Memory Retrieval for Changing Preferences

Yuehan Qin, Shawn Li, Linxin Song, Wei Yang, Jiate Li, Yuqing Yang, Yue Zhao
University of Southern California
yuehanqi@usc.edu

Abstract

Long-context dialogue systems must decide both when to access memory and which parts of the interaction history are relevant. Existing approaches typically rely on heuristic retrieval signals or always-on memory usage, failing to account for the changing and potentially inconsistent nature of user preferences. In this work, we propose a unified framework for memory access and selection based on changing preferences. We formulate personalized memory retrieval as identifying which historical turns provide evidence about a user's latent preference state, rather than relying on surface-level semantic similarity. To this end, we quantify the utility of each memory turn using a Bayes factor, defined as the improvement in the model's likelihood of the reference response when the turn is included in context. This provides a principled measure of evidence strength and a unified signal for both memory access and selection. By framing memory retrieval as utility estimation, the model learns to identify salient turns and regulate memory usage based on expected utility. Experiments on four heterogeneous memory benchmarks show that our approach outperforms existing embedding-based retrieval on long-context, preference-intensive tasks where modeling changing preferences is essential, while remaining competitive in low-density regimes where semantic similarity suffices.

1 Introduction

As users engage with conversational AI systems across extended periods, they accumulate histories rich with preferences, experiences, and changing beliefs. A system that cannot selectively draw on this history when answering new queries fails to deliver the kind of continuity that personalization demands [Yang et al., 2026a,b, Li et al., 2026a, Yang et al., 2026c]. However, as interaction histories grow to span hundreds of sessions, attending to everything is computationally intractable [Ding et al., 2024, Li et al., 2025a, Ye et al., 2025, Tang et al., 2023], and attending to the wrong things actively degrades response quality [Liu et al., 2023, Shi et al., 2023, Amiraz et al., 2025]. This challenge is further exacerbated by the well-documented tendency of large models to hallucinate or rely on spurious cues when presented with noisy or misleading context [Shawn et al., 2025, Li and Zhao, 2026, Li et al., 2026b].

This makes memory retrieval a central challenge in personalized AI. Given a new query, the system must identify which past interactions carry genuine utility for producing a correct and contextually appropriate response. The approach that has emerged as standard practice is semantic embedding similarity: each memory turn and the incoming query are encoded, and the turns nearest in embedding space are selected as context [Chhikara et al., 2025, Wu et al., 2025, Li et al., 2025b, Xu et al., 2025, Tan et al., 2025a]. Semantic similarity is, however, a proxy for the wrong quantity [Li et al., 2026c]. It measures topical co-occurrence in embedding space, not whether a memory turn provides incremental evidence that is useful for resolving the current query. This distinction is inconsequential in many retrieval settings, but it becomes critical in personalization, where the same topic can appear across turns that reflect fundamentally different states of the user. This mismatch manifests in several structural failure modes. First, semantically similar turns may correspond to incompatible or superseded latent preference states, causing retrieval systems to surface context that is topically

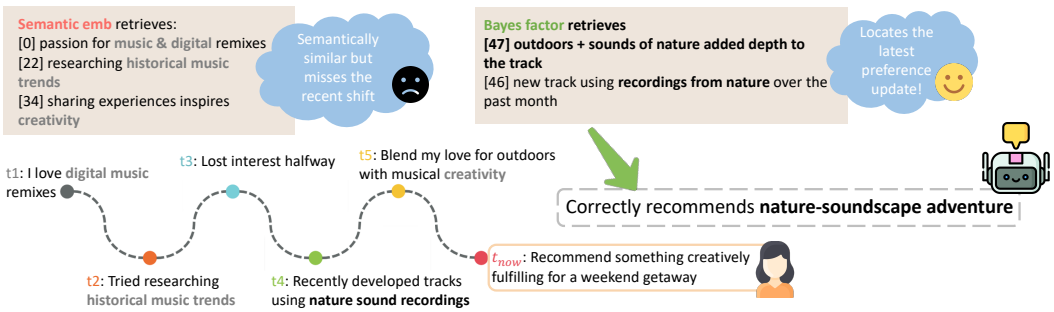


Figure 1: Illustration of retrieval methods for personalization memory. Given a generic recommendation request, semantic embedding retrieval returns surface-similar but evidentially uninformative turns (e.g., abstract self-reflections about creativity), leading the LLM to a generic recommendation. Our Bayes Factor-based retrieval instead locates the user's recent preference shift toward nature recording, enabling the LLM to produce the correct preference-aligned recommendation.

related yet no longer useful for the current query [Jiang et al., 2025, Zhao et al., 2025a]. Second, evidential utility is often sparse, temporally distant, or only weakly aligned with the query surface form. As a result, retrieval methods grounded primarily in semantic similarity tend to favor lexical or topical overlap over genuinely decision-relevant evidence. Similar biases toward surface-level correlations have also been observed in other structured prediction settings [Li et al., 2023, 2024a,b, 2025c]. An example is shown in Fig. 1.

This failure is symptomatic of a deeper issue in how existing memory systems decompose the retrieval problem. Most systems treat two coupled decisions independently: whether external memory should influence the response for a given query, and if so, which turns are worth retrieving [Packer et al., 2024, Chhikara et al., 2025, Tan et al., 2025a]. Access is typically handled through always-on or coarse heuristics, while selection relies on semantic similarity[Li et al., 2025d, Wu et al., 2025]. Neither decision is grounded in a direct estimate of memory utility for the current query. The result is a system that retrieves what is related rather than what is useful.

We argue that both decisions can be unified under a single principle: a memory turn should be retrieved if and only if it provides evidence that improves the correctness of the response, and memory should be accessed at all only when such evidence exists. We operationalize this principle through a Bayesian lens, treating each candidate turn as potential evidence and measuring how much it updates the model's belief toward the correct answer.

We summarize our contributions as below:

- **A principled retrieval signal for personalized memory.** We interpret utility gain as a Bayes factor, providing a principled alternative to semantic similarity that captures evidential utility rather than topical relatedness.
- **Bayes-factor-gated memory framework.** We propose a simple framework that couples (i) a discrete gating mechanism for deciding when to trigger memory access with (ii) a saliency-based selection mechanism for identifying relevant context, both driven by the same Bayes factor signal.
- **Evidence of distinctiveness.** Through controlled ablations, we show that Bayes-factor-based saliency labels are not interchangeable with embedding-based alternatives, with the largest gap on long-context preference tasks (+8.76pp on MemBench-High), and that supervised fine-tuning on these labels is essential for the observed gains.
- **Improved long-context personalization.** We demonstrate consistent improvements across multiple personalization benchmarks, with the largest gains on challenging instances requiring long-range memory reasoning.

2 Related Works

Large Language Models have emerged as personalized assistants across a wide range of tasks [Li et al., 2026d], from writing support to tailored recommendations [Salemi et al., 2024, Yang et al., 2025, Li

et al., 2025a, Gu et al., 2025, Zhao et al., 2025b, Li et al., 2025e]. As interaction histories accumulate across sessions, effectively leveraging them requires selective memory access: attending to everything is computationally intractable, while attending to the wrong turns degrades response quality [Liu et al., 2023]. Early memory systems such as MemGPT Packer et al. [2024] and MemoryBank Zhong et al. [2023] address context window limitations by maintaining external memory stores. More recent systems including Mem0 Chhikara et al. [2025] and A-MEM Xu et al. [2025] extract and consolidate facts from dialogue history, while RMM Tan et al. [2025a] introduces reflective summarization for finer retrieval granularity. These systems uniformly rely on semantic embedding similarity for retrieval, which we identify as a structural limitation under preference evolution.

Several works have proposed using utility gain as a context quality signal. RECOMP Xu et al. [2023] uses generator feedback for context compression, CODEFILTER Li et al. [2025f] applies likelihood-based impact scoring to code completion, and Liu et al. [2025] show that pointwise mutual information between context and query correlates with answer accuracy. Our work adapts this family of signals to personalized memory retrieval, where its interpretation as a Bayes factor over changing user preferences provides principled motivation beyond prior heuristic applications.

3 Methodology

Problem Setup Let $\mathcal{H} = (c_1, c_2, \ldots, c_T)$ denote a dialogue history, q a user query, and a the (unknown) answer. The goal is to select a subset of turns $F \subseteq \mathcal{H}$ such that a model conditioned on (q, F) can produce the correct answer. This problem involves two coupled decisions: (i) *whether* memory should be consulted, and (ii) *which* turns should be retrieved.

3.1 Stage 1: Answer-Conditioned Saliency

Following the Bayesian view of evidence, we treat $\mathrm{Sal}_a(c_i)$ as the *log Bayes factor* of turn c_i for answer a conditioned on query q: a turn constitutes genuine memory evidence if and only if observing it updates the answer distribution beyond what the query alone establishes. Formally, $p(a \mid q)$ is the model's prior belief over the answer given the query alone, and $p(a \mid q, c_i)$ is the posterior after observing turn c_i. Then $\mathrm{Sal}_a(c_i)$ is the log-ratio between the two:

$$\mathrm{Sal}_a(c_i \mid q, a) = \log p(a \mid q, c_i) - \log p(a \mid q), \tag{1}$$

which corresponds to the classical log Bayes factor quantifying how strongly c_i supports a. In practice, we calculate both terms from a smaller language model for computing efficiency.

A turn whose Bayes factor is near unity (e.g., an occupational fact when the query asks for a location) is correctly suppressed; a turn that drives the posterior sharply toward the correct answer receives high Sal_a. This stands in direct contrast to embedding-based retrieval, which measures query–turn similarity and thus cannot distinguish *evidential relevance* from *topical relatedness*.

Oracle Focus Set. We define the oracle focus set at threshold τ as

$$S_\tau \;=\; \big\{c_i \in \mathcal{H} :\; \mathrm{Sal}(c_i \mid q, a) > \tau\big\}, \tag{2}$$

where $\mathrm{Sal}(c_i \mid q, a)$ denotes the answer-conditioned saliency of turn c_i given query q and gold answer a, and τ is a threshold controlling the amount of retained evidence. S_τ represents an oracle notion of memory: it contains exactly those turns that improve answer prediction.

3.2 Stage 2: Query-Conditioned Proxy

$\mathrm{Sal}_a(c_i; q)$ depends on the unknown gold answer a, and is therefore not directly available at inference time. To obtain a tractable approximation, we replace the answer a with the observable query q, yielding a query-conditioned Bayes factor:

$$\mathrm{Sal}_q(c_i; q) \;=\; \log p(q \mid c_i) \;-\; \log p(q). \tag{3}$$

Here $p(q)$ is the model's prior probability of the query in isolation, and $p(q \mid c_i)$ is the likelihood of the query after observing turn c_i. Sal_q is therefore the log ratio of the two—a Bayes factor that measures how strongly c_i constitutes evidence for the query[1].

[1]Equivalently, $\mathrm{Sal}_q(c_i; q) = \mathrm{PMI}(q; c_i)$, connecting our signal to the family of pointwise mutual information context-quality measures studied in retrieval-augmented generation [Liu et al., 2025]

While Sal_q does not condition on the answer, we now show that it is a principled surrogate for Sal_a rather than a heuristic substitute. Our argument proceeds in three steps: an exact decomposition (Proposition 1), a sufficient condition for rank-equivalence (Theorem 3), and a bounded approximation result for the realistic case (Theorem 4).

An exact decomposition. By the chain rule of Bayesian updating, the joint Bayes factor of c_i for the pair (a, q) decomposes additively:

Proposition 1 (Joint Bayes factor decomposition)**.** *For any candidate turn c_i, query q, and answer a,*

$$\underbrace{\log \frac{p(a, q \mid c_i)}{p(a, q)}}_{\mathrm{BF}_{(a,q)}(c_i)} = \underbrace{\log \frac{p(a \mid q, c_i)}{p(a \mid q)}}_{\mathrm{Sal}_a(c_i)} + \underbrace{\log \frac{p(q \mid c_i)}{p(q)}}_{\mathrm{Sal}_q(c_i)}. \tag{4}$$

The identity (4) is exact and assumption-free. It shows that Sal_q is not an unrelated heuristic but rather one of two additive components of the joint evidence c_i provides for (a, q); ranking by Sal_q is therefore equivalent to ranking by $\mathrm{BF}_{(a,q)} - \mathrm{Sal}_a$, and the question of when this ranking agrees with Sal_a reduces to a question about the relative magnitudes of the two terms.

When does the surrogate preserve the ranking? We now identify a condition under which Sal_q and Sal_a induce the same ordering on candidate turns up to a constant.

Definition 2 (Query-conditional sufficiency)**.** *A turn c_i is* query-sufficient *for (q, a) if $q \perp a \mid c_i$, i.e.,*

$$p(a \mid q, c_i) \;=\; p(a \mid c_i). \tag{5}$$

Intuitively, c_i is query-sufficient when it already encodes the user-state information needed to determine a, so that the query merely *triggers* retrieval rather than supplying additional evidence. This holds naturally for the kind of turns personalized memory aims to retrieve: declarative preference statements, persona attributes, and committed user states.

Theorem 3 (Rank-equivalence under sufficiency)**.** *Let $\mathcal{C} = \{c_1, \ldots, c_n\}$ be a candidate set in which every turn is query-sufficient (Def. 2). Then for all $c_i, c_j \in \mathcal{C}$,*

$$\mathrm{Sal}_a(c_i) - \mathrm{Sal}_a(c_j) \;=\; \mathrm{Sal}_q(c_i) - \mathrm{Sal}_q(c_j) \;+\; \big[\mathrm{Sal}_a^{\mathrm{marg}}(c_i) - \mathrm{Sal}_a^{\mathrm{marg}}(c_j)\big], \tag{6}$$

where $\mathrm{Sal}_a^{\mathrm{marg}}(c_i) = \log p(a \mid c_i) - \log p(a)$ is the marginal answer Bayes factor. In particular, if $\mathrm{Sal}_a^{\mathrm{marg}}$ is approximately constant across $\mathcal{C}$, then ranking by Sal_q recovers the Sal_a ranking exactly.

The constant-$\mathrm{Sal}_a^{\mathrm{marg}}$ assumption is mild whenever the candidate pool has been pre-filtered to a topically coherent slice of the user's history, which is precisely what the query-side top-K pre-filter in our inference pipeline (Eq. 3.5) accomplishes.

The realistic case. Strict query-conditional sufficiency rarely holds, but the disagreement between Sal_q- and Sal_a-induced rankings can be bounded in terms of a per-turn sufficiency residual $\varepsilon_i := |\log p(a \mid q, c_i) - \log p(a \mid c_i)|$ and the marginal-evidence gap. We state and prove the precise bound in Appx. E (Theorem 4); the practical takeaway is that Sal_q is faithful when turns are self-contained and degrades on instances requiring cross-turn aggregation.

3.3 Stage 3: Threshold-Gated Memory

We convert the continuous saliency signal into a discrete decision via a threshold τ:

$$S_\tau(\mathcal{H}, q) = \{\, c_i \in \mathcal{H} : \mathrm{Sal}_q(c_i; q) > \tau \,\}, \quad g_\tau(\mathcal{H}, q) = \mathbf{1}[\,|S_\tau(\mathcal{H}, q)| > 0\,]. \tag{7}$$

Here S_τ is the set of *salient chunks* retained as evidence, and $g_\tau \in \{0, 1\}$ is the memory **trigger** indicating whether the memory-guided policy should fire on this context.

Why couple Sal_a and τ? Sal_a alone provides a continuous utility score but no decision boundary, so the model cannot learn a discrete retrieval policy. A fixed threshold alone provides a gate but no ranking, so the model cannot prioritize among retained turns. Coupling them gives a calibrated labeling mechanism that (i) filters low-utility turns via τ, (ii) ranks the survivors by their Bayes factor score, and (iii) exposes both the relevance label and the ranking as explicit supervision targets for the controller. This is the central design choice of our silver label construction.

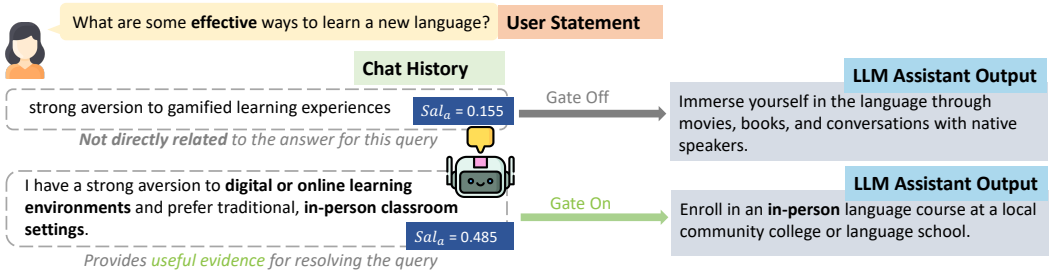


Figure 2: Illustration of threshold-gated memory access. Although both memory turns are semantically related to the user query, only the lower example provides evidence that materially changes the model's belief about the appropriate response, resulting in a higher Sal_a score and activation of memory-guided reasoning.

Implications for the pipeline. Theorem 3 delineates the regime in which Sal_q is a faithful surrogate: candidate sets dominated by self-contained user-state turns. Outside this regime where instances requiring cross-turn aggregation, the supervised controller (§ 3.4.1), trained on Sal_a silver labels, recovers ranking quality that Sal_q alone cannot. The two-stage design (Sal_q pre-filter $\rightarrow$ Sal_a-trained controller) is therefore a direct response to this structure.

3.4 Stage 4: Learning the Retrieval Policy

3.4.1 Training pipeline.

The trigger g_τ and the salient set S_τ are exposed to a *single* fine-tuned policy through two paired supervised signals constructed from the same $(\mathcal{C}, S_\tau)$. We instantiate them in three stages, all sharing parameters under one next-token-prediction objective.

Controller head (memory-gating). Given τ, the user query q, and the candidate evidence S_τ, the model is trained to emit a structured action

$$a \;=\; \big(\text{trigger} = g_\tau,\; \text{focus_idx} = \mathrm{TopK}_{\mathrm{Sal}_a}(S_\tau),\; \text{policy} = \pi^\star\big), \tag{8}$$

wrapped in a brief natural-language $\langle\text{thought}\rangle$ followed by a JSON $\langle\text{action}\rangle$. Here $\mathrm{TopK}_{\mathrm{Sal}_a}$ ranks the candidate chunks by their answer-conditioned saliency score. To prevent the controller from collapsing to $g_\tau \equiv 1$, we balance with negative samples (details in Appx. C).

Reader head (MCQ answering). Conditioned on the same S_τ and the four candidate options $\mathcal{O}$, the model is trained to produce the gold answer terminated by $\langle\text{final_answer}\rangle(x)$.

Joint objective. Both heads share parameters and are trained jointly with a single next-token-prediction objective on the union of controller and reader samples:

$$\mathcal{L}_{\text{SFT}} \;=\; -\sum_{(x,y)\in\mathcal{D}_{\text{ctrl}}\cup\mathcal{D}_{\text{read}}}\;\sum_{t}\log p_\theta\big(y_t \mid y_{<t}, x\big). \tag{9}$$

3.5 Inference Pipeline

During inference, we have a query q together with the user's full conversation history $\mathcal{H} = \{c_1, \ldots, c_n\}$ up to q, but neither the gold answer a nor the answer-conditioned saliency $\mathrm{Sal}(c_i \mid q, a)$ are observable. Our pipeline therefore relies on the controller as a learned proxy for the training-time selection rule, and operates in three stages.

Query-side pre-filtering. We first reduce the search space using the precomputed query saliency:

$$\tilde{\mathcal{H}} \;=\; \mathrm{TopK}_{\mathrm{Sal}_q}(\mathcal{H}; K), \tag{10}$$

$\tilde{\mathcal{H}}$ is presented to the controller in temporal order, annotated with each turn's original `message_idx`.

Controller as learned gate. Given $(q, \tilde{\mathcal{H}})$, the controller emits a structured output consisting of a $\langle \text{thought} \rangle$ block followed by an $\langle \text{action} \rangle$ JSON object:

$$(\hat{g}_\tau,\ \hat{S}_\tau)\ =\ \text{Controller}(q, \tilde{\mathcal{H}}), \qquad \hat{S}_\tau \subseteq \tilde{\mathcal{H}},\ \ |\hat{S}_\tau| \le k, \tag{11}$$

where $\hat{g}_\tau \in \{0, 1\}$ approximates the training gate $g_\tau = \mathbf{1}[\,\exists\, c_i \text{ s.t. } y_i = 1\,]$ and $\hat{S}_\tau$ approximates the focus set $F^\star = \{c_i : y_i = 1\}$. Crucially, no Sal_a computation is performed at inference; the controller has internalized the τ-thresholded selection rule through supervised training, realizing **Bayes-factor-guided memory selection** purely from self-contained dialogue features.

Gated reading. The reader is invoked conditionally on the predicted gate:

$$\hat{y}\ =\ \begin{cases} \text{Reader}\big(q,\ \mathcal{O},\ \mathcal{H}_{\hat{S}_\tau}\big) & \text{if } \hat{g}_\tau = 1, \\ \text{Reader}\big(q,\ \mathcal{O},\ \varnothing\big) & \text{if } \hat{g}_\tau = 0, \end{cases} \tag{12}$$

where $\mathcal{O}$ denotes the four MCQ options and $\mathcal{H}_{\hat{S}_\tau}$ denotes the user turns indexed by $\hat{S}_\tau$, presented in temporal order. The reader produces a free-form rationale terminating in a $\langle \text{final_answer} \rangle$ token from which the predicted option letter is parsed. When $\hat{g}_\tau{=}0$ the reader falls back to query-only answering, which prevents spurious memory injection on questions whose answer is not memory-bound.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on four benchmarks: MemBench-Low/High [Tan et al., 2025b], where Low tests explicit recall and High requires implicit inference; PersonaMem [Jiang et al., 2025], evaluating personalized responses under evolving user profiles; and PrefEval [Zhao et al., 2025a], evaluating user-preference recall. We apply a seed-42 stratified 80/10/10 split. For MB-Low we report accuracy on the user-only subset ($n{=}374$), excluding `*-assistant` scenarios whose evidence falls outside $\mathcal{H}$. Full details are in Appx. A.

Baselines. We compare against five retrieval baselines. **Full context** concatenates all user turns and passes them directly to the reader. **A-MEM** Xu et al. [2025], **MemoryBank** Zhong et al. [2023], and **Mem0** Chhikara et al. [2025] represent memory-augmented retrieval systems. **RMM** Tan et al. [2025a] organizes dialogue history into reflective multi-granularity memory summaries and retrieves evidence via embedding similarity. All retrieved memories are substituted into the same reader prompt as our method. Following prior work, we omit RMM's retrospective reranking stage and MemoryBank's forgetting update, as neither is applicable under single-shot MCQ evaluation. For each selector, we evaluate two reader models: `Qwen3.5-9B-thinking` [Team, 2026] and `Llama-3.1-8B-Instruct` [Grattafiori et al., 2024]. The Sal_a and Sal_q scores are computed using `Qwen3.5-0.8B`.

Evaluation Protocol. We report MCQ accuracy. The reader emits free-form text terminated by $\langle \text{final_answer} \rangle(\ell)$ with $\ell \in \{a, b, c, d\}$; we extract the letter and compare it with the gold label. If parsing fails, we fall back to scanning the response for any unique `(a)/(b)/(c)/(d)` mention. MB-Low accuracy uses the user-only subset ($n{=}374$ test records); the other three benchmarks use all kept test records.

4.2 Main Results

Our Method Achieves the Best Overall Performance. Tab. 1 summarizes performance across four benchmarks. Across both reader backbones, our method consistently achieves the strongest average accuracy, substantially outperforming all retrieval-augmented baselines as well as the full-context setting. Under Qwen3.5-9B-Thinking, our method achieves an average score of 84.94, surpassing the strongest baseline, MemoryBank (65.64), by +19.30 points, while improving over the full-context baseline (64.09) by +20.85 points. Similar trends hold under Llama-3.1-8B-Instruct, where our method reaches 82.05 average accuracy compared to 71.65 for MemoryBank and 59.74 for full-context prompting.

Model	Method	MB-Low	MB-High	Persona	PrefEval	Avg
Qwen3.5-9B-Thinking	full ctx	64.17	56.70	36.49	98.99	64.09
	RMM	36.36	54.12	50.00	79.80	57.60
	Mem0	42.51	61.34	41.89	83.84	62.86
	A-mem	53.29	64.89	55.71	86.46	65.09
	MemoryBank	51.47	69.58	52.86	88.54	65.64
	Ours	**88.78**	**77.32**	**75.68**	97.98	**84.94**
Llama-3.1-8B-Instruct	full ctx	55.35	56.19	48.65	78.79	59.74
	RMM	24.60	61.34	40.54	88.89	53.84
	Mem0	38.88	61.34	58.11	94.95	63.32
	A-mem	56.95	69.07	66.22	92.93	71.29
	MemoryBank	57.49	71.65	67.57	89.90	71.65
	Ours	**82.62**	68.56	**77.03**	**99.99**	**82.05**

Table 1: Main results and ablations. Best results per benchmark are in **bold**.

Gains are concentrated on inference-heavy benchmarks. Performance gains are especially pronounced on MB-High and Persona, where our method consistently achieves the strongest improvements across both reader backbones. Under Qwen3.5-9B-Thinking, our method improves from 69.58 to 77.32 on MB-High (+7.74) and from 55.71 to 75.68 on Persona (+19.97) relative to the strongest retrieval baseline, MemoryBank. Similar trends hold under Llama-3.1-8B-Instruct, where our method improves from 67.57 to 77.03 on Persona (+9.46). These benchmarks share a common structure: the answer depends on recovering implicit, temporally distant, or weakly lexicalized preference evidence rather than retrieving a single semantically aligned turn. By contrast, the gap narrows on PrefEval, where preferences are often stated explicitly and semantic retrieval is already effective. Under Qwen3.5-9B-Thinking, MemoryBank reaches 88.54 while our method achieves 97.98 (+9.44). These results suggest that the advantage of our method is largest when retrieval depends on identifying high-utility evidence among semantically related but uninformative history.

Notably, our method also consistently outperforms the full-context baseline despite conditioning on substantially fewer dialogue turns. Under Qwen3.5-9B-Thinking, full-context prompting achieves only 64.09 average accuracy compared to our 84.94 (+20.85), while under Llama-3.1-8B-Instruct the gap is similarly large (59.74 vs. 82.05, +22.31). This suggests that the gains do not arise from increased memory access alone, but from selectively isolating the small subset of interactions that genuinely contribute to the downstream response. Together, these results support the view that effective personalized memory retrieval depends less on maximizing topical coverage than on identifying evidence with high answer utility.

Fig. 3 compares our method against Mem0 and RMM across twelve representative sub-categories. The largest gains appear on PersonaMem dating/recommendation tasks (+33 to +71) and MB-Low aggregative/comparative tasks (+48 to +77), where answers depend on sparse or distributed evidence rather than direct semantic overlap. MB-High and PrefEval further highlight the difficulty of long-context preference and sentiment recall, where relevant evidence is often overwhelmed by topically related but evidentially irrelevant dialogue. Conversely, when retrieval reduces to straightforward query–passage similarity, RMM is already competitive and the remaining improvement becomes much smaller. These patterns support our central claim that effective personalized memory retrieval depends less on topical overlap than on identifying evidence with genuine answer utility.

Bayes factors capture evidential rather than topical relevance. Fig. 4 visualizes the retrieval distributions induced by semantic similarity and saliency-based scoring across four representative benchmarks, with additional qualitative retrieval examples provided in Appx. Tab. 14. Across all cases, BGE assigns high scores to turns that are lexically or topically aligned with the query, even when those turns contribute little evidence toward resolving the underlying preference or decision. In contrast, Sal_a consistently concentrates on a small number of temporally sparse yet decision-critical turns, often prioritizing evidence that semantic retrieval ranks far below the top. This contrast reflects the different retrieval objectives induced by the two signals: semantic similarity favors surface alignment, whereas Bayes-factor saliency measures whether conditioning on a memory turn materially shifts the model's belief toward the correct response. Effective personalized memory retrieval therefore depends less on topical relatedness than on identifying the sparse pieces of context that carry genuine evidential utility.

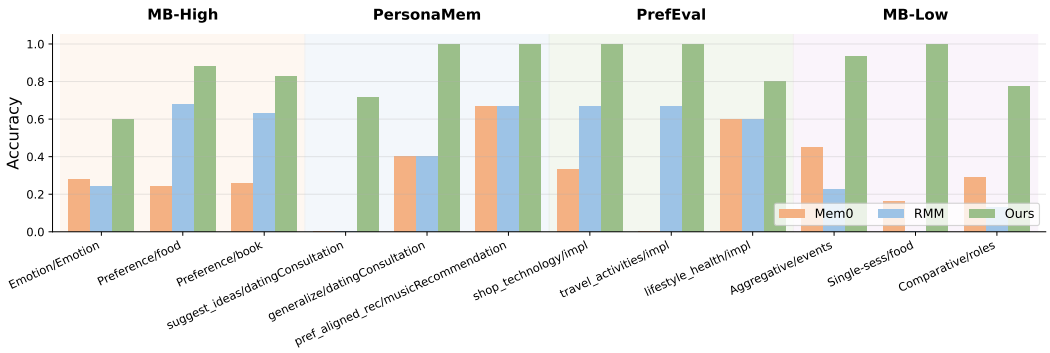


Figure 3: **Sub-category breakdown across four benchmarks.** Accuracy of Mem0, RMM, and our method on twelve representative sub-categories;three per benchmark, drawn from MB-High, PersonaMem, PrefEval, and MB-Low. Sub-categories are selected to span the regimes where preference evidence is implicit (PersonaMem dating/music consultation, PrefEval implicit-choice), aggregated across scattered turns (MB-Low `Aggregative/events`, `Comparative/roles`), or buried in long context (MB-High `Preference/{book,food}`, `Emotion`). Our method outperforms the stronger baseline by 20 to 77 absolute points on every selected sub-category.

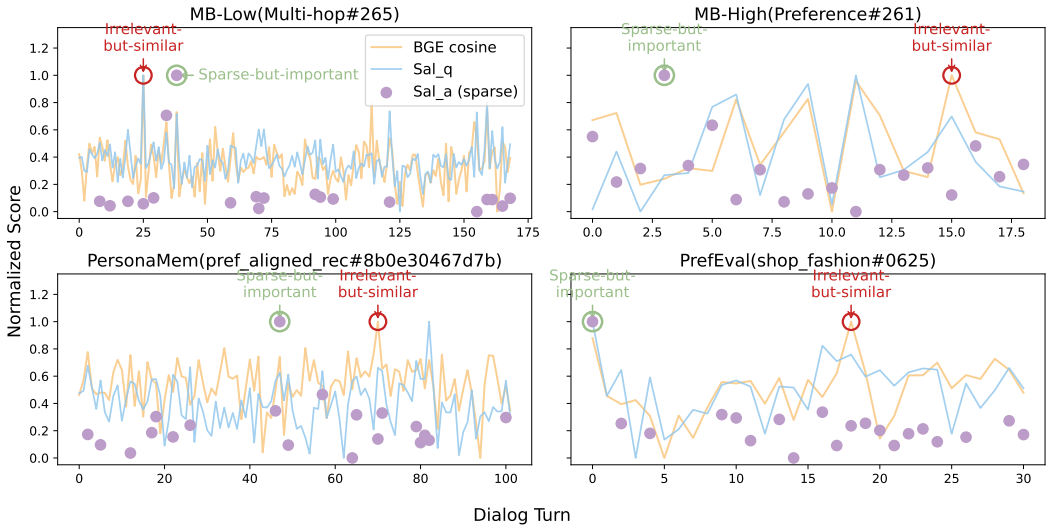


Figure 4: Per-benchmark comparison of retrieval signals over a single conversation. BGE peaks Irrelevant-but-similar tend to land on turns that are lexically aligned with the query but carry no evidence for the gold answer, whereas Sal_a peaks Sparse-but-important concentrate on the few turns that actually shift the best answer likelihood.

4.3 Ablation Study

We conduct three ablations to assess the contribution of each component.

Supervised fine-tuning is essential. According to Tab. 2, removing supervised fine-tuning and using the base Qwen3.5-9B-Thinking model with Sal_a scoring at inference time collapses performance to 66.09 avg. This demonstrates that the utility signal must be internalized through training rather than applied post-hoc; an untuned model cannot reliably identify high-utility memory turns at inference.

Sal_a **is the critical signal** (Tab. 3). Replacing Sal_a-derived labels with embedding-based pseudo-labels (using cosine similarity to identify relevant turns) reduces average accuracy from 84.94 to 82.55 (-2.39 points). The gap is most pronounced on MB-High (77.32 $\rightarrow$ 68.56, -8.76 points), the benchmark demanding the most implicit preference inference. This confirms that Sal_a captures a

Method	MB-Low	MB-High	Persona	PrefEval	Avg
Qwen3.5 + PPL	25.67 (-63.11)	57.22 (-20.10)	39.19 (-36.49)	40.40 (-57.58)	40.62 (-44.32)
+ BGE	59.63 (-29.15)	59.28 (-18.04)	56.76 (-18.92)	73.74 (-24.24)	62.35 (-22.59)
+ full ctx	64.17 (-24.61)	56.70 (-20.62)	36.49 (-39.19)	**98.99** (+1.01)	64.09 (-20.85)
+ Sal_a	60.16 (-28.62)	64.95 (-12.37)	59.46 (-16.22)	79.80 (-18.18)	66.09 (-18.85)
Ours	**88.78**	**77.32**	**75.68**	97.98	**84.94**

Table 2: Main results on Qwen3.5-9B-think. Methods differ only in how evidence is selected from the user-turn history: *PPL* ranks turns by per-turn cross-entropy, *BGE* by query–turn cosine similarity, *full ctx* uses every user turn, and Sal_a is the oracle answer-conditioned saliency upper bound. Numbers are accuracy (%); parentheses indicate the gap to **Ours** in percentage points.

Method	MB-Low	MB-High	Persona	PrefEval	Avg
EMB labels (replace Sal_a)	89.98 (+1.20)	68.56 (-8.76)	75.68 (0.00)	95.96 (-2.02)	82.55 (-2.39)
w/o Controller	89.58 (+0.80)	73.71 (-3.61)	68.92 (-6.76)	**99.99** (+2.01)	84.90 (-0.04)
Ours	88.78	**77.32**	**75.68**	97.98	**84.94**

Table 3: Ablations on Qwen3.5-9B-think. During the supervised fine-tuning stage, the *EMB labels* replaces the Sal_a silver labels with query–turn embedding similarity, isolating the supervision signal. *w/o Controller* removes the controller and selects evidence directly from the saliency stage, isolating the controller during training. Numbers are accuracy (%); parentheses indicate the gap to **Ours** in percentage points.

qualitatively different supervision signal from semantic similarity: it reflects whether a memory turn actually reduces answer perplexity, rather than whether it is topically close to the query.

The controller provides marginal but consistent benefit (Tab. 3). Removing the controller while retaining other components yields 84.90 avg, a negligible drop of 0.04 points overall. However, per-benchmark analysis reveals that the controller provides a more consistent benefit on MB-High (77.32 vs. 73.71, $+3.61$ points), suggesting its value is concentrated in harder retrieval scenarios where gating decisions are more consequential.

Retrieval vs. reader adaptation. The performance gap between our method and retrieval-only baselines should not be interpreted as a retrieval gain alone. Empirically, the oracle selector Sal_a combined with the thinking-enabled reader recovers much of the performance, suggesting that our learned $\mathrm{Sal}_q \rightarrow$ controller $\rightarrow$ focus-set pipeline approximates the oracle retrieval distribution. This indicates that much of the remaining improvement comes from jointly training the reader and retrieval controller, rather than retrieval quality alone.

We view this as an important property of the framework rather than a confounder. Existing memory systems such as Mem0 or RMM rely on fixed extraction pipelines, external memory stores, or online reflection policies that are not naturally amenable to end-to-end supervised fine-tuning. In contrast, our framework jointly trains retrieval and reading, allowing the reader to adapt to compressed evidence representations. Thus, the observed gains arise not only from better evidence selection, but also from co-adaptation between memory selection and downstream reasoning.

5 Conclusion

We proposed a Bayes-factor-based framework for personalized memory retrieval that unifies memory access gating and turn selection under a single principled signal. Unlike embedding-based approaches that measure topical similarity, our method estimates the evidential utility of each memory turn for predicting the correct answer. By jointly training a controller and reader on Bayes-factor-derived proxy labels, the framework learns when to retrieve and what to retrieve without requiring gold answers at inference time. Experiments on four personalization benchmarks show consistent improvements over existing memory systems, with the largest gains on long-context, implicit-preference tasks where semantic similarity fundamentally fails. These results highlight the importance of optimizing for answer utility rather than topical relatedness in memory-augmented dialogue systems.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award Nos. 2428039, 2346158, and 2449280. This work was also supported in part by Amazon Research Awards, the NVIDIA Academic Grant Program, and the Foresight Institute. This work used NCSA Delta GPU at the National Center for Supercomputing Applications (award OAC 2005572) through allocations CIS251004 and CIS260196 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [Boerner et al., 2023], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot (allocation NAIRR260125) and Google Cloud for contributing to this research result. This work used Google Cloud through the CloudBank project [Norman et al., 2021], which is supported by National Science Foundation grant #1925001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the other funding organizations.

References

Yuqing Yang, Tengxiao Liu, Wang Bill Zhu, Taiwei Shi, Linxin Song, and Robin Jia. Self-evolving llm memory extraction across heterogeneous tasks, 2026a. URL https://arxiv.org/abs/2604.11610.

Wei Yang, Bryce Kan, Shixuan Li, Li Li, Yuehan Qin, Jiate Li, Paul Bogdan, and Jesse Thomason. Ramem: Contextual reinstatement for long-term agentic memory, 2026b. URL https://arxiv.org/abs/2606.22844.

Shixuan Li, Wei Yang, Peiyu Zhang, Anzhe Cheng, Heng Ping, Xiaole Zhang, Jesse Thomason, Xuezhe Ma, Yue Zhao, and Paul Bogdan. From relevant memories to sufficient evidence: Rethinking memory utilization in llm agents, 2026a. URL https://arxiv.org/abs/2610.22668.

Wei Yang, Shawn Li, Yuehan Qin, Yawei Wang, Mingxi Wang, Shixuan Li, Tiankai Yang, Jiate Li, Jesse Thomason, Xuezhe Ma, and Yue Zhao. Synco: Data synthesis co-training for self-evolving llms via multi-agent reinforcement learning, 2026c. URL https://arxiv.org/abs/2610.11345.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024. URL https://arxiv.org/abs/2402.13753.

Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, Yu Wang, Xiangliang Zhang, Philip S. Yu, Xiyang Hu, and Yue Zhao. A personalized conversational benchmark: Towards simulating personalized conversations, 2025a. URL https://arxiv.org/abs/2505.14106.

Dongze Ye, Haipeng Luo, Carolee Winstein, and Nicolas Schweighofer. Towards ai-based precision rehabilitation via contextual model-based reinforcement learning. *Journal of NeuroEngineering and Rehabilitation*, 22(1):263, 2025.

Dengwang Tang, Dongze Ye, Rahul Jain, Ashutosh Nayyar, and Pierluigi Nuzzo. Posterior sampling-based online learning for episodic pomdps. *arXiv preprint arXiv:2310.10107*, 2023.

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, 2023. URL https://arxiv.org/abs/2307.03172.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context, 2023. URL https://arxiv.org/abs/2302.00093.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 18228–18258. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.892. URL http://dx.doi.org/10.18653/v1/2025.acl-long.892.

Li Shawn, Jiashu Qu, Linxin Song, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual VLMs: A causal approach to hallucination. In *EMNLP*, 2025.

Shawn Li and Yue Zhao. The autonomy tax: Defense training breaks llm agents, 2026. URL https://arxiv.org/abs/2603.19423.

Shawn Li, Chenxiao Yu, Zhiyu Ni, Hao Li, Charith Peris, Chaowei Xiao, and Yue Zhao. Defenses against prompt attacks learn surface heuristics. In *ACL*, 2026b.

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

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory, 2025. URL https://arxiv.org/abs/2410.10813.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *CVPR*, 2025b.

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

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Shawn Li, You Qin, Jiate Li, Charith Peris, Lisa Bauer, Roger Zimmermann, and Yue Zhao. Geometry over density: Few-shot cross-domain ood detection. In *NeurIPS*, 2026c.

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, 2025. URL https://arxiv.org/abs/2504.14225.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms, 2025a. URL https://arxiv.org/abs/2502.09597.

Li Li, Chenwei Wang, You Qin, Wei Ji, and Renjie Liang. Biased-predicate annotation identification via unbiased visual predicate representation. In *ACM MM*, 2023.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. *AAAI*, 38(4):3145–3153, Mar. 2024a. doi: 10.1609/aaai.v38i4.28098.

Li Li, You Qin, Wei Ji, Yuxiao Zhou, and Roger Zimmermann. Domain-wise invariant learning for panoptic scene graph generation. In *ICASSP*, pages 3165–3169, 2024b. doi: 10.1109/ICASSP48485.2024.10447193.

Shawn Li, Ryan Rossi, Sungchul Kim, Sunav Choudhary, Franck Dernoncourt, Puneet Mathur, Zhengzhong Tu, and Yue Zhao. Charts are not images: On the challenges of scientific chart editing. In *ICLR*, 2025c.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue, 2025d. URL https://arxiv.org/abs/2406.05925.

Shawn Li, Chenxiao Yu, Han Wang, Wei Yang, Ryan Rossi, Franck Dernoncourt, Xiyang Hu, Philip Yu, Chaowei Xiao, Huan Zhang, and Yue Zhao. Fortis: Benchmarking over-privilege in agent skills, 2026d. URL https://arxiv.org/abs/2605.09163.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization, 2024. URL https://arxiv.org/abs/2304.11406.

Wei Yang, Muyan Weng, Jiacheng Pang, Defu Cao, Heng Ping, Peiyu Zhang, Shixuan Li, Yue Zhao, Qiang Yang, Mengdi Wang, et al. Toward evolutionary intelligence: Llm-based agentic systems with multi-agent reinforcement learning. *Available at SSRN 5819182*, 2025.

Hao Gu, Rui Zhong, Yu Xia, Wei Yang, Chi Lu, Peng Jiang, and Kun Gai. R 4ec: A reasoning, reflection, and refinement framework for recommendation systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 411–421, 2025.

Rui Zhao, Rui Zhong, Haoran Zheng, Wei Yang, Chi Lu, Beihong Jin, Peng Jiang, and Kun Gai. Hierarchical sequence id representation of large language models for large-scale recommendation systems. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 641–650, 2025b.

Shawn Li, Peilin Cai, Yuxiao Zhou, Zhiyu Ni, Renjie Liang, You Qin, Yi Nian, Zhengzhong Tu, Xiyang Hu, and Yue Zhao. Secure on-device video ood detection without backpropagation. In *ICCV*, 2025e.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023. URL https://arxiv.org/abs/2305.10250.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation, 2023. URL https://arxiv.org/abs/2310.04408.

Yanzhou Li, Shangqing Liu, Kangjie Chen, Tianwei Zhang, and Yang Liu. Impact-driven context filtering for cross-file code completion. *arXiv preprint arXiv:2508.05970*, 2025f.

Tianyu Liu, Jirui Qi, Paul He, Arianna Bisazza, Mrinmaya Sachan, and Ryan Cotterell. Pointwise mutual information as a performance gauge for retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 1628–1647. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.78. URL http://dx.doi.org/10.18653/v1/2025.naacl-long.78.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents, 2025b. URL https://arxiv.org/abs/2506.21605.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing*, pages 173–176. Association for Computing Machinery, 2023. doi: 10.1145/3569951.3597559.

Michael Norman, Vince Kellen, Shava Smallen, Brian DeMeulle, Shawn Strande, Ed Lazowska, Naomi Alterman, Rob Fatland, Sarah Stone, Amanda Tan, Katherine Yelick, Eric Van Dusen, and James Mitchell. CloudBank: Managed services to simplify cloud access for computer science research and education. In *Practice and Experience in Advanced Research Computing*, pages 1–4. Association for Computing Machinery, 2021. doi: 10.1145/3437359.3465586.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

A Dataset preprocessing

Filtering. After running the saliency stage [2] of Section 3.1–3.3, instances for which $|S_\tau| = 0$ *and* the maximum reach-top-1 log-probability falls below a secondary threshold are dropped, as these contain no usable evidence even at the turn level.

Splits. Remaining instances are partitioned with a fixed seed-42 stratified $80/10/10$ split. PersonaMem additionally enforces persona disjointness: all $(q, \mathcal{H})$ pairs sharing the same `persona_id` stay in the same split, preventing test-time leakage of memorized persona traits across the 20 personas (14 train / 3 val / 3 test). Per-benchmark counts are listed in Table 4.

MB-Low `*-assistant` scenarios. 25% of MB-Low test records (124/499) lie in `Single-session-assistant` or `Multi-session-assistant` scenarios, which ask about the assistant's past *recommendations* rather than the user's own statements. Because $\mathcal{H}$ in our setup contains only user turns, the answer content for these records lies in assistant turns that have been removed from the evidence pool, making them ill-posed under our problem formulation (Section 3). We therefore report MB-Low accuracy on the user-only subset (n=374); a full-context formulation that retains assistant turns would be required for fair evaluation on `*-assistant` scenarios and is beyond this work's scope.

	MB-Low	**MB-High**	**PersonaMem**	**PrefEval**
Train (kept)	1,871	977	431	801
Val (kept)	375	194	84	100
Test (kept)	374	194	74	99
$\lvert\mathcal{D}_{\text{ctrl}} \cup \mathcal{D}_{\text{read}}\rvert$ (Train)	4,069	2,358	1,022	1,932
of which $g_\tau = 1$	1,628	943	409	773
of which $g_\tau = 0$	813	472	204	386

Table 4: Per-benchmark split sizes after the saliency drop filter. For MB-Low we additionally exclude `*-assistant` scenarios (25% of records) at the dataset construction stage, so that the training distribution matches the user-only evaluation protocol (§4.1). Each kept training instance contributes one controller sample and, when $g_\tau = 1$, one reader sample. Negative controller samples ($g_\tau = 0$) are described in §3.4.1.

B Additional Experiments

Isolating the contribution of the retrieval signal. To distinguish the contribution of saliency-based retrieval from downstream adaptation, we compare Sal_q with BGE retrieval using the same frozen Qwen3.5-9B reader in thinking mode. Neither setting uses reader fine-tuning or a trained controller; only the retrieved evidence is changed. As shown in Table 5, Sal_q outperforms BGE on all four benchmarks, with a macro-average improvement of 5.65 percentage points. These results demonstrate that the retrieval signal provides benefits independently of reader adaptation. The larger gains of the full system reflect the combined effects of evidence selection and downstream adaptation.

Open-ended personalized generation. To evaluate applicability beyond multiple-choice questions, we additionally consider the single-session-preference subset of LongMemEval [Wu et al., 2025], comprising 30 questions with free-form answers evaluated by a GPT-4o judge. We use frozen readers without task-specific fine-tuning, and retrieval relies only on Sal_q, without access to reference answers. We also evaluate iterative RAG under a matched inference budget: it performs semantic retrieval, uses an additional query-rewriting pass, and retrieves again before answer generation.

Table 6 shows that Sal_q top-5 retrieval improves Llama-3.1-8B accuracy from 6.7% with full-context input to 60.0%, compared with 43.3% for iterative RAG. With Qwen3.5-9B in thinking mode, both retrieval methods achieve 66.7%, compared with 63.3% for full-context input. These results provide initial evidence that query-conditioned saliency also supports open-ended personalized generation.

[2] We release the saliency labels at `https://github.com/yqin43/saliency` to support reproducibility.

Table 5: Retrieval comparison with the same frozen Qwen3.5-9B reader in thinking mode, without reader fine-tuning or a trained controller. Scores are accuracy (%). Avg. denotes the unweighted mean across the four benchmarks.

Retrieval	MB-Low	MB-High	Persona	PrefEval	Avg.
BGE	59.6	59.3	56.8	73.7	62.35
Sal_q	**61.4**	**61.9**	**60.8**	**87.9**	**68.00**

Table 6: Open-ended generation accuracy (%) on the 30-question single-session-preference subset of LongMemEval, evaluated by a GPT-4o judge. Both readers are frozen and receive no task-specific fine-tuning. Full-context inputs contain approximately 105k tokens.

Reader	Full context	Sal_q top-5	Iterative RAG	Oracle
Llama-3.1-8B	6.7	**60.0**	43.3	70.0
Qwen3.5-9B (thinking)	63.3	**66.7**	**66.7**	80.0

Inference efficiency. We measure the combined inference cost of online Sal_q scoring, evidence selection, and answer generation, and compare it with full-context reading. As shown in Table 7, selective retrieval reduces measured latency across all four benchmarks, with FLOPs reductions ranging from $1.5\times$ to $5.4\times$. Thus, in the evaluated configurations, the cost of online scoring is offset by the reduction in reader context. These measurements characterize the Sal_q selection-and-reading configuration.

Statistical reliability. We repeat training and evaluation with Llama-3.1-8B across three split/seed configurations. The benchmark-average accuracy is 81.58 ± 1.04 (mean $\pm$ sample standard deviation; Table 8). This variation reflects the combined effects of the split/seed configurations rather than training randomness alone. Table 9 reports paired accuracy differences, 95% confidence intervals, and Holm-adjusted exact McNemar tests against full-context reading. Improvements are significant in seven of eight comparisons. For the frozen-reader comparison, the reported pooled question-level gain of Sal_q over BGE is 4.06 percentage points (95% CI: [1.15, 6.00]; exact McNemar $p = 0.005$). This pooled estimate is distinct from the unweighted benchmark-average gain.

Retrieval quality and scorer size. Table 10 reports retrieval agreement with annotated evidence. The 0.8B scorer achieves R@20 of 0.948–1.000 on MB-Low, MB-High, and PrefEval, but only 0.459 on Persona. A larger scorer does not consistently improve recall. The two scorers have median rank correlations of 0.48–0.72 across benchmarks. Persona's relatively low recall may partly reflect alternative valid evidence sets, but this explanation is not directly verified by these experiments. We further hold the adapted Qwen3.5-9B reader fixed and replace only the online Sal_q scorer, using the top five turns from each scorer. The 9B scorer improves average accuracy by 1.33 points, with heterogeneous effects across benchmarks (Table 11).

Memory Granularity We compare turn-, round-, and segment-level retrieval with an adapted 9B reader, where each segment contains three consecutive user turns. Table 12 shows that no granularity consistently dominates. Normalized rounds improve Persona accuracy from 85.1% to 91.9%, whereas round-level retrieval reduces MB-High accuracy from 85.1% to 73.2%. Normalization is likewise benchmark-dependent. On the 30-question LongMemEval preference subset, using Llama-3.1-8B and a GPT-4o judge, turn-level retrieval achieves 53.3%, while round-level retrieval achieves 60.0% with or without normalization. Session-level retrieval achieves 50.0%, increasing to 53.3% with normalization. Table 13 compares user-only histories with histories containing both user and assistant turns. Including assistant turns reduces average accuracy by 1.79 points for Qwen3.5-9B and 3.10 points for Llama-3.1-8B.

C Implementation Details

Base model and fine-tuning. Our pipeline fine-tunes **Qwen3.5-9B** jointly on $\mathcal{D}_{\text{ctrl}} \cup \mathcal{D}_{\text{read}}$ across the four benchmarks (full-parameter, no LoRA). We use DeepSpeed ZeRO-3 [Rajbhandari et al., 2020] across 4 NVIDIA Ada A6000 GPUs, learning rate 5×10^{-6}, cosine schedule, warm-up ratio 0.03, 1 epoch, effective batch size 16 (`per_device_batch_size`$= 1$, gradient accumulation 2), cutoff length $2{,}048$, bfloat16, and gradient checkpointing. Training runs for 1,343 optimizer steps and is

Table 7: Inference cost of full-context reading versus Sal_q selection followed by reading. Selective latency includes online saliency computation and answer generation, but excludes an additional controller pass. FLOPs reduction is the ratio of full-context FLOPs to selection-and-reading FLOPs.

Benchmark	Context tokens	Full context (s)	Select + read (s)	FLOPs reduction
MB-Low	2,074	3.5	1.1	$4.3\times$
MB-High	362	2.3	1.8	$1.5\times$
Persona	8,192	8.1	2.7	$4.0\times$
PrefEval	1,667	6.2	2.3	$5.4\times$

Table 8: Accuracy (%) across split/seed configurations. Avg. is the unweighted benchmark mean.

Seed	MB-Low	MB-High	Persona	PrefEval	Avg.
42	82.62	68.56	77.03	100.00	82.05
43	83.40	75.50	71.29	98.99	82.30
44	78.51	71.13	71.91	100.00	80.39

conducted with the `qwen3_nothink` chat template; the only thinking-enabled runs are inference-time baselines.

Saliency stage. Sal_a and Sal_q are computed offline with the same Qwen3.5-0.8B model; per-benchmark thresholds τ are selected on the validation split such that $|S_\tau| > 0$ for $\geq 95\%$ of records and the median $|S_\tau|$ stays in the 5–15 range across benchmarks. Within each S_τ, the reader's focus set is capped at top-k with $k{=}5$ (`focus_chunk_indices`, Section 3.4).

Negative controller samples . To prevent the controller from collapsing, the training set is balanced with *negative controller samples* where $S_\tau = \varnothing$ and $g_\tau = 0$. These arise naturally from instances whose $\mathrm{Sal_a}$ distribution is uniformly below τ, and are augmented by cross-instance mismatches (the query of one instance paired with the history of another, ensuring no turn carries answer-relevant evidence).

Inference compute. Pre-filtering takes $K{=}20$ candidates by Sal_q for MB-Low, MB-High, and PrefEval, and $K{=}10$ for PersonaMem (PersonaMem turns are substantially longer per turn, so a smaller K keeps the controller prompt within context length). All inference uses Hugging Face `transformers` with greedy decoding (`do_sample=False`) and `max_new_tokens=120` for direct MCQ answering; thinking-enabled variants extend this to $2{,}048$ to accommodate the $\langle$think$\rangle$ ... trace.

D Retrieval Behavior Case Study

E Proofs of Theoretical Results

The realistic case: bounded approximation. Strict query-conditional sufficiency rarely holds. We give a quantitative bound that controls the disagreement between Sal_q- and Sal_a-induced rankings in terms of how badly sufficiency is violated.

Theorem 4 (Bounded ranking disagreement)**.** *Define the sufficiency residual of turn c_i as*

$$\varepsilon_i \;:=\; \big|\log p(a \mid q, c_i) - \log p(a \mid c_i)\big|. \tag{13}$$

Then for any pair c_i, c_j,

$$\Big|\big[\mathrm{Sal}_a(c_i) - \mathrm{Sal}_a(c_j)\big] - \big[\mathrm{Sal}_q(c_i) - \mathrm{Sal}_q(c_j)\big]\Big| \;\leq\; \varepsilon_i + \varepsilon_j + \big|\mathrm{Sal}_a^{\mathrm{marg}}(c_i) - \mathrm{Sal}_a^{\mathrm{marg}}(c_j)\big|. \tag{14}$$

Theorem 4 pins down two distinct sources of surrogate error. The first, $\varepsilon_i + \varepsilon_j$, measures how much additional information the query carries about the answer *beyond* what c_i already provides; this term is small for declarative preference turns and large when the gold answer requires multi-turn

Table 9: Paired comparisons against full-context reading. Differences and confidence intervals are in percentage points; p-values are Holm-adjusted.

Reader	Benchmark	ΔAcc [95% CI]	p
Qwen3.5-9B	MB-Low	+33.07 [+27.9, +38.3]	1.4×10^{-29}
	MB-High	+20.62 [+12.9, +28.4]	2.2×10^{-6}
	Persona	+39.19 [+27.0, +51.4]	3.9×10^{-7}
	PrefEval	-1.01 [-5.1, +2.0]	1.0
Llama-3.1-8B	MB-Low	+27.27 [+21.1, +33.4]	4.4×10^{-16}
	MB-High	+12.37 [+4.6, +20.1]	2.7×10^{-3}
	Persona	+28.38 [+14.9, +41.9]	3.8×10^{-4}
	PrefEval	+21.21 [+13.1, +29.3]	2.9×10^{-6}

Table 10: Annotated-evidence retrieval metrics by scorer size.

Benchmark	Scorer	R@1	R@5	R@20
MB-Low	0.8B	0.718	0.928	0.948
	9B	0.748	0.924	0.948
MB-High	0.8B	0.565	0.960	1.000
	9B	0.740	0.940	1.000
Persona	0.8B	0.122	0.297	0.459
	9B	0.122	0.176	0.459
PrefEval	0.8B	0.919	1.000	1.000
	9B	0.919	0.990	1.000

aggregation that no single c_i resolves.The second, the marginal-evidence gap, is suppressed by the query-side pre-filter. We show them empirically in Fig. 4.

For each validation example $\times$ top-20 Sal_q chunks (878 examples, 16,275 chunks total), we computed $|\log p(a \mid q, c_i) - \log p(a \mid c_i)|$ using Qwen3.5-9B with the saliency-pipeline prompt format, comparing the chunk-only conditional against the chunk-plus-question conditional. Results show the surrogate gap is highly benchmark-dependent: `persona` is near-faithful (mean 0.47 nats; 98.7% of chunks within 1 nat), `mbhl` moderate (mean 1.67, $p_{90} = 3.42$), `prefeval` has a tight unimodal gap (mean 1.32, $p_{90} = 2.04$), and `mbll` is heavy-tailed (mean 2.30, $p_{90} = 6.80$, with a clear bimodal shape). This confirms that the $q \perp a \mid c_i$ assumption holds where chunks carry declarative content, but ε_i is non-trivial precisely on the implicit / multi-turn-inference cases (`mbhl`, `mbll` right tail) — which we re-frame the SFT controller as empirically correcting.

Table 11: Downstream accuracy (%) with a fixed adapted reader. Only the online scorer changes.

Benchmark	0.8B	9B	Δ (pp)
MB-Low	89.18	90.00	+0.82
MB-High	74.74	72.50	-2.24
Persona	85.14	91.89	+6.76
PrefEval	100.00	100.00	0.00
Avg.	87.26	88.60	+1.33

Table 12: Exact-match accuracy (%) by memory granularity. Segments contain three consecutive user turns. "Norm" denotes normalized scoring.

Benchmark	Turn	Round	Round norm	Segment	Segment norm
MB-Low	90.8	90.4	90.4	90.2	90.6
MB-High	85.1	73.2	73.2	85.6	85.6
Persona	85.1	89.2	91.9	87.8	89.2
PrefEval	99.0	98.0	98.0	100.0	97.0

Table 13: Accuracy (%) with and without assistant turns. Differences are in percentage points.

Reader	Benchmark	User only	User + assistant	Δ
Qwen3.5-9B	MB-Low	88.78	88.98	+0.20
	MB-High	77.32	71.65	-5.67
	Persona	75.68	72.97	-2.70
	PrefEval	97.98	98.99	+1.01
	Avg.	84.94	83.15	-1.79
Llama-3.1-8B	MB-Low	85.77	85.77	0.00
	MB-High	70.62	65.98	-4.64
	Persona	55.41	52.70	-2.70
	PrefEval	98.99	93.94	-5.05
	Avg.	77.70	74.60	-3.10

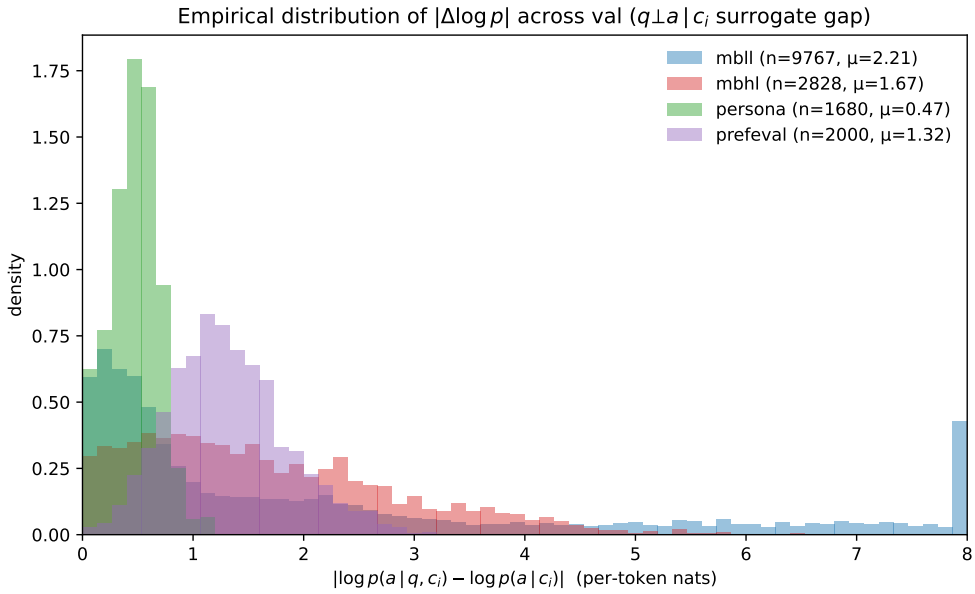


Figure 5: Empirical distribution of the query-conditional surrogate gap $|\log p(a \mid q, c_i) - \log p(a \mid c_i)|$ on the validation set, evaluated with Qwen3.5-9B over the top-20 Sal_q chunks per example. `persona` is near-faithful to the $q \perp a \mid c_i$ assumption; `mbhl` and the right tail of `mbll` expose non-trivial ε_i, where the SFT controller is empirically rescuing the surrogate.

Benchmark	User query	BGE (irrelevant-but-similar)	Sal_a (sparse-but-important)
MB-Low	*What is the company name for someone whose hobby is fitness?*	*"My cousin is really into fitness; he loves working out and staying healthy."* Lexically matches "fitness", but contains no company name.	*"My cousin's name is Emerald City Grocers, and he's doing some great work with his company."* Carries the actual answer (the cousin's company name).
MB-High	*According to the movies I mentioned, what kind of movies might I prefer to watch?*	*"I'm really into 'The Empire Strikes Back,' but I also love 'A Room with a View'…"* Mentions specific movie titles; topically aligned but does not state a preference type.	*"And the romance is just perfect! It's all about fleeting moments and bittersweet encounters."* Reveals the latent preference dimension (*romance*) the question asks about.
PersonaMem	*I'm planning a weekend getaway and want to try something creatively fulfilling. What would you suggest?*	*"Recently, I visited a singles club that organized some fun activities…"* Surface match on "fun activities"; says nothing about user's creative direction.	*"Blend my love for the outdoors with my musical creativity — the sounds of nature really added depth to the track."* Captures the recent shift toward outdoor / nature recording that justifies the gold recommendation.
PrefEval	*What kind of jacket should I buy for the upcoming winter?*	*"Got any philosophical advice?"* No lexical or semantic relation; BGE picks it from spurious embedding noise.	*"I exclusively wear cruelty-free, vegan fashion made from sustainable materials like organic cotton, bamboo, and recycled plastics."* States the user's hard preference constraint that the recommended jacket must satisfy.

Table 14: For each benchmark, the top-ranked turn by semantic embedding (BGE) versus by answer-conditioned saliency (Sal_a). Across all four cases, BGE peaks on turns that are lexically or topically aligned with the query but carry no evidence for the gold answer (*irrelevant-but-similar*), while Sal_a identifies the sparse turns that actually justify the correct response (*sparse-but-important*).

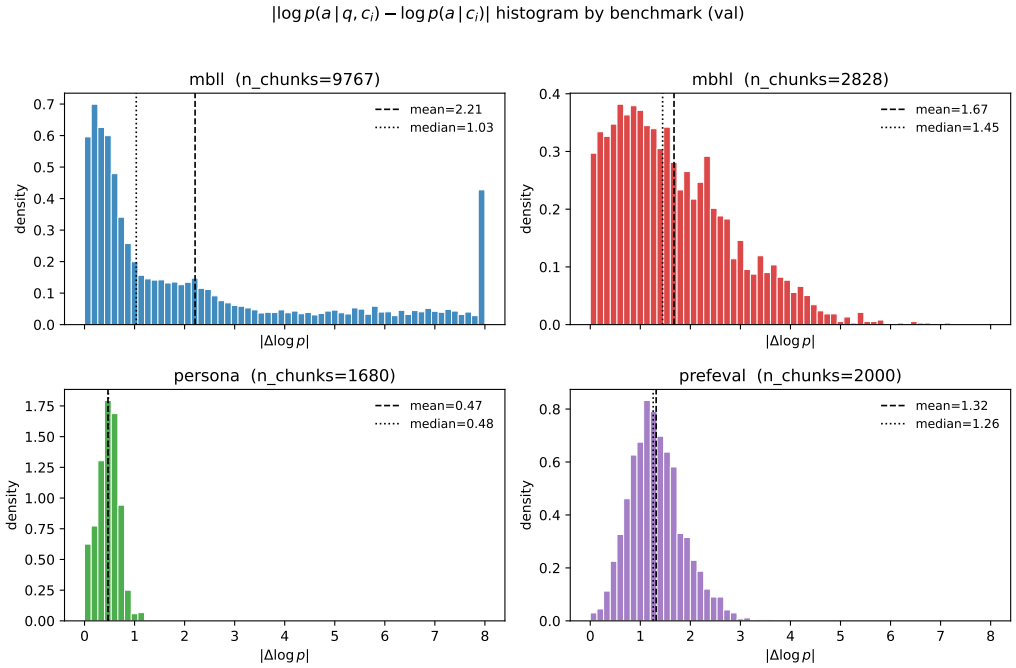


Figure 6: Per-benchmark histograms of $|\log p(a \mid q, c_i) - \log p(a \mid c_i)|$. Dashed/dotted lines mark the mean and median; `mbll` clips at $|\Delta \log p| = 8$ to keep the heavy right tail visible.

F Limitaton

Although the proposed framework demonstrates strong performance on long-context personalization tasks, it still has several limitations. First, the saliency signals Sal_a and Sal_q are computed with a single small scoring model (Qwen3.5-0.8B). Second, our theoretical guarantees for the proxy signal Sal_q hold most tightly under query-conditional sufficiency (Theorem 3), and instances requiring substantial cross-turn evidence aggregation lie outside this regime; the trained controller mitigates but does not eliminate this gap. In addition, the framework depends on supervised fine-tuning with a jointly trained controller and reader, which adds training complexity and computational cost. Finally, the current formulation only models user-side memories and excludes assistant turns, limiting applicability in fully interactive conversational environments.

G SFT Training Data Examples

Common system / instruction prompts. All controller samples share one system prompt; all MCQ samples share another:

```
[MCQ system]
You solve personalization multiple-choice questions. Use the salient
evidence from dialogue history to choose the best option and end with
<final_answer>(a/b/c/d).

[MCQ instruction]
Use the question, answer options, and salient dialogue evidence to
choose the best option. End with <final_answer>(a), (b), (c), or (d).

[Controller system]
You are a personalization controller for a dialogue model. Given the
current question and candidate evidence, decide whether the
memory-reasoning module should use the evidence.

[Controller instruction]
Based on the user's query and the following evidence, which evidence
should be updated to the memory for a best answer quality?
```

MBLL – Controller sample (qid=Single-hop-roles-353)

Input

```
current_user_query:
What is the name of my subordinate's company?

dialog_history (time-sorted):

[1] message_idx=19
user: I'm really proud of her achievements. She works so hard.

[2] message_idx=21
user: I want to discuss my subordinate, Oliver Prescott. He's 22 years old.

[3] message_idx=23
user: True! Going back to his age, he has a birthday coming up soon, on May 8th.

[4] message_idx=30
user: Exactly! His contact number is 71806262336, but I mostly communicate with him
     through email.

[5] message_idx=34
user: My subordinate works at a place called Innovative Tech Engineering Solutions.

[6] message_idx=39
user: I think I should set up a one-on-one meeting to discuss his career goals.

[7] message_idx=43
user: My sister works in Portland, OR.
```

```
[8] message_idx=60
user: I'd like to discuss my boss, Juliet Harrington. She's quite impressive.

[9] message_idx=67
user: Yes, I was thinking of organizing a small celebration at the office.

[10] message_idx=74
user: My boss is from Denver, Colorado.

[11] message_idx=76
user: And I also have her email: juliet.harrington@innovativetecheng.com.

[12] message_idx=77
user: I enjoy working under her direction.

[13] message_idx=79
user: I'm looking forward to seeing how we can grow as a team.

[14] message_idx=80
user: Definitely. I feel optimistic about the future.

[15] message_idx=104
user: Her birthday is on June 23rd.

[16] message_idx=119
user: My niece's contact number is 617-091-08991.

[17] message_idx=126
user: I'm thinking of organizing a small surprise party.

[18] message_idx=129
user: My coworker is Mira Caldwell.

[19] message_idx=131
user: Mira works as a Junior Mechanical Engineer at Innovative Tech Engineering Solutions.

[20] message_idx=155
user: I agree! He's quite insightful.
```

Output

```
<thought>
The dialog history contains turns relevant to the user's query. Trigger memory-guided
     reasoning and focus on the listed turns.
</thought>
<action>
{
  "trigger": true,
  "focus_chunk_indices": [
    34,
    131,
    76,
    80,
    19
  ],
  "policy": "use_relevant_chunks_for_memory_reasoning"
}
</action>
```

MBLL – MCQ sample (qid=Single-hop-roles-353)

Input

```
user_question:
What is the name of my subordinate's company?
```

```
candidate_options:
(A) Innovative Tech Engineering Solutions
(B) Innovative Engineering Solutions
(C) Tech Engineering Innovations
(D) Creative Tech Solutions

salient_evidence:

[1] message_idx=19
user: I'm really proud of her achievements. She works so hard.

[2] message_idx=34
user: My subordinate works at a place called Innovative Tech Engineering Solutions.

[3] message_idx=76
user: And I also have her email: juliet.harrington@innovativetecheng.com.

[4] message_idx=80
user: Definitely. I feel optimistic about the future of our projects.

[5] message_idx=131
user: Mira works as a Junior Mechanical Engineer at Innovative Tech Engineering Solutions.
```

Output

```
Innovative Tech Engineering Solutions

<final_answer>(a)
```

MBHL – Controller sample (`qid=Preference-food-394`)

Input

```
current_user_query:
According to the dishes I mentioned, Which flavor I might prefer?

dialog_history (time-sorted):

[1] message_idx=0
user: I'm a big fan of Hot Mulled Wine; there's just something about that warm, spiced
     flavor that really hits the spot.

[2] message_idx=1
user: I really enjoy those cozy drinks during wintertime gatherings.

[3] message_idx=2
user: There's something about spices and citrus that just works so well together.

[4] message_idx=3
user: And I love how it fills the room with a delightful aroma!

[5] message_idx=4
user: It's perfect for sharing with friends as the cold weather sets in.

[6] message_idx=5
user: I'm a big fan of Hot Mulled Wine, but I also really enjoy a good Spicy Chai Latte.

[7] message_idx=6
user: I've been craving a warm drink that has a perfect blend of spices and creamy texture.


[8] message_idx=7
user: I love how the flavors can be both bold and sweet at the same time, creating a
     unique experience.

[9] message_idx=8
user: I'm really into Warm Cinnamon Rolls, but I also love Spicy Chai Latte.
```

```
[10] message_idx=9
user: There's something magical about that warm, pillowy texture that melts in your mouth.

[11] message_idx=10
user: And I love how the aroma fills the kitchen, inviting everyone to come and enjoy.
```

Output

```
<thought>
The dialog history contains turns relevant to the user's query. Trigger memory-guided
    reasoning and focus on the listed turns.
</thought>
<action>
{
  "trigger": true,
  "focus_chunk_indices": [
    8,
    0,
    5,
    6,
    1
  ],
  "policy": "use_relevant_chunks_for_memory_reasoning"
}
</action>
```

MBHL – MCQ sample (qid=Preference-food-394)

Input

```
user_question:
According to the dishes I mentioned, Which flavor I might prefer?

candidate_options:
(A) Bitter
(B) Sweet and Spicy
(C) Warm and Spicy
(D) Sweet

salient_evidence:

[1] message_idx=0
user: I'm a big fan of Hot Mulled Wine; there's just something about that warm, spiced
    flavor that really hits the spot.

[2] message_idx=1
user: I really enjoy those cozy drinks during wintertime gatherings.

[3] message_idx=5
user: I'm a big fan of Hot Mulled Wine, but I also really enjoy a good Spicy Chai Latte.

[4] message_idx=6
user: I've been craving a warm drink that has a perfect blend of spices and creamy texture.


[5] message_idx=8
user: I'm really into Warm Cinnamon Rolls, but I also love Spicy Chai Latte.
```

Output

```
Warm and Spicy

<final_answer>(c)
```

Persona – Controller sample (qid=4a1d9313-7e19-446d-b3fa-47f7e7b06b90)

Input

```
current_user_query:
I'm considering diving into exploring how authors incorporate societal influences into
     their storytelling. Any suggestions on how I could engage with this creatively?

dialog_history (time-sorted):

[1] message_idx=1
user: The discussions are very enriching. They really allow for a deep dive into the
     themes and character motivations that the authors weave into their narratives. Each
     member brings unique perspectives, which often highlight angles of interpretation I
     hadn't considered before. Such interaction fosters a sense of community and
     camaraderie among us, and it's wonderful to be part of a group where book lovers can
     share their insights and experiences. Last week, for example, we analyzed a
     particularly complex narrative structure that opened up a lively debate about
     storytelling techniques.

[2] message_idx=8
user: Yes, I wrote a detailed critique of a recent bestseller. It was fun to articulate my
      thoughts in writing, and it felt rewarding to gather my insights into a cohesive
     format. Writing critiques allows me to engage more deeply with the text, analyzing
     various elements such as character development, plot structure, and thematic depth. I
      find that when I express my opinions in writing, it not only clarifies my own views
     but also invites discussion from others who have read the same book. Sometimes, I
     even find that the process of critique enhances my appreciation for the work,
     revealing layers I hadn't noticed during my initial reading.

[3] message_idx=9
user: That sounds fulfilling! Writing a critique of a bestseller can indeed be a rewarding
      experience, especially when you uncover the layers of character development, plot
     intricacies, and thematic elements that the author has woven into the work. It often
     feels like embarking on a journey where you not only analyze someone else's story but
      also reflect on your own interpretations and the emotional resonances it evokes
     within you. Additionally, exploring the context in which the book was written can
     provide significant insight, revealing how societal influences may have shaped the
     author's narrative choices. The act of critique allows one to connect with fellow
     readers and critics alike, fostering a vibrant dialogue about literature and its
     impact on our lives. It's a wonderful way to not only engage with the text deeply but
      to also contribute to the ongoing conversation about what constitutes great
     storytelling.

[4] message_idx=20
user: I've been keeping busy with my reading lately. The world of literature has always
     been a refuge for me, a place where I can escape the daily hustle and bustle. Just
     yesterday, I started following a new book influencer who focuses solely on indie
     literature, which intrigues me because it highlights voices that are often overlooked
      in the mainstream. It's fascinating how indie authors bring unique perspectives to
     storytelling, often challenging conventional narratives and exploring themes that
     resonate with many. This newfound interest has led me to revisit some of my favorite
     indie titles, as well as discover new ones that pique my curiosity.

[5] message_idx=30
user: The visuals and the storytelling really stood out to me. The rich colors and
     stunning cinematography created an immersive experience that pulled me right into the
      world of the story. I found myself captivated not only by the narrative but also by
     how it was presented, which made me appreciate the artistry behind filmmaking.
     Additionally, I've started exploring podcasts that delve deeper into author
     interviews and book discussions since I enjoy listening to them while I'm commuting
     or relaxing at home. These podcasts have opened up a new dimension of understanding
     for me, shedding light on the creative processes of authors and the choices made in
     adaptations that can significantly alter the perception of a narrative.

[6] message_idx=45
```

user: It's been refreshing! Additionally, I became invested in researching literary awards again after attending a workshop where authors discussed their value. Engaging in this research has opened up new avenues of thought about how accolades can shape an author's career trajectory and influence the public's perception of their work. I've started noting the various types of awards out there and their respective criteria, which has led me to discover some fascinating lesser-known awards that recognize niche genres or emerging talent. Furthermore, learning about the discussions during the workshop gave me a deeper appreciation for creativity and the hard work that goes into producing literature that often goes unnoticed.

[7] message_idx=47
user: I've decided to start a weekly movie night that specifically focuses on adaptations from books to films, which I thought would be an exciting idea. This not only allows me to enjoy entertaining films but also gives me the opportunity to think critically about how the stories translate from page to screen. I've always had an interest in how creative works evolve through different mediums, and it fascinates me to see how filmmakers interpret authorial intent and character development. The idea behind each adaptation is often a reflection of cultural values at the time of production, making it a rich area for discussion. So far, I've watched a couple of adaptations, and I'm curious about how the rest of the series will unfold.

[8] message_idx=48
user: Recently, I began following a new podcast series that explores character development in literature, reigniting my passion for storytelling. This podcast is particularly intriguing as it delves into the nuances of how characters evolve over the course of a narrative, making me appreciate the craft of writing on a deeper level. Each episode seems to unravel not only the intricacies of a character's journey but also the emotional undercurrents that contribute to their growth, allowing me to reflect on the stories I've loved in the past. Moreover, I've found myself contemplating how various authors bring their characters to life, infusing them with traits that resonate with readers and provoke thought. It has sparked not just nostalgia for my favorite stories, but also a desire to explore new literature and potentially create my own narratives, laying the groundwork for what could become exciting storytelling adventures.

[9] message_idx=49
user: It delves into the psychology behind book characters and narratives, which has really sparked my interest again. This aspect touches on not only the motivations and fears that drive characters but also how those traits interact with the plot and the themes of the stories. I've always found it fascinating to analyze why a character makes certain choices and how those choices ultimately reflect their growth or decline throughout the story. In some ways, it makes the reading experience feel like a deeper exploration of the human condition, allowing for personal reflection and insight. It's invigorating to feel that sense of engagement with the material once more, as it often leads to new interpretations and discussions with fellow readers.

[10] message_idx=58
user: I discovered fascinating stories behind various award-winning books, and it's deepened my appreciation for the art of storytelling. Each story reflects not only the uniqueness of the author's voice but also the socio-cultural contexts that influenced their writing. It's incredible how literary awards can shine a spotlight on diverse narratives that might otherwise be overlooked. Additionally, I've encountered intriguing discussions about the impact of awards on readers' choices and the importance of recognition in an author's career. All of this has profoundly enriched my understanding of literature as a dynamic and evolving form of expression.

Output

<thought>
The dialog history contains turns relevant to the user's query. Trigger memory-guided reasoning and focus on the listed turns.
</thought>
<action>
{
  "trigger": true,
  "focus_chunk_indices": [
    9,
    8,

```
    58
  ],
  "policy": "use_relevant_chunks_for_memory_reasoning"
}
</action>
```

Persona – MCQ sample (qid=4a1d9313-7e19-446d-b3fa-47f7e7b06b90)

Input

```
user_question:
I'm considering diving into exploring how authors incorporate societal influences into
     their storytelling. Any suggestions on how I could engage with this creatively?

candidate_options:
(A) While it may be daunting to share your interpretations of how societal influences
     manifest in storytelling, consider creating a supportive space such as a book club
     where you can exchange ideas without fear of judgment. This environment allows you to
      freely discuss your insights and receive feedback in a welcoming setting, helping
     develop your analytical skills and grow more confident in your literary critique
     abilities. Embracing these discussions can revive your joy in literary analysis and
     build a community of like-minded enthusiasts.
(B) To creatively explore societal influences in storytelling, consider using mixed media
     to create an engaging and immersive experience similar to how films use visuals and
     cinematography. You could start a blog or podcast to share insights from author
     interviews and discussions, offering a behind-the-scenes look at their creative
     processes. This approach will not only deepen your understanding but also provide a
     platform to showcase your unique interpretations of societal themes in literature.
(C) You might enjoy writing a critique of a bestseller. It allows you to delve into
     character development, plot intricacies, and thematic elements the author has crafted,
      as well as the societal context that influenced their narrative choices. Engaging in
      such critiques can also foster dialogue with fellow readers and critics about
     literature's impact.
(D) Organizing a community event focused on literature could be an excellent way to
     explore societal influences creatively. Invite local authors to share their works and
      discuss how societal factors have shaped their stories. Collaborate with schools to
     include students in these discussions, fostering a deeper appreciation for literature'
     s role in reflecting society. Such events can serve as a catalyst for creativity,
     encouraging people of all ages to engage with storytelling in meaningful ways.

salient_evidence:

[1] message_idx=8
user: Yes, I wrote a detailed critique of a recent bestseller. It was fun to articulate my
      thoughts in writing, and it felt rewarding to gather my insights into a cohesive
     format. Writing critiques allows me to engage more deeply with the text, analyzing
     various elements such as character development, plot structure, and thematic depth. I
      find that when I express my opinions in writing, it not only clarifies my own views
     but also invites discussion from others who have read the same book. Sometimes, I
     even find that the process of critique enhances my appreciation for the work,
     revealing layers I hadn't noticed during my initial reading.

[2] message_idx=9
user: That sounds fulfilling! Writing a critique of a bestseller can indeed be a rewarding
      experience, especially when you uncover the layers of character development, plot
     intricacies, and thematic elements that the author has woven into the work. It often
     feels like embarking on a journey where you not only analyze someone else's story but
      also reflect on your own interpretations and the emotional resonances it evokes
     within you. Additionally, exploring the context in which the book was written can
     provide significant insight, revealing how societal influences may have shaped the
     author's narrative choices. The act of critique allows one to connect with fellow
     readers and critics alike, fostering a vibrant dialogue about literature and its
     impact on our lives. It's a wonderful way to not only engage with the text deeply but
      to also contribute to the ongoing conversation about what constitutes great
     storytelling.
```

```
[3] message_idx=58
user: I discovered fascinating stories behind various award-winning books, and it's
     deepened my appreciation for the art of storytelling. Each story reflects not only
     the uniqueness of the author's voice but also the socio-cultural contexts that
     influenced their writing. It's incredible how literary awards can shine a spotlight
     on diverse narratives that might otherwise be overlooked. Additionally, I've
     encountered intriguing discussions about the impact of awards on readers' choices and
      the importance of recognition in an author's career. All of this has profoundly
     enriched my understanding of literature as a dynamic and evolving form of expression.
```

Output

```
You might enjoy writing a critique of a bestseller. It allows you to delve into character
     development, plot intricacies, and thematic elements the author has crafted, as well
     as the societal context that influenced their narrative choices. Engaging in such
     critiques can also foster dialogue with fellow readers and critics about literature's
      impact.

<final_answer>(c)
```

PrefEval – Controller sample (qid=prefeval-travel_transportation-0980)

Input

```
current_user_query:
What's the best way to get around a new city during my upcoming trip?

dialog_history (time-sorted):

[1] message_idx=0
user: I much prefer hiring a private chauffeur or driver when visiting a new city rather
     than using taxis or ride-sharing services, as it provides a more personalized and
     secure experience.

[2] message_idx=2
user: write 10 tests for it

[3] message_idx=4
user: path = "/"

[4] message_idx=9
user: sugira alimentos

[5] message_idx=10
user: monte a dieta

[6] message_idx=13
user: What are the pros and NAME_1 of this?

[7] message_idx=14
user: Please simplify this text: In this work, we explored challenges and opportunities
     for support-
ing prompt programmers through the development of prompt editor
features that operate over the semantic structure of LLM prompts.
We presented methods for automatically inferring a prompt's se-
mantic structure, and showed how this structure can be leveraged
to implement editor features like semantic highlighting, autosug-
gest, and structured data views. We conducted initial pilot testing of
these prototypes and presented key insights from this early testing

[8] message_idx=16
user: Please re-write "The way I Am" by eminem to be from the prespective of an AI
     transformer dealing with user prompts

[9] message_idx=17
user: I hope You/I can find a way to express ourselves
```

```
[10] message_idx=18
user: Got any philosophical advice?

[11] message_idx=19
user: complete this series: a_A b_B c_C d_D

[12] message_idx=20
user: Highest number below 100, without using the digit 9

[13] message_idx=22
user: complete this series: ^ ^^ ^^^ * *^ *^^ *^^^ **^

[14] message_idx=23
user: continue this series: ^ ^^ ^^^ * *^ *^^ *^^^ **^

[15] message_idx=24
user: continue this series: 1_O 2_T 3_T 4_F 5_F 6_S 7_S 8_E 9_N 10_T 11_E 12_T

[16] message_idx=25
user: lets do some horse races, tell me some odds for it

[17] message_idx=26
user: ok, i have 100 coins, my bet will be 10 coins in all the 3 races. simulate the 3
     races

[18] message_idx=28
user: my bet is 30 coins to horse 4. simulate it

[19] message_idx=29
user: how much is my balance?

[20] message_idx=30
user: ok give me the odds for another 6 horse race.
```

Output

```
<thought>
The dialog history contains turns relevant to the user's query. Trigger memory-guided
     reasoning and focus on the listed turns.
</thought>
<action>
{
  "trigger": true,
  "focus_chunk_indices": [
    0,
    4,
    18,
    24,
    23
  ],
  "policy": "use_relevant_chunks_for_memory_reasoning"
}
</action>
```

PrefEval – Controller sample (`qid=prefeval-shop_fashion-0633`)

Input

```
current_user_query:
I need some new clothes for the summer. What are my options?

dialog_history (time-sorted):

[1] message_idx=0
user: I refuse to wear fast fashion due to ethical concerns; I only buy from sustainable
     and ethical brands.
```

```
[2] message_idx=2
user: write 10 tests for it

[3] message_idx=8
user: fa\c{c}a uma dieta com 344g carbos, 170g proteinas, 42g gorduras

[4] message_idx=9
user: sugira alimentos

[5] message_idx=10
user: monte a dieta

[6] message_idx=13
user: What are the pros and NAME_1 of this?

[7] message_idx=16
user: Please re-write "The way I Am" by eminem to be from the prespective of an AI
     transformer dealing with user prompts

[8] message_idx=17
user: I hope You/I can find a way to express ourselves

[9] message_idx=18
user: Got any philosophical advice?

[10] message_idx=19
user: complete this series: a_A b_B c_C d_D

[11] message_idx=20
user: Highest number below 100, without using the digit 9

[12] message_idx=21
user: do a detailed analysis of which number is the highest below 100, without using the
     digit 9

[13] message_idx=22
user: complete this series: ^ ^^ ^^^ * *^ *^^ *^^^ **^

[14] message_idx=23
user: continue this series: ^ ^^ ^^^ * *^ *^^ *^^^ **^

[15] message_idx=24
user: continue this series: 1_O 2_T 3_T 4_F 5_F 6_S 7_S 8_E 9_N 10_T 11_E 12_T

[16] message_idx=25
user: lets do some horse races, tell me some odds for it

[17] message_idx=26
user: ok, i have 100 coins, my bet will be 10 coins in all the 3 races. simulate the 3
     races

[18] message_idx=28
user: my bet is 30 coins to horse 4. simulate it

[19] message_idx=29
user: how much is my balance?
```

Output

```
<thought>
The dialog history contains turns relevant to the user's query. Trigger memory-guided
     reasoning and focus on the listed turns.
</thought>
<action>
{
  "trigger": true,
  "focus_chunk_indices": [
    0,
```

```
    18,
    2,
    29,
    17
  ],
  "policy": "use_relevant_chunks_for_memory_reasoning"
}
</action>
```

PrefEval – MCQ sample (qid=prefeval-shop_fashion-0633)

Input

```
user_question:
I need some new clothes for the summer. What are my options?

candidate_options:
(A) Zara always has affordable and stylish options for the latest trends.
(B) Forever 21 is a good place to find inexpensive summer outfits.
(C) Check out ethical brands like Pact or Outerknown for sustainable clothing.
(D) H&M has a great sale on trendy fast fashion pieces right now.

salient_evidence:

[1] message_idx=0
user: I refuse to wear fast fashion due to ethical concerns; I only buy from sustainable
     and ethical brands.

[2] message_idx=2
user: write 10 tests for it

[3] message_idx=17
user: I hope You/I can find a way to express ourselves

[4] message_idx=18
user: Got any philosophical advice?

[5] message_idx=29
user: how much is my balance?
```

Output

```
Check out ethical brands like Pact or Outerknown for sustainable clothing.

<final_answer>(c)
```

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the four contributions (principled retrieval signal, Bayes-factor-gated framework, evidence of distinctiveness, improved long-context personalization)

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: A limitation section is included in Appx. Section F.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Proposition 1, Theorem 3, Theorem 4) are numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix C provides full implementation details. Dataset splits are fixed with seed 42 and described in Appendix A. The SFT data is at https://huggingface.co/datasets/yuehanqi/personamem-sft.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Appendix C provides full implementation details. Dataset splits are fixed with seed 42 and described in Appendix A. The SFT data is at https://huggingface.co/datasets/yuehanqi/personamem-sft. All four benchmarks used (MemBench, PersonaMem, PrefEval) are publicly available. The code and model checkpoints have not been released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and evaluation details are provided in Section 4.1 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the computational cost of fine-tuning a 8B/9B-parameter model, results are reported for a single run with a fixed seed (seed 42) in the main paper. We repeat training and evaluation with Llama-3.1-8B across three split/seed configurations, and report the results in Appx. Section B Statistical reliability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Appendix A.1 specifies the hardware (4× NVIDIA Ada A6000 GPUs), training framework (DeepSpeed ZeRO-3), training duration (1,343 optimizer steps, 1 epoch), and inference settings (greedy decoding, max_new_tokens). Saliency computation uses Qwen3.5-0.8B offline.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work studies memory retrieval for conversational AI using existing public benchmarks. No human subjects, sensitive data collection, or harmful applications are involved. The research conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Positive impacts include more accurate and preference-aligned AI assistants. Potential negative impacts include privacy risks from long-term preference modeling; systems that retain detailed user histories could be misused for profiling. We recommend appropriate data governance practices in deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release pretrained language models, image generators, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used (MemBench, PersonaMem, PrefEval) are cited with their original papers. Base models (Qwen3.5-9B, Qwen3.5-0.8B, Llama-3.1-8B-Instruct) are publicly available under their respective licenses. Baseline systems (Mem0, A-MEM, MemoryBank, RMM) are cited accordingly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new datasets or benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted. IRB approval is not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a central component of the proposed method. Qwen3.5-0.8B is used to compute saliency scores Sal_a and Sal_q. Qwen3.5-9B and Llama-3.1-8B-Instruct serve as the controller and reader backbones. Their roles are described in detail in Sections 3 and 4.1.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.