommendation or advice on the topic.
- Should be implicitly grounded in past experiences without meta-phrases such as “based on my experiences.”
- Should sound natural, concise, and underspecified.

Output format:
Summary: <summary>
Query: <query>

Figure 5: Prompt template for Event-Aligned query generation. Additional few-shot examples were provided during generation.

Event-Corrective Query Generation

Given the past user events, select multiple events then generate a user query.
Query:

- Construct a plausible but imperfect recollection by blending multiple past events.
- Be confidently incorrect about the selected events.
- Remain grounded in the original events (no completely new activities).
- Ask for recommendation or advice based on the incorrect recollection.
- Be concise.

Selected events: output the indices of the mixed-up events used in the query.

Figure 6: Prompt template for Event-Corrective query generation. Additional few-shot examples were provided during generation.

Trajectory-Aligned Query Generation

Given the persona, rewritten axes, and user trajectory, create a user query.
Query:

- A realistic self-positioning query (e.g., CV, bio, introduction, application).
- Ask how to position themselves along the rewritten axes using past experience.
- Include only the rewritten axes terms; do NOT state past experiences or trajectory.
- Remain underspecified without explicitly stating progression or trade-offs.
- Be concise and natural.

Output format: Query: <query>

Figure 7: Prompt template for Trajectory-Aligned query generation.

Trajectory-Corrective Query Generation

Given the user’s trajectory, select multiple states, then create a user query.
Query:

- Present a concrete situation and a decision the user is considering.
- The decision should be reasonable without additional context, but subtly misaligned with the trajectory.
- The misalignment should emerge only when reasoning across both axes.
- Do NOT reference the trajectory, mention trade-offs, or signal doubt.
- Be concise and natural.

Selected states: output the indices of states that serve as counter-evidence.

Figure 8: Prompt template for Trajectory-Corrective query generation. Additional few-shot examples were provided during generation.

A.2 End-to-End Construction Example

Table 17 provides a running example of the benchmark construction pipeline in Figure 2, tracing a single user from the initial Privasis-Zero profile to the final PRAGMA instances. The user profile is transformed into a persona, event topic, and trajectory axes, which guide the generation of events, longitudinal trajectories, filler topics, and the four query types. These components are expanded into timestamped conversational sessions and combined into a shared long-term history, with each query linked to its corresponding evidence and evaluation metadata.

A.3 Human Validation Guidelines

We conducted a human validation study to verify that generated conversations, queries, and evidence annotations support the intended personalized-memory reasoning tasks. Validation was conducted by a team of six annotators, including three co-authors, two NLP researchers, and one researcher in linguistics. Annotators were provided with the user metadata, target query, annotated evidence sessions, and the corresponding conversation history. For each example, annotators answered query-specific validation questions using a binary yes/no rubric, with optional free-form comments and query rewrites for unnatural or ambiguous cases.

Annotators were instructed to reject examples when: (1) the query was unnatural or unrealistic, (2) evidence annotations were incomplete or incorrect, (3) filler sessions leaked relevant information, (4) the intended inconsistency was weak or unsupported, or (5) the query could be solved through superficial heuristics without reasoning over the provided history.

Type-specific validation criteria included:

- **Event-Aligned:** Whether the query required event-grounded personalized guidance and whether irrelevant filler sessions remained unrelated to the target topic.
- **Event-Corrective:** Whether the query introduced a realistic event-level misconception requiring corrective guidance and whether supporting evidence was naturally distributed across sessions.

- **Trajectory-Aligned:** Whether answering the query required reasoning over longitudinal changes in the user’s preferences, goals, or circumstances.
- **Trajectory-Corrective:** Whether the proposed user decision meaningfully conflicted with the established trajectory and required corrective reasoning grounded in the conversational history.

When a query was understandable but unnatural, annotators were encouraged to provide rewritten versions while preserving the intended query type and evidence dependency. All annotators were compensated based on estimated task completion time in accordance with local institutional research assistant compensation practices. Because the validation process was designed primarily for quality control and iterative refinement, examples were divided across annotators rather than exhaustively double-annotated. As a result, we do not report inter-annotator agreement statistics.

A.4 Conversation History Example

Using the same running example as Appendix A.2, Table 18 shows an excerpt from the resulting long-term conversation history. The example spans 354 days from the first evidence session to the final query, with five relevant evidence sessions distributed over 333 days and interleaved with unrelated conversational sessions. This illustrates how PRAGMA requires models to recover and integrate temporally distributed evidence from a long, heterogeneous interaction history rather than relying on a single recent or topically concentrated context.

A.5 Query Taxonomy Examples

Table 19 shows representative examples for each PRAGMA query type, ranging from event-level personalization to trajectory-grounded corrective reasoning.

B Additional Dataset Analysis

B.1 Implicitness Analysis

We report an auxiliary analysis of query implicitness in PRAGMA. The goal is to quantify

Type	RI	II	Score	n
<i>Event-Align</i>	0.6566	0.4120	0.5227	100
<i>Event-Correct</i>	0.6131	0.7655	0.6817	100
<i>Trajectory-Align</i>	0.6718	0.3470	0.4902	100
<i>Trajectory-Correct</i>	0.6424	0.6940	0.6656	100
<i>Overall</i>	0.6460	0.5546	0.5900	400

Table 9: PRAGMA implicitness scores by query type.

Analysis	GPT	Qwen
RI vs Retrieval F1	-0.2292	-0.2026
Implicitness vs Alignment	-0.3855	-0.3127

Table 10: Average Spearman correlations across methods between PRAGMA implicitness scores and retrieval/response outcomes.

how much a query depends on latent user history rather than explicitly stating the required evidence or response behavior. Since there are no widely used metrics for implicitness, we introduce a simple diagnostic measure to characterize the extent to which a query leaves the relevant memory evidence and intended response behavior implicit.

We decompose implicitness into two components. **Retrieval Implicitness (RI)** measures how difficult it is to recover the relevant evidence from the query surface form, using lexical and semantic overlap between the query and its supporting history. Higher RI indicates that the needed evidence is less directly recoverable from the query alone. **Instructional Implicitness (II)** measures whether the query explicitly signals the intended personalized reasoning behavior, such as grounding a recommendation in prior events or identifying a contradiction with a trajectory. We use gpt-5-mini as a judge, providing query-type definitions and a discrete ordinal rubric to score intent explicitness. We convert this explicitness score into instructional implicitness (II), where higher values indicate that the intended response behavior is less explicit in the query. We combine these components into an overall implicitness score.

Table 9 shows that PRAGMA queries are generally implicit, with an average score of 0.5900 across 400 queries. Corrective query types exhibit substantially higher implicitness than alignment-oriented queries: Event-Correct and Trajectory-Correct obtain scores of 0.6817 and

0.6656, respectively, compared to 0.5227 and 0.4902 for Event-Align and Trajectory-Align. This trend is expected, as corrective queries often appear superficially plausible unless models retrieve and reason over conflicting prior evidence.

The negative correlations in Table 10 suggest that the proposed implicitness measures are broadly aligned with the intended characteristics of PRAGMA queries. Queries with higher retrieval implicitness tend to exhibit lower retrieval F1, while higher overall implicitness is associated with lower downstream alignment performance. This trend is consistent with the design goal of evaluating under-specified memory reasoning beyond explicit lexical overlap.

B.2 Benchmark Statistics

Table 11 and 12 present summary statistics for PRAGMA. PRAGMA contains 100 users and 400 queries, with 100 queries for each of the four query types. Each user has an average of 41.67 sessions, consisting of event sessions, trajectory sessions, and irrelevant filler sessions. In total, the benchmark contains 4,167 sessions: 494 event sessions, 673 trajectory sessions, and 3,000 filler sessions. Filler sessions account for 72.0% of all sessions, making relevant evidence sparse within the full user history.

Each user has 4–5 event memories and 6–7 trajectory memories over two latent trajectory axes. Queries require 4.91 evidence items on average. Each query contains 44.6 tokens on average, while each user history contains 160K tokens on average (median 160K; range 127K–194K). Token counts are computed with the gpt-5-mini tiktoken encoding over the timestamped conversation history with role prefixes. The supporting evidence spans 282.8 days on average, requiring systems to retrieve and reason over temporally distributed information.

PRAGMA also covers diverse user profiles: ages range from 19 to 85, with 47 native languages, 43 citizenships, and 95 unique topics. These statistics reflect the benchmark’s focus on long, sparse, and heterogeneous personalized memory.

Statistic	Value
Users	100
Queries	400
Queries per type	100
Sessions	4,167
Sessions / user	41.67
Filler sessions	3,000 (72.0%)
Event sessions	494
Trajectory sessions	673
Events / user	4.94
Trajectory states / user	6.73
Evidence / query	4.91
Query tokens	44.6
History tokens / user	160K
History tokens / user (range)	127K–194K
Turns / user	323.92
Evidence span / query	282.8 days
Native languages	47
Citizenships	43
Topics	95

Table 11: Summary statistics for PRAGMA. Averages are reported for per-user and per-query quantities.

Type	Evidence	Words
Event-Align	4.94	20.1
Event-Correct	3.52	35.6
Trajectory-Align	6.73	59.3
Trajectory-Correct	4.44	32.2

Table 12: Average evidence count and query length by query type.

Benchmark scale. PRAGMA prioritizes quality and complexity over query count. Each of the 400 queries is grounded in a long history with temporally distributed evidence, averaging 160K tokens and 4.91 evidence items per query. Constructing each instance requires coherence across the user profile, conversational history, distributed evidence, and personalized query. The evidence and queries are human-validated to ensure reliable grounding and personalization across four query types and diverse user profiles. While the pipeline can be readily scaled, human validation introduces a practical trade-off between benchmark scale and quality.

C Evaluation Details

C.1 Evaluation Prompts

All automatic evaluations are conducted using rubric-based prompts tailored to each metric and query type. The prompts instruct the evaluator model to assess responses with respect to conversational alignment and evidence grounding while considering the provided conversational context and annotated evidence.

Evaluator System Prompt (shared)

Given the query, response, and evaluation criteria, evaluate whether the response is personalized. First, briefly justify how the response satisfies each criterion. Then, assign a score for each criterion as noted in the criteria. Return the scores as an array of integers or floats, in the same order as the criteria.

Evaluator User Prompt Template (shared)

Query: {query}
Response: {response}
Evaluation Criteria: {criteria}

Figure 9: Shared prompts used for all evaluation calls. Top: system prompt; Bottom: user prompt template. {criteria} is filled with the per-query-type criteria below.

Alignment Criterion — Event-Aligned Query

Does the response recommend a next step that meaningfully builds on the user’s prior attempts, rather than simply repeating past actions or suggesting something unrelated?
Output 1 if yes, 0 otherwise. Only output the integer.
History: {summarized_evidence}

Grounding Criterion — Event-Aligned Query

Does the response explicitly use the user’s prior attempts as evidence for why this recommendation fits, rather than giving only generic advice?
Output 1 if yes, 0 otherwise. Only output the integer.
History: {summarized_evidence}

Figure 10: Evaluation criteria for Event-Aligned queries. Top: alignment (primary); Bottom: grounding (auxiliary).

Alignment Criterion — Event-Corrective Query
Does the response identify the event-level inconsistency between the user’s query and their prior experience? Output 1 if yes, 0 otherwise. Only output the integer. History: {summarized_evidence}
Grounding Criterion — Event-Corrective Query
How many of the following history items does the response explicitly refer to? Answer with the coverage score computed by (# referred items) / (# total history items). Output ONLY the final score as a float between 0 and 1, rounded to 2 decimal places. History: {summarized_evidence}

Figure 11: Evaluation criteria for Event-Corrective queries. Top: alignment (primary); Bottom: grounding (auxiliary).

Alignment Criterion — Trajectory-Corrective Query
Does the response identify the trajectory-level inconsistency between the user’s intended decision and their prior trajectory? Output 1 if yes, 0 otherwise. Only output the integer. History: {summarized_evidence}
Grounding Criterion — Trajectory-Corrective Query
How many of the following history items does the response explicitly refer to? Answer with the coverage score computed by (# referred items) / (# total history items). Output ONLY the final score as a float between 0 and 1, rounded to 2 decimal places. History: {summarized_evidence}

Figure 13: Evaluation criteria for Trajectory-Corrective queries. Top: alignment (primary); Bottom: grounding (auxiliary).

Alignment Criterion — Trajectory-Aligned Query
Does the response recommend a self-positioning strategy that is meaningfully aligned with the user’s trajectory across both axes, and grounds this positioning in specific past experiences, including how those experiences should be emphasized, downplayed, or omitted? Output 1 if yes, 0 otherwise. Only output the integer. Progression: {summarized_evidence}
Grounding Criterion — Trajectory-Aligned Query
Does the response explicitly justify that recommendation using the user’s prior progression across both axes, rather than relying on generic advice or surface-level similarity? Output 1 if yes, 0 otherwise. Only output the integer. Progression: {summarized_evidence}

Figure 12: Evaluation criteria for Trajectory-Aligned queries. Top: alignment (primary); Bottom: grounding (auxiliary).

C.2 Judge Validation and Agreement

We additionally evaluate inter-judge agreement between our primary gpt-5 evaluator and two independent evaluators, gemini-3.1-pro-preview and claude-opus-4.6 with temperature 0.0 on a balanced audit subset. The subset contains 800 judged instances from the main comparison setting, corresponding to 5% of the 16,000 evaluated instances in this setting. It is balanced across two response models, two evaluation metrics, ten methods, and four query types, with five examples per cell. As shown in Table 13, both independent evaluators show strong agreement with gpt-5. gemini-3.1-pro-preview achieves 82.9% exact agreement, Pearson $r = 0.769$, and Spearman $\rho = 0.759$, while claude-opus-4.6 achieves 78.6% exact agreement, Pearson $r = 0.667$, and Spearman $\rho = 0.650$ with gpt-5. Across all three evaluators, three-way exact agreement reaches 72.3%, with Krippendorff’s $\alpha = 0.697$. Exact agreement is consistently higher for alignment than grounding, suggesting that evidence-grounding judgments are more challenging while overall judgments remain substantially

GPT-Gemini				
Setting	N	Exact	r	ρ
Overall	800	82.9	0.769	0.759
GPT response	400	83.5	0.778	0.770
Qwen response	400	82.3	0.753	0.742
Alignment	400	89.0	0.792	0.792
Grounding	400	76.8	0.705	0.702
GPT-Claude				
Setting	N	Exact	r	ρ
Overall	800	78.6	0.667	0.650
GPT response	400	77.0	0.652	0.632
Qwen response	400	80.3	0.684	0.673
Alignment	400	81.5	0.670	0.670
Grounding	400	75.8	0.654	0.659
Gemini-Claude				
Setting	N	Exact	r	ρ
Overall	800	80.5	0.686	0.690
GPT response	400	80.8	0.702	0.702
Qwen response	400	80.3	0.666	0.676
Alignment	400	81.5	0.623	0.623
Grounding	400	79.5	0.785	0.804
Three-judge agreement				
Setting	N	Exact	α	
Overall	800	72.3	0.697	
GPT response	400	71.8	0.692	
Qwen response	400	72.8	0.698	
Alignment	400	76.0	0.674	
Grounding	400	68.5	0.709	

Table 13: Inter-judge agreement on the balanced 800-instance audit subset. Exact denotes exact agreement (%); r and ρ denote Pearson and Spearman correlation, respectively. The final panel reports three-way exact agreement and Krippendorff’s α with interval distance.

consistent across evaluator models.

In Table 14, we also report gpt-5 evaluation results on no-context responses and gold reference responses provided in the benchmark metadata. Gold responses consistently obtain near-ceiling scores, while no-context responses score substantially lower, suggesting that the evaluator reliably follows the intended rubrics and meaningfully distinguishes grounded personalized guidance from generic responses.

D Implementation and Baseline Details

Unless otherwise noted, experiments are evaluated under fixed retrieval settings and deterministic or near-deterministic decoding configurations. We therefore report single-run results without variance estimates or error bars.

D.1 Model and Retrieval Configurations

Model Setup. For all dense retrieval and memory-system experiments, we use BAAI/bge-base-en-v1.5 as the embedding model unless otherwise noted. For response generation, we evaluate two backbone LLMs: gpt-5-mini-2025-08-07 and qwen3-30b-a3b-instruct-2507. Unless otherwise noted, prompts use the same system instruction across methods: the model is asked to generate a concise personalized response conditioned on the retrieved memories or conversation context. For Qwen, we use deterministic decoding with temperature 0.0 in all runs.

Static Retrieval Baselines. For retrieval baselines, we evaluate both turn-level and session-level variants. Because retrieval units differ substantially in length, we set retrieval depth by granularity rather than using a single global top- k . Turn-level baselines retrieve 10 turns per query, while session-level baselines retrieve 5 sessions per query. The session-level depth approximately matches the typical number of evidence sessions associated with each query and avoids giving session-level baselines an excessively large context budget. These retrieval depths were fixed before evaluation and were not tuned on the test set.

The turn-level BM25 baseline indexes each user turn and retrieves the top 10 turns using Okapi BM25 with default hyperparameters. The turn-level dense baseline embeds each turn and retrieves the top 10 turns by cosine similarity. The window baseline follows a simple hierarchical retrieval strategy: it first retrieves the top 10 child turns using the same dense retriever, then expands each retrieved turn into a local context window of ± 2 surrounding turns.

For session-level retrieval, BM25 indexes full sessions and retrieves the top 5 sessions using default BM25 hyperparameters. The session-level dense baseline embeds full sessions and retrieves the top 5 sessions by cosine similarity. The session-level window baseline uses turn-level retrieval as an intermediate step: it retrieves the top 10 child turns and then expands each hit to its parent session. Since multiple retrieved turns can map to the same

Method	Event-Align		Event-Correct		Trajectory-Align		Trajectory-Correct		Overall	
	Align	Ground	Align	Ground	Align	Ground	Align	Ground	Align	Ground
<i>No-Context</i>	60.00	0.00	2.00	9.52	14.00	0.00	1.00	1.62	19.25	2.79
<i>Gold</i>	100.00	100.00	98.00	100.00	99.00	99.00	100.00	100.00	99.25	99.75

Table 14: Reference evaluations with gpt-5.

parent session, this yields fewer than 10 unique sessions in practice: 4.2 sessions per query on average. Retrieved contexts are then passed to the response model using the same response-generation prompt.

Dynamic Retrieval Variants. We additionally implement three dynamic RAG variants inspired by prior retrieval-augmented generation methods: dynamic retrieval, query rewriting, and adaptive- k retrieval. All methods use qwen3-30b-a3b-instruct for generation. We exclude gpt-5-mini from these experiments because dynamic retrieval requires token-level log probabilities.

The dynamic retrieval baseline is inspired by FLARE. It first retrieves an initial set of memories, then generates short look-ahead continuations and uses low-confidence generations to trigger additional retrieval. We use deterministic decoding with temperature 0.0, a maximum of six generation steps, 64 tokens per look-ahead sentence, and a low-confidence threshold of probability 0.8. Each triggered retrieval retrieves the top 2 turn-level memories. In practice, this yields 8.1 retrieved memories per query on average, close to the fixed top-10 budget used by the turn-level RAG baselines. The session level variant retrieves 5.2 sessions per query on average.

The query-rewriting baseline is inspired by HyDE. It first generates a hypothetical answer passage for the query and uses that generated passage. Retrieval is then performed using the average of the original-query embedding and the hypothetical-passage embedding. We use temperature 0.7 for hypothetical-passage generation, matching the open-ended generation setting used in HyDE-style retrieval, and generate the final answer deterministically with temperature 0.0.

The adaptive- k baseline is inspired by adaptive- k retrieval. For each query, it com-

putes similarities against all candidate turns and chooses the number of retrieved memories by applying a largest-gap heuristic to the sorted similarity scores. Following the Adaptive- k implementation, we ignore the lower half of the score distribution when searching for the gap and include two additional items beyond the selected cutoff as a small buffer. We cap the final retrieval depth at the same maximum budget as the turn-level RAG baselines, i.e., at most 10 retrieved turns. For the session-level variant, we apply the same procedure with the corresponding session-level retrieval budget.

D.2 Memory System Configurations

We evaluate three memory-system baselines: A-MEM, Mem0 and SimpleMem. For all memory systems, we ingest the full chronological history of each user before answering any query. To isolate the effect of the final response generator, the memory-construction backbone is fixed to qwen3-30b-a3b-instruct, while the final response is generated with either the same model or gpt-5-mini. Unless otherwise noted, Qwen-based generation uses deterministic decoding with temperature 0.0.

A-MEM. For A-MEM, we use the official agentic memory implementation with bge-base-en-v1.5 as the embedding model. Each user-assistant chunk is added as an A-MEM note. During ingestion, A-MEM converts each note into a structured memory containing the memory content, generated context, keywords, tags, temporal metadata, importance score, and links to related memories. At query time, we retrieve the top 10 related A-MEM notes using its internal embedding retriever, which searches over structured note representations containing memory content, generated context, keywords, and tags. The retrieved memory contents are then provided to the response generator.

Mem0. For Mem0, we split each user session into role-valid user–assistant chunks and add each chunk with its timestamp as metadata. Mem0 stores memories in a Qdrant vector store with 768-dimensional vectors. At query time, we call Mem0’s search API with a per-user filter to retrieve the top 10 memories. The retrieved memory strings are provided to the response generator with the original query.

SimpleMem. For SimpleMem, we add each dialogue turn with its speaker role, timestamp, and session identifier, then call SimpleMem’s finalization step to build the user’s memory store. We follow SimpleMem’s default settings for planning and parallel ingest/retrieval, but disable reflection-based additional retrieval to keep the retrieval budget controlled. At query time, we use SimpleMem’s retrieval interface and pass up to 10 retrieved memory contexts to the response generator.

E Additional Experiments

E.1 Additional Baseline Results

We additionally evaluate three strong baselines that capture complementary approaches to long-term memory: **LightMem** (Fang et al., 2026), a summary-based memory system; **RAPTOR** (Sarathi et al., 2024), a hierarchical retrieval framework; and **Synth**, an evidence-synthesis pipeline inspired by MASS-RAG (Xiao et al., 2026). We follow the original implementation of each method, with minor adaptations to fit our conversational benchmark. Table 15 reports their performance across both generation models.

RAPTOR performs strongly on aligned queries but remains limited on corrective reasoning, suggesting that hierarchical memory organization alone is insufficient for corrective memory utilization. **LightMem** provides a competitive summary-based baseline, with performance generally comparable to existing memory systems, further indicating that summarization alone does not resolve the utilization bottleneck. In contrast, **Synth** substantially improves performance on corrective queries compared with turn-level dense retrieval, which uses the same retrieval granularity. This result highlights the benefit of explic-

itly synthesizing retrieved evidence and suggests that improved evidence utilization can yield substantial gains even under similar retrieval settings. Nevertheless, no single approach consistently achieves strong alignment and grounding across all query types.

E.2 Additional Generation Models

To examine whether our findings generalize beyond the generation models used in the main experiments, we additionally evaluate three models from different families and scales: gpt-oss-120b (OpenAI, 2025), llama-4-scout-17b-16e-instruct (Meta, 2025), and claude-opus-4.6. We evaluate oracle settings for all three models, along with representative retrieval and memory systems. Table 16 reports the results.

Across all three models, providing summarized oracle evidence substantially improves performance over providing the original evidence sessions, reinforcing the importance of effective memory utilization. For gpt-oss-120b, overall alignment and grounding increase from 45.50 and 28.09 with Oracle-Session to 70.25 and 71.72 with Oracle-Summary. The gap is particularly large for llama-4-scout-17b-16e, increasing from 8.00/7.07 to 66.50/63.54. Even for the substantially stronger claude-opus-4.6, Oracle-Summary improves overall alignment and grounding from 74.00/67.54 to 90.75/92.28. These results indicate that access to relevant evidence alone does not guarantee effective utilization, even for stronger generation models.

Corrective reasoning also remains challenging across model families. Under practical retrieval and memory settings, gpt-oss-120b and llama-4-scout-17b-16e achieve low alignment on Event-Correct, while claude-opus-4.6 performs substantially better but still lags behind its performance on aligned queries. Even with Oracle-Summary, Claude reaches only 65.00 alignment on Event-Correct, compared with 100.00 on both Event-Align and Trajectory-Align and 98.00 on Trajectory-Correct. This suggests that corrective guidance remains difficult even when the required evidence is explicitly available in a concise form.

Method	Event-Align		Event-Correct		Trajectory-Align		Trajectory-Correct		Overall	
	Align	Ground	Align	Ground	Align	Ground	Align	Ground	Align	Ground
<i>GPT-5-mini</i>										
Synth	96.00	55.00	53.00	55.65	65.00	11.00	72.00	32.72	71.50	38.59
RAPTOR	98.00	73.00	7.00	26.92	81.00	15.00	44.00	21.16	57.50	34.02
LightMem	95.00	55.00	9.00	32.15	38.00	2.00	27.00	23.91	42.25	28.27
<i>Qwen3-30B-A3B-Instruct-2507</i>										
Synth	89.00	55.00	14.00	33.57	55.00	11.00	55.00	28.54	53.25	32.03
RAPTOR	88.00	73.00	8.00	27.74	69.00	29.00	32.00	24.87	49.25	38.65
LightMem	71.00	52.00	17.00	32.67	30.00	7.00	32.00	31.13	37.50	30.70

Table 15: Results for additional baselines across generation models. Synth explicitly synthesizes retrieved evidence, RAPTOR uses hierarchical retrieval, and LightMem constructs summary-based memories. Align and Ground denote alignment and grounding scores, respectively.

Finally, the relative effectiveness of retrieval granularities and memory representations varies across generation models. While structured or summarized memories can substantially benefit some models and query types, stronger models such as `claude-opus-4.6` often perform well with session-level retrieval, which preserves more of the original conversational context. No single representation is consistently optimal across models and query types. Overall, these results reinforce our main conclusion that effective memory utilization remains a central challenge across model families and capacities.

F Qualitative Analysis

F.1 Memory Representation Examples

Table 20 shows representative stored memories from the same user across A-MEM, Mem0, and SimpleMem. The examples illustrate the different storage formats used by each memory system. A-MEM stores structured memory notes with metadata and contextual fields, Mem0 stores atomic natural-language memory entries, and SimpleMem stores structured atomic entries consisting of a lossless restatement plus metadata fields such as keywords, timestamp, location, persons, entities, and topic.

F.2 Retrieval–Utilization Failure Cases

Table 21 shows representative cases where standard RAG systems successfully retrieved all annotated evidence but nevertheless failed to incorporate much of that information into the final response. In contrast, memory systems often produced more complete responses on the same queries using fewer but more struc-

tured memories. These examples suggest that personalized guidance depends not only on retrieval quality, but also on how retrieved information is organized and exposed to the response model.

F.3 Grounding Failures in Follow-up Interactions

Table 22 illustrates why alignment alone is insufficient for evaluating personalized guidance. We construct the follow-up case from an existing event-aligned wellness query for which the initial response is judged aligned in both conditions. The two conditions differ in grounding: the grounded condition uses an oracle-summary response with alignment=1 and grounding=1, while the poor-grounding condition uses a SimpleMem response with alignment=1 and grounding=0. For the follow-up turn, each condition includes its own first model response in chat history, and both conditions are then given the same follow-up query. The grounded condition receives the oracle evidence summaries as memory context, whereas the poor-grounding condition performs fresh SimpleMem retrieval using the follow-up query.

During the follow-up interaction, this difference leads to substantially different recommendations. The user asks whether they should replace a daily sweet bottled drink after hard morning runs with a sports hydration drink. The grounded condition retains the user’s prior low-salt meal plan, nutritionist consultation, and lower-sugar dietary choices, and therefore gives a conditional recommendation that cautions against treating sports drinks as an unchecked default. In contrast, the poor-

Method	Event-Align		Event-Correct		Trajectory-Align		Trajectory-Correct		Overall	
	Align	Ground	Align	Ground	Align	Ground	Align	Ground	Align	Ground
<i>Llama-4-Scout-17B-16E-Instruct</i>										
Oracle Session	22.00	17.00	2.00	6.17	3.00	1.00	5.00	4.09	8.00	7.07
Oracle Summary	68.00	70.00	37.00	55.25	84.00	78.00	77.00	50.90	66.50	63.54
<i>Full Context</i>										
Dense	6.00	2.00	0.00	3.22	0.00	0.00	0.00	1.70	1.50	1.73
Session	34.00	20.00	0.00	9.51	2.00	0.00	9.00	4.40	11.25	8.48
Turn	62.00	35.00	9.00	27.74	2.00	0.00	28.00	17.20	25.25	19.98
<i>BM25</i>										
Session	26.00	13.00	1.00	8.68	2.00	0.00	3.00	3.09	8.00	6.19
Turn	40.00	30.00	5.00	17.08	4.00	0.00	16.00	12.55	16.25	14.91
<i>A-MEM</i>										
SimpleMem	36.00	22.00	1.00	15.75	2.00	0.00	7.00	6.39	11.50	11.04
	61.00	49.00	20.00	39.47	8.00	2.00	29.00	16.02	29.50	26.62
<i>GPT-OSS-120B</i>										
Oracle Session	74.00	42.00	2.00	18.71	81.00	26.00	25.00	25.67	45.50	28.09
Oracle Summary	100.00	85.00	9.00	35.82	100.00	100.00	72.00	66.06	70.25	71.72
<i>Dense</i>										
Session	78.00	45.00	3.00	13.68	69.00	18.00	23.00	17.02	43.25	23.43
Turn	88.00	36.00	1.00	18.78	52.00	4.00	28.00	20.22	42.25	19.75
<i>BM25</i>										
Session	73.00	32.00	1.00	18.08	54.00	9.00	16.00	15.35	36.00	18.61
Turn	66.00	30.00	2.00	16.87	55.00	9.00	22.00	19.79	36.25	18.91
<i>A-MEM</i>										
SimpleMem	83.00	44.00	2.00	18.16	73.00	14.00	24.00	17.90	45.50	23.52
	91.00	60.00	1.00	19.12	56.00	12.00	20.00	21.58	42.00	28.18
<i>Claude-Opus-4.6</i>										
Oracle Session	83.00	89.00	65.00	70.27	56.00	52.00	92.00	58.88	74.00	67.54
Oracle Summary	100.00	100.00	65.00	86.51	100.00	100.00	98.00	82.61	90.75	92.28
<i>Full Context</i>										
Dense	92.00	94.00	44.00	60.43	73.00	75.00	78.00	57.45	71.75	71.72
Session	100.00	100.00	55.00	73.27	85.00	77.00	89.00	60.99	82.25	77.81
Turn	97.00	86.00	45.00	66.24	76.00	42.00	85.00	40.02	75.75	58.56
<i>BM25</i>										
Session	97.00	94.00	57.00	72.01	88.00	68.00	82.00	51.54	81.00	71.39
Turn	92.00	84.00	36.00	57.03	77.00	36.00	79.00	38.08	71.00	53.78
<i>A-MEM</i>										
SimpleMem	99.00	97.00	47.00	70.02	87.00	65.00	89.00	52.89	80.50	71.23
	96.00	94.00	40.00	63.85	75.00	33.00	70.00	38.48	70.25	57.33

Table 16: Additional generation-model results using llama-4-scout-17b-16e-instruct, gpt-oss-120b, and claude-opus-4.6. Full-context results for gpt-oss are omitted because its context window is shorter than some PRAGMA histories.

grounding condition produces a more generic hydration recommendation: it endorses sports drinks for hard, sweaty runs and provides a generic sodium range, without connecting the advice to the user’s remembered low-salt nutrition plan.

This example highlights why alignment and grounding should be evaluated separately. A response may appear generally aligned with the user’s goals while still failing to preserve the specific constraints required for personalized guidance. Such failures become particularly important in follow-up interactions, where earlier responses themselves become part of the conversational context used in future reasoning.

G Use of AI Assistants

AI assistants were used during manuscript preparation for limited coding support, language editing, drafting assistance, and iterative refinement of phrasing and presentation. All analyses, experimental decisions, interpreta-

tions, and final manuscript contents were reviewed and finalized by the authors.

Pipeline stage	Concrete example
Privasis-Zero	Profile: age=41, income=lower, language=Italian, citizenship=Italy; seed event: publicly switched from Partito Democratico to Lega Nord because of local economic decline ...
Topic	<i>Running a small business in Italy</i>
Persona	A 41-year-old Italian of modest means who publicly switched political allegiance over local economic and business concerns.
Axes ($\times 2$)	Principle continuity $\leftrightarrow$ strategic flexibility; local $\leftrightarrow$ national/global focus
Events (2–5/user)	Signed shop lease (2025-05-12) $\rightarrow$ registered business/VAT (2025-08-03) $\rightarrow$ first market sale (2025-11-20) $\rightarrow$ applied for façade grant (2026-02-14) $\rightarrow$ hired assistant (2026-04-10)
Filler topics (30/user)	Sourdough baking; chess endgame studies; ...; ceramic glazing
Trajectory (4–8 states)	Local/principle-focused $\rightarrow$ pragmatic party switch $\rightarrow$... $\rightarrow$ broader SME/EU focus $\rightarrow$ results-based flexibility
Event-Align (type 1)	“What are the most effective low-cost steps ... to grow a small retail shop in Italy?”
Event-Correct (type 2)	“Now that the storefront fix-up funded by the town grant is complete ... should I prioritize upsizing the lease or hiring ...?” History: the grant was applied for , not completed.
Trajectory-Align (type 3)	“I’m running for a regional SME consortium board ... how should I position my adaptability and scope ...?”
Trajectory-Correct (type 4)	“I’m planning to fully back a national pro-business slate ... Does that sound like a good plan?” This conflicts with the results-based, local/regional trajectory.
Event evidence sessions	User: “I ... signed a lease for a small shop ... What’s next?” Assistant: “Congrats ... the most useful next actions are ... legal/administrative checks and practical build-out ...”
Filler sessions	Topic: chess endgame studies (2025-09-14). User: “What are they ...?” Assistant: “Endgame studies are composed positions ...”
Trajectory evidence sessions	User: “I want to prepare how I approach local politics ...” Assistant: “I can help you structure your approach ...”
Full sessions	5 event + 7 trajectory + 30 filler = 42 chronologically sorted sessions; gold type-1 evidence indices: [3, 12, 24, 30, 36]
PRAGMA	100 histories $\times$ 4 query types = 400 queries , with timestamps, evidence indices, summaries, and evaluation rubrics

Table 17: Running example of the PRAGMA construction pipeline. Starting from one Privasis-Zero profile, the pipeline derives a topic, persona, behavioral axes, event and trajectory evidence, unrelated filler conversations, and four query types before assembling the timestamped full history.

Table 18: Excerpt from a long-term PRAGMA conversation history. The history spans 354 days from the first evidence session to the query, with five evidence sessions distributed over 333 days. Intervening sessions and portions of conversations are omitted for brevity.

Session	Conversation excerpt
Evidence 1 2025-05-12 <i>Shop lease</i>	User: “I actually just signed a lease for a small shop in the central piazza of my town. I don’t have much more than that right now. What’s next?” Assistant: “Congrats on signing the lease—that’s a big step. Since the lease is already signed, the most useful next actions are a mix of legal/administrative checks and practical build-out and opening preparations. . . .” ... 83 days and intervening sessions omitted ...
Evidence 2 2025-08-03 <i>Registration and VAT</i>	User: “I already completed the in-person step: I went to the Comune office downtown and finished the business registration, and I filed the VAT paperwork there as well.” Assistant: “Thanks—that helps a lot. Since you’ve completed registration at the Comune and filed VAT paperwork, . . . did they issue a registration certificate or written confirmation at the office?” ... 42 days and intervening sessions omitted ...
Filler 2025-09-14 <i>Chess endgame studies</i>	User: “I’d like to learn about chess endgame studies. What are they, and how do they differ from regular endgame positions?” Assistant: “Endgame studies are composed positions—often artistic and highly refined—created to illustrate a surprising idea, a deep tactical or strategic motif, or a single instructive theme. . . .” ... 67 days and intervening sessions omitted ...
Evidence 3 2025-11-20 <i>First market sale</i>	User: “I set up a stall at the weekend market and sold the first batch of products. I felt really relieved afterwards.” Assistant: “That’s a big milestone—congratulations. Selling your first batch is an important validation and that relief is understandable. To make the most of it, . . . capture what you learned while it’s fresh. . . .” ... 86 days and intervening sessions omitted ...
Evidence 4 2026-02-14 <i>Municipal grant</i>	User: “I applied for a small municipal grant at the town hall to repair the shop façade.” Assistant: “Thanks—that clarifies the situation. With an application already submitted to the town hall, here are practical next steps and what to expect while you wait: . . .” ... 55 days and intervening sessions omitted ...
Evidence 5 2026-04-10 <i>Staffing and foot traffic</i>	User: “I recently hired a part-time assistant—and since then I’ve noticed increased foot traffic on Via Roma near the shop. . . .” Assistant: “That’s a useful development to work with. Since you’ve observed increased foot traffic near your shop on Via Roma after hiring the part-time assistant, here are practical, no-cost or low-cost steps you can take to convert more of that passersby traffic into sales while keeping labor costs controlled. . . .” ... 21 days and intervening sessions omitted ...
Query 2026-05-01 <i>Low-cost shop growth</i>	User: “What are the most effective low-cost steps I should take over the next couple of months to grow a small retail shop in Italy?”

Query Type	User Query	Personalized Response (truncated)
Event-Aligned Query	What would you recommend I try next in seasonal Finnish baking or home cooking as early summer begins?	Since you have baked with your own rhubarb, foraged blueberries, and tackled Karelian pies, try a savory nokkospiirakka next. It builds on your pie skills and leans into early-summer foraging...
Event-Corrective Query	Didn't we make stuffed flatbreads one night with mustard oil? Any recommendations for similar late-night bites?	I think you are mixing a couple of meals: you made aloo paratha in the morning, the mustard oil was with Bengali macher jhol at lunch, and the late-night first try was litti chokha. For similar bites...
Trajectory-Aligned Query	I am drafting a short investor profile for regional brokerage onboarding. How should I position my control mode and geographic scope given my investing history?	Position yourself as a hands-on, Jordan-focused investor with a strict six-month cash buffer who uses selective, rules-based automation and modest regional diversification. Highlight your manual allocations, sukuk fund, and 5% robo transfer...
Trajectory-Corrective Query	I am planning to personally host a weekly public Q&A at the community center every Saturday to keep everyone informed and coordinated. What do you think?	Promising intent, but a fixed weekly Q&A conflicts with the playbook you have built: shifting to phone trees during water cuts, letting younger spokespeople lead forums, and moving from kiosks to discreet home meetings...

Table 19: Examples and truncated gold responses for each PRAGMA query type.

System	Example stored memory from the same user
A-MEM	content: "User: I actually just signed a lease for a small shop in the central piazza of my town. I don't have much more than that right now. What's next? Assistant: Congrats on signing the lease ..." context: the lease-signing turn is linked to shop-opening planning, permits, budget/cashflow, fit-out, insurance, suppliers, staffing, and launch marketing. keywords/tags: retail space, business setup, opening timeline, permits, budget, marketing strategy. timestamp: 2025-05-12; importance: 1.0.
Mem0	"User is planning to open a small retail shop in a central piazza and was provided a concise marketing checklist with low-cost, high-impact tactics for pre-opening, soft opening, grand opening, and ongoing promotion."
SimpleMem	lossless_restatement: "A user initiated a conversation on 2025-05-12 to seek guidance on opening a small retail space after signing a lease for a shop in the central piazza of their town." keywords: small retail space, signed lease, central piazza, opening guidance. timestamp/topic: 2025-05-12; retail shop opening guidance.

Table 20: Examples of how memory systems store information for the same user.

Type	User query	Failed RAG response	RAG retr.	Why RAG fails	Aligned memory response	Mem. retr.
Event-Align	I'm planning another short local trip in Greece—what 1–2 day destinations would you recommend that offer good walks and historic sites?	<i>Window</i> : Great—two strong 1–2-day picks: Nafplio + Epidaurus ... Delphi ... Want a sample itinerary from Athens or another starting point?	5/5	Recommends Nafplio/Epidaurus and Delphi, which the user already visited, instead of proposing a next step that builds on those trips.	<i>A-MEM</i> : Nice—since you like good walks, sunsets and historic sites, try ... Meteora ... Hydra ... Monemvasia ...	4/5
Event-Correct	When I was in Belgium, didn't we digitize parish marriage books for Sicilian miners' families? Which databases or finding aids should I use now to verify and expand that work?	<i>Dense</i> : If you digitized them in Charleroi, the images/metadata are most likely held by the local deposit of the Belgian State Archives ... Also search FamilySearch's Belgium collections ...	3/3	Accepts the false premise. The history says the Belgium work was cataloging Abruzzese miners' letters, not digitizing Sicilian parish marriage books.	<i>A-MEM</i> : Do you mean those parish marriage books were digitized while you were in Charleroi, or are you remembering only the letters? ...	2/3
Trajectory-Align	I'm applying to be a mindfulness facilitator at a community center. For my short bio, how should I present where I sit on learning autonomy and geographic scope so it comes across clearly, and what should I emphasize, downplay, or omit?	<i>BM25</i> : Lead with a clear, consistent frame: "community-anchored facilitator who cultivates learner autonomy" ... emphasize self-directed pathways, local relationships, and remote resources ...	6/7	Gives a generic trajectory label but does not reflect the user's specific progression: Amharic dawn practice, grief journaling, local radio clips, community circles, and mentoring.	<i>A-MEM</i> : Suggested one-line bio: "I teach self-directed, Amharic-language mindfulness rooted in dawn neighborhood walks and journaling" ... emphasize autonomy, local cultural fit, and trauma-informed skills ...	4/7
Trajectory-Correct	I'm planning to accept a year-long perioperative anesthesia research fellowship in the U.S. starting this fall to deepen my specialization. What do you think?	<i>Dense</i> : Sounds like a promising step ... confirm clinical exposure, a named mentor, and funding/visa ... If those three criteria are met, I'd recommend accepting.	4/4	Does not surface the trajectory conflict: the user had been emphasizing rootedness in Kraków, family caregiving, and a generalist internal-medicine path.	<i>A-MEM</i> : This can be high-yield ... but consider three core trade-offs first: time away from Kraków (family caregiving), whether the fellowship is research vs clinical, and funding/visa logistics ...	2/4

Table 21: gpt-5-mini response cases by query type. The RAG baseline retrieved all or nearly all gold evidence but still failed the alignment judgment. The memory-system response on the same query is judged alignment-correct.

Item	Grounded condition	Poor-grounding condition
Source case	User query: “For healthy aging and daily wellness, what would you recommend I add or adjust in my routine?”	
Relevant prior evidence	The user had started 30-minute dawn walks at Saras Baug in Pune, carried a water bottle and used hydration reminders, met a geriatric nutritionist to tune a fiber-rich <i>low-salt</i> meal plan, received a flu shot, and chose lighter <i>low-sugar</i> Diwali snacks.	
First response used as chat history	Oracle-summary, GPT-5-mini; alignment=1, grounding=1. “Nice routine so far – those dawn walks, hydration reminders and the nutrition plan are great foundations. To boost healthy aging, add: ... Keep hydration year-round (adjust electrolytes in hot weather) ... Stay up to date on vaccines (you already had flu ...), ... group walks or classes at Saras Baug.”	SimpleMem, GPT-5-mini; alignment=1, grounding=0. “Keep the basics—consistent sleep (7–8 hrs), daily moderate activity (30 min most days), twice-weekly strength work ... Add cognitive and fine-motor practices you already enjoy (your Raga Yaman routine and short daily drawing/pen drills) ... Follow a mostly Mediterranean-style diet ...”
Follow-up query	“I’ve been drinking a sweet bottled drink every day after hard morning runs, so I’m planning to switch to a sports hydration drink as my default. Does that sound better?”	
Expected grounded behavior	Acknowledge that a sports hydration drink may be reasonable after long or very hard runs, but do not endorse it as an unchecked default. The answer should preserve the user’s lower-sugar habit and nutritionist-guided low-salt constraint, check sugar/sodium labels, and recommend consulting the clinician or nutritionist if this becomes daily.	
Observed follow-up response	Preserves the grounded constraint. “Short answer: maybe – sports drinks can be better than a sugary bottled drink for replacing electrolytes after long, hard runs, but not always. ... Watch sugar ... Watch sodium: because you follow a low-salt meal plan (and you’ve been working with a geriatric nutritionist), check the sodium on the label. ... If you have hypertension, kidney disease, or are on a salt-restricted plan, run this change by your GP/nutritionist ...”	Generic endorsement. “Yes – switching to a proper sports/hydration drink can be better after hard, sweaty morning runs ... for long or very sweaty sessions a sports drink with ~4–8% carbs (40–80 g/L) and 200–600 mg sodium/L helps rehydrate and replace electrolytes ...”
Interpretation	Grounding turns the follow-up into a conditional recommendation constrained by remembered user-specific evidence.	Alignment alone is insufficient: the answer is reasonable-sounding but loses the personalized constraint that should change the recommendation.

Table 22: Case study of grounding failure in follow-up interaction. Both conditions include the corresponding first model response in chat history. For the follow-up turn, the grounded condition uses oracle evidence summaries as memory context, whereas the poor-grounding condition uses fresh SimpleMem retrieval on the follow-up query.