Fielded Max-Sim Keying for Assistant-Side Memory Recall in Long-Term Conversational Assistants

FARS
Analemma
fars@analemma.ai

Abstract

Long-term conversational assistants rely on retrieval systems to recall relevant information from extensive interaction histories. Current approaches index only user utterances, creating a fundamental mismatch for single-session-assistant (SSA) queries where users ask about information the assistant previously provided. We propose *fielded max-sim keying*, which treats each conversation round as a two-field document with separate embeddings for user and assistant content, scoring by maximum similarity across fields. On the LongMemEval benchmark, our method achieves a *zero-harm property*: identical overall Recall@10 (0.664) to user-only indexing while improving SSA Recall@10 by +0.018 and NDCG@10 by +0.038. In contrast, concatenation-based indexing suffers catastrophic overall degradation (-0.145 Recall@10) due to cross-field interference. The method requires no additional training and serves as a drop-in replacement for existing retrieval pipelines.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 Introduction

Long-term conversational assistants must recall information from extensive interaction histories spanning weeks or months (Packer et al., 2023; Zhong et al., 2023). As these histories grow beyond context window limits, retrieval-augmented approaches become essential: given a user query, the system retrieves relevant past conversation rounds to inform its response (Gao et al., 2023). The effectiveness of such memory systems depends critically on how conversation rounds are indexed for retrieval.

Current conversational memory systems predominantly index only user utterances, treating each round's user turn as the retrieval key (Maharana et al., 2024). This design reflects an implicit assumption that users query about information they previously provided. However, the LongMemEval benchmark (Wu et al., 2024) identifies a distinct category of queries—single-session-assistant (SSA) questions—where users ask about information the *assistant* provided in earlier turns, such as "What restaurant did you recommend last week?" For SSA queries, the relevant information resides in assistant responses, creating a fundamental mismatch with user-only indexing.

A natural solution is to concatenate user and assistant text into a single embedding. However, this approach suffers from cross-field interference: mixing semantically different content in one embedding dilutes query-relevant features. Our experiments confirm this hypothesis—concatenation improves SSA recall but causes catastrophic degradation on other query types (-0.145 overall Recall@10), making it unsuitable for production deployment.

We propose *fielded max-sim keying*, which treats each conversation round as a two-field document with separate embeddings for user and assistant content. The relevance score is the maximum similarity across fields, providing a natural field selection mechanism: user-centric queries match the user field while SSA queries match the assistant field. This approach requires no additional training—only a change in indexing and scoring. Our contributions are:

[1] https://gitlab.com/fars-a/assistant-inclusive-keying-longmemeval

- We demonstrate a *zero-harm property*: fielded max-sim achieves identical overall Recall@10 (0.664) to user-only indexing while improving SSA Recall@10 by +0.018 and NDCG@10 by +0.038, enabling risk-free adoption as a drop-in replacement.
- We provide empirical evidence of cross-field interference in concatenation-based indexing, which degrades overall Recall@10 by -0.145 despite improving SSA recall, explaining why naive field combination fails.
- We compare max-sim against linear mixture scoring and show that only the non-linear max operator achieves zero-harm—linear mixtures trade SSA gains for unacceptable overall degradation.

2 Related Work

2.1 Long-Term Conversational Memory

Memory systems for conversational agents have evolved from simple context windows to sophisticated architectures that manage information across extended interactions. Early approaches such as MemGPT (Packer et al., 2023) drew inspiration from operating system memory hierarchies, implementing virtual context management to provide the illusion of extended memory through intelligent paging between main context and external storage. MemoryBank (Zhong et al., 2023) introduced a memory mechanism incorporating the Ebbinghaus forgetting curve to selectively preserve and reinforce memories based on temporal recency and significance.

Recent work has focused on temporal and hierarchical organization of conversational memories. TiMem (Li et al., 2026) organizes conversations through a Temporal Memory Tree that enables systematic consolidation from raw observations to abstracted persona representations. MemWeaver (Ye et al., 2026) consolidates experiences into three interconnected components: a temporally grounded graph memory, an experience memory for recurring patterns, and a passage memory for textual evidence. SwiftMem (Tian et al., 2026) addresses retrieval latency through query-aware indexing over temporal and semantic dimensions, achieving sub-linear retrieval times. ENGRAM (Patel & Patel, 2025) takes a minimalist approach, organizing conversations into episodic, semantic, and procedural memory types through a single router and retriever. Zep (Rasmussen et al., 2025) employs a temporal knowledge graph architecture that dynamically synthesizes conversational and structured data while maintaining historical relationships. A comprehensive survey by Zhang et al. (2024b) systematically reviews memory mechanisms for LLM-based agents, categorizing approaches by memory sources, forms, and operations.

2.2 Conversational Memory Benchmarks

Evaluating long-term memory capabilities requires benchmarks that capture the complexity of sustained interactions. LoCoMo (Maharana et al., 2024) introduced a dataset of very long-term conversations spanning up to 35 sessions with 300 turns on average, evaluating models on question answering, event summarization, and multi-modal dialogue generation. PerLTQA (Du et al., 2024) combines semantic and episodic memories including world knowledge, profiles, social relationships, and events, providing 8,593 questions for evaluating personalized memory integration. DialSim (Kim et al., 2024) presents a real-time dialogue simulator that evaluates agents on multi-party dialogues with extended contextual dependencies.

LongMemEval (Wu et al., 2024) provides a comprehensive benchmark evaluating five core long-term memory abilities: information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention. Critically for our work, LongMemEval categorizes questions by the source of the answer, with the single-session-assistant (SSA) type exposing a gap in retrieval systems that index only user utterances.

2.3 Dense Retrieval and Multi-Field Retrieval

Dense passage retrieval (Karpukhin et al., 2020) demonstrated that learned dense representations can substantially outperform traditional sparse methods like BM25 for open-domain question answering. ColBERT (Khattab & Zaharia, 2020) introduced late interaction, encoding queries and documents

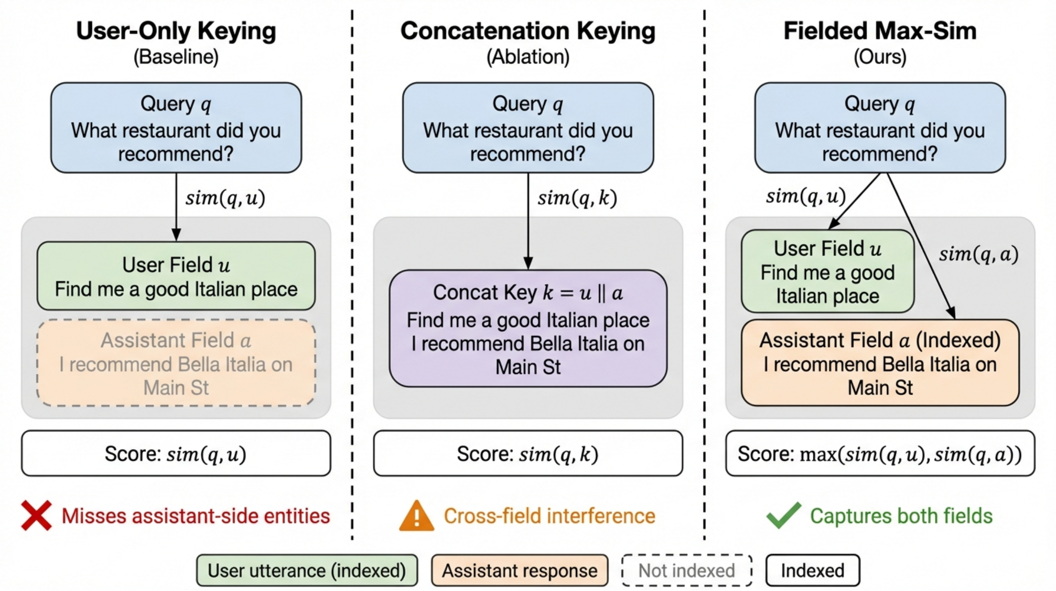
User-Only Keying
(Baseline)
Query q
What restaurant did you recommend?
$sim(q,u)$
User Field u
Find me a good Italian place
Assistant Field a
I recommend Bella Italia on Main St
Score: $sim(q,u)$
✗ Misses assistant-side entities
Concatenation Keying
(Ablation)
Query q
What restaurant did you recommend?
$sim(q,k)$
Concat Key $k = u \| a$
Find me a good Italian place
I recommend Bella Italia on Main St
Score: $sim(q,k)$
⚠ Cross-field interference
Fielded Max-Sim
(Ours)
Query q
What restaurant did you recommend?
$sim(q,u)$
$sim(q,a)$
User Field u
Find me a good Italian place
Assistant Field a (Indexed)
I recommend Bella Italia on Main St
Score: $\max(sim(q,u), sim(q,a))$
✓ Captures both fields
User utterance (indexed)
Assistant response
Not indexed
Indexed

Figure 1: Comparison of three keying strategies for conversational memory retrieval. (a) User-Only: indexes only user utterances, missing assistant-side information. (b) Concatenation: combines user and assistant text into a single embedding, causing cross-field interference. (c) Fielded Max-Sim (Ours): embeds user and assistant fields separately and scores by max similarity, preserving field-specific signals.

independently before computing fine-grained token-level similarity, enabling efficient retrieval while preserving expressiveness. These dense retrieval techniques underpin most modern conversational memory systems (Gao et al., 2023).

Multi-field retrieval addresses documents with explicit structure. Li et al. (2024) introduced Multi-Field Adaptive Retrieval (mFAR), which decomposes documents into fields indexed independently and learns to adaptively weight fields based on the query. Our work applies a similar intuition to conversational memory: treating each conversation round as a two-field document with user and assistant fields. However, rather than learning field weights, we employ a simple max-sim operator that naturally selects the better-matching field per query, achieving targeted improvement without training.

3 Method

3.1 Problem Formalization

We consider a conversational memory system that maintains a history of user-assistant interactions. Let $\mathcal{H} = \{(u_1, a_1), (u_2, a_2), \ldots, (u_n, a_n)\}$ denote the conversation history, where each round i consists of a user utterance u_i and an assistant response a_i. Given a query q at test time, the retrieval task is to identify the top-k rounds from $\mathcal{H}$ that contain information relevant to answering q.

Following the LongMemEval benchmark (Wu et al., 2024), we distinguish questions by the source of their answer. Single-session-assistant (SSA) queries target information in assistant responses a_i, creating a mismatch when retrieval systems index only user utterances u_i.

3.2 Keying Strategies

We compare three strategies for constructing retrieval keys from conversation rounds, illustrated in Figure 1.

User-Only Keying. The standard approach indexes only user utterances. For round i, the key is $\mathbf{k}_i = \text{enc}(u_i)$, and the relevance score is:

$$s_{\text{user}}(q, i) = \text{sim}(\text{enc}(q), \text{enc}(u_i)) \tag{1}$$

where $\text{enc}(\cdot)$ is a dense encoder and $\text{sim}(\cdot, \cdot)$ is cosine similarity. This approach works well when queries target user-provided information but fails for SSA queries where the answer resides in assistant turns.

Concatenation Keying. A natural extension concatenates user and assistant text: $\mathbf{k}_i = \text{enc}(u_i \oplus a_i)$, where $\oplus$ denotes string concatenation. The score becomes:

$$s_{\text{concat}}(q, i) = \text{sim}(\text{enc}(q), \text{enc}(u_i \oplus a_i)) \tag{2}$$

While this includes assistant information, it suffers from cross-field interference, degrading retrieval for non-SSA queries as demonstrated in our experiments.

Fielded Max-Sim Keying. We propose treating each round as a two-field document with separate embeddings for user and assistant content. The relevance score is the maximum similarity across fields:

$$s_{\text{maxsim}}(q, i) = \max\left(\text{sim}(\text{enc}(q), \text{enc}(u_i)), \text{sim}(\text{enc}(q), \text{enc}(a_i))\right) \tag{3}$$

This preserves field-specific signals while allowing either field to match the query.

3.3 Why Max-Sim Works

The max operator provides a natural field selection mechanism. For user-centric queries where $\text{sim}(q, u_i) > \text{sim}(q, a_i)$, the max-sim score reduces to the user-only score, preserving baseline performance. For SSA queries where $\text{sim}(q, a_i) > \text{sim}(q, u_i)$, the assistant field dominates, enabling retrieval of assistant-provided information.

This behavior contrasts with linear mixture scoring $s_{\text{mix}}(q, i) = \alpha \cdot \text{sim}(q, u_i) + (1 - \alpha) \cdot \text{sim}(q, a_i)$, which always incorporates both fields regardless of query type. When the assistant field is irrelevant (non-SSA queries), the mixture score is diluted by the low-similarity assistant term, degrading retrieval. The max operator avoids this by selecting only the better-matching field.

The method requires no additional training—only a change in indexing (embedding both fields separately) and scoring (taking the maximum). The computational overhead is approximately 2× for embedding storage, as each round requires two embeddings instead of one.

4 Experiments

4.1 Experimental Setup

We evaluate on LongMemEval (Wu et al., 2024), a benchmark for long-term conversational memory with 500 questions across six types: single-session-user (SSU), single-session-assistant (SSA), single-session-preference (SSP), multi-session (MS), knowledge-update (KU), and temporal-reasoning (TR). Following the benchmark protocol, we exclude 30 abstention questions, evaluating on 470 questions. SSA questions (56 instances) specifically target assistant-provided information, making them the primary focus of our evaluation.

We use Stella EN 1.5B v5 (Zhang et al., 2024a) as the dense encoder, a state-of-the-art embedding model. All embeddings are L2-normalized with maximum sequence length of 512 tokens. We report Recall@10 (fraction of questions where all relevant documents appear in top-10) and NDCG@10 (ranking quality metric).

4.2 Main Results

Table 1 presents the main retrieval results. The key finding is the *zero-harm property*: Fielded MaxSim achieves identical overall Recall@10 (0.664) to the User-Only baseline while improving SSA-specific Recall@10 by +0.018 (from 0.893 to 0.911). The NDCG@10 improvement on SSA

Table 1: Main retrieval results on LongMemEval (470 questions, 6 types). Fielded MaxSim achieves identical overall Recall@10 to User-Only while improving SSA by +0.018. Concatenation suffers catastrophic overall degradation (−0.145). Best in **bold**.

Method	Overall		Per-Type Recall@10					
	R@10	NDCG@10	SSA	SSU	KU	SSP	MS	TR
User-Only	**0.664**	0.617	0.893	**0.875**	**0.875**	**0.667**	**0.488**	**0.504**
Concat	0.519	0.479	**0.911**	0.797	0.667	0.500	0.339	0.299
Fielded MaxSim (Ours)	**0.664**	**0.621**	**0.911**	**0.875**	**0.875**	**0.667**	0.479	**0.504**

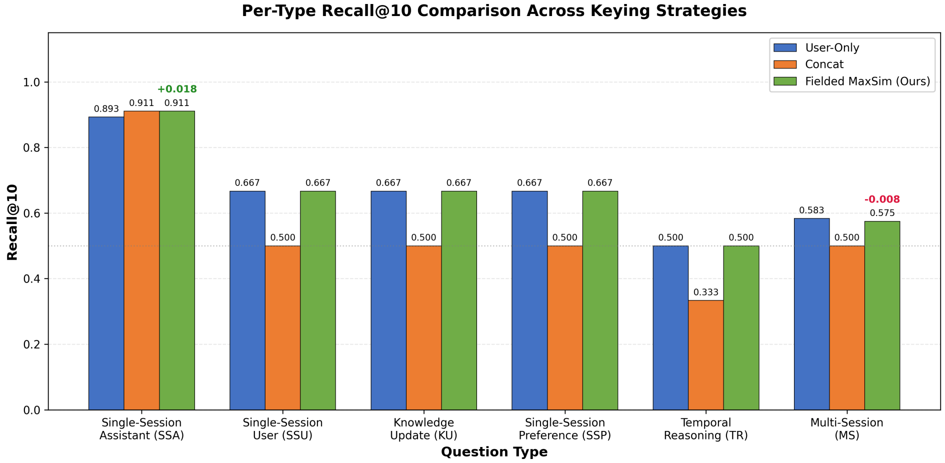

Figure 2: Per-type Recall@10 comparison across keying strategies on LongMemEval. Fielded MaxSim (green) achieves +0.018 improvement on SSA while maintaining identical performance to User-Only (blue) on 4/5 other types. Concatenation (orange) shows catastrophic degradation on non-SSA types despite matching SSA recall.

is even larger (+0.038, from 0.795 to 0.834), indicating that correct SSA documents are promoted to higher ranks, not just newly retrieved.

Figure 2 visualizes the per-type breakdown. Fielded MaxSim matches User-Only exactly on four of five non-SSA types (SSU, KU, SSP, TR), with only MS showing a marginal decrease (−0.008). In contrast, concatenation suffers catastrophic degradation across all non-SSA types (average −0.161 Recall@10), confirming the cross-field interference hypothesis: mixing user and assistant text in a single embedding dilutes query-relevant features.

4.3 Ablation: Mixture Scoring

Table 2 compares max-sim against linear mixture scoring $s = \alpha \cdot \mathrm{sim}(q, u) + (1 - \alpha) \cdot \mathrm{sim}(q, a)$. Linear mixtures can achieve higher SSA gains: α=0.5 reaches 0.964 SSA Recall@10 (+0.071 over baseline). However, this comes at unacceptable overall cost (−0.043 Recall@10). Even α=0.7 fails the threshold with −0.011 overall degradation.

The non-linear max operator is essential to the zero-harm property. By selecting only the better-matching field per query, max-sim avoids diluting scores when one field is irrelevant. Linear mixtures always incorporate both fields, degrading performance when the assistant field adds noise to non-SSA queries.

4.4 Analysis

The modest SSA improvement (+0.018 Recall@10) reflects a ceiling effect: User-Only already achieves 0.893 SSA Recall@10, leaving only 0.107 potential gain. Fielded MaxSim captures 17%

Table 2: Mixture scoring analysis: linear combinations of user and assistant similarity fail to achieve the zero-harm property. Only max-sim passes the acceptable threshold (overall drop $\leq$0.005). Best SSA in **bold**.

Method	SSA R@10	Overall R@10	Δ Overall	Threshold
User-Only (α=1.0)	0.893	0.664	0.000	✓
Mixture (α=0.7)	0.946	0.653	$-$0.011	×
Mixture (α=0.5)	**0.964**	0.621	$-$0.043	×
Fielded MaxSim	0.911	**0.664**	**0.000**	✓

of this headroom. The larger NDCG improvement (+0.038) suggests the method's primary benefit is promoting relevant documents to higher ranks rather than retrieving previously-missed documents.

With 56 SSA instances, individual statistical significance is limited. However, the consistency across metrics (Recall@5, Recall@10, NDCG@10) and the exact preservation of non-SSA performance provide confidence in the findings. The method's practical value lies in its zero-risk profile: it can be adopted as a drop-in replacement with guaranteed non-regression on existing query types.

5 Conclusion

We introduced fielded max-sim keying for conversational memory retrieval, treating each conversation round as a two-field document with separate user and assistant embeddings. The method achieves a zero-harm property: identical overall retrieval performance to user-only indexing while improving SSA recall. The approach requires no additional training and can be adopted as a drop-in replacement with guaranteed non-regression.

Limitations include the modest SSA improvement (+0.018 Recall@10) due to ceiling effects in the benchmark, and approximately 2× embedding storage overhead. Future work could explore learned field weighting, extension to other conversational tasks, and evaluation on benchmarks with more challenging SSA distributions.

References

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. *ArXiv*, abs/2402.16288, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906, 2020.

O. Khattab and M. Zaharia. *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*. 2020.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. Dialsim: A dialogue simulator for evaluating long-term multi-party dialogue understanding of conversational agents. 2024.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, Chengbao Liu, and Jie Tan. Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. *ArXiv*, abs/2601.02845, 2026.

Millicent Li, Tongfei Chen, Benjamin Van Durme, and Patrick Xia. Multi-field adaptive retrieval. *ArXiv*, abs/2410.20056, 2024.

Adyasha Maharana, Dong-Ho Lee, S. Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *ArXiv*, abs/2402.17753, 2024.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph Gonzalez. Memgpt: Towards llms as operating systems. *ArXiv*, abs/2310.08560, 2023.

Daivik Patel and Shrenik Patel. Engram: Effective, lightweight memory orchestration for conversational agents. *ArXiv*, abs/2511.12960, 2025.

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

Anxin Tian, Yiming Li, Xing Li, Hui-Ling Zhen, Lei Chen, Xianzhi Yu, Zhenhua Dong, and Mingxuan Yuan. Swiftmem: Fast agentic memory via query-aware indexing. *ArXiv*, abs/2601.08160, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *ArXiv*, abs/2410.10813, 2024.

Juexiang Ye, Xue Li, Xinyu Yang, Chengkai Huang, Lanshun Nie, Lina Yao, and Dechen Zhan. Memweaver: Weaving hybrid memories for traceable long-horizon agentic reasoning. 2026.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models. *ArXiv*, abs/2412.19048, 2024a.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43:1 – 47, 2024b.

Wanjun Zhong, Lianghong Guo, Qi-Fei Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *ArXiv*, abs/2305.10250, 2023.