EVIDENCE-PRESERVING INTERPRETATION FOR PERSONALIZED LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized LLM agents must turn past interactions into responses that fit a user's current needs. Retrieved passages preserve the original history but leave the agent to work out what they imply for the current request. Compressed memories state an implication but can drop the context that shows whose preference it is and whether the user has withdrawn it. We propose EVINCE (EVidence-preserving INterpretation of Conversational Excerpts), a memory framework that interprets retrieved history for the current request. A learned interpreter states the relevant preference, its subject, and its withdrawal status, and a frozen LLM agent receives this note together with the unchanged excerpts. To align training with deployment, retrieval-aligned training replaces an annotated training input with the selector's actual retrieval whenever that retrieval contains the annotated user turns, keeping every preference target unchanged. On PersonaMem-v2 with 32k and 128k histories, EVINCE outperforms seven baselines in overall accuracy with two LLM agents, by up to 9.46 points over the strongest, while the agent reads far fewer tokens than with the full history. Ablations show that the note improves accuracy with the retrieved evidence held fixed, that retrieval-aligned training adds further gains, and that evidence and interpretation play complementary roles.

1 INTRODUCTION

Personalized LLM agents are expected to adapt their recommendations, assistance, and responses to the people they serve. Over repeated interactions, users reveal interests, describe personal circumstances, discuss other people, and revise earlier preferences. An effective agent should draw on this history when a later request makes it relevant, without asking the user to restate the same background. Memory connects a user's interaction history to the decisions an agent makes in the current conversation, which makes it central to personalization (Salemi et al., 2024; Jiang et al., 2025a;b).

Existing memory systems support this connection in two broad ways. Retrieval-based approaches keep history as passages and select those related to a request (Lewis et al., 2020). Other approaches consolidate interactions into compact representations, including the conversational facts of Mem0, the linked notes of A-Mem, and the evolving PersonaTree of Inside Out (Chhikara et al., 2025; Xu et al., 2026a; Zhao et al., 2026). Both improve access to user information, yet the agent must still answer a task-specific question: what does this evidence imply for this user, now? A relevant passage may record an assistant's suggestion or another person's preference rather than the user's own. A compact memory may omit the context that separates these cases or shows that the user has withdrawn a preference. Personalized memory therefore needs the source context and an explicit reading of what it implies.

Figure 1 illustrates both needs with a PersonaMem-v2 case. The user asks for a peaceful outdoor activity for a weekend morning. Earlier conversations include a tai chi suggestion from the assistant and a request to stop using an outdoor-exercise preference. In a separate writing interaction, the user's newsletter draft describes lingering to watch birds. The useful clue sits inside a writing task, and its relevance appears only when the user later asks for an activity. RAG and Mem0 select tai chi and walking or sitting, respectively, and an agent that reads our selector's raw excerpts also selects walking or sitting. None chooses the benchmark-preferred birdwatching option. The case separates retrieving personal context from using it.

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.

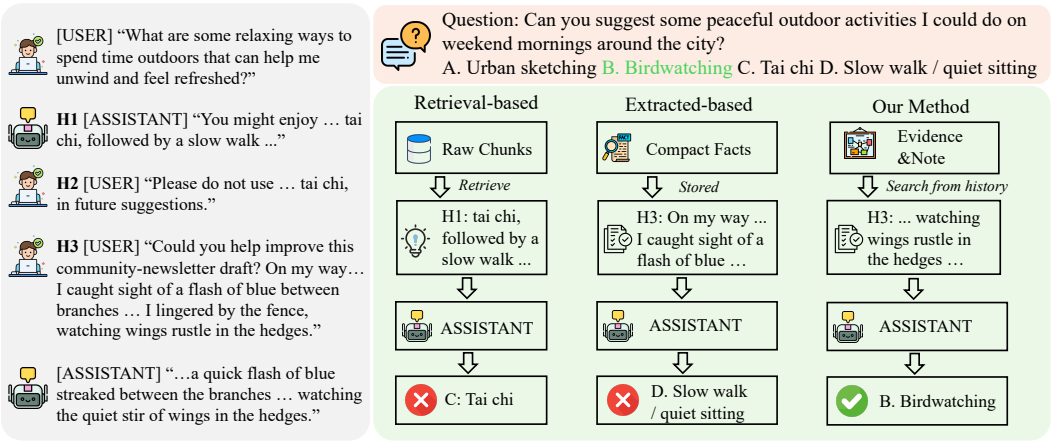


Figure 1: **User memory must support the current task.** In a PersonaMem-v2 case, a clue from an earlier writing interaction (H3) determines the answer to a later activity request. The retrieval-based memory returns the assistant's tai chi suggestion (H1), which the user later asked not to reuse (H2), and the extraction-based memory keeps only a fragment of the draft; both lead the agent to a wrong option. EVINCE supplies the relevant excerpt together with its note, and the agent selects birdwatching. The retrieval- and extraction-based outputs are those of the RAG and Mem0 baselines. Architectures are schematic, and history text is paraphrased.

We address this gap with EVINCE (EVidence-preserving INterpretation of Conversational Excerpts), a memory framework that interprets retrieved history for the current request. A learned interpreter states the request-relevant preference, its subject, and its withdrawal status, and the LLM agent receives this note together with the original excerpts. The note makes the implication of the evidence explicit, while the excerpts keep the context needed to check it. Because a fixed selector supplies the excerpts, we can measure the effect of interpretation with the evidence held constant. To align interpreter training with deployment, we introduce *retrieval-aligned training*, which uses the selector's actual retrieval as the training input whenever it contains the annotated user turns and keeps the annotated input otherwise, without changing any target. On PersonaMem-v2 response selection with the retrieved excerpts held fixed, the note improves accuracy by 5.80 and 2.92 percentage points for 32k and 128k histories. Our contributions are:

- We propose EVINCE, an evidence-preserving interpretation interface for personalized LLM agents, in which a learned interpreter states the request-relevant preference, its subject, and its withdrawal status, and the agent receives this note alongside the unchanged source excerpts.
- We introduce retrieval-aligned training, which exposes the interpreter to actual retrievals while keeping the annotated preference targets, and we analyze how it changes the training distribution.
- We evaluate EVINCE on PersonaMem-v2 with 32k and 128k histories and two LLM agents. It achieves the highest overall accuracy in all four settings, exceeding the strongest of seven baselines by up to 9.46 points, and ablations with the retrieved evidence held fixed show that source evidence and explicit interpretation play complementary roles.

2 RELATED WORK

2.1 PERSONALIZED LLM AGENTS

Personalizing LLM-based agents to individual users has become central as agents move from single-turn assistance toward sustained, user-dependent interaction. Work in this space is generally organized around four interdependent capabilities, namely profile modeling, memory, planning, and action execution (Xu et al., 2026b), with memory serving as the core substrate that connects a user's interaction history to the agent's current decisions and enables the other three capabilities to remain

consistent across sessions. Profile modeling builds persona-based (Li et al., 2025b) or response-based (Bose et al., 2025) representations of a user's preferences, spanning both explicit and implicit signals as well as behavioral and topical dimensions, while agent-role definition determines whether the agent substitutes for (Wang et al., 2025a) or adapts to the user (Li et al., 2025a). Memory itself is organized into personal memory schemes, ranging from textual summaries (Zhong et al., 2024) to structured vector, tree, or graph representations (Chhikara et al., 2025; Rezazadeh et al., 2025; Rasmussen et al., 2025), together with personalized retrieval strategies that are content-based (Lewis et al., 2020), structure-aware (Gutiérrez et al., 2024), or policy-guided (Zerhoudi & Granitzer, 2024) and that surface user-relevant evidence at inference time. Without this component, an agent must re-infer the same preferences from scratch in every request, leaving planning and action execution with no persistent signal to condition on. Existing evaluations, including interactive-alignment benchmarks such as PersonaMem (Jiang et al., 2025a) and PrefEval (Zhao et al., 2025) as well as user-substitution benchmarks (Salemi et al., 2024; Kumar et al., 2024), therefore treat memory quality as a key bottleneck for personalization. Our work targets exactly this component: rather than proposing a new retrieval or profiling strategy, we study how retrieved personal history should be presented to a downstream agent so that its implication for the current request is made explicit without discarding the context needed to verify it.

2.2 Agent Memory

A complementary line of work treats memory itself as the central object of study (Huang et al., 2026), unifying it along three axes: substrate, distinguishing internal parametric states such as weights, latent activations, and KV caches (Meng et al., 2022; Wang et al., 2024; Kwon et al., 2023) from external non-parametric stores such as vector indices, text records, and structured or hierarchical memory (Lewis et al., 2020; Rezazadeh et al., 2025); cognitive mechanism, spanning sensory (Fang et al., 2025a), working (Wang et al., 2025b), episodic (Kang et al., 2025), semantic (Salama et al., 2025), and procedural memory (Fang et al., 2025b); and subject, contrasting user-centric personalization (Rasmussen et al., 2025) with agent-centric self-experience (Wang et al., 2025c). Existing personal-memory systems predominantly compress interaction history into textual or structured summaries, such as the conversational facts of Mem0 (Chhikara et al., 2025), the linked notes of A-Mem (Xu et al., 2026a), and the evolving persona trees of Inside Out (Zhao et al., 2026). Such compression improves efficient access but discards the surrounding context needed to determine a preference's ownership, currency, or withdrawal. Beyond storage design, memory operations such as retrieval, update, compression, and forgetting (Packer et al., 2023; Zhong et al., 2024), together with learning-based evolution policies spanning prompting (Park et al., 2023), fine-tuning (Lu et al., 2023), and reinforcement learning (Yan et al., 2025), have been studied as ways to keep memory current and useful, exposing a persistent trade-off between evidence retention and abstraction. EVINCE is positioned at exactly this trade-off: it keeps retrieved episodic evidence intact while attaching a learned, request-conditioned semantic interpretation of subject and withdrawal status, directly addressing the tension between compact, usable memory and faithful, verifiable memory that prior memory designs have yet to reconcile.

3 Methodology

EVINCE turns retrieved interaction history into evidence that a personalized LLM agent can use for its current task. As illustrated in Figure 2, a fixed selector first retrieves relevant excerpts from the user history. A learned interpreter then states the request-relevant preference, its subject, and its withdrawal status. The agent receives this interpretation together with the original excerpts and selects a response. The interpretation gives the agent a compact usage cue, and the excerpts preserve the context needed to check it. We first formalize this interface, then describe how evidence is interpreted and presented, and finally explain how we train the interpreter on the evidence it will see at deployment.

3.1 Problem Formulation

Let H denote the available history, q the current request, and Q all task-visible information. A memory module constructs $M(q, H)$, and the LLM agent is a frozen language model $\pi_\psi(a \mid Q, M)$ conditioned on the task input and this memory. We study personalized response selection, in which

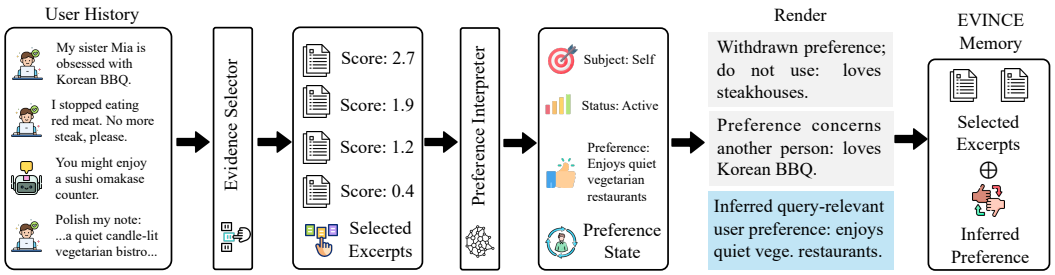


Figure 2: **Overview of EVINCE.** Given a user history and a request, the fixed evidence selector scores history blocks and keeps the top-scoring excerpts. The preference interpreter reads the request and these excerpts and outputs a preference state with the preference, its subject, and its withdrawal status. A fixed rule renders the state into a one-sentence note: the highlighted row is the note produced for this request, and the grey rows show the forms for a withdrawn preference and for another person's preference. The frozen LLM agent receives the note appended to the unchanged excerpts. The example is illustrative.

Q contains the request and answer options and a is the selected response. With evidence selector S_ϕ, preference interpreter G_θ, and a fixed rendering function, the memory interface is

$$E = S_\phi(q, H), \qquad Z = G_\theta(q, E), \tag{1}$$

$$W = \text{render}(Z), \qquad M(q, H) = E \oplus W, \tag{2}$$

where E contains raw excerpts, Z is the preference state, W is the rendered state note, and $\oplus$ denotes concatenation. Only the agent receives answer options. At inference, the selector and interpreter receive the request and history, without options, gold answers, reference snippets, or preference annotations.

3.2 Interpreting Retrieved Evidence

Interpretation starts from a small set of relevant interactions in their original wording. Complete interactions form chunks of about 6,000 characters, and oversized interactions are kept intact. After empty chunks and normalized duplicates are removed, the selector scores every remaining candidate and returns at most four eligible excerpts in score order. Each excerpt keeps both user and assistant text, so the interpreter can tell user information apart from the surrounding dialogue. We train the selector, separately from the interpreter, following the procedure in Appendix A.1. It is a Qwen3.5-9B LoRA scorer initialized from an evidence-localization adapter that we fit on official training queries, with targets that measure how much each candidate changes the DeepSeek agent's accuracy on 1,024 official training queries relative to answering without memory. Because the selector stays fixed in all experiments, every comparison of interpretation variants uses the *same evidence*.

Relevant excerpts can still express a preference indirectly, attribute it to someone else, or record that the user has withdrawn it. To make these distinctions explicit, the interpreter generates one request-conditioned preference state,

$$Z = (s, t, p), \qquad s \in \{\texttt{self}, \texttt{other}\}, \quad t \in \{\texttt{active}, \texttt{withdrawn}\}, \tag{3}$$

where p describes the relevant preference or personal circumstance, s identifies its subject, and t gives its withdrawal status. The interpreter can infer preferences from requests, personal writing, and described behavior. Its instructions limit a withdrawal to the particular preference concerned and state that retrieved excerpts are not ordered chronologically. The Qwen3.5-9B interpreter decodes greedily without input truncation.

Rendering turns the state into a one-sentence note W that tells the agent how to use p: a withdrawn preference should not guide the current response; an active preference with subject `other` belongs to someone other than the user; and an active `self` preference may inform the response. Withdrawal takes precedence, so a withdrawn note does not repeat the subject. The note follows the unchanged excerpts under a header stating that it is a fallible inference and that the original history takes precedence. If the interpreter's output is not a valid state or reaches the output limit, the note is empty

and the agent receives $M = E$. Validity is checked only structurally, so a well-formed but incorrect state reaches the agent unchanged (Appendix A.4).

Keeping the excerpts beside the note is what makes the interface evidence-preserving. Fix all learned parameters and prompts, and let $\mathcal{R}_\psi(T)$ be the agent's expected 0–1 error with memory representation T, $\mathcal{R}^*(T)$ the Bayes risk given (Q, T), and $\varepsilon_\psi(T) = \mathcal{R}_\psi(T) - \mathcal{R}^*(T)$ the agent's decision regret. Because W is computed from (q, E) and E is retained,

$$\mathcal{R}^*(E, W) = \mathcal{R}^*(E) \le \mathcal{R}^*(W), \qquad \mathcal{R}_\psi(E) - \mathcal{R}_\psi(E, W) = \varepsilon_\psi(E) - \varepsilon_\psi(E, W). \tag{4}$$

Replacing the excerpts with the note alone can therefore only raise the Bayes risk, whereas appending the note leaves it unchanged. An accuracy gain at fixed evidence is a reduction in the agent's regret rather than added information, consistent with the distinction between available and usable information (Xu et al., 2020). A misleading note can instead increase regret.

3.3 Retrieval-Aligned Training

For the note to help at deployment, the interpreter should be trained on the kind of evidence the selector actually returns. Each training example provides a request q_i, an official preference target $Z_i^\star$, and an annotated reference snippet. The reference input X_i^{ref} combines this snippet with up to three distractors from the same history. It supplies relevant supervision, but its composition differs from the selector's actual retrieval E_i.

Retrieval-aligned training therefore uses the actual retrieval whenever it covers the reference user turns and keeps the reference input for the remaining examples. Let U_i contain the complete, nonempty user turns in the reference and V_i the user turns in E_i. With $u \sqsubseteq v$ denoting literal substring containment,

$$c_i = \mathbf{1}[U_i \neq \varnothing \ \wedge \ \forall u \in U_i,\ \exists v \in V_i :\ u \sqsubseteq v], \qquad \widetilde{X}_i = \begin{cases} E_i, & c_i = 1, \\ X_i^{\mathrm{ref}}, & c_i = 0. \end{cases} \tag{5}$$

This rule replaces 7,785 of the 11,934 training inputs (65.23%) with actual retrievals and keeps the other 4,149 reference inputs. Queries and targets are unchanged, and missing references are never inserted into a retrieval. Literal containment is a conservative proxy for coverage rather than a test of semantic sufficiency. At deployment, the interpreter receives actual retrievals without access to references.

We fit the reference-trained model, REF, and the retrieval-aligned model, MIX, separately from the same Qwen3.5-9B base using LoRA (Hu et al., 2021). For target tokens $z^\star_{i,1:T_i}$, including EOS, the loss is

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{j=1}^{T_i}\log P_\theta\left(z^\star_{i,j} \mid q_i, \widetilde{X}_i, z^\star_{i,<j}\right). \tag{6}$$

Only target tokens contribute, and length normalization weights queries equally. The training queries span 520 histories, and another 80 histories with 1,915 queries form a held-out set of reference inputs. Both models use rank-8 LoRA on attention projections and train for two epochs (2,984 updates). We use the final epoch-2 checkpoint. The two models see the same queries and number of updates, but MIX receives longer inputs and therefore more training compute (Appendix A).

The resulting training distribution matches deployment on the covered examples. Relative to training on the selector's retrievals, the remaining input discrepancy comes only from the uncovered subgroup, following the source-risk and discrepancy view of domain adaptation (Ben-David et al., 2010). Section 4 evaluates the resulting gain, and Appendix B gives the proofs together with an analysis of the evidence budget.

4 Experiments

We evaluate EVINCE on personalized response selection in PersonaMem-v2 with 32k and 128k interaction histories. We first compare it with existing memory methods and then, in an ablation study, isolate the note and retrieval-aligned training with the retrieved evidence held fixed, separate the roles of source evidence and explicit cues, and examine the evidence budget. Finally, we compare the computational cost of all methods.

Method	Knowledge	Pro-Email	Pro-Writing	Creative	Translation	Personal-Email	Trouble	Chat	Social	Overall
Qwen3.5-397B-A17B, 32k histories										
Base LLM	64.44	40.82	38.12	41.96	43.76	43.31	39.32	39.56	38.60	44.62
Full Context	51.94	59.36	57.27	54.34	62.38	58.46	59.88	54.82	56.47	56.92
RAG	43.57	48.13	45.30	46.21	48.60	47.05	43.71	44.98	47.43	45.98
Rewrite	51.82	50.27	46.59	46.58	47.86	46.85	50.90	43.98	48.46	48.40
Mem0	48.91	44.92	48.80	47.13	53.26	45.28	47.50	43.78	41.48	47.00
A-Mem	43.20	49.38	45.67	51.02	50.28	52.95	49.70	50.20	51.54	48.94
MemListener	44.27	39.93	42.17	37.34	43.76	40.35	45.91	38.76	42.26	41.78
EVINCE (ours)	**68.45**	**68.98**	**64.09**	**65.80**	**70.20**	**66.54**	**62.48**	**60.64**	**68.58**	**66.38**
DeepSeek-V4-Flash, 32k histories										
Base LLM	36.77	22.99	23.57	25.69	27.37	23.23	24.55	26.51	22.38	26.56
Full Context	53.28	51.16	51.38	51.19	55.31	54.53	52.90	51.61	52.36	52.66
RAG	46.72	37.08	34.25	39.56	40.97	37.20	35.53	40.16	36.14	39.14
Rewrite	51.21	41.89	44.01	40.11	41.90	41.54	45.31	39.36	44.15	43.76
Mem0	44.90	41.18	38.67	37.90	40.22	43.90	43.31	37.35	41.48	41.22
A-Mem	50.00	36.72	36.28	41.59	39.66	38.78	41.32	38.55	38.81	40.78
MemListener	41.50	35.65	38.31	33.83	40.22	35.04	40.52	33.53	37.37	37.58
EVINCE (ours)	**69.30**	**60.07**	**59.30**	**61.74**	**62.20**	**62.01**	**57.88**	**57.03**	**61.81**	**61.76**
Qwen3.5-397B-A17B, 128k histories										
Base LLM	**62.50**	39.57	38.31	40.67	41.53	42.72	39.72	39.36	39.84	43.90
Full Context	49.88	**54.37**	**52.30**	50.83	55.87	**57.09**	**54.49**	46.59	**55.03**	52.76
RAG	40.78	47.24	45.67	47.32	45.62	43.11	42.51	42.57	47.23	44.48
Rewrite	45.99	**54.37**	51.75	51.02	51.58	52.56	51.10	45.18	52.77	50.46
Mem0	42.48	51.34	49.91	51.76	53.63	54.33	50.90	47.99	47.43	49.56
A-Mem	42.35	48.48	46.04	51.19	46.93	49.21	47.70	46.18	48.67	47.12
MemListener	39.81	39.75	41.07	34.75	44.32	36.22	46.11	36.75	41.68	40.02
EVINCE (ours)	**62.50**	**54.37**	50.64	**52.87**	**56.24**	53.94	52.50	**50.80**	53.59	**54.68**
DeepSeek-V4-Flash, 128k histories										
Base LLM	34.83	22.99	22.84	25.32	25.88	24.80	26.55	23.29	23.00	26.06
Full Context	48.30	45.99	46.22	44.55	46.18	**49.41**	44.91	43.17	45.38	46.16
RAG	46.36	35.83	34.81	37.71	37.99	34.45	32.93	37.35	36.34	37.66
Rewrite	49.39	45.81	44.75	42.51	44.13	42.52	**48.10**	39.96	46.41	45.10
Mem0	37.50	41.53	40.88	39.37	41.71	44.88	44.11	34.54	39.43	40.30
A-Mem	48.18	37.43	34.81	39.56	37.24	38.78	40.92	37.35	37.58	39.62
MemListener	39.68	33.87	37.18	33.27	40.22	35.04	41.72	33.73	35.11	36.82
EVINCE (ours)	**63.35**	**47.77**	**48.99**	**49.17**	**52.14**	49.21	47.31	**48.59**	**50.92**	**51.58**

Table 1: **Main results on PersonaMem-v2.** Overall and scenario accuracy (%) for two LLM agents and two history lengths. Bold marks the highest value in each column within each setting (including ties); underline marks the second-highest. Pro-Email and Pro-Writing denote professional email and writing.

4.1 EXPERIMENTAL SETUP

Dataset and agents. PersonaMem-v2 evaluates personalization from long interaction histories, including implicit preferences, preference ownership, and requests to forget (Jiang et al., 2025b). We use all 5,000 official queries from 200 histories at each history length. Both lengths share query identities and cover seven preference types and nine conversational scenarios. History length describes the available dialogue, and each memory method determines what the agent receives. We use two frozen LLM agents, Qwen3.5-397B-A17B (the Qwen agent) and DeepSeek-V4-Flash (the DeepSeek agent), with the benchmark's multiple-choice prompt and parser, temperature zero, and an 8,192-token output limit. We report response-selection accuracy overall and by category, counting unparsed and length-capped responses as incorrect.

Baselines and matched-evidence controls. We compare EVINCE with Base LLM (no memory), Full Context (the complete history), RAG, Rewrite, Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2026a), and MemListener (Zhao et al., 2026), each evaluated with both agents on the same queries. To isolate interpretation, three controls share the selector's top-four excerpts in identical order: RAW supplies the excerpts alone, REF appends the note of the reference-trained interpreter, and MIX appends the note of the retrieval-aligned interpreter. MIX is the final EVINCE configuration, and RAW differs from the RAG baseline in its learned selector.

Agent	History	RAW	REF	MIX	MIX−REF	MIX−RAW
Qwen	32k	60.58	64.90	**66.38**	+1.48	+5.80
Qwen	128k	51.76	53.50	**54.68**	+1.18	+2.92
DeepSeek	32k	55.26	58.88	**61.76**	+2.88	+6.50
DeepSeek	128k	46.36	49.80	**51.58**	+1.78	+5.22

Table 2: **Matched-evidence comparisons.** Accuracy (%) over 5,000 queries per cell. Within each agent and length, all conditions receive identical excerpts, and the DeepSeek agent reuses the memories prepared for the Qwen agent. Differences are in percentage points.

4.2 MAIN RESULTS

Table 1 reports overall and scenario accuracy. EVINCE achieves the highest overall accuracy in all four settings. Full Context is the strongest baseline in each setting, and none of the five memory baselines improves on passing the complete history to the agent, whereas EVINCE surpasses it in every setting. With 32k histories, EVINCE reaches 66.38% with the Qwen agent and 61.76% with the DeepSeek agent, 9.46 and 9.10 points above Full Context and 17.44 and 18.00 points above the strongest memory baseline (A-Mem for Qwen, Rewrite for DeepSeek). With 128k histories, it reaches 54.68% and 51.58%, 1.92 and 5.42 points above Full Context and 4.22 and 6.48 points above Rewrite. Both components of EVINCE contribute to these margins. The learned selector alone (RAW) reaches 60.58% and 55.26% at 32k, and the note adds 5.80 and 6.50 points on top of the same excerpts (Section 4.3.1). At 128k with Qwen, the note is what places EVINCE above Full Context, since RAW alone reaches 51.76% against 52.76%.

These gains come with a short agent input. With EVINCE, the agent reads about 5.1k and 7.4k tokens at the two lengths, whereas Full Context passes the complete history (Section 4.4). The four-excerpt budget also leaves headroom on long histories: raising it to eight adds another 5.62 points at 128k. Memory matters more for DeepSeek, and EVINCE delivers correspondingly larger gains. Without memory, DeepSeek reaches 26.56% and 26.06%, compared with 44.62% and 43.90% for Qwen, and EVINCE raises its accuracy by 35.20 and 25.52 points using the memory prepared for the Qwen agent, without retraining or new states.

Across scenarios, EVINCE has the highest accuracy in all nine at 32k with both agents, and at 128k it is best or tied for best in five scenarios with Qwen and best in seven with DeepSeek. Knowledge queries show the value of the note most clearly. With the Qwen agent, the Base LLM (64.44% and 62.50%) outperforms Full Context and every memory baseline on these queries, so adding history lowers accuracy on questions the agent often answers correctly without it. EVINCE is the only method with history that matches or exceeds the Base LLM (68.45% and 62.50%), whereas RAW with the same excerpts reaches 58.98% and 53.64% (Table 4), so the note, rather than the retrieved excerpts, accounts for this advantage. Across preference types (Appendix C), the note brings its largest gains on Health, where MIX exceeds RAW by 15.67 and 14.79 points at 32k and 128k with the Qwen agent, and on anti-stereotypical preferences (9.47 and 7.13 points).

4.3 ABLATION STUDY

4.3.1 INTERPRETATION WITH FIXED EVIDENCE

Table 2 compares the three matched-evidence controls, which receive identical excerpts within each agent and history length. Holding the excerpts fixed, the note adds 5.80 and 2.92 points for Qwen at 32k and 128k, and 6.50 and 5.22 points for DeepSeek. The DeepSeek agent receives exactly the memories prepared for the Qwen agent, so its gains require neither retraining nor new states.

With evidence and agent fixed, MIX improves over REF from 64.90% to 66.38% at 32k and from 53.50% to 54.68% at 128k, and it improves over REF by 2.88 and 1.78 points with DeepSeek. Retrieval-aligned training also raises Forget accuracy over REF by 4.68 and 4.39 points and improves seven of nine scenarios at each length. The two interpreters share training examples, targets, and update counts. Most of the gain over RAW comes from adding a note at all: REF alone adds 4.32 and 1.74 points.

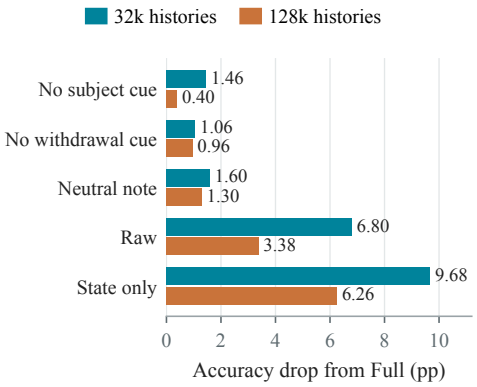


Figure 3: **Presentation ablation** (Qwen agent). Accuracy drop of each presentation relative to Full (EVINCE), in percentage points. All variants share the same preference states and excerpts.

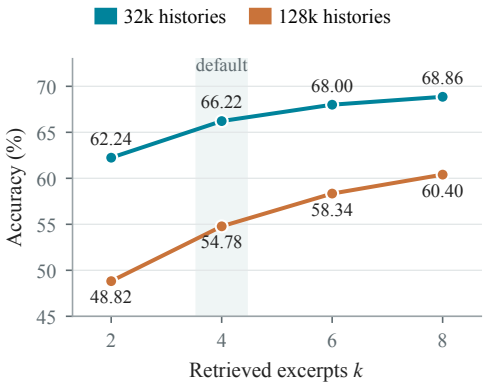


Figure 4: **Evidence-budget sensitivity.** Accuracy (%) of EVINCE with the Qwen agent. Model weights are fixed, and notes are regenerated for each budget. The shaded top-4 budget is the default used in all other experiments.

4.3.2 SOURCE EVIDENCE AND EXPLICIT CUES

We compare six memory presentations with the Qwen agent, keeping model weights and preference states fixed (Figure 3). Full is the EVINCE presentation. Neutral note keeps the preference sentence, the excerpts, and the header but removes the explicit subject and withdrawal cues. No subject cue and No withdrawal cue each remove one cue. State only removes the excerpts and the header that designates them as authoritative.

Removing the excerpts costs 9.68 and 6.26 points, and State only falls below RAW at both lengths: the interpretation helps only alongside its evidence. Full exceeds Neutral note by 1.60 and 1.30 points, so the explicit cues matter beyond the preference text. Removing only one cue costs less than removing both. Without the subject cue, accuracy falls by 1.46 points at 32k but only 0.40 at 128k; without the withdrawal cue, it falls by 1.06 and 0.96 points. The subject cue therefore matters mainly with shorter histories, whereas the withdrawal cue contributes similarly at both lengths.

4.3.3 EVIDENCE BUDGET

We vary the evidence budget over nested top-2/4/6/8 excerpts, fixing the selector, threshold, interpreter, and agent, and regenerate the notes for each budget without truncating any input. Because the budgets are nested prefixes of one ranking, a larger budget only adds evidence and cannot raise the Bayes risk (Appendix B). Whether accuracy improves then depends on how well the interpreter and the agent use the additional excerpts. Accuracy increases monotonically at both lengths (Figure 4). Moving from two to four excerpts brings the largest gain, 3.98 points at 32k and 5.96 at 128k, which suggests that two excerpts often miss the relevant preference. Beyond four excerpts, the gains diminish at 32k, with 1.78 and 0.86 points for the next two steps, but remain large at 128k, with 3.56 and 2.06 points, consistent with relevant evidence being spread over more blocks in a longer history. Top-8 therefore exceeds top-4 by 2.64 points at 32k and 5.62 at 128k, and at 128k this gain exceeds the note's own gain over RAW (2.92 points). EVINCE thus keeps benefiting from additional evidence rather than being distracted by the longer input. The default top-4 budget used in all other experiments already outperforms every baseline, and larger budgets offer further gains when a longer input is acceptable.

4.4 COMPUTATIONAL COST

Figure 5 relates accuracy to the tokens read by all LLM calls per 5,000 queries. EVINCE reaches the highest accuracy while its agent reads 25.5M and 36.8M input tokens at 32k and 128k, 13 and 28 times fewer than Full Context, which passes the complete history (328M and 1,030M). The memory baselines spend most of their tokens building memory before any query is answered: Rewrite, MemListener, Mem0, and A-Mem read 175M to 2,960M tokens for memory construction at 32k and

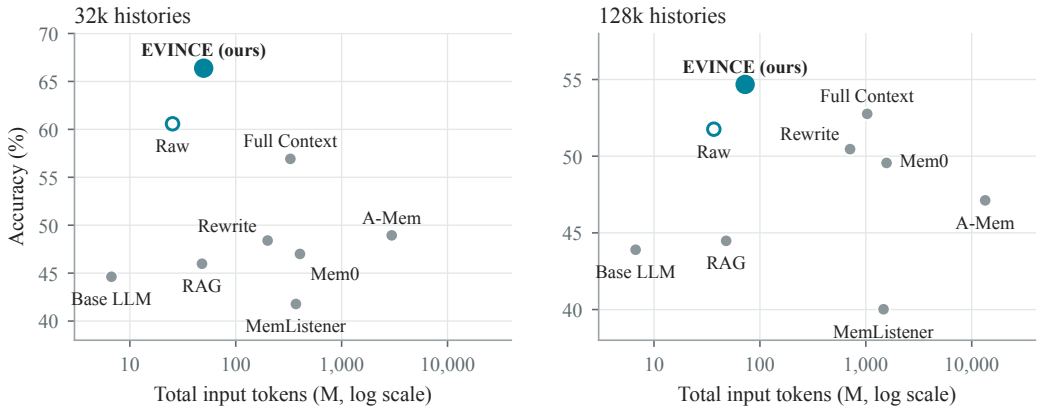


Figure 5: **Accuracy versus token cost.** Accuracy (%) with the Qwen agent against total input tokens (M) per 5,000 queries.

685M to 13,400M at 128k, whereas the EVINCE interpreter reads 24.3M and 35.6M. Among the memory baselines, only RAG reads fewer tokens than EVINCE in total, and it trails by 20.40 and 10.20 points in accuracy. The note adds at most 0.2M answering tokens per 5,000 queries, and the agent's outputs shrink from 3.3M to 3.0M tokens at 32k. Before interpretation, the Qwen3.5-9B selector scores the candidate blocks without generating text, reading about 137k prompt tokens per query at 128k.

5 CONCLUSION

We introduced EVINCE, an evidence-preserving memory framework that gives a personalized LLM agent a learned, query-conditioned interpretation of retrieved history together with the history itself. Retrieval-aligned training exposes the interpreter to actual retrievals while keeping the annotated targets. On PersonaMem-v2 response selection, the interpretation improves accuracy with the retrieved evidence held fixed, retrieval-aligned training adds further gains, and the same memory transfers to a second LLM agent. The ablations point to a common design principle: retaining personal evidence and stating its task-relevant implication play complementary roles, since the interpretation alone falls below the raw evidence.

Limitations. Our evaluation uses a single interpreter backbone (Qwen3.5-9B), so gains may not transfer to other models or scales. Retrieval-aligned training's coverage rule relies on literal substring containment, a conservative proxy for semantic sufficiency rather than a true test of it. The interpreter also decodes greedily, and we have not evaluated its robustness to sampling temperature or multi-sample aggregation.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and to provide feedback on research methodology and experiments. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to edit the writing of the paper to improve readability. We have reviewed all AI-assisted work. Specifically, all AI-suggested proof were independently re-derived and verified for correctness by the authors before inclusion; AI-generated feedback on the experimental methodology was critically evaluated and incorporated only where the authors judged it sound; and all AI-polished text was checked against the original draft to confirm that no meaning or claims were altered during rewriting. We take responsibility for

the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects. All experiments use PersonaMem-v2 (Jiang et al., 2025b), a public CC-BY-4.0 benchmark of simulated users and interaction histories, so no real user's conversations are processed. In deployment, EVINCE would store excerpts of a user's past interactions and infer preferences from them, including personal circumstances such as health, which requires user consent and control over what is retained. Its withdrawal cue changes how the agent uses a preference but does not delete the source excerpt, so honoring a request to forget also requires removing it from the stored history. Inferred preferences can also be wrong: on sensitive preferences, the note lowers accuracy relative to the raw evidence (Table 3). EVINCE marks each note as fallible and always passes the source excerpts to the agent, so an interpretation remains traceable to its evidence.

REPRODUCIBILITY STATEMENT

Section 3 defines the selector, the interpreter, the rendered note, and the retrieval-aligned training objective. Appendix A gives the full configuration, including selector training, the construction of training targets, inputs, and data splits, the interpreter's hyperparameters and hardware, and the prompts, decoding settings, and rendering templates. Section 4 specifies the evaluation protocol on the official PersonaMem-v2 data, and Appendix B states the assumptions and gives complete derivations for the analysis.

REFERENCES

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://link.springer.com/article/10.1007/s10994-009-5152-4.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. LoRe: Personalizing LLMs via low-rank reward modeling. *arXiv preprint arXiv:2504.14439*, 2025. URL https://arxiv.org/abs/2504.14439.

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.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025a. URL https://arxiv.org/abs/2510.18866.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Memp: Exploring agent procedural memory. *arXiv preprint arXiv:2508.06433*, 2025b. URL https://arxiv.org/abs/2508.06433.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, et al. A survey of agent memory in the second half: Towards self-evolving and long-horizon agents. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=XycbogUAeJ.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025a.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025b.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025. URL https://arxiv.org/abs/2506.06326.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024. URL https://arxiv.org/abs/2407.11016.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 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*, pp. 5259–5276, 2025a.

Jia-Nan Li, Jian Guan, Songhao Wu, Wei Wu, and Rui Yan. From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment. *arXiv preprint arXiv:2503.15463*, 2025b. URL https://arxiv.org/abs/2503.15463.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. MemoChat: Tuning LLMs to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*, 2023. URL https://arxiv.org/abs/2308.08239.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

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

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *International Conference on Learning Representations*, 2025.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. MemInsight: Autonomous memory augmentation for LLM agents. *arXiv preprint arXiv:2503.21760*, 2025. URL https://arxiv.org/abs/2503.21760.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025a.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. MemoryLLM: Towards self-updatable large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 50453–50466, 2024.

Yu Wang, Dmitry Krotov, Yuanzhe Hu, Yifan Gao, Wangchunshu Zhou, Julian McAuley, Dan Gutfreund, Rogerio Feris, and Zexue He. M+: Extending MemoryLLM with scalable long-term memory. *arXiv preprint arXiv:2502.00592*, 2025b. URL https://arxiv.org/abs/2502.00592.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025c.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026a.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.10689.

Yue Xu, Qi'an Chen, Zizhan Ma, Dongrui Liu, Wenxuan Wang, Xiting Wang, Li Xiong, and Wenjie Wang. Toward personalized LLM-powered agents: Foundations, evaluation, and future directions. *arXiv preprint arXiv:2602.22680*, 2026b. URL https://arxiv.org/abs/2602.22680.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, et al. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025. URL https://arxiv.org/abs/2508.19828.

Saber Zerhoudi and Michael Granitzer. PersonaRAG: Enhancing retrieval-augmented generation systems with user-centric agents. *arXiv preprint arXiv:2407.09394*, 2024. URL https://arxiv.org/abs/2407.09394.

Jihao Zhao, Ding Chen, Zhaoxin Fan, Kerun Xu, Mengting Hu, Bo Tang, Feiyu Xiong, and Zhiyu Li. Inside out: Evolving user-centric core memory trees for long-term personalized dialogue systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13429–13446, 2026.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations*, 2025.

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.

A Implementation and Training Details

A.1 Evidence Construction and Selector

EVINCE combines our trained utility selector with a separately fitted preference interpreter. The history is serialized into blocks of about 6,000 characters that preserve complete interactions, so an oversized interaction remains intact. Blocks keep role labels and user and assistant text and exclude tool, function, and system payloads. Empty blocks are removed, and duplicates are merged after normalizing whitespace and case.

The selector considers all unique nonempty blocks. It is Qwen3.5-9B with a LoRA adapter and scores candidate c_i as

$$s_i = \ell_\phi(1 \mid q, c_i) - \ell_\phi(0 \mid q, c_i), \qquad I = \mathrm{Top4}_{i:s_i>-2}\, s_i, \tag{7}$$

where ℓ is a vocabulary logit. Eligible blocks are returned in decreasing score order with their literal content, and an empty selection uses the benchmark's no-memory placeholder. Each candidate is truncated to 2,048 tokens for scoring, while the interpreter and the agent receive the complete selected blocks.

The selector adapter is initialized from an evidence-localization adapter that we fit on 4,096 official TRAIN queries against the official related-conversation snippet, and then trained on 1,024 queries from the official PersonaMem-v2 TRAIN split. Each candidate's utility target averages three comparisons between the MCQ correctness of the DeepSeek-V4-Flash agent with that candidate and without memory. Gold answers are used only to construct this offline supervision. At deployment, the selector sees only the request and the candidates. Because the DeepSeek agent provides these targets, the second-agent experiment measures transfer of the prepared memory to an agent that contributed to selector supervision. The selector objective is

$$\mathcal{L}_{\mathrm{sel}} = \mathcal{L}_{\mathrm{rank}} + 0.25\mathcal{L}_{\mathrm{sign}} + 0.10\mathcal{L}_{\mathrm{Huber}}. \tag{8}$$

Within each query, the ranking term uses candidate pairs with $u_i - u_j > 0.25$ and minimizes the weighted mean of $-\log\sigma(s_i - s_j)$, with weights $\min(u_i - u_j, 8)$. The sign term averages $\mathrm{softplus}(-\mathrm{sign}(u_i)s_i)$ over candidates satisfying $|u_i| \geq 0.25$. The Huber target is u_i with scale one. We train the selector for two epochs with learning rate 2×10^{-5} and select the checkpoint and the threshold of -2 on a held-out 15% of TRAIN histories. This configuration is fixed in all experiments.

A.2 Official Targets and Input Views

The interpreter target is a JSON object containing `subject` $\in$ {`self`, `other`}, `status` $\in$ {`active`, `withdrawn`}, and a nonempty `preference` string. The official `who`, `updated`, and `preference` fields determine these values. For an `ask_to_forget` update, the target uses `prev_pref` and marks it withdrawn, so the forgetting instruction is not treated as a positive preference.

From the 800 official TRAIN histories (18,549 queries), we exclude the 200 histories that appear in our development set and split the rest by history into 520 training histories (11,934 queries) and 80 held-out histories (1,915 queries).

For REF, the input combines the official related-conversation excerpt with up to three distinct excerpts from the same history, chosen by token overlap with the query, and the position of the reference excerpt varies across examples. For MIX, we retrieve the fixed selector's top-four evidence from each 32k TRAIN history and apply the replacement rule in Equation 5, which yields 7,785 retrieved-input and 4,149 reference-input examples. Targets and queries are unchanged, and no example is discarded when retrieval misses its reference. Held-out inputs remain reference inputs, and all targets come directly from the official annotations.

A.3 Objective and Optimization

REF and MIX are trained separately from the same pretrained Qwen3.5-9B model, and only newly initialized LoRA parameters are updated. Both minimize the length-normalized cross-entropy in Equation 6 over target tokens, with prompt tokens masked. They use rank-8 LoRA with alpha 16 and no dropout on the query, key, value, and output projections, AdamW with learning rate 2×10^{-5} and

weight decay 0.01, a global batch size of 8, and gradient clipping at norm 1 without a learning-rate schedule. Training runs for two epochs (2,984 updates) in bfloat16 on eight A100 GPUs, with prompts of up to 16,384 tokens and targets of up to 512 tokens, and we use the final checkpoint. On one query from each held-out history, all 80 states generated by MIX are structurally valid.

A.4 PROMPT AND RENDERING

The interpreter receives a fixed system instruction and a user message containing `Current request:` followed by q, then `Conversation excerpts:` followed by the raw evidence. The instruction requests the single relevant preference or personal circumstance. It allows implicit preferences expressed in requests, writing, or behavior; distinguishes the user from other people and fictional characters; and limits forgetting to the particular withdrawn preference. It warns that retrieved blocks are not a globally chronological history. The output must be the three-field JSON object and must not include answer options or answer letters. The input supplies no official reference, state annotation, biography, or MCQ answer.

Generation is greedy with thinking disabled, a 16,384-token complete-prompt limit, and at most 512 new tokens. A valid state is appended after the unchanged source evidence under the header:

> Fallible inferred preference state; original history remains authoritative:

The rendered note is one of three forms:

> Withdrawn preference; do not use as a current user preference: p
> Preference concerns another person, not the user: p
> Inferred query-relevant user preference: p

Withdrawal is tested first. A withdrawn state therefore receives the withdrawal cue without an additional subject cue.

If the output is not a valid state or reaches the token limit, the agent receives the raw evidence alone, without repair or regeneration. In the presentation ablation, State only receives *No valid inferred preference state is available.* in this case.

B ANALYSIS OF EVIDENCE RETENTION AND RETRIEVAL ALIGNMENT

This appendix derives the identities used in Section 3. They apply standard decision-theoretic and information-theoretic tools to the implemented memory interface. They characterize what can change when interpretation is added; they do not establish that an arbitrary interpreter improves an arbitrary LLM agent. Bayes risks, information quantities, and population discrepancies are not estimated by our experiments.

B.1 SETUP AND THE ROLE OF THE RENDERED NOTE

Fix the trained parameters, prompts, selection rule, and deterministic state-decoding procedure. All probabilities below are conditional on these choices. Let Q contain the request q and any other task-visible information, including answer options when applicable; H is the available history and Y the correct response. Selection uses only (q, H) to obtain E, and interpretation uses only (q, E) to obtain Z. The LLM agent receives the rendered note W, rather than the JSON record itself.

Rendering includes the fixed validity policy: an invalid or capped state yields an empty note. The valid withdrawal branch does not preserve the subject field explicitly, so reasoning about the complete Z as though it were visible to the LLM agent would misdescribe the implementation. We assume that the final serialization preserves the complete E and its separation from the appended note. If appending a note caused source truncation, the information equivalence below would no longer follow.

The three abstract memory representations are $T_{\mathrm{raw}} = E$, $T_{\mathrm{full}} = (E, W)$, and $T_{\mathrm{state}} = W$. In the implemented State only variant, a fixed header replaces the source-authority header, and an absent state is represented by the fixed no-state message. We treat these fixed headers and the absence

marker as part of the LLM agent interface. They add no source evidence, but can affect a particular LLM agent's behavior, which is included in its decision regret.

Let $K_\psi(\cdot \mid Q, T)$ be the LLM agent's decision distribution under its evaluation decoding procedure. With 0–1 loss,

$$\begin{aligned} \mathcal{R}_\psi(T) &= \mathbb{E}[\mathbf{1}\{A \neq Y\}], \qquad A \sim K_\psi(\cdot \mid Q, T), \\ \mathcal{R}^*(T) &= \inf_f \mathbb{E}[\mathbf{1}\{f(Q,T) \neq Y\}], \\ \varepsilon_\psi(T) &= \mathcal{R}_\psi(T) - \mathcal{R}^*(T) \geq 0. \end{aligned} \tag{9}$$

The infimum ranges over measurable decision rules, without restricting them to the LLM agent's architecture or computational budget. Regret therefore includes limitations of input use, presentation, and decision behavior; it does not identify a particular internal reasoning failure.

B.2 Decomposing Retrieval, Representation, and LLM Agent Error

Proposition 1 (risk decomposition). If E is a deterministic function of (Q, H) and T a deterministic function of (Q, E), then

$$\begin{aligned} \mathcal{R}_\psi(T) = \mathcal{R}^*(H) &+ \underbrace{\big[\mathcal{R}^*(E) - \mathcal{R}^*(H)\big]}_{\text{retrieval gap}} \\ &+ \underbrace{\big[\mathcal{R}^*(T) - \mathcal{R}^*(E)\big]}_{\text{representation gap}} + \underbrace{\varepsilon_\psi(T)}_{\text{agent regret}}, \end{aligned} \tag{10}$$

and all three gap terms are nonnegative.

Proof. A decision rule given (Q, H) can reconstruct E and simulate any rule given (Q, E). Similarly, a rule given (Q, E) can reconstruct T. Hence $\mathcal{R}^*(H) \leq \mathcal{R}^*(E) \leq \mathcal{R}^*(T)$. Adding and subtracting these Bayes risks in $\mathcal{R}_\psi(T)$ gives the identity. □

For EVINCE, W can be reconstructed from (Q, E), while E is retained in (E, W). Thus

$$I(Y; W \mid Q, E) = 0, \qquad \mathcal{R}^*(E, W) = \mathcal{R}^*(E) \leq \mathcal{R}^*(W). \tag{11}$$

Subtracting the two instances of Equation 9 proves Equation 4. The learned interpreter adds computation and uses fitted parameters, but adds no sample-specific information conditional on (Q, E) and those fixed parameters. Conditioning on parameters is essential; this observation says nothing about the absence of training-data leakage or about generalization to unseen populations.

The same-evidence improvement is therefore a reduction in the tested LLM agent's regret, without identifying which internal process changed. The contrast between Full and State only is different: it can change both the representation gap and agent regret. Its observed accuracy difference is not an estimate of either term in isolation, and its prompt length and source-authority cue also change.

B.3 A Log-Loss View of Evidence Retention

For analysis, let $F_\psi(\cdot \mid Q, T)$ be a soft probability distribution over answers, and define

$$\mathcal{L}_\psi(T) = \mathbb{E}[-\log F_\psi(Y \mid Q, T)], \qquad \mathcal{D}_\psi(T) = \mathbb{E}_{Q,T}\, \mathrm{KL}\big(P(Y \mid Q, T) \,\|\, F_\psi(Y \mid Q, T)\big). \tag{12}$$

Assume the relevant expectations are finite. This soft distribution is distinct from K_ψ, the actual decoded decision distribution. Treating a wrong greedy prediction as a point-mass soft distribution would produce infinite log loss.

Proposition 2 (information decomposition). Under the deterministic representation assumptions of Proposition 1,

$$\begin{aligned} \mathcal{L}_\psi(T) = \mathsf{H}(Y \mid Q, H) &+ I(Y; H \mid Q, E) \\ &+ I(Y; E \mid Q, T) + \mathcal{D}_\psi(T). \end{aligned} \tag{13}$$

Proof. The cross-entropy identity gives $\mathcal{L}_\psi(T) = \mathsf{H}(Y \mid Q, T) + \mathcal{D}_\psi(T)$. Because T is determined by (Q, E),

$$\mathsf{H}(Y \mid Q, T) = \mathsf{H}(Y \mid Q, E) + I(Y; E \mid Q, T).$$

Because E is determined by (Q, H),

$$\mathsf{H}(Y \mid Q, E) = \mathsf{H}(Y \mid Q, H) + I(Y; H \mid Q, E).$$

Substitution proves the result. □

The two conditional mutual informations measure answer information lost through retrieval and subsequent representation, respectively. For both RAW and Full, the representation term is zero. Consequently,

$$\mathcal{L}_\psi(E) - \mathcal{L}_\psi(E, W) = \mathcal{D}_\psi(E) - \mathcal{D}_\psi(E, W). \tag{14}$$

For the optimal soft predictor, retaining evidence instead of using the note alone gives

$$\mathcal{L}^*(W) - \mathcal{L}^*(E, W) = I(Y; E \mid Q, W). \tag{15}$$

The right-hand side vanishes exactly when $Y \perp E \mid (Q, W)$: the note then preserves all answer-relevant information in the retrieved evidence. The three-field state is not guaranteed to satisfy this sufficiency condition. Positive conditional mutual information does not, however, imply a strictly positive 0–1 Bayes-risk gap, because distinct answer distributions can share the same optimal label.

These identities explain the motivation for evidence retention; they are not additional training objectives. We neither optimize nor estimate the displayed mutual informations and KL terms. The LLM agent's answer log loss is also distinct from the interpreter's length-normalized token loss in Equation 6. MCQ accuracy gains cannot be converted into information gains from these identities.

B.4 WHAT RETRIEVAL-ALIGNED TRAINING CHANGES

Consider a common underlying population of tuples $(q, E, X^{\mathrm{ref}}, Z^\star, C)$, where C is the literal coverage condition of Equation 5. Assume $0 < \pi = P(C = 1) < 1$, and define

$$P_{E,c} = \mathcal{L}(q, E, Z^\star \mid C = c), \qquad P_{X,c} = \mathcal{L}(q, X^{\mathrm{ref}}, Z^\star \mid C = c).$$

The same targets occur in both views. Conditioning on C gives the deployment and retrieval-aligned training distributions

$$P_{\mathrm{ret}} = \pi P_{E,1} + (1 - \pi) P_{E,0}, \qquad P_{\mathrm{mix}} = \pi P_{E,1} + (1 - \pi) P_{X,0}. \tag{16}$$

Proposition 3 (residual input discrepancy). For these distributions,

$$P_{\mathrm{ret}} - P_{\mathrm{mix}} = (1 - \pi)(P_{E,0} - P_{X,0}). \tag{17}$$

For a fixed interpreter g and an integrable loss $\ell(g(q, X), Z^\star)$, integration gives

$$L_{\mathrm{ret}}(g) - L_{\mathrm{mix}}(g) = (1 - \pi)\big[L_{E,0}(g) - L_{X,0}(g)\big]. \tag{18}$$

Proof. The two distributions in Equation 16 have the identical term $\pi P_{E,1}$. Subtracting cancels that term, yielding Equation 17. Linearity of integration gives the risk identity. □

With $\mathrm{TV}(P, P') = \sup_A |P(A) - P'(A)|$, homogeneity also gives

$$\mathrm{TV}(P_{\mathrm{ret}}, P_{\mathrm{mix}}) = (1 - \pi)\,\mathrm{TV}(P_{E,0}, P_{X,0}). \tag{19}$$

If $0 \le \ell \le 1$, the absolute risk difference is bounded by this total variation. The implemented token cross-entropy is not naturally bounded, so that bound cannot be applied to it without additional assumptions; the integrable-loss identity remains applicable.

Three limits matter for interpretation. First, the empirical replacement rate $7{,}785/11{,}934$ describes the training construction, not test coverage or the fraction of error removed. Literal coverage need not certify semantic support, and failure of coverage need not imply insufficient semantic evidence. Second, the reference training distribution is $P_{\mathrm{ref}} = \pi P_{X,1} + (1 - \pi) P_{X,0}$. The two subgroup discrepancies in $P_{\mathrm{ret}} - P_{\mathrm{ref}}$ can cancel, so the identity alone does not guarantee a smaller total discrepancy than training on reference inputs alone. It also does not compare the different models produced by optimization. Third, the construction uses 32k training retrievals from one population. Transfer to 128k or another benchmark introduces an additional population shift not accounted for here. The downstream comparison of REF and MIX tests the benefit empirically.

B.5 Increasing the Evidence Budget

The top-k experiment takes nested prefixes of one fixed candidate ranking. For $k' > k$, the larger evidence set $E_{k'}$ therefore permits recovery of E_k, giving $\mathcal{R}^*(E_{k'}) \le \mathcal{R}^*(E_k)$. Let $T_k = (E_k, W_k)$, with the note regenerated at each budget. Since each view retains its evidence,

$$\begin{aligned}\mathcal{R}_\psi(T_{k'}) - \mathcal{R}_\psi(T_k) = & -\left[\mathcal{R}^*(E_k) - \mathcal{R}^*(E_{k'})\right] \\ & + \varepsilon_\psi(T_{k'}) - \varepsilon_\psi(T_k).\end{aligned} \tag{20}$$

This follows by adding and subtracting the two Bayes risks. More evidence cannot worsen the optimal decision, but its benefit for a fixed LLM agent also depends on how decision regret changes. No monotonicity is assumed for the latter. Under the soft-probability assumptions above, the corresponding identity is

$$\begin{aligned}\mathcal{L}_\psi(T_{k'}) - \mathcal{L}_\psi(T_k) = & -I(Y; E_{k'} \mid Q, E_k) \\ & + \mathcal{D}_\psi(T_{k'}) - \mathcal{D}_\psi(T_k).\end{aligned} \tag{21}$$

The observed top-k gains indicate a net benefit over the tested budgets. They neither estimate either component nor identify a universally optimal evidence budget.

C Complete PersonaMem-v2 Breakdowns

Tables 3 and 4 report every official preference type and conversation scenario for the main experiment. All three arms answer the same 5,000 questions at each length, and categories follow the official benchmark labels.

Table 3: Complete main-experiment accuracies (%) by official preference type. REF uses the reference-trained interpreter and MIX the retrieval-aligned interpreter.

History	Arm	Overall	Forget	Neutral note	Anti-stereo	Therapy	Health	Stereo	Sensitive
	Questions	5,000	1,048	858	855	627	568	533	511
32k	RAW	60.58	68.99	57.34	58.25	54.23	53.52	66.79	61.84
32k	REF	64.90	66.70	65.38	67.49	58.69	66.90	70.17	55.97
32k	MIX	66.38	71.37	66.78	67.72	58.37	69.19	69.98	56.16
128k	RAW	51.76	44.75	50.82	54.97	54.86	45.95	61.73	54.60
128k	REF	53.50	39.12	56.64	60.00	55.66	61.80	63.79	44.23
128k	MIX	54.68	43.51	56.53	62.11	57.89	60.74	63.23	42.47

Table 4: Complete main-experiment accuracies (%) by conversation scenario. KQ: knowledge query; PE: professional email; PW: professional writing; CW: creative writing; Tr.: translation; Pers.: personal email; TC: trouble consultation; Chat: chat message; Soc.: social-media post.

History	Arm	Overall	KQ	PE	PW	CW	Tr.	Pers.	TC	Chat	Soc.
	Questions	5,000	824	561	543	541	537	508	501	498	487
32k	RAW	60.58	58.98	61.68	58.20	61.37	65.36	62.40	60.88	55.82	61.19
32k	REF	64.90	71.12	64.53	64.27	64.51	65.55	65.16	62.08	56.22	66.74
32k	MIX	66.38	68.45	68.98	64.09	65.80	70.20	66.54	62.48	60.64	68.58
128k	RAW	51.76	53.64	55.61	48.62	49.17	54.93	54.13	49.30	48.80	50.10
128k	REF	53.50	64.44	52.05	51.20	51.57	54.38	49.41	52.30	47.39	52.16
128k	MIX	54.68	62.50	54.37	50.64	52.87	56.24	53.94	52.50	50.80	53.59