LEARNING TO RETRIEVE LONG-TERM MEMORIES WITH UNLABELED CONVERSATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory in LM agents relies on retrieving relevant information from external memory stores. Despite training on massive query-document pairs, general-purpose embedding models struggle with long-term memory retrieval, as query–memory relevance labels are rare and costly to obtain. In this work, we propose adding self-supervised retriever learning on unlabeled conversations to the training recipe to fill this gap. Specifically, we introduce MemRevela, which learns retrieval directly from the dependencies between conversational turns. An auxiliary language model predicts conversational fragments using retriever-weighted context, allowing the prediction loss to train the retriever. MemRevela then jointly optimizes this objective with standard MS MARCO contrastive learning. We evaluate five model sizes from 0.1B to 8B on LMEB's 22 datasets across four memory types. Without any task instructions, `MemRevela-8B` achieves the best performance in our comparison with dataset-averaged nDCG@10 **63.24**, exceeding `NV-Embed-v2` by **3.43** absolute points. Our method also improves downstream question answering over long conversations compared with two proprietary embedding APIs. Without annotated labels, MemRevela injects long-term memory knowledge into the retriever: it spreads conversational turns in embedding space about as widely as general passages, whereas general-purpose models cluster them tightly. We show that conversational self-supervision and contrastive learning are complementary, each strengthening different memory types. Finally, the memory gains do not come at the expense of general retrieval: MemRevela stays comparable to contrastive-only baselines on BEIR, making conversational self-supervision a promising addition to standard retriever training.

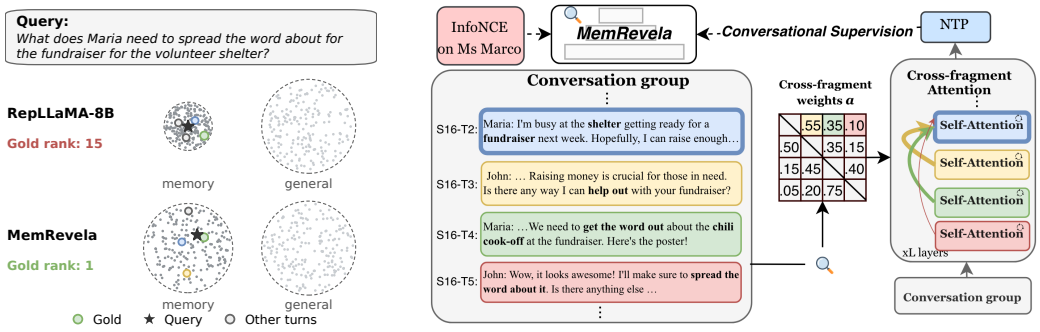
1 INTRODUCTION

Language agents keep past exchanges, facts, and experiences in external memory (Chhikara et al., 2025; Xu et al., 2025), whose usefulness depends on retrieving evidence that supports the current task. This selection is hard because many memories share the same entities and topics: a turn on the right topic may describe a different event, and the evidence a query needs is often spread across several fragments of a conversation.

Strong general-purpose embedding models handle this setting poorly despite training on massive text corpora. Performance on general embedding benchmarks such as MTEB (Muennighoff et al., 2023) does not reliably transfer to memory retrieval (Zhao et al., 2026b). We find that conversational memory is out of distribution for these models: they place conversational turns in a much tighter region of embedding space than ordinary passages, so turns that share a topic become hard to tell apart and the evidence a query needs is ranked below its neighbours (Figure 1a; §5.1). Learning representations that separate supporting evidence from related memories is therefore central to memory retrieval in conversational or agentic systems.

The natural remedy is to train on conversational text itself, but conversations rarely come with retrieval labels. Query–memory relevance annotations are costly to create. Unlabeled conversations, by contrast, are abundant and already contain useful dependencies: participants elaborate on earlier facts, clarify references, explain decisions, and revisit solutions. These dependencies resemble the cross-fragment relationships that long-term memory retrieval must recover. We therefore ask whether they can supervise a retriever without explicit query–memory pairs.

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



(a) Memory retrieval (b) MemRevela training

Figure 1: **Memory retrieval and training.** (a) General-purpose and contrastive-only retrievers place conversational turns in a much tighter region of embedding space than general passages, so turns on the same topic crowd around the query and outrank the evidence. In this LoCoMo example with 663 candidates, RepLLaMA-8B (same backbone and labels, no conversations) ranks the gold turn 15th, while MemRevela spreads memory turns about as widely as general passages and ranks it first. (b) Retriever scores weight the fragments an auxiliary LM uses to predict each conversational fragment, so the prediction loss trains the retriever without labels; MemRevela then trains jointly with MS MARCO contrastive learning, and the auxiliary LM is discarded at inference.

Our approach, MemRevela, learns from these dependencies through cross-fragment prediction (Figure 1b). Retriever scores weight the fragments available to an auxiliary language model, allowing prediction gradients to train the retriever through those weights. The mechanism builds on retriever-weighted language modeling (Cai et al., 2026a); our focus is learning memory retrieval representations from unlabeled conversations. Beyond this signal, MemRevela needs only standard MS MARCO query–passage pairs: it first learns from conversations and then adds contrastive learning on these pairs, with no memory-specific annotations. At inference, the auxiliary model is discarded, and memories are encoded independently for standard dense retrieval.

We evaluate MemRevela at five sizes, from 0.1B to 8B parameters, on LMEB's 22 datasets spanning episodic, dialogue, semantic, and procedural memory (Zhao et al., 2026b). Without task instructions, MemRevela-8B is the strongest retriever in our comparison at 63.24 dataset-averaged nDCG@10, and MemRevela-1B already surpasses the much larger NV-Embed-v2 (Lee et al., 2025). At the small end, MemRevela-0.6B is the best sub-billion retriever and overtakes Qwen3-Embedding-4B and 8B (Zhang et al., 2025). Against a contrastive-only retriever trained on the same backbone and labels, MemRevela improves at every size from 0.6B to 8B. On LoCoMo (Maharana et al., 2024), the 8B retriever also improves downstream question answering over supervised models tuned on massive data and two proprietary APIs.

Beyond benchmark scores, analysis of the learned representations shows that MemRevela spreads conversational turns about as widely as general passages (Figure 1a) and, within a conversation, separates supporting evidence from nearby turns more sharply than the evaluated general-purpose embedding models. Evaluation on 13 BEIR datasets (Thakur et al., 2021) shows that general retrieval is broadly preserved. Ablations further show that conversational initialization and contrastive supervision are complementary: each is stronger on different memory types, and combining them outperforms either alone on nearly every type.

Our contributions are three-fold as follows:

- We identify a representational mismatch in memory retrieval: general-purpose embedding models place conversational turns in a much tighter region of embedding space than ordinary passages, making same-topic turns hard to separate from the evidence a query needs.
- To mitigate this mismatch, we introduce MemRevela, which trains the retriever on unlabeled conversations by having an auxiliary LM predict each fragment from retriever-weighted other fragments and then trains it jointly with contrastive learning, without any query–memory pairs.
- From 0.1B to 8B parameters, our method achieves the best LMEB score in every size group we compare. It spreads conversational turns about as widely as general passages while preserving general retrieval on BEIR, and its conversational and contrastive supervision are complementary.

2 Related Work

Memory Systems and Memory Retrieval Research on long-term agent memory spans memory construction, organization, and retrieval. Mem0 extracts and consolidates salient information from conversations, while A-MEM organizes memories into interconnected notes through dynamic linking and revision (Chhikara et al., 2025; Xu et al., 2025). At the retrieval level, LMEB evaluates embeddings across episodic, dialogue, semantic, and procedural memories, showing that strong performance on conventional text embedding benchmarks does not necessarily transfer to long-horizon memory retrieval (Zhao et al., 2026b). Recent methods improve different components of memory access. Nano-Memory retrieves conversational evidence using maximum query–turn similarity and prunes redundant context, while lexical–dense fusion combines turn-level dense scores with BM25 without additional retriever training (Wu et al., 2026; Lysenstøen, 2026). EvoEmbedding conditions embeddings on an evolving latent memory to incorporate preceding context and temporal order (Nie et al., 2026). These methods change how memories are accessed or condition each embedding on an evolving memory state. MemRevela instead changes only how the retriever is trained: memories remain independently encoded and compatible with standard offline indexing.

LM-guided Retriever Learning Language models can supply feedback for retriever training through their use of retrieved context. RAA and Atlas couple document selection to a reader's downstream prediction (Jiang et al., 2022; Izacard et al., 2023), while REPLUG uses a frozen LM's likelihood of a desired continuation to derive retrieval targets (Shi et al., 2024). Revela and ElicitR establish retriever-weighted language modeling as a mechanism for learning cross-text dependencies: Revela jointly trains a retriever and LM on raw text, and ElicitR incorporates the mechanism into supervised fine-tuning on query–document groups (Cai et al., 2026a;b). ElicitR applies this prediction within labeled query–positive–negative groups, as a regularizer of contrastive fine-tuning for reasoning-intensive retrieval. MemRevela instead keeps the prediction on unlabeled conversations, a data stream separate from the labeled pairs, in both training stages; the dependencies it learns therefore come from conversational structure rather than from query–document groups, and the labeled supervision stays plain contrastive learning on MS MARCO.

Generalization and Domain Adaptation of Dense Retrievers Transferring retrieval models across tasks requires accommodating differences in document distributions and relevance criteria. BEIR evaluates this challenge across heterogeneous datasets, revealing limitations in the zero-shot generalization of dense retrievers (Thakur et al., 2021). GPL adapts retrievers using queries generated from target-domain documents and relevance scores supplied by a cross-encoder (Wang et al., 2022a), while Promptagator uses a few task-specific examples to guide query generation and train specialized retrievers (Dai et al., 2023). Both rely on a generator model to synthesize queries as training data. MemRevela generates no training data: it learns from dependencies already present in unlabeled conversations, and its conversational corpus is separate from every evaluation collection.

3 Method

MemRevela learns memory retrieval from predictive dependencies in unlabeled conversations: a dense retriever weights the fragments used by an auxiliary language model (LM) to predict another fragment, and prediction gradients train the retriever. We define the retrieval representations in Section 3.1, describe conversational supervision and its combination with contrastive learning in Section 3.2, and present the training schedule and inference procedure in Section 3.3.

3.1 Preliminaries

Given a query q and a memory collection $\mathcal{M} = \{m_1, \ldots, m_N\}$, a shared encoder R_θ maps queries and memories to dense representations. We use the final hidden state at an appended end-of-sequence token and apply ℓ_2 normalization:

$$r_\theta(x) = \frac{R_\theta(x)_{\mathrm{EOS}}}{\|R_\theta(x)_{\mathrm{EOS}}\|_2}, \qquad s_\theta(q, m) = r_\theta(q)^\top r_\theta(m). \tag{1}$$

Retrieval returns the k highest-scoring memories. Each memory is encoded independently, allowing the collection to be indexed in advance. During training, an auxiliary LM M_ϕ uses the same retriever similarities to weight contexts for next-token prediction.

3.2 Retrieval Learning with Conversational Supervision

We combine two objectives that share the retriever: contrastive learning on labeled query–passage examples and generative supervision from unlabeled conversations. The conversational objective follows the retriever-weighted conditioning framework of Revela (Cai et al., 2026a), using conversational fragments to supply predictive relationships beyond the labeled retrieval examples.

Supervised contrastive learning. For a query batch $\mathcal{Q}_\mathcal{B}$, let $d^+(q)$ denote the positive passage for q and $\mathcal{D}_\mathcal{B}$ the positive and negative passages gathered across the distributed retrieval batch. We optimize

$$\mathcal{L}_{\text{ret}} = -\frac{1}{|\mathcal{Q}_\mathcal{B}|} \sum_{q \in \mathcal{Q}_\mathcal{B}} \log \frac{\exp(s_\theta(q, d^+(q))/\tau_{\text{ret}})}{\sum_{d \in \mathcal{D}_\mathcal{B}} \exp(s_\theta(q, d)/\tau_{\text{ret}})}, \tag{2}$$

where τ_{ret} is the retrieval temperature. This objective directly supervises query–passage relevance.

Learning conversational dependencies. We split conversational text into fragments and place consecutive fragments into groups $\mathcal{G} = \{x_1, \ldots, x_K\}$. Groups preserve fragment order, although a group may cross conversation boundaries. Each fragment is predicted using its own prefix and the other fragments in its group. Group membership alone supplies no relevance label; the retriever determines the contribution of each candidate context through

$$\alpha_{ij} = \frac{\exp(s_\theta(x_i, x_j)/\tau_{\text{mem}})}{\sum_{u \neq i} \exp(s_\theta(x_i, x_u)/\tau_{\text{mem}})}, \quad j \neq i, \qquad \alpha_{ii} = 0, \tag{3}$$

where τ_{mem} controls the concentration of the weights. The weights are computed from complete observed fragments and shared across LM layers and token positions.

Let $x_{-i} = \mathcal{G} \setminus \{x_i\}$ and $\mathcal{T}_i$ be the valid next-token prediction positions in x_i. The conversational objective is

$$\mathcal{L}_{\text{mem}}(\mathcal{G}) = -\frac{1}{\sum_i |\mathcal{T}_i|} \sum_{i=1}^{K} \sum_{t \in \mathcal{T}_i} \log P_{\theta,\phi}(x_{i,t} \mid x_{i,<t}, x_{-i}; \boldsymbol{\alpha}_i)\,. \tag{4}$$

The LM learns to use information across fragments, while prediction gradients reach the retriever through $\boldsymbol{\alpha}_i$. Thus, predictive usefulness provides a training signal for fragment similarities without generated queries or query–memory annotations.

Retriever-weighted attention. We implement cross-fragment conditioning as an additional attention contribution in each transformer layer of M_ϕ. The LM first processes every fragment independently with standard causal attention, caching its layer-wise keys and values. A second forward pass predicts each fragment using both its local prefix and the cached representations of the other fragments. Both passes use the same LM parameters.

For one attention head at layer ℓ, let $(K_j^{e,\ell}, V_j^{e,\ell})$ be the cached projections for fragment x_j, and $(Q_i^{h,\ell}, K_i^{h,\ell}, V_i^{h,\ell})$ the projections in the prediction pass for x_i. The local attention contribution is

$$S_i^\ell = \text{softmax}\left(\frac{Q_i^{h,\ell}(K_i^{h,\ell})^\top}{\sqrt{d_H}}\right) V_i^{h,\ell}, \tag{5}$$

where d_H is the head dimension. Cross-fragment attention queries each other fragment separately and combines the resulting outputs with the retriever weights:

$$B_i^\ell = \sum_{j \neq i} \alpha_{ij}\ \text{softmax}\left(\frac{Q_i^{h,\ell}(K_j^{e,\ell})^\top}{\sqrt{d_H}}\right) V_j^{e,\ell}. \tag{6}$$

Local attention is causally masked, while cross-fragment attention accesses all non-padding source tokens; masks are omitted from the notation. The retriever weights control which fragments contribute, and token-level attention selects information within each fragment. For each head a, the two contributions are added before the standard output projection:

$$A_i^\ell = \mathrm{Concat}_{a=1}^{H}\big(S_{i,a}^\ell + B_{i,a}^\ell\big)\, W_O^\ell. \tag{7}$$

The LM's residual, normalization, and feed-forward operations remain unchanged. The independently computed caches ensure that source representations do not themselves incorporate cross-fragment information. This construction lets the prediction objective train the retriever through a differentiable mixture of contexts.

Joint objective. For a batch of conversational groups $\mathcal{B}_{\mathrm{mem}}$, we average the group losses and combine them with contrastive learning:

$$\mathcal{L}_{\mathrm{joint}} = c\,\mathcal{L}_{\mathrm{ret}} + \frac{1-c}{|\mathcal{B}_{\mathrm{mem}}|}\sum_{\mathcal{G}\in\mathcal{B}_{\mathrm{mem}}} \mathcal{L}_{\mathrm{mem}}(\mathcal{G}), \tag{8}$$

where $c \in [0,1]$ balances the two signals. Labeled retrieval examples and conversational groups are sampled separately, with no required correspondence between them. Both losses update the retriever; only conversational prediction updates the auxiliary LM. This allows explicit relevance supervision and unlabeled conversational dependencies to shape a shared representation space.

3.3 Training and Inference

Both the retriever and auxiliary LM start from pretrained LM checkpoints, and we train LoRA adapters in both models through two stages. We first optimize only $\mathcal{L}_{\mathrm{mem}}$ (Eq. 4) on unlabeled conversations, adapting the auxiliary LM to cross-fragment conditioning while learning the retriever weights that govern it (stage 1). Starting from these adapters, we then introduce labeled query–passage examples and optimize $\mathcal{L}_{\mathrm{joint}}$ (Eq. 8), retaining conversational supervision alongside supervised retrieval learning (stage 2). Backbone choices and hyperparameters are specified in Section 4. At inference, we discard the auxiliary LM and use the trained retriever to encode queries and memory entries independently, allowing memories to be indexed in advance and retrieved by standard vector search using Eq. 1.

4 Experiments

4.1 Experimental Setup

Benchmark and metrics. We evaluate memory retrieval on LMEB (Zhao et al., 2026b), covering 22 datasets grouped into episodic, dialogue, semantic, and procedural memory. We report nDCG@10 and Recall@10 without task instructions, following the organization of the benchmark's main results table. For datasets containing multiple subsets, we first average the subset scores to obtain a dataset score. We then report the mean over all 22 datasets (Dataset Mean), the mean within each memory type, and the unweighted average of the four type scores (Type Mean). Dataset-averaged nDCG@10 is our primary metric.

Training configuration. We train retrievers at 1B, 3B, and 8B from Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B, and a 0.6B retriever from Qwen3-0.6B-Base (Yang et al., 2025) with the same recipe. We additionally evaluate a 0.1B retriever initialized from `HuggingFaceTB/SmolLM2-135M`. Both stages update LoRA adapters of rank 256 on all attention and MLP projections; stage 1 starts from the pretrained backbone, and stage 2 starts from the stage-1 adapters of both the retriever and the auxiliary LM. The auxiliary LM is Llama-3.2-1B for all sizes. Stage 2 uses $c = 0.95$ for all reported sizes. The 8B checkpoint uses a second-stage learning rate of 2×10^{-5} and a contrastive passage length limit of 160 tokens; the 0.1B, 0.6B, 1B, and 3B checkpoints use 5×10^{-5} and 196 tokens. Appendix C provides the full two-stage configuration.

Baselines. We compare with 15 general-purpose embedding models from 0.3B to 12B parameters, including the `Qwen3-Embedding` family, `NV-Embed-v2`, `E5-Mistral-7B-Instruct`, and `BGE`, `KaLM`,

Table 1: Results of LMEB retrieval without task instructions (×100). N@10/R@10 denote nDCG@10/Recall@10; Dataset/Type Mean average 22 datasets/four memory types. **Bold**/underline indicate best/second-best within each size group. Per-dataset results: Appendix D.

		Episodic		Dialogue		Semantic		Procedural		Dataset Mean		Type Mean	
Retriever	Size (B)	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10
<1B													
MemRevela-0.1B	0.1	48.66	57.48	49.48	58.29	50.72	66.71	46.30	53.38	48.99	59.94	48.79	58.96
EmbeddingGemma-300M	0.3	**68.19**	**74.85**	53.96	61.52	53.67	69.32	57.31	65.40	56.06	66.63	58.28	67.77
bge-large-en-v1.5	0.3	55.93	65.17	53.12	61.41	52.55	67.92	52.59	60.68	53.03	63.92	53.55	63.79
KaLM_Embedding_V1.5	0.5	56.60	65.97	41.31	49.29	47.47	64.03	54.84	63.27	48.63	59.98	50.06	60.64
KaLM_Embedding_V2.5	0.5	61.86	69.23	48.93	58.07	52.02	68.14	52.97	62.18	52.33	63.87	53.95	64.41
bge-m3	0.6	66.98	73.16	55.62	63.26	56.28	71.25	55.35	62.57	56.82	66.88	58.56	67.56
Qwen3-Embedding-0.6B	0.6	57.63	65.40	50.41	58.41	52.03	67.78	57.12	65.26	53.49	64.32	54.30	64.22
snowflake-arctic-embed-l-v2.0	0.6	63.74	70.33	50.34	59.44	52.36	68.20	54.90	62.46	53.54	64.44	55.34	65.11
harrier-oss-v1-0.6b	0.6	61.98	69.00	44.84	52.81	49.15	66.50	**58.95**	**66.53**	51.81	63.00	53.73	63.71
MemRevela-0.6B	0.6	61.53	67.78	**56.90**	**64.98**	**59.38**	**74.59**	58.14	64.88	**58.56**	**68.70**	**58.99**	**68.05**
1–7B													
Qwen3-Embedding-4B	4	58.32	67.12	41.20	48.87	53.76	70.36	56.34	65.82	51.45	62.97	52.41	63.04
F2LLM-v2-4B	4	68.01	73.78	53.83	62.21	56.54	72.03	60.25	**67.62**	57.86	68.31	59.66	68.91
MemRevela-1B	1	65.79	72.08	**58.86**	66.04	60.25	74.59	59.41	66.65	60.15	69.87	61.08	69.84
RepLLaMA-3B	3	66.96	72.32	58.62	65.46	60.68	75.17	59.51	66.14	60.37	69.80	61.44	69.77
MemRevela-3B	3	**69.37**	**74.69**	58.69	**66.34**	**62.05**	**75.64**	**60.42**	66.83	**61.36**	**70.61**	**62.63**	**70.87**
7–12B													
e5-mistral-7b-instruct	7	57.22	66.37	48.51	57.98	46.00	63.66	52.99	63.04	49.61	62.19	51.18	62.76
NV-Embed-v2	8	70.43	76.99	56.48	65.43	59.14	73.69	60.49	68.64	59.81	70.36	61.63	71.19
Qwen3-Embedding-8B	8	61.03	68.39	49.01	57.59	55.53	71.94	56.97	66.57	54.65	66.24	55.64	66.12
bge-multilingual-gemma2	9	54.78	62.19	40.70	49.82	40.70	59.93	52.19	60.52	45.11	57.54	47.09	58.12
KaLM-Embedding-Gemma3	12	66.99	74.12	50.87	59.73	47.83	65.88	60.72	**69.11**	53.91	65.83	56.60	67.21
RepLLaMA-8B	8	68.07	73.86	59.92	66.08	61.35	75.52	59.99	67.15	61.20	70.51	62.33	70.65
MemRevela-8B	8	**72.26**	**77.30**	**61.15**	**67.83**	**63.69**	**76.54**	**61.72**	68.60	**63.24**	**72.07**	**64.70**	**72.57**

and `EmbeddingGemma` models. As a controlled baseline, we reproduce `RepLLaMA` on the same backbones at 0.6B, 1B, 3B, and 8B, trained on MS MARCO only, which isolates the effect of conversational training. All table entries are aggregated from our local evaluation outputs covering all 22 datasets, and instruction-tuned baselines are also evaluated without task instructions. Appendix B gives references, pooling choices, and `RepLLaMA` training details.

Training data. We use WildChat-1M (Zhao et al., 2024) and the `train_sft` split of `HuggingFaceH4/ultrachat_200k` (Ding et al., 2023) as conversational text, without constructing query–memory relevance labels. Preprocessing selects English, non-toxic WildChat conversations and applies conversation-length, short-fragment, boilerplate, and jailbreak-pattern filters, together with repeated-prompt filtering and within-conversation near-duplicate removal. Each turn is split independently at sentence boundaries with a target budget of 120 words; a sentence longer than this budget is retained intact. The remaining fragments stay in conversation order and are grouped contiguously, so a group may cross conversation boundaries. The resulting training files contain 3,632,176 fragment instances: 227,011 groups of 16 for initialization and 454,022 groups of 8 for joint training, obtained by halving each initialization group. The conversational objective therefore reuses the same text in both stages, while the contrastive objective uses `Tevatron/msmarco-passage-aug`. The conversational files are constructed from these external corpora rather than the evaluation memory collections.

4.2 Main Results

Table 1 reports results at five backbone sizes, with per-dataset scores in Appendix D. MemRevela-8B is the strongest retriever in the table on both dataset-averaged metrics, reaching 63.24 nDCG@10. It is 3.43 points ahead of NV-Embed-v2 and further still ahead of KaLM-Embedding-Gemma3, which has half again as many parameters; the nDCG@10 advantage over NV-Embed-v2 holds across all four memory types. MemRevela-3B is second overall at 61.36, ahead of every baseline in the 7–12B group, including the 8B RepLLaMA reproduction, at well under half their size.

Table 2: LoCoMo QA with two fixed readers ($\times 100$): answer F1 for non-adversarial questions and abstention accuracy otherwise, averaged over 1,986 questions. Avg.: averages the three memory formats. Column bests are in **bold**.

	Llama-3.1-8B-Instruct				GPT-4o-mini			
Retriever	Dialog	Obs.	Summ.	Avg.	Dialog	Obs.	Summ.	Avg.
text-embedding-3-large	50.13	**48.18**	34.63	44.31	52.51	50.66	36.47	46.55
Voyage-3-large	49.17	47.20	36.79	44.38	52.17	49.61	37.84	46.54
E5-Mistral-7B-Instruct	49.82	47.21	36.27	44.44	51.78	49.40	36.85	46.01
MemRevela-8B	**51.15**	48.09	**37.91**	**45.72**	**53.41**	**50.95**	**38.60**	**47.65**

The pattern repeats lower down. MemRevela-0.6B is the best sub-billion retriever in the table at 58.56 nDCG@10, 1.74 points above bge-m3, and it also overtakes the 4B and 8B members of the Qwen3-Embedding family by 7.11 and 3.91 points. MemRevela-0.1B, the smallest entry, matches KaLM-Embedding-V1.5 with roughly a quarter of its parameters.

Against MS MARCO-only training on the same backbone and labels, MemRevela improves at every size from 0.6B to 8B (Figure 3a), across all four memory types except episodic memory at 0.6B.

4.3 Downstream Generation on LoCoMo

Setup. We evaluate whether retrieved memories support downstream question answering on LoCoMo (Maharana et al., 2024). We use the same 8B checkpoint as in the LMEB evaluation and compare it with three external embedding retrievers: E5-Mistral-7B-Instruct, OpenAI text-embedding-3-large, and Voyage-3-large. The evaluation contains 1,986 questions across all five question categories. We retrieve from three memory formats: dialogue turns, extracted observations, and session summaries. The reader receives up to 10 context items for dialogue and observation retrieval, or up to 2 session summaries. Memory content, context construction, and prompting are shared across retrievers within each format. We evaluate Llama-3.1-8B-Instruct and GPT-4o-mini as fixed readers. We follow the evaluation code released with LoCoMo,[1] which uses answer F1 for non-adversarial questions and binary abstention correctness for adversarial questions, with a choice-based prompt for the latter. We average question scores within each format and report the unweighted mean across the three formats.

Results. Table 2 reports absolute QA scores for MemRevela and three external retrievers across memory formats and readers. Averaged over the three formats, MemRevela scores 45.72 with Llama-3.1-8B-Instruct and 47.65 with GPT-4o-mini, ahead of all three baselines under both readers by between one and two points. The gain is positive in 17 of the 18 per-format comparisons, the exception being observation retrieval against text-embedding-3-large with the Llama reader. Averaged over the three baselines and two readers, the gain is largest on summary retrieval and smallest on observation retrieval. The positive average gains with both fixed readers indicate that the benefit is not confined to a specific reader and should be general.

5 Analysis

5.1 MemRevela Reshapes Conversational Memory Representations

Long-term conversations keep returning to the same people and events, so a retriever must pick out the turn that supports a query from many turns on the same topic. We test this by restricting retrieval to the query's own conversation. The LoCoMo probe has 1,529 queries and 2,337 query–gold-turn pairs, scored against the same candidates; inputs keep the speaker-tagged text but drop date, session and turn titles (details in Appendix A). Because cosine scales differ across models, we standardize each query's scores over the turns of its conversation, $z(q, m) = (s_\theta(q, m) - \mu_q)/\sigma_q$, so the gold z-score gives the evidence's distance above the conversation average in standard deviations. We then compare how widely each model spreads conversational turns relative to general passages.

[1] https://github.com/snap-research/locomo

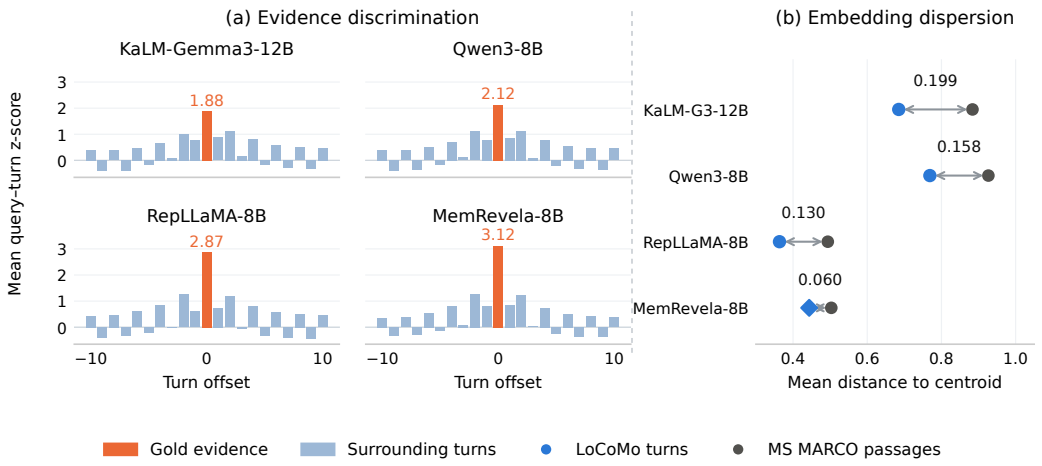


Figure 2: **MemRevela best separates evidence and spreads conversational turns more widely.** RepLLaMA-8B shares MemRevela's backbone and labels. (a) Mean query–turn z-score by offset from the gold turn (orange; zig-zag from speaker alternation, Appendix A). (b) Mean distance to the centroid of L2-normalized embeddings; arrows mark the MS MARCO-minus-LoCoMo gap.

MemRevela separates evidence from same-topic turns. Figure 2a aligns each gold turn at offset zero and averages the standardized scores of the turns around it, excluding other gold turns. MemRevela reaches the highest mean gold z-score (3.12, against 1.88–2.87 for the baselines), while the surrounding turns look alike across models. The gain carries over to ranking the full conversation, where MemRevela also leads on top-1 accuracy and MRR (Table 3).

MemRevela narrows the dispersion gap between memory and general text. Figure 2b examines the geometric mismatch between conversational memory and general passages discussed in §1. For each model, we measure embedding dispersion as the mean distance of L2-normalized embeddings to their corpus centroid, using all LoCoMo turns and sampled MS MARCO passages. Conversational turns occupy a tighter region than general passages for all four models. The MS MARCO-minus-LoCoMo gap is 0.13–0.20 for the baselines but only 0.06 for MemRevela. Conversational supervision thus brings the spread of memory representations closer to that of general text, alongside the stronger evidence discrimination in panel (a). RepLLaMA-8B provides a same-backbone comparison, and §5.3 further examines the contribution of conversational supervision.

5.2 MemRevela Preserves General Retrieval

Training on conversations could in principle specialize the retriever away from general retrieval. We test this on 13 BEIR datasets (Thakur et al., 2021), comparing MemRevela at each size with the RepLLaMA reproduction of the same backbone. Figure 3 shows mean LMEB and BEIR scores side by side; Appendix E lists per-dataset scores. Across the four model sizes, MemRevela maintains BEIR performance comparable to the contrastive-only baselines while improving LMEB at every size, so conversational supervision benefits memory retrieval without sacrificing general retrieval. Since our contrastive stage uses only MS MARCO, scaling it to larger and more diverse corpora may further improve both, which we leave to future work.

5.3 Ablations

To examine how each stage and training signal contributes, Figure 4 compares five configurations at 1B and 3B: conversational stage 1 alone, MS MARCO contrastive training alone, the joint stage-2 objective applied directly to the base model, stage 1 followed by contrastive training, and the full two-stage recipe (Section 3.3).

Conversational and contrastive supervision are complementary. All three combinations outperform either source alone on Dataset Mean at both sizes (Figure 4). Stage 1 alone trails con-

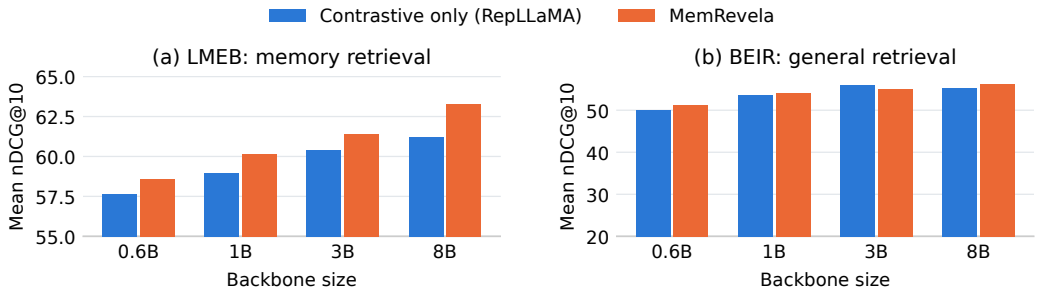


Figure 3: **MemRevela improves memory retrieval at every size while preserving general retrieval.** Mean nDCG@10 across (a) 22 LMEB and (b) 13 BEIR datasets. Tables 1 and 11 report the main scores; the 0.6B/1B contrastive-only LMEB scores are 57.61/58.94.

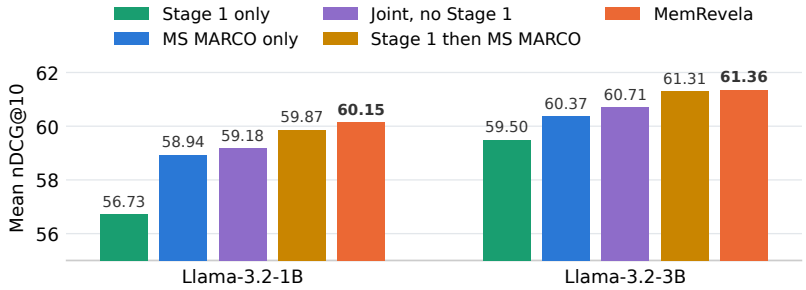


Figure 4: **Conversational and contrastive supervision are complementary, and conversational initialization contributes most of the gain.** LMEB mean nDCG@10 for five configurations at 1B and 3B; per-type results in Appendix F.

trastive training overall but performs better on episodic memory, whereas contrastive training alone is stronger on dialogue, semantic, and procedural memory at both sizes (Appendix F).

Conversational initialization is important. Starting from the stage-1 retriever matters more than the objective used afterwards. Initializing from stage 1 and then training contrastively improves over contrastive training from the base model by 0.93 and 0.94 nDCG@10 points at 1B and 3B, whereas the joint objective without stage 1 improves by only 0.24 and 0.34 points. Retaining conversational prediction in stage 2 on top of the initialization adds a further 0.28 and 0.05 points; semantic memory improves at both sizes, while other type-level changes are mixed.

Training efficiency is a target for future optimization. At 1B, on the same eight-A100 node with 3,836 steps and the same effective batch size, joint stage-2 training takes 33 hours and contrastive-only stage 2 takes 13 hours. Combining self-supervised and contrastive learning more efficiently, for example by interleaving the objectives or applying conversational supervision during only part of training, is a natural goal for future work.

6 Conclusion

Language agents depend on retrieving the right memories, yet general-purpose embedding models treat conversational memory as out of distribution, placing its turns in a much tighter region of embedding space than ordinary passages. MemRevela learns retrieval from the dependencies in unlabeled conversations and adds only standard contrastive learning on MS MARCO. MemRevela-8B outperforms NV-Embed-v2 on LMEB and improves the quality of downstream question answering; it also spreads conversational turns in embedding space about as widely as general passages while preserving general retrieval on BEIR. Future work includes combining the two signals more efficiently and scaling contrastive learning beyond MS MARCO.

AI USE STATEMENT

We used AI tools to assist with drafting and polishing the manuscript, including improving clarity and organization; retrieving and discovering relevant literature; and research ideation and execution, including discussing experimental designs, developing and debugging analysis and visualization code, and exploring interpretations of the results. The reported experimental results were obtained from model training and evaluation runs. The authors are responsible for the final methodological choices, interpretation of the findings, accuracy of the references, and all content of the paper.

REFERENCES

Fengyu Cai, Tong Chen, Xinran Zhao, Sihao Chen, Hongming Zhang, Sherry Tongshuang Wu, Iryna Gurevych, and Heinz Koeppl. Revela: Dense retriever learning via language modeling. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Fengyu Cai, Iryna Gurevych, and Heinz Koeppl. Elicitr: Unlocking latent reasoning in dense retrievers via generative regularization. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=xWHzF6WRGE.

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.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. Kalm-embedding: Superior training data brings a stronger embedding model, 2025. URL https://arxiv.org/abs/2501.01028.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.

Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan, and Graham Neubig. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2336–2349, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.149. URL https://aclanthology.org/2022.emnlp-main.149/.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Christian Lysenstøen. Training-Free Lexical-Dense Fusion for Conversational-Memory Retrieval, 2026. URL https://arxiv.org/abs/2606.04194.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval, 2023. URL https://arxiv.org/abs/2310.08319.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Chang Nie, Chaoyou Fu, Junlan Feng, and Caifeng Shan. EvoEmbedding: Evolvable Representations for Long-Context Retrieval and Agentic Memory, 2026. URL https://arxiv.org/abs/2606.21649.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL https://aclanthology.org/2024.naacl-long.463/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2345–2360, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.168. URL https://aclanthology.org/2022.naacl-main.168/.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022b.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, Bangkok, Thailand, August

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Yuqian Wu, Wei Chen, Zhengjun Huang, Junle Chen, Qingxiang Liu, Kai Wang, Xiaofang Zhou, and Yuxuan Liang. Back to Basics: Let Conversational Agents Remember with Just Retrieval and Generation, 2026. URL https://arxiv.org/abs/2604.11628.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings, 2024. URL https://arxiv.org/abs/2309.07597.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic Memory for LLM Agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model, 2026a. URL https://arxiv.org/abs/2506.20923.

Xinping Zhao, Xinshuo Hu, Jiaxin Xu, Danyu Tang, Xin Zhang, Mengjia Zhou, Yan Zhong, Yao Zhou, Zifei Shan, Meishan Zhang, Baotian Hu, and Min Zhang. LMEB: Long-horizon Memory Embedding Benchmark, 2026b. URL https://arxiv.org/abs/2603.12572.

A PROBE DETAILS

Table 3 reports the full probe results. Inputs keep the speaker-tagged turn text but drop date, session, and turn titles. Gold z-score and margin average over 2,337 query–gold pairs; top-1 and MRR average over 1,529 unique queries. The latter avoids overweighting queries with multiple annotated evidence turns. Top-1 and reciprocal rank use the highest-ranked gold turn for each query. The margin is the standardized score of a gold turn minus that of the strongest non-gold candidate. MemRevela and RepLLaMA use EOS pooling and a 512-token limit, while Qwen3 and KaLM use last-token pooling and a 1,024-token limit, without task instructions.

The candidate corpus contains 5,882 turns across 10 conversations. Queries come from the single-hop, multi-hop, temporal, and open-domain subsets. The geometry comparison uses all LoCoMo turns and 1,500 sampled MS MARCO positive passages. We L2-normalize each model's embeddings and take the embedding dispersion of a corpus to be the mean Euclidean distance of its items to the corpus centroid. Formally, for normalized embeddings $v_1, \ldots, v_n$, we use $D = n^{-1} \sum_i \|v_i - \bar{v}\|_2$, where $\bar{v} = n^{-1} \sum_i v_i$. Smaller values indicate greater concentration around the centroid. The LoCoMo dispersion is taken over the full corpus rather than within individual conversations. We report the MS MARCO dispersion minus the LoCoMo dispersion within each model.

The alternating profile in Figure 2a is associated with LoCoMo's two-speaker structure. Turns two positions from the gold turn usually share its speaker, whereas immediate neighbors usually belong to the other speaker. Queries often name a participant, and turn text includes speaker names, so speaker matching can contribute to the scores alongside contextual relevance. The pattern occurs across the evaluated models; nearby turns are therefore not interchangeable controls for evidence relevance.

Table 3: Within-conversation probe results on LoCoMo. Top-1/MRR are query-averaged ($\times 100$); gold z-score/margin are pair-averaged. Δ_D is the MS MARCO-minus-LoCoMo dispersion gap.

Model	Gold z-score	Margin	Top-1	MRR	Δ_D
Qwen3-Embedding-8B	2.12	-0.91	33.29	45.53	0.158
KaLM-Embedding-Gemma3-12B	1.88	-1.16	26.23	37.58	0.199
RepLLaMA-8B	2.87	-0.51	52.32	62.78	0.130
MemRevela-8B	3.12	-0.38	53.89	65.83	0.060

B BASELINE DETAILS

Table 1 compares with the Qwen3-Embedding family (Zhang et al., 2025), NV-Embed-v2 (Lee et al., 2025), E5-Mistral-7B-Instruct (Wang et al., 2024; 2022b), BGE models (Xiao et al., 2024), KaLM models (Hu et al., 2025; Zhao et al., 2026a), EmbeddingGemma-300M (Vera et al., 2025), snowflake-arctic-embed-l-v2.0,[2] harrier-oss-v1-0.6b,[3] and F2LLM-v2-4B.[4] We use CLS pooling for snowflake-arctic-embed-l-v2.0 and last-token pooling for harrier-oss-v1-0.6b.

Encoding settings. Table 4 records the pooling and length limits used in our local LMEB runs. These are the settings of the reported instruction-free comparison, rather than an evaluation of each model under every recommended prompting configuration. Embeddings are L2-normalized for cosine scoring. MemRevela and RepLLaMA retain their fixed `query:` and `passage:` prefixes; task-specific instructions are omitted. The shorter limits used in the within-conversation probe are specified separately in Appendix A; BEIR uses a 2,048-token limit for all paired models.

RepLLaMA reproductions. We reproduce RepLLaMA (Ma et al., 2023) at 0.6B, 1B, 3B, and 8B from Qwen3-0.6B-Base, Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B, each trained on

[2] https://huggingface.co/Snowflake/snowflake-arctic-embed-l-v2.0

[3] https://huggingface.co/microsoft/harrier-oss-v1-0.6b

[4] https://huggingface.co/codefuse-ai/F2LLM-v2-4B

Table 4: LMEB encoding settings. Length limits apply to queries and passages. All runs omit task instructions.

Retriever	Pooling	Maximum length
MemRevela/ RepLLaMA	EOS	2,048
Qwen3-Embedding (0.6B, 4B, 8B)	Last token	1,024
NV-Embed-v2	Mean	1,024
E5-Mistral-7B-Instruct	Last token	1,024
KaLM-Embedding-Gemma3	Last token	1,024
KaLM_Embedding_V1.5 / V2.5	Mean	1,024
EmbeddingGemma-300M	Mean	1,024
bge-large-en-v1.5	CLS	512
bge-m3	CLS	1,024
bge-multilingual-gemma2	Last token	1,024
snowflake-arctic-embed-l-v2.0	CLS	1,024
harrier-oss-v1-0.6b	Last token	1,024
F2LLM-v2-4B	Last token	1,024

MS MARCO only. The reproductions use LoRA rank 256 at 0.6B, 1B, and 3B; for RepLLaMA-8B, we use rank 64 to accommodate GPU memory constraints during training. Prior work reports comparable retrieval performance with LoRA and full fine-tuning (Ma et al., 2023).

C Training Configuration

Table 5 summarizes the training recipe for the five main checkpoints. Both stages use eight devices and four gradient-accumulation steps, with one conversational group per device in stage 1 and four labeled queries per device in stage 2. Each stage-2 query is accompanied by a group from the separate conversational file, cycling through that file when necessary; this pairing encodes no relevance relationship. Thus, effective batches contain 32 conversational groups in stage 1 and 128 queries with 128 conversational groups in stage 2. Conversational group size and the number of contrastive passages per query are distinct settings.

Table 5: Training settings for the main checkpoints. Length limits are in tokens; stage-2 exceptions for 8B are shown explicitly.

Setting	Stage 1	Stage 2
Objective	Conversational prediction	Joint ($c = 0.95$)
Learning rate (0.1B, 0.6B, 1B, 3B)	10^{-4}	5×10^{-5}
Learning rate (8B)	10^{-4}	2×10^{-5}
Conversational temperature τ_{mem}	10^{-4}	10^{-3}
Contrastive temperature τ_{ret}	—	0.01
Fragments per conversational group	16	8
Conversational fragment length limit	156	144
Contrastive query length limit	—	32
Contrastive passage length limit	—	196 (8B: 160)
Passages per labeled query	—	16
Effective batch	32 groups	128 queries + 128 groups
Training epochs	1	1 (MS MARCO)
Optimizer updates	7,094	3,836
Warmup updates	100	100

Both stages use bf16 precision and train LoRA adapters on the attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and MLP projections (`gate_proj`, `up_proj`, `down_proj`). LoRA rank is 256, scaling parameter is 64, and dropout is 0.1. The auxiliary LM is Llama-3.2-1B for every retriever size. Retriever representations use EOS pooling and L2 normalization. Cached attention values are not additionally normalized. The supervised data are `Tevatron/msmarco-passage-aug`, revision `8ec0b7ad`; stage 2 samples one positive and

15 negatives per query and uses other passages in the distributed batch as additional negatives. Training query and passage prefixes are `query:` and `passage:`.

D LMEB RESULTS BY DATASET

Tables 6–9 list the per-dataset nDCG@10 and Recall@10 behind Table 1, following the layout of the LMEB paper. All scores are from our local evaluation runs without task instructions; where a dataset has several subsets, the subset scores are averaged first.

Table 6: LMEB results (datasets 1–11) without instructions, N@10. Models/groups follow Table 1; bold/underline mark column bests/second-bests within each group.

Retriever	EPBench	KnowMeBench	LoCoMo	LongMemEval	REALTALK	TMD	MemBench	ConvoMem	QASPER	NovelQA	PeerQA
<1B											
MemRevela-0.1B	70.95	26.36	41.64	68.33	38.57	16.30	66.07	65.96	51.07	21.91	29.83
EmbeddingGemma-300M	**89.23**	**47.14**	42.15	77.03	41.64	**33.94**	70.54	58.43	45.16	26.27	29.83
bge-large-en-v1.5	70.90	40.96	43.77	74.26	**47.22**	24.00	71.10	58.39	45.41	24.81	27.57
KaLM_Embedding_V1.5	75.41	37.79	32.52	42.36	33.56	21.58	66.43	51.40	36.67	19.39	28.76
KaLM_Embedding_V2.5	82.87	40.85	40.73	68.63	38.29	18.07	69.63	58.25	39.75	25.14	25.63
bge-m3	87.47	46.50	46.55	**78.41**	45.00	28.09	71.23	64.41	51.53	32.20	29.73
Qwen3-Embedding-0.6B	75.22	40.05	38.50	67.61	38.91	25.97	70.46	61.00	40.64	24.91	26.83
snowflake-arctic-embed-l-v2.0	83.85	43.64	44.11	68.78	39.37	20.17	63.96	65.66	43.89	27.08	27.84
harrier-oss-v1-0.6b	80.24	43.71	36.30	44.04	34.46	27.56	68.60	58.07	43.61	22.24	26.63
MemRevela-0.6B	81.88	41.18	**51.87**	73.43	47.14	22.93	**74.65**	**71.38**	**56.22**	**33.97**	**33.20**
1–7B											
Qwen3-Embedding-4B	74.00	42.64	35.54	34.16	36.33	21.05	66.37	53.73	44.98	26.66	28.24
F2LLM-v2-4B	**91.77**	44.25	41.86	78.89	39.19	**28.84**	68.83	65.35	**56.53**	28.40	**34.22**
MemRevela-1B	84.97	46.61	54.57	77.61	47.88	27.89	74.57	**70.65**	56.03	35.06	33.06
RepLLaMA-3B	87.35	46.56	54.98	**79.23**	48.78	27.39	71.44	69.93	54.64	35.97	32.94
MemRevela-3B	88.61	**50.14**	**55.69**	78.97	**49.64**	23.17	**75.44**	69.21	55.80	**39.22**	33.50
7–12B											
e5-mistral-7b-instruct	68.55	45.89	39.86	69.57	35.61	18.99	67.11	59.91	40.63	25.72	24.78
NV-Embed-v2	84.54	**56.32**	49.69	80.30	46.18	20.81	72.14	69.74	51.07	35.31	30.98
Qwen3-Embedding-8B	76.57	45.49	41.27	70.38	41.63	20.14	69.30	51.34	43.15	28.69	31.13
bge-multilingual-gemma2	68.94	40.62	40.84	46.64	32.58	18.69	57.87	47.57	32.31	22.55	12.17
KaLM-Embedding-Gemma3	86.34	47.64	37.59	75.67	38.03	21.12	72.74	60.05	32.77	23.65	24.17
RepLLaMA-8B	88.73	47.40	55.60	**80.65**	49.86	**32.47**	72.11	68.81	55.06	37.66	33.95
MemRevela-8B	**90.33**	54.19	**58.72**	78.91	**52.01**	29.21	**74.94**	**73.10**	**56.27**	**42.42**	**34.47**

Table 7: LMEB results (datasets 12–22) without instructions, N@10. Models/groups follow Table 1; bold/underline mark column bests/second-bests within each group.

Retriever	Covid-QA	ESG-Reports	MLDR	LooGLE	SciFact	Gorilla	ToolBench	ReMe	Proced_mem_bench	MemGovern	DeepPlanning
<1B											
MemRevela-0.1B	73.01	32.64	76.17	45.54	75.59	24.38	36.10	56.18	43.95	76.87	40.33
EmbeddingGemma-300M	70.92	43.42	76.75	58.93	78.06	37.01	56.39	**63.23**	51.66	86.46	49.11
bge-large-en-v1.5	72.68	41.38	80.57	54.09	73.90	33.48	40.13	62.82	49.83	84.01	45.29
KaLM_Embedding_V1.5	58.66	39.34	73.09	50.25	73.59	36.97	51.95	58.65	45.59	84.08	51.82
KaLM_Embedding_V2.5	69.96	43.12	78.83	56.03	77.71	34.24	53.98	61.14	47.75	72.14	48.57
bge-m3	79.42	40.74	**81.33**	58.95	76.36	34.32	46.29	61.40	51.39	85.98	52.74
Qwen3-Embedding-0.6B	69.79	41.40	78.15	56.13	78.39	**41.59**	54.48	60.51	51.04	82.27	52.85
snowflake-arctic-embed-l-v2.0	71.98	42.75	73.09	54.73	77.50	35.17	46.12	62.70	49.53	86.71	49.18
harrier-oss-v1-0.6b	63.15	42.49	80.72	57.19	57.15	41.17	**59.32**	62.61	**52.78**	85.41	52.42
MemRevela-0.6B	**82.80**	**47.43**	78.91	**60.01**	**82.49**	37.95	53.03	62.33	51.79	**90.53**	**53.20**
1–7B											
Qwen3-Embedding-4B	74.09	44.93	79.75	59.94	71.52	**44.91**	51.40	63.10	50.33	84.40	43.92
F2LLM-v2-4B	72.39	46.67	79.88	56.69	77.53	44.27	**63.51**	**67.30**	52.29	86.24	47.92
MemRevela-1B	83.80	48.09	80.47	63.00	82.50	39.16	55.41	61.99	**55.47**	90.89	53.52
RepLLaMA-3B	83.46	**51.80**	79.26	63.84	**83.53**	40.64	55.65	63.17	54.10	90.17	53.35
MemRevela-3B	**85.72**	50.90	**82.37**	**65.94**	82.98	40.38	59.36	63.76	53.39	**91.51**	**54.13**
7–12B											
e5-mistral-7b-instruct	50.81	39.37	66.47	40.12	80.07	37.38	55.83	58.35	49.30	80.83	36.25
NV-Embed-v2	76.91	49.52	82.06	64.18	83.10	43.99	60.29	**65.21**	50.45	91.19	51.82
Qwen3-Embedding-8B	75.42	52.11	79.51	59.45	74.82	**46.81**	48.52	63.63	49.16	86.14	47.56
bge-multilingual-gemma2	39.70	**54.96**	61.23	40.73	61.94	31.84	56.14	50.27	52.40	69.02	53.47
KaLM-Embedding-Gemma3	62.10	44.21	72.34	49.42	73.95	46.40	**63.30**	63.22	51.44	87.25	52.73
RepLLaMA-8B	84.91	52.15	80.60	64.93	81.53	41.57	56.92	63.71	52.08	89.69	55.99
MemRevela-8B	**85.60**	54.07	**84.96**	**67.73**	**84.00**	42.74	58.85	63.85	**54.93**	**92.33**	**57.61**

Comparison with ElicitR. Table 10 compares MemRevela with ElicitR (Cai et al., 2026b), which applies retriever-weighted prediction within MS MARCO query–document groups, at the same 1B

Table 8: LMEB results (datasets 1–11) without instructions, R@10. Models/groups follow Table 1; bold/underline mark column bests/second-bests within each group.

Retriever	EPBench	KnowMeBench	LoCoMo	LongMemEval	REALTALK	TMD	MemBench	ConvoMem	QASPER	NovelQA	PeerQA
<1B											
MemRevela-0.1B	82.34	32.62	54.43	82.30	47.62	7.17	77.25	80.98	74.63	30.95	38.43
EmbeddingGemma-300M	**93.99**	**55.72**	57.67	89.03	51.58	**14.70**	80.56	75.57	68.84	36.50	39.26
bge-large-en-v1.5	81.94	48.39	57.99	86.54	**58.22**	10.37	80.51	74.82	67.61	35.06	36.02
KaLM_Embedding_V1.5	86.02	45.92	46.16	51.00	42.46	8.27	78.00	69.82	57.60	27.55	41.15
KaLM_Embedding_V2.5	89.94	48.53	55.37	82.54	47.53	8.26	79.97	74.74	61.35	34.91	37.46
bge-m3	92.65	53.68	60.71	**90.06**	53.93	12.47	82.86	79.53	74.22	42.61	40.61
Qwen3-Embedding-0.6B	83.24	47.57	52.84	78.62	48.67	10.79	81.47	78.10	64.33	34.48	36.55
snowflake-arctic-embed-l-v2.0	89.86	50.80	58.74	80.73	49.30	9.26	77.26	81.35	67.21	37.13	36.28
harrier-oss-v1-0.6b	86.45	51.56	49.98	54.28	43.88	12.12	81.31	75.29	66.43	32.23	36.01
MemRevela-0.6B	88.18	47.37	**66.31**	86.92	57.18	10.52	**83.56**	**85.39**	**79.22**	**44.44**	**45.09**
1–7B											
Qwen3-Embedding-4B	83.39	50.84	48.42	42.64	43.67	8.19	79.15	71.15	68.72	36.93	38.99
F2LLM-v2-4B	**95.24**	52.32	56.86	**91.52**	49.45	12.37	80.03	83.04	76.68	39.74	**43.68**
MemRevela-1B	90.80	53.36	68.47	89.72	57.56	12.53	83.28	**84.70**	78.64	45.46	42.53
RepLLaMA-3B	91.80	52.83	67.06	89.41	57.76	**12.55**	82.22	83.79	77.89	46.67	43.00
MemRevela-3B	92.90	**56.48**	**70.17**	89.66	**59.28**	11.29	**83.63**	84.03	**78.66**	**49.93**	42.35
7–12B											
e5-mistral-7b-instruct	78.82	53.92	54.38	82.66	45.61	8.77	78.19	78.25	63.58	36.34	35.17
NV-Embed-v2	90.20	**63.79**	66.28	90.67	57.26	10.52	83.21	84.66	75.64	46.35	42.62
Qwen3-Embedding-8B	84.31	52.48	54.80	81.42	50.75	7.98	82.02	68.56	65.78	40.02	42.67
bge-multilingual-gemma2	76.66	47.72	54.51	55.51	40.47	9.41	73.44	65.60	56.43	30.65	18.33
KaLM-Embedding-Gemma3	92.17	56.08	53.25	87.66	47.50	9.33	82.36	78.26	54.30	32.82	33.23
RepLLaMA-8B	93.11	54.61	68.09	**91.00**	58.38	**14.23**	82.16	82.63	78.37	48.33	42.64
MemRevela-8B	**94.27**	60.33	**72.35**	89.97	**61.05**	13.87	**83.30**	**86.42**	**79.78**	**53.21**	**43.72**

Table 9: LMEB results (datasets 12–22) without instructions, R@10. Models/groups follow Table 1; bold/underline mark column bests/second-bests within each group.

Retriever	Covid-QA	ESG-Reports	MLDR	LooGLE	SciFact	Gorilla	ToolBench	ReMe	Proced_mem_bench	MemGovern	DeepPlanning
<1B											
MemRevela-0.1B	90.46	51.07	95.00	61.10	92.08	41.30	41.86	77.19	31.39	88.82	39.69
EmbeddingGemma-300M	90.91	52.78	**99.00**	72.45	94.83	57.33	64.10	**85.22**	**38.16**	95.44	52.18
bge-large-en-v1.5	90.19	56.57	97.00	68.27	92.64	51.36	48.74	84.98	36.60	93.83	48.55
KaLM_Embedding_V1.5	84.43	50.42	94.00	66.17	90.92	59.50	59.82	81.85	31.70	94.16	52.60
KaLM_Embedding_V2.5	90.19	60.05	97.00	69.75	94.44	55.17	61.61	83.71	36.17	86.15	50.28
bge-m3	93.79	57.96	97.00	71.77	92.05	54.37	52.95	82.60	37.76	94.75	52.96
Qwen3-Embedding-0.6B	90.37	56.53	96.00	69.58	94.41	62.17	62.96	83.13	36.70	93.28	53.34
snowflake-arctic-embed-l-v2.0	89.74	54.54	98.00	68.65	94.05	55.75	53.37	83.79	35.85	95.27	50.72
harrier-oss-v1-0.6b	86.14	56.11	98.00	70.94	86.17	**63.13**	**66.53**	84.28	37.11	94.74	**53.38**
MemRevela-0.6B	**95.50**	**68.01**	96.00	**72.58**	**95.85**	58.95	59.71	82.99	37.25	**97.11**	53.23
1–7B											
Qwen3-Embedding-4B	92.17	62.45	**98.00**	72.61	92.98	**69.05**	61.02	84.58	38.23	94.34	47.73
F2LLM-v2-4B	90.73	61.43	**98.00**	70.98	94.96	63.05	**71.48**	**87.69**	39.14	94.99	49.39
MemRevela-1B	96.13	64.12	**98.00**	74.69	97.15	61.12	62.15	83.11	**41.38**	97.25	**54.87**
RepLLaMA-3B	95.41	**67.36**	**98.00**	75.58	**97.45**	58.36	63.39	83.97	40.41	97.05	53.66
MemRevela-3B	**96.22**	65.51	**98.00**	**77.09**	97.36	59.03	66.97	84.37	38.46	**97.58**	54.56
7–12B											
e5-mistral-7b-instruct	76.60	49.91	95.00	57.07	95.62	61.56	64.46	80.67	37.07	92.38	42.08
NV-Embed-v2	92.62	60.51	**99.00**	76.28	96.49	65.01	68.36	**87.04**	38.02	97.51	55.89
Qwen3-Embedding-8B	93.25	68.47	**99.00**	72.38	93.98	**70.97**	58.97	85.86	37.23	95.15	51.23
bge-multilingual-gemma2	68.41	**69.54**	96.00	59.54	80.56	51.83	64.33	72.90	38.85	80.95	54.27
KaLM-Embedding-Gemma3	89.29	57.41	**99.00**	66.71	94.29	69.65	**71.79**	85.53	37.40	95.87	54.41
RepLLaMA-8B	96.40	67.13	98.00	76.49	96.78	62.26	65.04	84.08	37.24	96.94	57.36
MemRevela-8B	**96.49**	65.05	**99.00**	**78.16**	**96.90**	63.97	66.48	85.11	**39.70**	**97.86**	**58.46**

Table 10: LMEB comparison with ElicitR at matched backbones, without task instructions (nDCG@10 $\times 100$). Column bests within each model size are bold.

Retriever	Episodic	Dialogue	Semantic	Procedural	Dataset Mean	Type Mean
Llama-3.2-1B						
RepLLaMA	63.98	57.62	58.66	58.95	58.94	59.80
ElicitR	64.78	**59.26**	59.05	56.57	58.95	59.92
MemRevela	**65.79**	58.86	**60.25**	**59.41**	**60.15**	**61.08**
Llama-3.2-3B						
RepLLaMA	66.96	58.62	60.68	59.51	60.37	61.44
ElicitR	66.72	**59.21**	60.65	58.72	60.29	61.33
MemRevela	**69.37**	58.69	**62.05**	**60.42**	**61.36**	**62.63**

and 3B backbones, with RepLLaMA as the contrastive-only reference. ElicitR matches RepLLaMA on Dataset Mean (58.95 vs. 58.94 at 1B; 60.29 vs. 60.37 at 3B) and is the strongest of the three on dialogue memory, while MemRevela leads on the other three memory types and on both summary columns at both sizes.

E BEIR Results by Dataset

Table 11 lists the per-dataset nDCG@10 behind the BEIR means in Figure 3. All models are evaluated with the same pipeline (maximum length 2048, bf16, test split).

Table 11: BEIR test-set nDCG@10 across 13 datasets; Δ is MemRevela minus RepLLaMA. Pairs share backbones and MS MARCO supervision; 8B LoRA ranks are 256/64, respectively.

	0.6B			1B			3B			8B		
Dataset	RepLLaMA	MemRevela	Δ	RepLLaMA	MemRevela	Δ	RepLLaMA	MemRevela	Δ	RepLLaMA	MemRevela	Δ
ArguAna	53.65	48.89	−4.76	50.46	53.03	+2.57	52.41	51.20	−1.21	50.97	53.96	+2.99
ClimateFEVER	29.58	25.61	−3.97	28.82	26.22	−2.60	31.82	24.62	−7.20	28.46	25.56	−2.90
DBPedia	39.69	38.67	−1.02	42.42	42.82	+0.40	44.94	44.44	−0.50	45.72	44.99	−0.73
FEVER	73.88	77.66	+3.79	79.75	81.45	+1.70	82.81	83.84	+1.03	81.83	83.34	+1.51
FiQA-2018	38.26	38.48	+0.22	40.32	41.02	+0.70	43.09	44.69	+1.60	44.98	47.26	+2.28
HotpotQA	60.29	61.15	+0.86	66.53	67.63	+1.10	69.99	70.00	+0.01	70.99	71.64	+0.65
NFCorpus	34.59	34.53	−0.07	35.73	36.20	+0.47	37.18	37.38	+0.20	37.99	38.66	+0.67
NQ	53.62	53.88	+0.26	57.82	58.51	+0.69	62.05	61.91	−0.14	63.57	63.18	−0.39
Quora	80.29	86.10	+5.81	84.39	86.90	+2.51	86.53	86.39	−0.14	84.30	87.97	+3.67
SCIDOCS	16.78	16.36	−0.42	17.79	17.83	+0.04	18.66	18.06	−0.60	17.98	18.44	+0.46
SciFact	70.59	69.53	−1.06	74.47	73.37	−1.10	74.14	74.97	+0.83	75.17	74.82	−0.35
Touché-2020	31.45	33.58	+2.13	35.42	34.39	−1.03	39.13	33.35	−5.78	35.30	37.17	+1.87
TREC-COVID	68.96	81.23	+12.27	83.23	83.91	+0.68	83.85	83.40	−0.45	82.71	84.72	+2.01
Mean	50.13	**51.21**	+1.08	53.63	**54.10**	+0.47	**55.89**	54.94	−0.95	55.38	**56.29**	+0.90

F Training Ablation by Memory Type

Table 12 lists the per-memory-type scores behind Figure 4, together with the two summary columns. All ten runs are evaluated with the same pipeline as the rest of the paper, without task instructions.

Table 12: Training-stage ablations on LMEB without task instructions (nDCG@10 $\times 100$). Dataset/Type Mean average 22 datasets/four memory types. Column bests within each model size are bold.

Training	Episodic	Dialogue	Semantic	Procedural	Dataset Mean	Type Mean
Llama-3.2-1B						
Stage 1 only	65.87	55.33	56.66	55.19	56.73	58.26
MS MARCO only	63.98	57.62	58.66	58.95	58.94	59.80
Joint, no Stage 1	63.57	57.65	59.04	59.44	59.18	59.93
Stage 1 $\rightarrow$ MS MARCO	**66.16**	58.39	59.60	**59.60**	59.87	60.94
MemRevela	65.79	**58.86**	**60.25**	59.41	**60.15**	**61.08**
Llama-3.2-3B						
Stage 1 only	68.14	56.71	59.86	58.92	59.50	60.91
MS MARCO only	66.96	58.62	60.68	59.51	60.37	61.44
Joint, no Stage 1	68.58	57.79	61.31	60.20	60.71	61.97
Stage 1 $\rightarrow$ MS MARCO	**69.79**	**59.48**	61.60	59.91	61.31	**62.70**
MemRevela	69.37	58.69	**62.05**	**60.42**	**61.36**	62.63