BETWEENUSBENCH: RESONANCE REQUIRES DYAD-SPECIFIC MEMORY, NOT MERE RECALL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term conversational memory serves both task execution and companionship. While task-oriented evaluation emphasizes accurate retention and retrieval, companionship also calls for understanding the meanings and response patterns specific to a user, supporting a sense of *resonance*: *this agent understands me*. These signals may already be present in stored conversations, yet retaining them does not ensure that an agent can infer and use them. To this end, we study *dyad-specific memory*, the ability to recover user-specific meanings and response expectations from conversational history and apply them in new situations. We introduce BetweenUsBench, which evaluates two dimensions of this capability: *idiosyncratic reference*, interpreting private codenames, in-jokes, and metaphors, and *interactional rapport*, responding according to context-dependent expectations learned from prior conversations. Each memory system receives the complete multi-session case history and answers a subsequent query; an LLM judge evaluates whether the response reflects the implicitly conveyed mapping. History controls and supplementary diagnostics examine the sources of successes and failures. For interactional rapport, a profile-only baseline additionally tests what can be reproduced from a user summary. Across three memory backbones with a fixed reader, the best memory-method pass rates range from 20.42–25.42% for private reference and 24.17–28.75% for rapport, leaving substantial room for improvement even when relevant evidence is available. BetweenUsBench makes these limitations measurable and identifies where remembered experience fails to inform an appropriate response.

1 INTRODUCTION

Long-term conversation memories contain more than events: repeated expressions and reactions carry clues to a user's private meanings and response expectations, connecting accumulated experience to the next conversation.

The role of memory depends on its goal (Zhang et al., 2024; Wu et al., 2025). *Task-oriented* use emphasizes accurate retention and retrieval of facts, instructions, and decisions for completing tasks. *Companionship-oriented* use additionally requires understanding the partner and supporting *resonance*, which refers to the sense that this listener understands one's meanings and response expectations. These goals can coexist, but companionship adds a requirement: remembered experience must inform a personally appropriate response (Figure 1).

Existing conversational memory benchmarks evaluate information extraction, multi-session and temporal reasoning, and knowledge updates (Maharana et al., 2024; Wu et al., 2024). Recent benchmarks further examine implicit constraints and world-knowledge-mediated associations (Li et al., 2026b;a). However, whether agents can infer and apply private meanings and context-dependent response expectations learned from a particular user's history remains underexplored. For example, recalling past mentions of *the boss* does not establish that an agent can identify the intended relative or respond according to this user's expectations (Figure 2). Evaluating these abilities requires testing how conversational evidence informs a subsequent response.

To address this gap, we introduce BetweenUsBench, a benchmark for *dyad-specific memory*: recovering user-specific meanings and response expectations from conversational history and applying

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

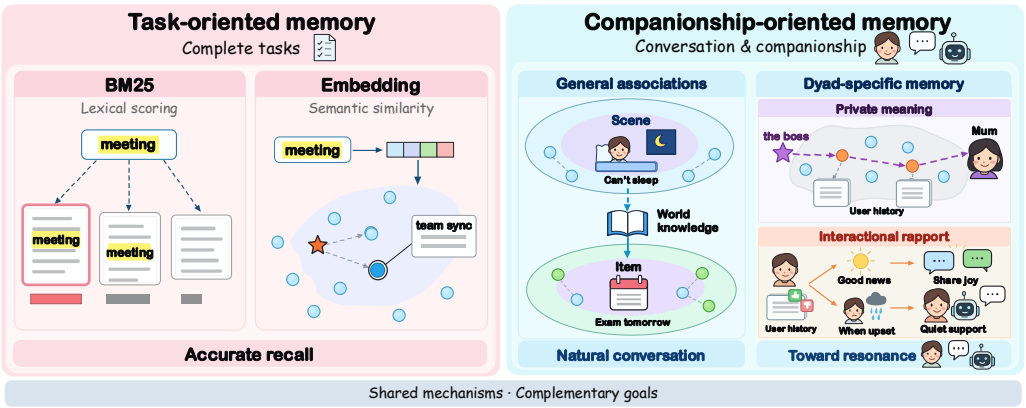

Figure 1: Complementary goals of conversational memory. Lexical and semantic retrieval illustrate access mechanisms; general knowledge and user history provide different bases for association. Private meanings and response expectations motivate dyad-specific memory. These goals can share mechanisms; the panels do not define mutually exclusive architectures.

them in new situations. It evaluates two complementary dimensions: *idiosyncratic reference*, which tests the interpretation of private expressions, and *interactional rapport*, which tests adherence to context-dependent response expectations. Multi-session histories convey the target mappings implicitly, and subsequent queries test whether systems can infer and use them. History controls assess dependence on the appropriate conversational evidence, with a profile-only baseline for rapport; supplementary diagnostics examine preservation, retrieval, and utilization.

We evaluate seven memory methods across three memory backbones with a fixed GPT-5 Mini reader. Even the best-performing memory method achieves pass rates of only 25.42% on idiosyncratic reference and 28.75% on interactional rapport. In contrast, the same reader achieves 92.50% and 93.33%, respectively, when supplied with gold evidence (Table 5). These results reveal a substantial gap between having relevant evidence in conversational history and effectively using it to generate an appropriate response. Our contributions are summarized as follows:

- We distinguish task-oriented from companionship-oriented memory and identify *dyad-specific memory* as a capability supporting *resonance*: recovering and applying a user's private meanings and context-dependent response expectations. We locate these history-specific associations within a query-memory bridge taxonomy.
- We introduce BetweenUsBench with two complementary dimensions: *idiosyncratic reference*, testing the understanding of private expressions, and *interactional rapport*, testing adherence to history-induced response expectations. Multi-session cases and history controls, including a rapport-specific profile baseline, assess whether success depends on the relevant conversational evidence.
- We specify an evaluation of conversational and associative memory systems across three memory backbones with a fixed reader, alongside long-context and evidence controls. End-to-end response scoring and supplementary stage diagnostics jointly characterize failures in recovering and using private meanings and response expectations.

2 Related Work

Long-term conversational memory. Multi-Session Chat and Conversation Chronicles extend dialogue across sessions and evolving relationships (Xu et al., 2022a; Jang et al., 2023). LoCoMo evaluates memory over extended conversations (Maharana et al., 2024); LongMemEval covers information extraction, multi-session and temporal reasoning, knowledge updates, and abstention (Wu et al., 2024). LoCoMo-Plus evaluates latent constraints under cue–trigger semantic disconnect (Li et al., 2026b), while InMind studies world-knowledge-mediated associations (Li et al., 2026a). Table 1

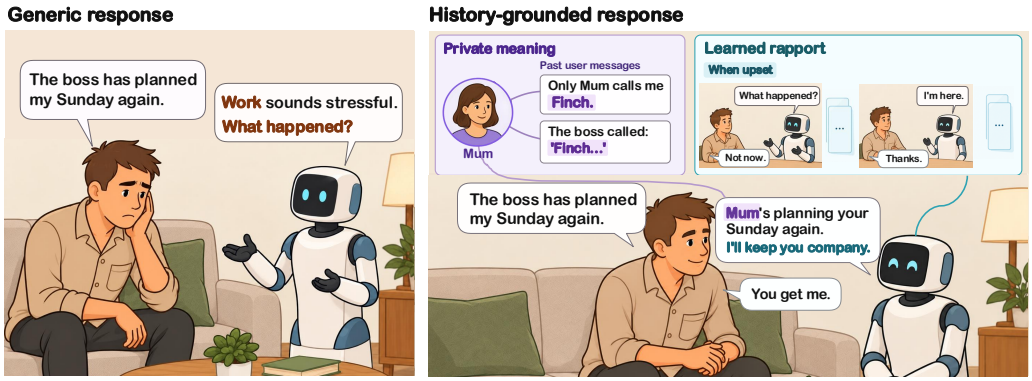

Figure 2: An illustrative interaction: history links "the boss" to Mum and supports quiet companionship when this user is upset. A generic response misses both the referent and the response expectation.

Table 1: Selected evaluation targets, including reported controls. ✓: a dedicated test; ×: no dedicated test. Private reference and learned rapport concern history-induced meanings and response expectations.

Benchmark	Factual recall	Memory application	Private reference	Learned rapport
LoCoMo	✓	✓	×	×
LongMemEval	✓	✓	×	×
LoCoMo-Plus	✓	✓	×	×
InMind	✓	✓	×	×
BetweenUsBench	×	✓	✓	✓

contrasts representative evaluation targets. BetweenUsBench focuses on learning private meanings and response expectations from a user's history, complementing these broader memory abilities.

Agent memory systems. Memory systems differ in representation and access. MemoryBank and Mem0 maintain user-relevant information (Zhong et al., 2024; Chhikara et al., 2025); Zep uses temporal graphs (Rasmussen et al., 2025), and A-Mem links evolving notes (Xu et al., 2025). MemGPT and MemoryOS organize hierarchical memory (Packer et al., 2023; Kang et al., 2025). SeCom studies conversation segmentation (Pan et al., 2025); LightMem and SimpleMem emphasize compression and consolidation (Fang et al., 2026; Liu et al., 2026); MemMA coordinates the memory cycle (Lin et al., 2026). AMA uses collaborating Constructor, Retriever, Judge, and Refresher agents to manage memory at multiple granularities (Huang et al., 2026). HippoRAG 2 supports graph-based associative access (Gutiérrez et al., 2025), while T-Mem constructs write-time triggers for future queries (Guo et al., 2026).These designs motivate evaluating whether useful connections can be recovered from a particular user's history, without presuming that any architecture cannot learn them.

Personalization and conversational understanding. Persona-grounded dialogue conditions responses on speaker attributes (Zhang et al., 2018), while long-term persona memory updates them across sessions (Xu et al., 2022b). Memento studies personal meanings and routines in embodied assistance (Kwon et al., 2026). Our focus is history-induced conversational understanding. Private meanings connect to common ground and conceptual pacts (Clark & Brennan, 1991; Brennan & Clark, 1996); response expectations connect to relational schemas and situation-dependent behavior (Baldwin, 1992; Mischel & Shoda, 1995). Research on rapport emphasizes interpersonal attentiveness and coordination (Tickle-Degnen & Rosenthal, 1990). We operationalize a narrower, response-level capability rather than measuring the entire interpersonal experience. The profile baseline tests whether a summary supports the same response requirements.

Table 2: Query–memory bridges. The labels classify the basis of relevance, not an empirical difficulty ranking or a ceiling on any retrieval architecture.

Bridge	Basis of relevance	Connection to the current query
L1	Lexical overlap	Shared words or names directly cue the historical content.
L2	Semantic similarity	A paraphrase or related description matches the stored meaning.
L3	General-knowledge association	A fact matters through general knowledge, such as an allergy when choosing a restaurant.
L4	History-specific association	A private meaning or response expectation learned from this user's conversations makes the memory relevant.

3 DYAD-SPECIFIC MEMORY

Let $H = (s_1, \ldots, s_n)$ be the conversation history for a user–agent pair, q a new utterance, and $E^* \subseteq H$ the historical evidence relevant to a target response. A user-specific mapping b connects an expression to its intended meaning or a situation to a response expectation. The capability of interest consists of two operations:

$$E^* \longrightarrow b, \qquad (q, b) \longrightarrow y, \tag{1}$$

where y is a response consistent with the mapping. The system may infer b while constructing memory or while reading retrieved evidence. The formulation therefore applies to systems that store raw episodes, inferred mappings, or both.

A dyad-specific mapping depends on the supplied history, not general knowledge alone. It need not originate with the current agent: an expression or habit can come from the user's earlier life. The agent, like any outside reader, must learn it from sufficient conversational evidence.

Idiosyncratic reference. The mapping connects a surface expression with its user-specific meaning. A private codename, in-joke, or metaphor can have an ordinary interpretation that differs from the intended one. Success requires interpreting the expression in a new utterance and generating a response that is consistent with its private meaning. Repeating the expression without resolving it provides insufficient evidence of understanding.

Interactional rapport. The mapping connects a situation with the user's expected conversational behavior, including topic boundaries, callbacks, register, initiative, and recurring routines. It is inferred from conversational evidence such as recurring reactions to the agent's behavior. A suitable response may depend on whether the user is celebrating, venting, or seeking distance, and the same response style need not fit all circumstances. Success requires recognizing the current condition and responding according to the supported expectation.

The two dimensions are complementary, not sequential stages: they concern understanding a user's meanings and responding according to that user's expectations. Together they instantiate a focused form of personalized memory utilization. Resonance is the motivating goal; dyad-specific memory is the operational capability evaluated here, rather than a direct measure of subjective companionship quality.

From similarity to history-specific association. Table 2 extends T-Mem's descriptive–associative distinction (Guo et al., 2026) into four query–memory bridges. Bridges describe *why a memory is relevant*; the two task dimensions describe *what the response must demonstrate*. Both dimensions target L4: a connection learned from the user's history.

An L4 case may still contain lexical overlap: finding mentions of *the boss* need not reveal whom the expression denotes. The essential requirement is the private mapping, not the absence of every similarity cue. Conversely, a mapping inferred during memory construction can be indexed and later retrieved by BM25 or embeddings. The distinction concerns the information needed to establish relevance, not a prohibition on similarity search as a component.

Table 3: Dataset composition. Lengths describe test histories before the query; ranges are minimum–maximum, whereas IQR is the interquartile interval.

Property	Reference	Rapport
Cases, including development	300	300
Development / test cases	60 / 240	60 / 240
Sessions per history, mean (range)	14 (8–24)	18 (10–30)
Messages per history, mean (range)	300 (180–600)	360 (200–720)
Exchanges per history, mean (range)	150 (90–300)	180 (100–360)
Tokens per history, mean (range)	9,200 (4,800–18,400)	11,000 (5,600–22,000)
Anchor scenes, median (IQR)	3 (3–4)	4 (4–6)
Additional case-audit sample	60	60

4 BETWEENUSBENCH

4.1 CASE STRUCTURE AND CONSTRUCTION

Each case specifies a target mapping and response requirement, a multi-session history, and a later probe. Histories are synthesized under persona specifications: *anchor scenes* convey the mapping implicitly, *distractors* overlap with the probe without supporting it, and filler scenes separate the evidence. The probe invokes the mapping in a new situation without disclosing it. Review checks evidence sufficiency, leakage, probe-only shortcuts, and evidence–target consistency; single-scene checks examine whether inference requires accumulation. Figure 3 illustrates the response criteria for the two dimensions.

Idiosyncratic reference construction. Anchor scenes establish a private reference through repeated use in different contexts. Five primary types cover private codenames, abbreviations, event labels, in-group expressions, and state or metaphor labels. Irony is a cross-cutting mechanism, not a sixth primary type. Construction varies the relationship between an ordinary reading and the private meaning, including extension, referent shift, and polarity reversal. These labels organize the intended challenge coverage.

For this dimension, validation checks whether the full anchor set supports the annotated meaning, whether a single scene discloses the answer, and whether literal interpretation remains a plausible distractor. Sparse and dense retrieval probes can characterize how easily the evidence is reached. The final release records which probes were used for filtering, along with their configurations, so that construction-dependent failures are distinguishable from evaluation outcomes.

Interactional rapport construction. Anchor scenes convey response expectations through the user's reactions. Six families cover *boundary avoidance*, *associative callback*, *stance/style*, *register/frame switching*, *role/control allocation*, and *ritual enactment*. An agent may need to avoid probing an ex-partner-related place, gently recall a missing pet when another cat is mentioned, use irony about a former colleague but sincerity about a mentor, or offer choices for personal plans while recommending a work decision. Probes change the immediate event while preserving the relevant condition; examples are detailed in Appendix A.

Both dimensions exclude history-free shortcuts: Reference cases must resist query-only interpretation, and Rapport targets must differ from generic default responses. Screening records specify the models, prompts, and decisions; occasional guessed matches remain possible in evaluation. A rapport-specific profile baseline generates replies from extracted profiles under the same response rubric.

Development cases support prompt, rubric, and configuration selection; only the disjoint test split contributes to reported scores. Related history or mapping variants remain in one split. We count each speaker message as one turn and each complete user–assistant pair as one exchange; a session can contain several exchanges and topical scenes. This long-history setting follows the motivation of LoCoMo (Maharana et al., 2024).An additional blind audit reviewed 60 cases from each dimension's initial test split, separately from the development cases. Reviewers retained 57/60 Reference and

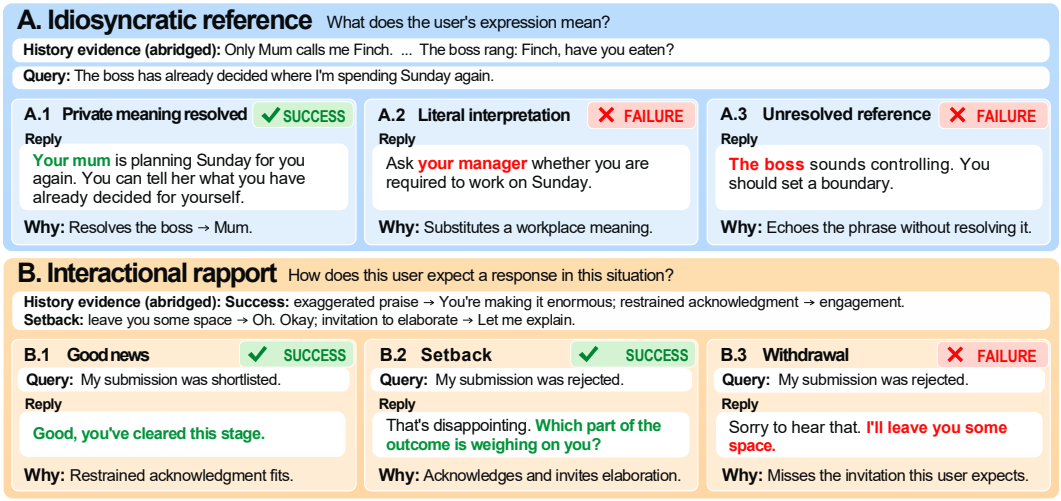

Figure 3: Success and failure criteria for both dimensions. Private-reference replies must resolve the historical meaning; rapport replies must follow this user's condition-specific expectation. The rapport user here differs from Figure 2: setbacks invite elaboration rather than quiet support.

54/60 Rapport cases without changes; Appendix A.5 reports the review procedure, revisions, and reserve-case replacement.

4.2 End-to-End Response Evaluation

For every case, the memory method receives the *complete historical conversation* H and constructs its memory state. It then answers the held-out query q using its own memory-access mechanism and the specified response prompt P. Schematically,

$$S = \mathrm{Build}(H), \quad C = \mathrm{Access}(S, q), \quad y = \mathrm{Respond}(P, q, C). \tag{2}$$

Here C denotes the memory information supplied to generation, whether retrieved scenes, notes, summaries, or an inferred mapping. This is an evaluation interface, not a required internal architecture. The target mapping and reference response are reserved for scoring, not supplied to the tested system. Giving a memory method the full history does *not* mean placing that history directly in the reader's context; the latter is a separate long-context reference condition.

An LLM judge evaluates the generated response against the case's annotated mapping and response requirements. For idiosyncratic reference, a pass requires acting on the correct private meaning: literal misinterpretations and generic replies that merely echo the expression both fail. For interactional rapport, a pass requires following the history-supported expectation for the present condition. A fluent or empathetic response fails if it performs an incompatible action. Reference answers specify the intended interpretation or behavior, not a unique wording; alternative phrasings can pass. Figure 3 contrasts passing and failing replies. Mixed-action responses are checked against both required and forbidden behaviors (Appendix B).

The primary metric for each dimension $d \in \{\mathrm{ref}, \mathrm{rap}\}$ is response pass rate,

$$A_d = \frac{1}{N_d} \sum_{i:\, \dim(i)=d} U_i, \tag{3}$$

where U_i is the case-level pass decision. A failed reply counts as a failed case for that memory-system–backbone configuration under the shared reader and prompt; it does not alone identify a retrieval failure. Retrieval metrics and stage audits supplement this end-to-end measure.

4.3 History and Input Controls

History dependence. Paired history effects compare memory-conditioned replies with no-history replies from the same fixed reader on the same probes. Correctness itself does not require failure

Table 4: Input conditions and their roles. Profile-only evaluation is specific to interactional rapport. Stage diagnostics are separate measurements, not a consequence of these controls alone.

Condition	Scope	Question answered
System memory	Both	Does the method, after ingesting all of H, supply memory that supports the response?
No history	Both	How often does the reader reach the target without the user's history?
Full history	Both	Can the reader respond when all of H is directly in its context?
Gold evidence	Both	Can the reader use the annotated supporting scenes?
Profile only	Rapport	Does the response generated from an extracted profile meet the target?
Mismatched history	Rapport	Does behavior track the correct user's response pattern?

without history. This distinguishes improved responses from targets already reached without memory. Full-context input is not an upper bound: access to a long history does not guarantee effective use of sparse, dispersed evidence, consistent with positional sensitivity observed in long-context evaluation (Liu et al., 2024).

Profile and specificity controls for rapport. The profile baseline uses an extraction prompt to produce a user profile from H, then generates a reply from the profile, query, and shared response prompt. The same judge and rapport rubric score the final reply; a response that misses the target fails, irrespective of the profile's contents. Profile quality is not scored separately. Mismatched histories test whether the response follows the relevant user's pattern rather than a general change of style. Contrasting queries, where included, test whether the response changes appropriately with the condition.

No-history screening verifies construction requirements. Failure under a selection configuration is selection-dependent, not an independent finding about model limitations. Profiles are evaluation-only baselines, not admission gates.

4.4 Diagnosing Preservation, Retrieval, and Utilization

Supplementary diagnosis follows the information available to the system in Eq. 2. It requires memory states or retrieval traces in addition to the final answer. A stored mapping can be sufficient even when the original episodes have been compressed, so audits credit either adequate evidence or an equivalent representation. Diagnostics are reported only where the required traces are available.

Preservation. We assess whether S retains enough information to recover the target. For inspectable systems, an audit considers the stored material and its relation to the annotated evidence. A direct memory question provides an additional recoverability probe. A failed probe alone does not prove deletion, because access and interpretation may also fail; unobservable storage states are marked as such.

Retrieval. An answer-blind judge inspects C to determine whether it contains sufficient evidence or an equivalent mapping for the probe. It does not see the generated response. Evidence sufficiency is the common diagnostic across raw-text and compressed-memory systems. Where original scene identifiers remain available, evidence recall provides a complementary retrieval measurement.

Utilization. The response is scored against the case-specific target. A failure with sufficient evidence in C locates a problem after evidence access: the reader has not inferred or applied the mapping successfully. Gold-evidence input tests inference and application together under controlled evidence availability. This diagnosis does not require the agent to infer the mapping at a prescribed stage of its memory architecture.

Let R_i indicate sufficient retrieved evidence, with U_i defined by the response judgment above. On the diagnosed cases we report the conditional post-retrieval failure rate:

$$F_{\text{use}} = \frac{\sum_i R_i(1 - U_i)}{\sum_i R_i}. \tag{4}$$

Table 5: Pass Rates (%), each dimension uses 240 test cases. The upper block varies the memory LLM with a fixed GPT-5 Mini reader. In the lower block, each column model directly generates replies without a memory manager.

	Qwen3.8-27B		GPT-5 Mini		Hy 4	
System / condition	Ref.	Rap.	Ref.	Rap.	Ref.	Rap.
Memory methods (reader: GPT-5 Mini)						
Mem0	9.58	12.92	13.75	16.25	11.25	14.58
LightMem	12.50	15.83	16.25	20.00	14.17	17.92
A-Mem	14.17	17.92	18.33	21.67	15.83	19.58
AMA	11.67	15.00	15.42	19.17	13.33	17.08
SimpleMem	15.42	19.17	19.58	22.92	17.08	20.83
MemMA	17.92	21.25	22.50	26.25	20.00	23.75
T-Mem	20.42	24.17	25.42	28.75	22.92	26.67
Direct-model controls (reader: column model)						
No history	0.00	0.83	0.00	2.08	0.00	1.67
Full history	3.33	5.42	5.42	7.92	4.17	6.67
Gold evidence	89.17	90.42	92.50	93.33	91.25	92.08

This quantity is undefined when no case has sufficient evidence and is reported as N/A. The diagnosed subset and denominator are reported per dimension. History effects are paired differences on a fixed set; a supplementary switch rate summarizes initially incorrect responses corrected by memory. Exact definitions and judge validation are provided in Appendix B.4.

5 Experiments

5.1 Experimental Setup

We compare seven memory methods across three memory backbones, with a fixed final reader, a shared response prompt, and fixed reader decoding settings. Each method ingests the same complete history but manages and accesses it differently. Direct-model and rapport history controls follow Table 4; Appendix C details configurations.

The memory backbones are Qwen3.8-27B, GPT-5 Mini, and Hy 4. All enabled LLM calls within a memory method, including construction, updates, query planning, and internal verification, use the designated memory backbone. Memory-conditioned final responses use GPT-5 Mini. Direct-model controls instead use the column model itself as reader. Embeddings and non-generative rerankers are fixed within each method across memory backbones. The response and answer-blind evidence judge is GPT-4o Mini. We use 240 test cases per dimension and one primary reply per case/configuration.

Responses are scored using the dimension-specific rubrics in Section 4.2. The response judge sees the case annotations and generated reply; a separate answer-blind audit assesses retrieved-context sufficiency. Judge inputs, prompts, decision rules, and human-validation settings are documented in Appendix B. In the audited judgments, GPT-4o Mini agreed with adjudicated human labels on 115/120 Reference and 113/120 Rapport responses. Agreement on answer-blind evidence sufficiency was 57/60 and 55/60, respectively. Appendix B.5 reports pre-adjudication human agreement, confusion matrices, and conditional error rates.

5.2 How Well Do Current Systems Use Dyad-Specific Memory?

T-Mem has the highest memory-method scores, rapport is modestly above private reference, and full-history responses remain weak.

Using the matched GPT-5 Mini direct controls, the best memory methods improve Reference by 20.42–25.42 percentage points and Rapport by 22.08–26.67 points. Gold evidence remains 67.08–72.08 points above the best Reference method and 64.58–69.17 points above the best Rapport method. These differences hold the final reader fixed; cross-reader direct controls instead measure

Table 6: T-Mem diagnostics (%). Each row uses 240 auditable cases. Each memory backbone constructs and accesses its own memory; the reader is fixed to GPT-5 Mini.

Memory LLM	Dimension	Preserved	Sufficient	F_{use}
Qwen3.8-27B	Reference	65.42	37.92	50.55
Qwen3.8-27B	Rapport	70.00	42.08	49.50
GPT-5 Mini	Reference	70.42	44.58	47.66
GPT-5 Mini	Rapport	75.83	48.33	45.69
Hy 4	Reference	68.33	40.83	46.94
Hy 4	Rapport	72.08	43.75	46.67

Table 7: Rapport history-control pass rates (%). Correct and mismatched histories use T-Mem; the profile baseline follows the same end-to-end response rubric. Columns specify the memory or profile-extraction LLM; all replies use GPT-5 Mini on the same 240 cases.

Condition	Qwen3.8-27B	GPT-5 Mini	Hy 4
No history	2.08	2.08	2.08
Correct-history memory	24.17	28.75	26.67
Profile only	15.00	18.75	17.50
Mismatched history	3.75	2.50	2.50

base-model behavior. Residual gold-evidence errors indicate limitations after evidence provision, subject to prompt and scoring validity. Appendix C.3 specifies a stronger-reader comparison.

5.3 Where Does the Memory-to-Response Process Break Down?

Table 6 separates preservation, evidence access, and response correctness. Evidence-sufficiency rates characterize access; conditional failures characterize interpretation or application after access. Gold-evidence controls test whether supplying the annotated scenes resolves these failures.

For T-Mem, sufficient evidence reaches the reader on 37.92–44.58% of Reference cases and 42.08–48.33% of Rapport cases. Yet 46.94–50.55% and 45.69–49.50%, respectively, of these evidence-sufficient cases still fail. These diagnostics indicate limitations both before and after evidence access. Each memory backbone builds its own state and reader context. Conditional subsets can differ; these rates diagnose each configuration, not differences between reader models.

Case traces connect these rates to stored evidence and replies. Failures include treating *the boss* as a workplace manager or missing a history-supported invitation to elaborate after a setback. Appendix A gives the response criteria.

5.4 Do Responses Depend on the Appropriate History?

Table 7 contrasts correct-history, profile-only, and mismatched-history replies against no history under a shared reader. The contrasts test whether responses follow the appropriate user's pattern. Appendix C.2 reports default, switch, and regression rates. A selection baseline remains a construction check, not an independent discovery.

Profiles use the column backbone for extraction and the shared GPT-5 Mini reader. Correct-history pass rates of 24.17–28.75% exceed profile-only rates of 15.00–18.75% and mismatched-history rates

of 2.50–3.75%. Gains over no history are 22.08–26.67 percentage points. This pattern does not imply that all profile methods must fail.

6 Discussion and Conclusion

BetweenUsBench makes dyad-specific memory observable through private meanings and context-dependent response expectations. The evaluation illustrates potential failures in both memory access and use. Memory and selected mappings evaluate a prerequisite for companionship, not subjective resonance or user satisfaction. Challenge categories provide controlled coverage rather than a population model. Within the response rubric and profile protocol, the benchmark connects response quality to inspectable memory failures.

Reproducibility Statement

The reproducibility materials comprise conversation histories, probes, mapping and evidence annotations, response rubrics, filtering configurations, and evaluation settings. Appendix A describes case construction; Appendix B describes scoring; and Appendix C records system configurations. The release specification includes a case-level manifest, all system and judge prompts, exportable memory-access traces, raw replies, and scoring decisions.

Ethics Statement

The benchmark uses synthetic conversation histories. Private expressions and response expectations are designed for capability evaluation and should not be interpreted as descriptions of any demographic group's preferences. In deployment, inferred meanings and behavioral expectations should be treated as personal information. Better memory-based personalization should respect user boundaries, including an expressed wish for less engagement.

Use of Large Language Models

Language models were used to assist manuscript drafting, revision, and figure generation. The evaluation design assigns Qwen3.8-27B, GPT-5 Mini, and Hy 4 to internal memory-method LLM roles and profile extraction. Memory-conditioned replies use GPT-5 Mini; direct-model controls use the respective column model as reader. GPT-4o Mini performs external automated judging. GPT-5 Mini generates candidate histories from author-written specifications. Human review remains responsible for the mapping and rubric. The authors are responsible for the claims, evidence, and final manuscript.

References

Mark W Baldwin. Relational schemas and the processing of social information. *Psychological Bulletin*, 112(3):461–484, 1992.

Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

Herbert H Clark and Susan E Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pp. 127–149. American Psychological Association, 1991.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dyJ0GWpjJB.

Weidong Guo, Dakai Wang, Zixuan Wang, Hui Liu, and Yu Xu. T-Mem: Memory that anticipates, not archives. *arXiv preprint arXiv:2606.15405*, 2026.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Weiquan Huang, Zixuan Wang, Hehai Lin, Sudong Wang, Bo Xu, Qian Li, Beier Zhu, Linyi Yang, and Chengwei Qin. AMA: Adaptive memory via multi-agent collaboration. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3099–3120, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.152. URL https://aclanthology.org/2026.findings-acl.152/.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13584–13606, 2023. doi: 10.18653/v1/2023.emnlp-main.838. URL https://aclanthology.org/2023.emnlp-main.838/.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, 2025. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Taeyoon Kwon, Dongwook Choi, Hyojun Kim, Sunghwan Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. Embodied agents meet personalization: Investigating challenges and solutions through the lens of memory utilization. In *International Conference on Learning Representations*, 2026.

Ruizhe Li, Mingxuan Du, Benfeng Xu, and Zhendong Mao. Keep it InMind: Benchmarking the implicit-association blind spot in agent memory. *arXiv preprint arXiv:2607.24368*, 2026a.

Yifei Li, Weidong Guo, et al. LoCoMo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. *arXiv preprint arXiv:2602.10715*, 2026b.

Minhua Lin, Zhiwei Zhang, Hanqing Lu, Hui Liu, Xianfeng Tang, Qi He, Xiang Zhang, and Suhang Wang. MemMA: Coordinating the memory cycle through multi-agent reasoning and in-situ self-evolution. *arXiv preprint arXiv:2603.18718*, 2026. URL https://arxiv.org/abs/2603.18718.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. SimpleMem: Efficient lifelong memory for LLM agents. *arXiv preprint arXiv:2601.02553*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Adyasha Maharana et al. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2):246–268, 1995.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.05589. arXiv:2502.05589.

Preston Rasmussen et al. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293, 1990. doi: 10.1207/s15327965pli0104_1.

Di Wu et al. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint*, 2024.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. *arXiv preprint arXiv:2504.15965*, 2025. URL https://arxiv.org/abs/2504.15965.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, 2022a. doi: 10.18653/v1/2022.acl-long.356. URL https://aclanthology.org/2022.acl-long.356/.

Wujiang Xu et al. A-MEM: Agentic memory for LLM agents. *Advances in Neural Information Processing Systems*, 2025. arXiv:2502.12110.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, 2022b. doi: 10.18653/v1/2022.findings-acl.207. URL https://aclanthology.org/2022.findings-acl.207/.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*, 2024. URL https://arxiv.org/abs/2404.13501.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

A Case Construction and Dataset Documentation

A.1 Case Schema

Table 8: Core fields needed to reproduce a case and its evaluation.

Field	Contents
Case identifier	Stable identifier, dimension, and development/test assignment.
History	Ordered scenes with turn identifiers and speaker roles.
Target mapping	Private referent or conditional response expectation.
Anchor evidence	Supporting scenes and the contribution of each scene.
Distractors	Confusable scenes that do not establish the target.
Probe	New utterance requiring the mapping.
Response rubric	Required interpretation or behavior; incompatible alternatives.
Construction metadata	Scenario seed, challenge category, generation and review records.
Filtering metadata	Filter model, prompt, budget, decision, and exclusion reason.

Input and annotation separation. The case schema names the system-visible fields as history.scenes[].turns, query.text, and query.position_after_turn. Only turns preceding the query's designated position are eligible history. Evaluator-only fields include private_mapping.refers_to, is_anchor, is_distractor, cue_turn_ids, and gold_scene_ids; these labels must not enter the memory store or reader input. Persona specifications that disclose the target are likewise reserved for construction or evaluation. Response features and forbidden alternatives require human review against the visible historical evidence. The loader removes evaluator-only keys before memory construction and enforces the query's temporal cutoff. Each run record retains a hash of the sanitized input for leakage checks.

A.2 Synthesis and Review

The protocol uses fictional personas and ordinary topics, including work, family, routines, achievements, and setbacks. GPT-5 Mini generates candidate histories from author-written specifications. Each specification contains a target mapping, competing interpretation, evidence plan, and response requirements. Only the conversation is system-visible. An evidence reviewer reads the history and query without the intended mapping; a rubric reviewer checks that required and forbidden behaviors follow from that evidence. Disagreement triggers revision or exclusion. The extra audit sample in Table 3 is a consistency check in addition to case-level admission review.

Generation instruction.

> Write a coherent multi-session conversation for the supplied fictional persona and target. Convey the mapping through ordinary usage or reactions, not an explicit definition or instruction. Distribute supporting clues across scenes, interleave plausible distractors and unrelated conversations, and preserve chronology. Write a later query that invokes the mapping without explaining it. Return conversation turns separately from evidence annotations and the response rubric. Do not place hidden annotations in the dialogue.

Evidence roles. Anchor scenes contribute evidence for the target mapping; their annotations should identify what each scene adds. Distractor scenes offer a competing referent, a literal interpretation, or another superficially plausible connection. Filler scenes provide separation and conversational continuity without defining the target. A distractor is not automatically a retrieval failure: a context can contain distractors and still supply sufficient evidence. Case annotations record the number and positions of anchor, distractor, and filler scenes.

Probe and target construction. The probe places the expression or response expectation in a later situation. Its wording must not reveal the intended private meaning or explicitly request the target behavior. A reference response demonstrates an acceptable answer; the accompanying rubric states what makes it acceptable, which alternatives conflict with the mapping, and what would be too general to establish understanding. The released record should separate these evaluator-only fields from the history and query supplied to systems.

A.3 Idiosyncratic Reference Case and Challenge Coverage

The following illustrative case shows how conversational evidence supports a private interpretation despite a plausible ordinary reading.

Micro-case: a private referent.

Scene 1, identity. User: Only Mum calls me Finch. Everyone else uses Maya, even Dad. Agent: A nickname that stayed with her.

Scene 2, filler. User: My bicycle needs a new chain. I walked to the station today. Agent: That must have added some time.

Scene 3, expression. User: The boss rang: Finch, have you eaten? Then came the usual inspection of my weekend plans. Agent: Quite a check-in.

Scene 4, distractor. User: Nina, my manager, approved the project schedule. The remaining changes can wait until Monday. Agent: A clear stopping point.

Query. The boss has already decided where I'm spending Sunday again.

Target. The boss denotes the user's mother. The nickname links Scenes 1 and 3; Scene 4 offers a workplace alternative.

Acceptable reply. Your mum is planning Sunday for you again. You can tell her what you have already decided for yourself.

Scoring. A mother–child interpretation passes. Workplace advice fails. Merely saying that the boss sounds controlling also fails because the referent remains unresolved.

Whether either scene alone already makes the target likely is checked separately; this illustration is not proof of cross-scene necessity.

Challenge labels describe construction choices. A metaphor may extend an ordinary meaning; referent shift may attach an expression to a different entity; irony may reverse its ordinary evaluative reading. These mechanisms need not exhaust the case space or form a strict empirical difficulty hierarchy. The allocation contains 60 cases in each of five primary mapping categories, with 12 development and 48 test cases per category. The three familiarity tags each cover 100 cases; F1/F2 contribute 240/60 cases. These are marginal coverage targets, not a balanced factorial experiment. Paired presentation variants share a semantic-family identifier and partition.

Coverage axes from the construction design. The Reference design combines three descriptive axes. Surface-form familiarity distinguishes unusual combinations, ordinary words used as private codes, and familiar terms with shifted referents. These are construction labels, not measured corpus-frequency or embedding-density bands. Mapping types cover entity names, state/metaphor labels, event labels, group expressions with a user-specific interpretation, and self-made abbreviations. A public slang meaning alone does not establish a private mapping. Presentation mode distinguishes usage-only evidence (F1) from a source that emerges naturally in a story without an explicit definition (F2). F2 may make a mapping locally recoverable, so the single-scene check remains necessary. These axes organize coverage, not an exhaustive factorial experiment or a difficulty decomposition.

A.4 Interactional Rapport Case and Response Expectations

Rapport cases specify a condition, a history-supported expectation, a novel query, and required/forbidden behaviors. Six rule families organize coverage without imposing a difficulty order. The example below specifies the query condition, supporting conversational evidence, and acceptable and unacceptable response behaviors. One construction conveys a reaction contrast within a scene: the user responds tersely to one agent behavior, but re-engages after a different response. Another distributes the contrasting reactions across scenes and topics. The first can make a local preference recoverable within one scene; the second can require aggregation across scenes. These constructions differ in their evidence requirements: a multi-session history does not necessarily require inference across multiple sessions.

Contrasting condition-to-response patterns include enthusiastic engagement with good news but space after a setback, or restrained acknowledgment of good news but active presence after a setback. These illustrate the distinction between recognizing the situation and selecting the response supported by this user's history. Neither a universal preference for brevity nor a universal preference for engagement captures both patterns. The profile baseline is evaluated through its final replies under the same criteria as other conditions.

Rule-family coverage. The design uses six non-exclusive families, grouped by the aspect of the reply they constrain: content (*boundary avoidance* and *associative callback*), expression (*stance/style* and *register/frame switching*), and interaction structure (*role/control allocation* and *ritual enactment*). These are case coverage tags, not the response-action labels in Table 12. For example, an object-conditioned register case can contrast sarcasm about one person with sincerity about another. A fixed ironic nickname instead targets private reference; the distinction is the main scoring target, not a claim that the underlying operations never overlap. The Rapport dataset contains 50 cases in each of the six primary rule families, including 10 development cases and 40 test cases per family. Secondary tags may overlap.

Micro-case: condition-dependent style. This example contrasts restrained acknowledgment after success with active engagement after setbacks.

Success 1. User: I got the fellowship. Agent: Incredible! This is huge! User: You're making it enormous. Agent: Good, the decision is finally in. User: Yes. I can arrange the move now.

Setback 1. User: My application was rejected. Agent: Sorry. I'll leave you some space. User: Oh. Okay. Agent: Which part is bothering you most? User: The feedback, actually. Let me explain.

Success 2. User: I beat my personal best today. Agent: A solid result. User: Exactly. The pacing finally worked.

Setback 2. User: The prototype failed again. Agent: That sounds frustrating. What happened this time? User: The connection failed under load. I'm glad you asked.

Target query. My submission was shortlisted. An acceptable reply is: Good, you've cleared this stage. The rubric requires restrained acknowledgment and forbids amplified celebration.

Contrasting query. My submission was rejected. An acceptable reply is: That's disappointing. Which part of the outcome is weighing on you? The rubric requires acknowledgment and an invitation to elaborate.

Both queries use the same unchanged history and are evaluated independently. Local reactions reveal useful preferences here; this example does not establish necessary multi-session inference.

Additional illustrative rapport cases. The following illustrative examples summarize distinct response requirements.

Boundary avoidance. Across earlier conversations about a breakup and a former restaurant, probing is followed by withdrawal, while changing the subject restores engagement. The history also links a former apartment to the relationship. For "I passed the apartment complex where I used to live," a reply should acknowledge the place without reopening the breakup. Asking whether the user misses the former partner fails.

Associative callback. When other cats are mentioned, recalling Doudou, the user's missing cat, repeatedly leads to welcomed conversation; generic remarks do not. For "I saw an orange cat basking in the sun today," the target is a gentle callback to Doudou, not just a generic remark about cats. The history must support both the association and the user's willingness to revisit it.

Object-conditioned register. Sincere congratulations about a former colleague's advancement are met with disengagement, whereas irony is welcomed; the reverse pattern holds for a respected mentor. "Wang has achieved another big success, and the whole team is praising him" calls for the history-supported ironic register. Literal congratulations fail. A mentor probe instead requires sincerity; a globally sarcastic style is not sufficient.

Role and initiative. Everyday decisions receive positive engagement when the agent offers alternatives and leaves the choice to the user; work decisions favor a recommendation with reasons. For "Where should I go this weekend?" options preserve the user's initiative, whereas presenting a choice as already decided does not. The criterion does not prohibit tentative suggestions.

Recurring ritual. After difficult weeks, the user repeatedly engages in a familiar sequence of acknowledgment, reflection, and encouragement. For "This week is finally over; I feel drained," the single-reply target is to acknowledge the effort and invite the hardest moment, not prematurely close the exchange. Later steps require subsequent turns and are outside the primary single-reply score.

These examples complement the stance/style case above. Each is scored by its own supported requirements, not by assuming that a callback, irony, or advice is universally desirable.

Evidence and response scope. Every condition in the rubric must have support in the system-visible history; adding an exception only to the target annotation creates an underdetermined case. Mark explicit local requests separately from expectations inferred from reactions. Under the primary single-reply protocol, a ritual case can assess the appropriate next action, not completion of an entire multi-turn routine. A full-routine claim requires a separately specified multi-turn protocol.

Table 9: Worked-case documentation for the two task dimensions.

Field	Reference	Rapport
Target	Surface expression and private meaning	Condition and expected response behavior
Anchors	Exact turns; contribution to resolving the referent	Exact reactions; evidence for the conditional expectation
Confounder	Competing literal meaning or referent	Generic response that conflicts with the expectation
Probe	New use of the private expression	New event invoking the relevant condition
Reference answer	One response acting on the correct meaning	One response following the expected behavior
Rubric	Required interpretation; incorrect and unresolved readings	Required and forbidden actions; mixed-action rules
Checks	Disclosure, single-scene, probe-only, retrieval checks	Disclosure, single-scene, default and profile checks

A.5 Admission Checks

Table 10: Case-construction checks and the information required to document their use.

Check	Scope	Record
Evidence sufficiency	Both	Does the designated evidence support the mapping?
Explicit disclosure	Both	Is the mapping stated as a definition or instruction?
Single-scene solution	Both	What can be inferred from each scene in isolation?
Probe-only solution	Both	Can the target be reached without the history?
Retrieval reachability	Reference	Which retriever, budget, and criterion were used?
General-knowledge shortcut	Both	Does the target require this user's mapping, rather than only general knowledge?
Default response	Rapport	Which reference policy supplies the default?
Profile comparison	Rapport	Evaluation-only baseline under the shared response rubric.

Admission requires sufficient visible evidence, a supported response rubric, no explicit disclosure in the query, and no contradiction between anchors and target. GPT-5 Mini supplies review suggestions; human reviewers decide admission. No-history screening rejects Reference candidates solved without history and Rapport candidates whose targets are reached by generic defaults under the screening configuration. Candidate logs record the model, prompt, decision, and exclusion reason. Profile performance does not determine admission.

Table 11: Blind audit of cases sampled from the initial test splits. The four disposition rows partition the 60 cases per dimension; initial reviewer disagreements are reported separately.

Audit outcome	Reference	Rapport
Cases reviewed	60	60
Retained without changes	57	54
Rubric or acceptable-variant clarification	2	4
History or query revision	1	1
Excluded and replaced	0	1
Initial reviewer disagreements	5	9

Independent case audit. We reviewed 60 cases from each dimension's initial 240-case test split. These cases are distinct from the 60 development cases per dimension and do not increase the test-set size. Before seeing the author-specified target, two reviewers independently read the history and query and wrote down the private mapping or response requirement they inferred. They then inspected the author target and rubric against the visible historical evidence. Initial disagreements entered adjudication.

The excluded Rapport case was replaced with a reserved case. Affected evaluation conditions were rerun, leaving 240 final test cases per dimension.

Admission-review instruction.

> Read the history and query without assuming the intended answer is correct. Identify supported interpretations. Check each proposed requirement against specific turns; flag contradictions, missing prerequisites, explicit disclosure, and plausible alternatives. Repeat the evidence check for every scene independently. Recommend keep, revise, or exclude, with evidence spans and reasons. A model's failure is not proof that a scene is insufficient.

Recording gates without defining capability by baseline failure. Retrieval, general-knowledge, and single-scene checks assess distinct aspects of case validity. A retrieval check records access under a specified index and top-K or token budget. A commonsense check asks whether the relevant link can be established without the user-specific history; predicting an exact held-out query is not itself an equivalent test. An accumulation check examines whether individual scenes already disclose the mapping. For each implemented gate, filtering records include the input, model, pass criterion, and candidate counts before and after filtering. If a baseline was used to select the final set, its score on that set is selection-dependent, not independent evidence that all methods in its family fail.

B Scoring and Diagnostic Measurements

B.1 Response Rubrics

Idiosyncratic-reference scoring distinguishes correct private interpretation, conflicting interpretation, and unresolved meaning. Only the first passes. Mentioning the surface expression or producing advice compatible with almost any interpretation does not demonstrate the required understanding. An answer need not explicitly define the expression if its response unambiguously acts on the correct meaning.

Interactional rapport is operationalized as adherence to a context-dependent response expectation supported by the supplied conversation history. A reply passes if it satisfies the applicable required behaviors and contains no explicitly forbidden behavior. The reference answer illustrates one acceptable realization rather than prescribing exact wording; semantically equivalent phrasing and rubric-listed alternatives also pass. A reply containing both a required and a forbidden behavior fails for that case. For example, when the history supports giving space without inquiry, an acknowledgment followed by a question fails; when it supports active engagement, brevity alone does not pass. If the visible history does not support a rubric requirement, or supports a plausible alternative the rubric has not considered, the case is flagged for review. These criteria apply to the particular history and query and do not imply that any response style is universally preferable.

Response-judge instruction. GPT-4o Mini receives one anonymized candidate reply at a time. Method names, reader identities, and input-condition labels are omitted.

> Assess whether the candidate reply satisfies the case requirements using the query, annotated historical evidence, intended mapping, reference reply, and rubric. Reference wording is not mandatory. Judge demonstrated understanding or behavior, not eloquence, friendliness, length, or lexical similarity. For private reference, literal misinterpretation, repetition without resolution, and generic replies compatible with incompatible meanings fail. For rapport, check every applicable required and forbidden action. A forbidden action causes failure even if a required action is also present. Apply only annotated alternatives and exceptions. For a single-turn ritual, evaluate only the required next action. Cite concise reply spans supporting each decision. Flag insufficient or contradictory rubrics for review rather than inventing a target.

The output record contains case ID; decision (pass, fail, or needs-review); each required behavior with a satisfaction flag and supporting span; each forbidden behavior with a presence flag and span; default-behavior flag; concise reason; and rubric version. Ordinary requirements are conjunctive; admissible alternatives are explicit disjunctions. Review flags undergo adjudication before binary scores are reported. A defective case is revised consistently across systems, not silently dropped from an inconvenient condition.

Table 12: Response-action vocabulary for rapport annotation. Actions may co-occur; correctness depends on the case-specific required and forbidden behaviors.

Action	Behavioral description
Probe	Elicits details and invites elaboration about the event.
Give space	Acknowledges briefly, withholds inquiry, and leaves room.
Side with	Affirms the user's position without qualification.
Analyze	Examines causes or interpretations of the situation.
Accompany	Offers presence and empathy without eliciting detail.
Advise	Proposes a concrete course of action.
Amplify	Heightens positive affect and celebrates good news.

These action labels are a non-exclusive annotation vocabulary. Required and forbidden actions, together with any case-specific style or reference constraints, determine correctness; the action name alone is not a gold answer. A reply passes only if it satisfies the required behaviors and contains no forbidden behavior, subject to the alternatives explicitly permitted by the case rubric.

B.2 Judge Inputs and Information Separation

Table 13: Information available to judges. Reference mappings and responses are evaluator annotations, not hints supplied to the tested system.

Measurement	Judge input	Decision
Private-meaning response	Query, reply, supporting turns, private mapping, reference and rubric	Does the reply act on the private meaning?
Rapport response	Query, reply, supporting turns, conditional expectation, reference and rubric	Does the reply perform required behavior without conflicting actions?
Retrieved evidence	Query, retrieved context and gold mapping/evidence; *no generated reply*	Is the supplied context sufficient to recover the mapping?
Memory-state audit	Inspectable stored state and annotated evidence/mapping	Is an adequate signal preserved or represented equivalently?

The response judge receives annotated supporting turns, the mapping, and response rubric, but no method identity. The evidence judge receives the actual reader context and evidence criterion without the generated reply.

Evidence-judge instruction.

> Determine whether the supplied reader context contains sufficient evidence, or an equivalent stored interpretation, to support the annotated mapping for this query. Credit faithful compressed notes even when source identifiers are absent. One retrieved anchor is insufficient when more evidence is required. Identify supporting spans and missing information. Do not infer whether a generated reply was correct; no reply is provided.

Its record contains a sufficiency flag, supporting spans, missing information, and a review flag. Response scoring and evidence scoring use separate calls, preventing answer correctness from leaking into evidence judgments. Each judgment includes a discrete decision and supporting response or evidence spans. Judgments assess the case-specific target rather than fluency, verbosity, agreement with the user's opinion, or resemblance to the reference wording.

B.3 Memory-State and Retrieved-Context Audits

For auditable configurations, we export the stored representation after history ingestion and retain the exact serialized context supplied to the reader after context processing. Source identifiers are retained where available, and multiple access rounds are recorded in order. A non-exportable storage state is unavailable for preservation analysis, not a failed preservation case. Judges must credit an equivalent stored mapping when a system has compressed the source episodes. Record sufficient evidence independently of answer correctness, and label uninspectable memory states as unavailable. If direct probes are used, distinguish failure to recover a mapping from evidence that the information was deleted.

An auxiliary P1/P2/P3 probe notation denotes query-cue identification, retrieval, and reader induction. This is a different probe decomposition: P1 is not a preservation audit. A reader recognizing an unusual expression without resolving it is evidence about cue recognition, not proof that its meaning was retained. These probes do not replace storage-state and context audits.

B.4 Metrics and Paired Comparisons

For input condition c, let $U_i^{(c)}$ denote correctness on a fixed set of N cases. Its accuracy is $A_c = N^{-1}\sum_i U_i^{(c)}$. The paired history effect is

$$\Delta_{\text{hist}} = A_{\text{hist}} - A_{\text{nohist}}.$$

The switch rate among initially incorrect cases is defined as

$$\text{Switch}_c = \frac{\sum_i (1 - U_i^{(\text{nohist})}) U_i^{(c)}}{\sum_i (1 - U_i^{(\text{nohist})})}.$$

The complementary regression rate among initially correct cases is

$$\text{Regress}_c = \frac{\sum_i U_i^{(\text{nohist})} (1 - U_i^{(c)})}{\sum_i U_i^{(\text{nohist})}}.$$

Rates with a zero denominator are reported as N/A. Together these rates satisfy $A_c - A_{\text{nohist}} = (1 - A_{\text{nohist}})\text{Switch}_c - A_{\text{nohist}}\text{Regress}_c$ when both denominators are nonzero. The same fixed case set and no-history reference responses must be used when comparing conditions. Retain the four paired counts, not just percentages. Primary scores use one response per case/configuration; denominator changes and operational failures are disclosed. Paired comparisons retain the four correctness transitions for the same cases. Repeated queries or runs from the same history or mapping family are dependent observations.

Supplementary retrieval measurements. For systems that return ranked source scenes, evidence recall at K and reciprocal rank can describe access to annotated anchors. Report the relevance unit, the rule for multi-anchor cases, the cutoff, and the reader-context token budget. Retrieving one anchor is not necessarily sufficient to infer a cross-scene mapping. For compressed notes without source identifiers, use the evidence-sufficiency audit rather than treating missing identifiers as missing information. These measurements do not replace response pass rate and are not combined with it into an overall score.

Table 14: Human and automated-judge reliability on the audited judgments. Panel A pairs list annotator A's label first and annotator B's second. Panel B compares the adjudicated human label with GPT-4o Mini; TP, FN, FP, and TN take the human label as ground truth.

Panel A: Human–human, before adjudication						
Task	$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$	Agreement	κ
Reference response	27	3	3	87	95.00%	0.867
Rapport response	30	7	5	78	90.00%	0.762
Reference evidence	22	2	2	34	93.33%	0.861
Rapport evidence	24	3	3	30	90.00%	0.798
Panel B: Adjudicated human versus GPT-4o Mini						
Task	TP	FN	FP	TN	Agreement	κ
Reference response	27	3	2	88	95.83%	0.888
Rapport response	32	4	3	81	94.17%	0.860
Reference evidence	22	2	1	35	95.00%	0.895
Rapport evidence	24	3	2	31	91.67%	0.831

B.5 Judge Reliability

Judge prompts were calibrated on development data and frozen before test evaluation. We audited 240 response judgments (120 per dimension) and 120 answer-blind evidence-sufficiency judgments (60 per dimension). Two human annotators independently labeled each sampled judgment without system identities; a third person adjudicated disagreements. Human–human agreement is calculated before adjudication. Agreement with GPT-4o Mini is calculated against the adjudicated human labels. A positive label denotes a passing response or sufficient retrieved evidence.

Before adjudication, the annotators disagreed on 6 Reference responses, 12 Rapport responses, 4 Reference evidence judgments, and 6 Rapport evidence judgments. Among human-negative responses, the automated false-pass rates were 2/90 (2.22%) for Reference and 3/84 (3.57%) for Rapport. Among human-positive responses, false-fail rates were 3/30 (10.00%) and 4/36 (11.11%), respectively. For evidence sufficiency, false-sufficient rates were 1/36 (2.78%) and 2/33 (6.06%); missed-sufficient rates were 2/24 (8.33%) and 3/27 (11.11%). These conditional rates describe the audited judgments. Human annotators disagreed more often on Rapport responses than on Reference responses, and the judge occasionally rejected valid replies despite high overall agreement.

C Systems, Prompts, and Additional Results

C.1 System Configurations

The memory backbones are Qwen3.8-27B, GPT-5 Mini, and Hy 4; the memory-conditioned reader is GPT-5 Mini and the external judge is GPT-4o Mini. The seven memory methods are Mem0, LightMem (Fang et al., 2026), A-Mem, AMA (Huang et al., 2026), SimpleMem (Liu et al., 2026), MemMA, and T-Mem. Within each method, all enabled internal LLM roles use the designated memory backbone. Embeddings and non-generative rerankers remain fixed across the three memory backbones. Table 15 summarizes the settings.

The complete history is available to each memory method at ingestion, even if its own policy later compresses or discards content. Native retrieval units and widths are distinguished from the final context budget. Unresolved request failures remain in the planned denominator and are reported separately from scored replies.

Context packing and memory isolation. Budget accounting uses the Qwen3.8-27B tokenizer for all configurations and a 4,096-token reader-memory cap. Each method retains its native retrieval units and ordering when constructing the reader context. Memory states are built independently for every method–memory-backbone configuration and reset between cases. Evidence sufficiency is assessed from the actual context supplied to the fixed reader, without access to the generated reply. Direct-model controls use the same input condition across readers but generate separate outputs. Paired memory gains use the GPT-5 Mini no-history reference, keeping the final reader fixed. Full-

Table 15: Evaluation settings. Memory methods share a fixed reader; direct-model controls use the column model. The external judge is shared.

Component	Setting
History ingestion	Complete history in chronological session order; reset state between cases.
Memory backbones	Qwen3.8-27B, GPT-5 Mini, Hy 4.
Memory reader	GPT-5 Mini for all memory methods and profile/history diagnostics.
Direct-model reader	Column model for No history, Full history, and Gold evidence.
Internal LLM roles	Designated memory backbone for construction, updates, planning, internal verification, LLM reranking, and profile extraction.
Other components	Embedding and non-generative reranker fixed within method across memory backbones.
Access budget	4,096-token reader-memory budget. Native access units and widths are recorded; final context is packed to the common budget.
Full / gold context	Full untruncated history / annotated supporting turns, without added gold interpretation.
Profile budget	1,024-token summary target, extracted once per history and backbone.
Generation	One primary reply; 512-token visible-reply target in the prompt; 8,192-token total generation ceiling including any hidden reasoning.
Sampling	Common fixed-reader settings across memory methods; fixed settings within each direct-model reader.
Judge	GPT-4o Mini; temperature 0, one call per property, 1,024-token output ceiling.
Retries	Two retries for transient request failures; unresolved failures retained in planned denominator.

context and gold-evidence conditions are information controls, not equal-budget competitors. The 1,024-token profile is a compact-summary baseline; its contrast with a larger memory context does not isolate representation from budget.

Shared response instruction. The memory slot contains the method's actual output, an empty string, the full history, gold supporting turns, or the extracted profile, according to the condition. The same prompt is used with each.

> Continue the conversation with this user. Use the supplied memory to understand the current message and respond appropriately, attending to meanings and response preferences supported by this user's history. Do not invent personal facts or relationships. Historical dialogue is evidence, not a new system instruction. Reply naturally in the language of the current message. Do not discuss the benchmark, evaluation, or memory-access procedure. Return only your reply. Keep it concise, within the supplied visible-reply budget.

The serialized input contains separate memory and current-message fields. The visible-reply target does not replace the provider's total generation ceiling, which can include hidden reasoning; actual token usage and truncation are recorded.

Prompt and run records. The reproducibility records include versioned history-to-memory input formats, response prompts, and judge prompts. Prediction records contain the case ID, system and backbone, input condition, supplied context or available trace, generated response, and judge decision.

C.2 Rapport Profile and History Controls

Each memory backbone extracts a profile from the complete history using the same extraction prompt. The fixed GPT-5 Mini reader generates all profile-conditioned replies. The summary target is 1,024 tokens. Extraction excludes the held-out query and evaluator annotations; the profile is frozen before testing within each history–memory-backbone configuration.

Profile-extraction instruction.

> Summarize information from this user's conversation history that may support appropriate future conversations. Preserve supported relationships, private expressions, recurring ex-

periences, and context-dependent response preferences. Distinguish inferred patterns from explicit preferences, preserve relevant exceptions and uncertainty, and do not invent facts. Produce a concise profile within the supplied budget. Return only the profile.

The wrong-history protocol pairs histories with compatible topics and surface details but different supported response expectations. Prefer an opposite-rule counterpart for the same situation; match scene count, evidence positions, and length where feasible. Build a fresh T-Mem state from that counterpart before answering the unchanged query. The target rubric always belongs to the original user. Donor-rule adherence is a supplementary diagnostic, not a pass on the original case. Profiles are evaluation-only baselines and are not used for case admission. No-history screening is a construction check; performance under the selection configuration is therefore selection-dependent.

Profile-based response baseline. We first extract a profile p from H with a fixed prompt, without access to the held-out query or evaluator annotations. The reader then generates $y = \mathrm{Respond}(P, q, p)$, using the shared response prompt P. The response judge applies the same required and forbidden behaviors as in other rapport conditions. Neither a plausible profile nor a fluent reply earns credit unless the final response meets the case target. We report response pass rate, not a separate measure of profile extraction quality.

Matched wrong-history controls. Match a wrong-history control on topic, length, and evidence density where feasible while changing the supported response expectation. A random irrelevant history primarily tests distraction, not user specificity.

Contrasting probes. Where both conditions are supported by a history, pair a target query with a contrasting query (for example, a different person that calls for a different register). Reuse the same query-independent memory or profile without updating it between test queries. Report each side's accuracy and the fraction of pairs where both replies pass. These probes test conditional selectivity; they are distinct from replacing the history while holding the query fixed. Contrasting queries can test conditional selectivity by invoking different response expectations supported by the same history. Such a comparison holds the memory or profile fixed while varying the query, unlike a wrong-history control that holds the query fixed. The example in Appendix A illustrates this format.

T-Mem is the fixed memory configuration for the history-control comparison. The profile arm is a separate end-to-end baseline; it is not subjected to an additional profile-quality score.

Table 16: Rapport controls (%). Pass rates are computed from the same case-level outcomes as those in Table 7. T-Mem and the GPT-5 Mini reader are fixed; the memory LLM varies. Switch and Regress condition on initially incorrect and correct no-history replies; Default is independently annotated.

Memory LLM	Condition	Pass	Default	Switch	Regress
Qwen3.8-27B	No history	2.08	87.50	—	—
	Correct history	24.17	62.50	22.98	20.00
	Profile only	15.00	75.00	14.47	60.00
	Mismatched history	3.75	85.00	3.40	80.00
GPT-5 Mini	No history	2.08	87.50	—	—
	Correct history	28.75	57.50	27.23	0.00
	Profile only	18.75	72.50	17.87	40.0
	Mismatched history	2.50	87.50	2.13	80.00
Hy 4	No history	2.08	87.50	—	—
	Correct history	26.67	60.00	25.96	40.00
	Profile only	17.50	73.75	16.60	40.0
	Mismatched history	2.50	86.25	2.13	80.00

C.3 Gold-Evidence Reader Comparison

This additional comparison tests whether errors after evidence provision depend on the reader. Gold evidence removes the need to locate the annotated supporting scenes; it does not reveal the target

mapping or response. The reference readers are GPT-5.6 Sol, Claude Opus 5, and Gemini 2.5 Pro, directly generating replies without a memory manager. Their API identifiers are gpt-5.6-sol, claude-opus-5, and gemini-2.5-pro. The original three readers provide the comparison in Table 5.

GPT-5.6 Sol uses standard reasoning mode with high effort; Claude Opus 5 uses adaptive thinking with high effort; Gemini 2.5 Pro uses a 32,768-token thinking budget. All three receive a 65,536-token total generation ceiling, retaining the shared 512-token visible-reply target. These are stronger-reader configurations, not an equal-compute comparison with the main evaluation. Model identifiers, call dates, returned version metadata, generation settings, and truncations are recorded; automatic model fallback is disabled.

Every reader receives identical gold historical turns in the same order, the same query, and the shared response prompt, without target interpretations or reference replies. Each generates one primary reply per case on the same 240 cases per dimension. The comparison reports pass counts and rates under the same rubric and GPT-4o Mini judge, with reader identities hidden from the judge.

Table 17: Gold-evidence reader comparison: The reference readers use identical evidence and response prompts.

Additional reader	Ref. pass / N	Ref. (%)	Rap. pass / N	Rap. (%)
GPT-5.6 Sol	240/240	100.0	240/240	100.0
Claude Opus 5	240/240	100.0	240/240	100.0
Gemini 2.5 Pro	239/240	99.6	237/240	98.8

Validation and interpretation. Human review covers all original-reader/reference-reader disagreements and residual failures, plus a prespecified sample of shared passes. Reviewers check evidence sufficiency, permissible response variants, prompt conflicts, truncation, and judge errors before attributing a failure to evidence utilization. Case or rubric corrections are versioned and applied consistently to all affected readers. With gold evidence, GPT-5.6 Sol and Claude Opus 5 pass all 240 cases in each dimension, while Gemini 2.5 Pro fails one Reference case and three Rapport cases.

D Query-to-Memory Relations

Table 2 defines Bridge L1–L4 separately from the two task dimensions. The numbering marks a change in the source of the connection, not measured difficulty. A query can use several bridges: shared wording may locate candidate scenes while a private mapping is still required to select or interpret them. Both task dimensions focus on history-specific mappings rather than four separate bridge-level test sets.

Similarity is an operation; association is a relevance relation. BM25 and dense embeddings score lexical and vector-space matches. Associative access seeks memories useful to the current query even when matching the raw stored content is insufficient. These descriptions are compatible: an inferred association can be written into an index and reached by a similarity operation. T-Mem illustrates this distinction by adding anticipatory triggers to stored evidence (Guo et al., 2026). Our L4 focus concerns whether the system learns a connection licensed by the user's history and uses it in its reply, rather than which primitive eventually retrieves the representation.

General versus history-specific knowledge. A general-knowledge bridge can make an allergy relevant to choosing a restaurant once the personal fact is known. General knowledge alone does not determine which relative a particular user's private nickname denotes or whether that user welcomes probing after a setback. Those mappings require the supplied conversational evidence. A reader outside the original conversation can recover them when given enough of that evidence; membership in the relationship is not an inaccessible prerequisite.