NEAR: NEIGHBORHOOD-AWARE RESIDUALS FOR MEMORY RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory retrieval aims to identify evidence from stored records to answer a query. Stored records with different factual details may share topics, entities, or events. Such shared contents would make records without the required evidence more likely to receive high similarity scores and outrank the correct evidence. In this paper, we use the information shared by these records to construct local references for query–memory matching. We propose **NE**ighborhood-**A**ware **R**esiduals (NEAR), a training-free framework that uses these references to construct residual representations that complement the original embeddings. Specifically, each memory derives a reference direction from its neighbors, and each query obtains its reference from a precomputed table. Using these references, NEAR constructs projection residuals for both queries and memories. These residuals are combined with the original embeddings for a single approximate nearest-neighbor search over a fixed memory index. Experiments on eight long-context retrieval benchmarks show consistent improvements across nine embedding models. Downstream evaluations show that NEAR improves question-answering performance.

1 INTRODUCTION

Memory retrieval enables intelligent assistants to find evidence in past conversations and stored documents to answer current queries (Maharana et al., 2024; Wu et al., 2025). A common approach is embedding-based retrieval, which encodes queries and memories separately and ranks memories by vector similarity (Morris & Rush, 2025). Memory embeddings can be computed and indexed in advance, allowing new queries to search the memory bank without re-encoding its records (Ashok et al., 2026). This design separates offline memory processing from online query retrieval.

In a memory bank, records with different factual details may share topics, entities, or events (Zhang et al., 2026). When ranking relies only on query–memory similarity, such overlap would make records that do not provide the required evidence more likely to outrank correct evidence (Ashok et al., 2026). However, these similar records are not only competing candidates. We argue that they also provide a local neighborhood that can serve as a reference for each memory. We turn this potential source of confusion into useful matching information by constructing memory-specific local references from the relationships among similar memories, as illustrated in Figure 1.

In this paper, we propose **NE**ighborhood-**A**ware **R**esiduals (NEAR), a training-free framework that uses these local references for query–memory matching. Figure 2 gives an overview of NEAR. For each memory, NEAR builds a reference from its nearest neighbors and removes the embedding component along this direction. A small precomputed table provides references for new queries without an initial search over the memory bank. The same projection yields a residual representation for each query. NEAR combines the original and residual representations so that the final matching score is a single inner product. Offline indexing enables retrieval through a single approximate nearest-neighbor (ANN) search over the fixed memory index.

We evaluate NEAR on eight long-context retrieval benchmarks using nine frozen embedding models. Across all nine models, NEAR improves mean Recall@10 and NDCG@10 over the original embeddings and the evaluated same-backbone baselines. Downstream evaluations further show that NEAR improves question-answering accuracy by 9.6 percentage points on LoCoMo with Qwen3-Embedding-4B as the retrieval model. Controlled comparisons show that local reference construction and memory–reference correspondence contribute to the retrieval gains. Additional analyses

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

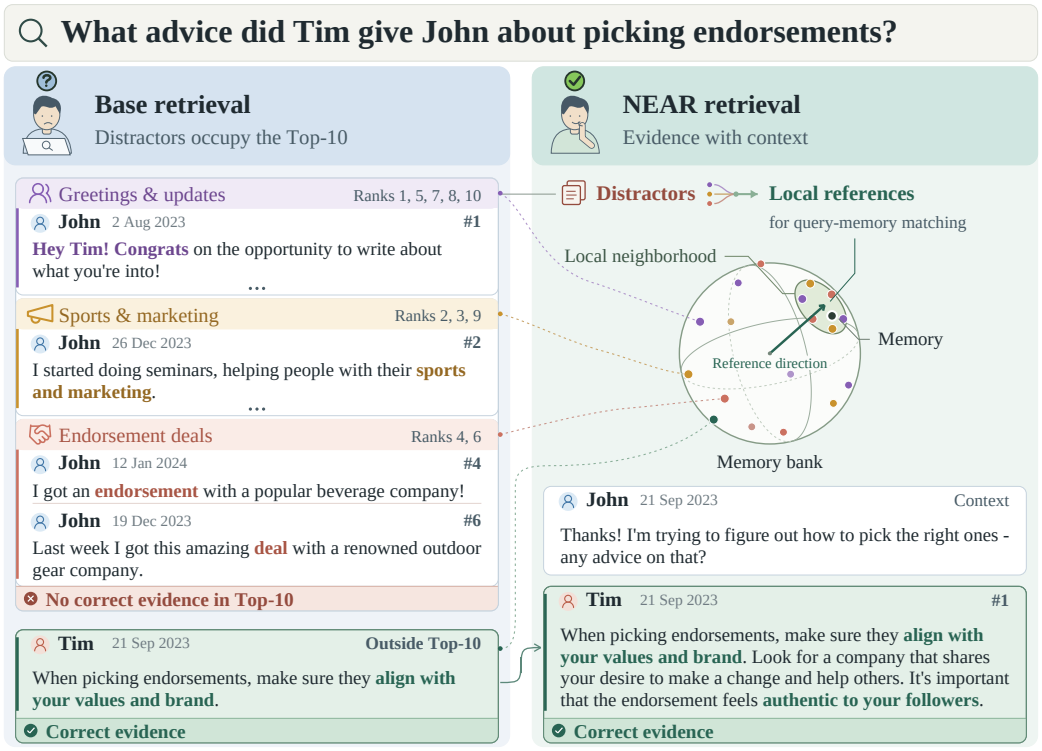


Figure 1: **Turning distractors into local references.** In this LoCoMo example (Qwen3-Embedding-4B), Base's Top-10 contains records about the same people and related topics but misses Tim's advice on choosing endorsements. However, these shared contents also provide useful information for distinguishing the candidates. NEAR uses this information to construct local references for query–memory matching and ranks the missed evidence first.

characterize parameter sensitivity and the trade-offs between retrieval quality, storage, and query latency, including the benefits of index compression.

Our main contributions are as follows:

- We propose NEAR, a training-free method that constructs memory-specific local references from shared contents and uses the resulting residual representations to complement the original embeddings for query–memory matching.
- We construct a compact query-reference table and a fixed memory index, enabling new queries to retrieve memories through a single ANN search without preliminary candidate retrieval.
- We demonstrate improvements in mean Recall@10 and NDCG@10 across eight retrieval benchmarks and nine embedding models, together with downstream question-answering gains.

2 Related Work

Long-term memory and retrieval. Long-term memory research explores how to organize, update, and retrieve past information to support question answering over ongoing interactions. A-Mem organizes structured memory notes into a dynamically linked network, while MemoryOS manages memories at different time scales through a hierarchical architecture (Xu et al., 2025; Kang et al., 2025). LightMem reduces memory-processing costs through information compression, topic-based consolidation, and offline updates (Fang et al., 2026). For retrieval, AssoMem uses an associative memory graph and combines relevance, importance, and temporal signals for ranking (Zhang et al., 2026). EvoEmbedding instead learns context-dependent retrieval representations using a continuously updated latent memory (Nie et al., 2026). We focus on scoring queries against a given memory bank, without changing its content or updating the pretrained encoder.

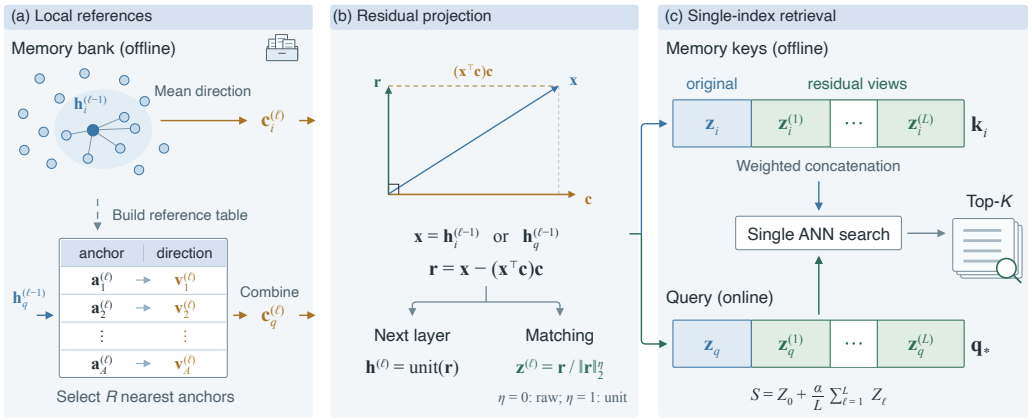


Figure 2: **Overview of NEAR.** Memory neighborhoods yield local references, while a table supplies query references. Original and residual views form static keys for one ANN search.

Embedding adaptation and similarity calibration. Contextual embedding methods incorporate surrounding text or other documents during encoding (Morris & Rush, 2025; Conti et al., 2025). Existing embeddings can be adapted for retrieval through supervised mappings or training-free corrections. Search-Adaptor learns transformations of pretrained embeddings from relevance supervision while keeping the encoder fixed (Yoon et al., 2024). Training-free methods use neighborhoods or reference collections, but differ in whether they adjust scores or vectors. LC corrects similarity scores using local affinity estimates to reduce hubness (Hara et al., 2015). For cross-modal retrieval, NNN estimates matching biases from a reference database, while DBSN uses query and target reference banks to calibrate score distributions (Chowdhury et al., 2024; Pan et al., 2025). Rather than correcting scores directly, InvGC transforms embedding vectors through inverse graph convolution with global or local reference sets (Jian & Wang, 2023).

Residual representations and reference-based matching. Residual-based methods compare or encode representations after removing a reference component. Local residual similarity constructs neighborhood-derived residuals for image re-ranking (Sun et al., 2017). PRQ-KMeans instead recursively removes selected-centroid projections to construct semantic IDs for generative retrieval and recommendation (Luo et al., 2026). For first-stage memory retrieval, memory-specific references can be computed offline, but a new query also needs a reference before any candidates have been retrieved. NEAR addresses this requirement by combining each memory's neighborhood reference with a precomputed query-reference mapping. It retains original similarity alongside residual matching and expresses their combined score as an inner product over a fixed memory index, enabling a single ANN search without preliminary candidate retrieval.

3 METHOD

Given a text query q and a bank of memory records $\mathcal{M} = \{m_i\}_{i=1}^{N}$, we use a frozen encoder to obtain unit-length embeddings $\mathbf{z}_q, \mathbf{z}_i \in \mathbb{R}^d$. Original retrieval ranks memories by their cosine similarity $\mathbf{z}_q^\top \mathbf{z}_i$. NEAR complements this score with residual representations constructed from memory neighborhoods. It builds memory representations and a query-reference table offline. A new query uses the table to form its residuals before a single ANN search.

3.1 NEIGHBORHOOD-BASED MEMORY REPRESENTATIONS

Initialize $\mathbf{h}_i^{(0)} = \mathbf{z}_i$. At layer ℓ, let $\mathcal{N}_i^{(\ell)}$ contain the $\min(k, N-1)$ nearest other memories to $\mathbf{h}_i^{(\ell-1)}$ by cosine similarity. Their mean direction defines the local reference

$$\mathbf{c}_i^{(\ell)} = \mathrm{unit}\left(\frac{1}{|\mathcal{N}_i^{(\ell)}|} \sum\nolimits_{j \in \mathcal{N}_i^{(\ell)}} \mathbf{h}_j^{(\ell-1)} \right),$$

where $\mathrm{unit}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$ for nonzero vectors. Removing the projection onto this direction gives the residual

$$\mathbf{r}_i^{(\ell)} = \mathbf{h}_i^{(\ell-1)} - \left((\mathbf{h}_i^{(\ell-1)})^\top \mathbf{c}_i^{(\ell)}\right) \mathbf{c}_i^{(\ell)}. \tag{1}$$

For subsequent layers, we repeat this procedure using the normalized residuals $\mathbf{h}_i^{(\ell)} = \mathrm{unit}(\mathbf{r}_i^{(\ell)})$. The matching representation either retains the residual length or is normalized:

$$\mathbf{z}_i^{(\ell)} = \begin{cases} \mathbf{r}_i^{(\ell)}, & \eta = 0, \\ \mathrm{unit}(\mathbf{r}_i^{(\ell)}), & \eta = 1, \end{cases} \tag{2}$$

where η is selected on development data and shared by memories and queries. We retain the original embeddings and the matching representations from all L layers for score fusion.

3.2 Precomputed Query References

At each layer, NEAR precomputes a small table that maps queries to local reference directions. Spherical K-means (Dhillon & Modha, 2001) clusters the input memory states, with cluster centers serving as anchors $\mathbf{a}_h^{(\ell)}$. Each anchor stores $\mathbf{v}_h^{(\ell)}$, the normalized mean reference direction of the memories in its cluster.

Initialize the query state $\mathbf{h}_q^{(0)} = \mathbf{z}_q$. At layer ℓ, the query selects the $R = \min(R_{\max}, A)$ most similar anchors from the table's A entries, with index set $\mathcal{H}_q^{(\ell)}$. Similarity-based softmax weights combine their stored directions into a query reference:

$$p_h^{(\ell)}(q) = \frac{\exp\{\beta(\mathbf{h}_q^{(\ell-1)})^\top \mathbf{a}_h^{(\ell)}\}}{\sum_{j\in\mathcal{H}_q^{(\ell)}} \exp\{\beta(\mathbf{h}_q^{(\ell-1)})^\top \mathbf{a}_j^{(\ell)}\}}, \qquad \mathbf{c}_q^{(\ell)} = \mathrm{unit}\left(\sum\nolimits_{h\in\mathcal{H}_q^{(\ell)}} p_h^{(\ell)}(q)\mathbf{v}_h^{(\ell)}\right),$$

where β controls weight concentration. The same projection and normalization rules yield the query matching representation $\mathbf{z}_q^{(\ell)}$ and next-layer state $\mathbf{h}_q^{(\ell)}$, and Appendix A lists the parameter settings.

3.3 Score Fusion and Single-Index Retrieval

Let $\mathbf{z}_x^{(0)} = \mathbf{z}_x$ denote the original embedding and $s_\ell(q, m_i) = (\mathbf{z}_q^{(\ell)})^\top \mathbf{z}_i^{(\ell)}$ the matching score at layer ℓ. NEAR combines original and residual matching as

$$S(q, m_i) = \frac{s_0(q, m_i)}{\tau_0} + \frac{\alpha}{L} \sum\nolimits_{\ell=1}^{L} \frac{s_\ell(q, m_i)}{\tau_\ell}, \tag{3}$$

where $\alpha \geq 0$ controls the total residual weight independently of depth. Each scale τ_ℓ is computed offline from the standard deviations of memory–memory inner products and shared across queries for the same bank (Appendix C). We write $Z_\ell = s_\ell/\tau_\ell$ for the scaled score.

The fused score can be evaluated as one inner product by concatenating representations into a static memory key and a query vector:

$$\mathbf{k}_i = \left[\mathbf{z}_i;\ \sqrt{\alpha/L}\,\mathbf{z}_i^{(1)};\ \ldots;\ \sqrt{\alpha/L}\,\mathbf{z}_i^{(L)}\right],$$
$$\mathbf{q}_* = \left[\mathbf{z}_q/\tau_0;\ \frac{\sqrt{\alpha/L}}{\tau_1}\mathbf{z}_q^{(1)};\ \ldots;\ \frac{\sqrt{\alpha/L}}{\tau_L}\mathbf{z}_q^{(L)}\right], \tag{4}$$

where semicolons denote concatenation. Their inner product equals the fused score exactly: $\mathbf{q}_*^\top \mathbf{k}_i = S(q, m_i)$. Memory keys can therefore be constructed and indexed offline. A new query forms $\mathbf{q}_*$ using the reference tables and performs a single inner-product ANN search to retrieve memories.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate retrieval on eight benchmarks: LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2025), and REALTALK (Lee et al., 2025) for conversational memory, and

Table 1: **Retrieval quality across three Qwen scales on eight LMEB tasks (%).** Cells report capped Recall@10 / NDCG@10. Overall reports the unweighted mean across the eight tasks.

Method	LoCoMo	LongMemEval	REALTALK	ESG-Reports	QASPER	PeerQA	CovidQA	MLDR	Overall
Standalone retrieval systems									
BM25	34.7/25.3	73.4/62.7	27.7/21.7	49.8/34.3	55.1/32.2	26.7/20.3	89.8/73.6	99.0/83.4	57.0/44.2
ColBERTv2	63.0/50.0	87.5/80.0	54.3/45.8	56.9/45.0	75.5/52.0	45.3/36.3	95.2/84.4	96.0/85.4	71.7/59.9
XTR	43.5/31.6	89.4/79.1	42.7/33.1	62.9/51.9	71.2/48.8	38.8/29.8	93.9/83.3	98.0/88.2	67.5/55.7
Qwen3-Embedding-0.6B									
Base	46.4/33.9	84.8/72.8	46.4/37.4	**52.8**/40.1	69.7/47.0	36.2/26.7	91.4/73.1	**95.0**/76.4	65.3/50.9
LC	50.4/37.3	85.9/73.8	51.1/**41.0**	46.1/36.4	69.6/47.2	37.1/27.3	90.8/73.5	**95.0/76.9**	65.7/51.7
InvGC	48.1/35.0	**88.7**/77.0	49.3/39.7	45.6/36.2	69.8/47.1	36.7/27.4	92.2/74.5	**95.0**/76.4	65.7/51.7
DBSN	46.6/34.1	86.7/75.9	46.7/37.7	48.1/35.3	69.7/47.0	38.1/28.4	92.5/74.4	**95.0**/76.4	65.4/51.2
Search-Adaptor	51.9/38.0	88.4/**77.3**	49.3/39.3	41.4/30.8	**73.7/51.6**	38.6/28.2	**93.1/76.1**	**95.0**/76.7	66.4/52.2
NEAR	**55.7/41.5**	86.2/75.1	**51.2**/40.8	49.4/**40.5**	69.4/47.4	**39.8/29.0**	90.5/73.6	**95.0**/76.1	**67.2/53.0**
Qwen3-Embedding-4B									
Base	41.4/29.7	59.0/47.1	42.5/34.5	60.7/41.2	73.1/50.2	35.2/26.7	93.5/76.6	96.0/77.0	62.7/47.9
LC	44.6/32.2	59.0/47.1	44.1/34.7	56.2/41.5	73.0/50.1	37.4/28.0	93.1/76.7	96.0/77.7	62.9/48.5
InvGC	49.3/35.6	91.3/81.2	46.8/37.6	59.1/44.7	73.2/51.2	38.3/27.9	93.5/78.1	96.0/**78.5**	68.4/54.4
DBSN	51.7/38.1	91.7/80.5	**49.5/40.2**	58.4/42.6	73.1/50.2	**39.6**/29.1	93.4/76.7	96.0/77.0	69.2/54.3
Search-Adaptor	50.1/37.4	**93.6/83.1**	43.7/35.5	52.9/38.3	**75.7/54.8**	37.9/28.8	**94.0/79.1**	96.0/78.3	68.0/54.4
NEAR	**56.1/42.0**	89.7/78.5	47.2/37.7	**62.7/47.8**	72.8/50.9	38.9/**30.3**	93.8/77.6	**97.0**/77.8	**69.8/55.3**
Qwen3-Embedding-8B									
Base	49.2/36.2	87.9/77.4	49.8/40.7	62.2/43.0	70.4/47.4	40.8/30.1	94.0/79.5	95.0/77.4	68.6/53.9
LC	50.7/37.1	87.9/77.3	50.5/40.2	63.1/44.2	70.3/47.3	42.1/30.2	94.0/79.3	**96.0**/78.1	69.3/54.2
InvGC	52.0/38.8	**94.7**/84.2	51.9/**42.4**	56.2/37.9	71.6/48.3	38.2/28.6	94.1/**80.0**	**96.0**/77.0	69.4/54.6
DBSN	51.7/39.2	**94.7/84.7**	51.3/42.3	62.3/42.4	70.4/47.4	41.4/30.7	93.7/78.9	95.0/77.4	70.1/55.4
Search-Adaptor	54.7/41.3	93.1/83.0	50.6/41.0	52.9/37.2	**74.9/52.1**	41.2/30.8	**94.6**/79.9	**96.0/78.7**	69.7/55.5
NEAR	**59.9/45.3**	90.7/81.6	**52.6**/42.3	**68.0/49.9**	69.8/47.1	**43.1/32.1**	93.9/79.1	**96.0**/78.5	**71.8/57.0**

ESG-Reports (Macé et al., 2025), QASPER (Dasigi et al., 2021), PeerQA (Baumgärtner et al., 2025), CovidQA (Möller et al., 2020), and MLDR (Chen et al., 2024) for document retrieval. The document benchmarks cover financial reports, scientific papers, biomedical text, and long documents. We use the candidate banks and relevance labels released by LMEB, with 5,873 queries in total (Zhao et al., 2026b). We further evaluate downstream question answering on LoCoMo and LongMemEval to assess whether retrieval improvements lead to better answers (Maharana et al., 2024; Wu et al., 2025).

Baselines. We use cosine similarity between the original embeddings as Base. We compare NEAR with LC's local-affinity correction (Hara et al., 2015), InvGC's reference-based embedding transformation (Jian & Wang, 2023), and DBSN's similarity normalization (Pan et al., 2025). Search-Adaptor provides a supervised comparison by learning embedding mappings from relevance labels (Yoon et al., 2024). These same-backbone methods use the same frozen embeddings, data splits, and development-selection settings. We also include BM25 (Robertson & Zaragoza, 2009) as a lexical baseline, alongside ColBERTv2 (Santhanam et al., 2022b) and XTR (Lee et al., 2023) as dedicated multi-vector retrievers. The latter use their native encoders and retrieval configurations on the same candidate banks.

Implementation Details. We evaluate nine frozen embedding models (Zhang et al., 2025; Vera et al., 2025; Zhao et al., 2026a; Nussbaum et al., 2025; Moreira et al., 2024; Babakhin et al., 2025). Following the LMEB evaluation settings, we report Recall@10 and NDCG@10, with Overall averaged equally across the eight tasks. NEAR uses HNSW for retrieval (Malkov & Yashunin, 2020). Residual depth, normalization, and fusion weight are selected using development NDCG@10. Model configurations, fixed hyperparameters, and evaluation splits are detailed in Appendices A and B.

For downstream question answering, we compare Base with NEAR using Qwen3-30B-A3B (Yang et al., 2025) to generate answers and GPT-4o-mini (OpenAI, 2024) to judge correctness. We report Accuracy and answer-text F1 at $K = 10$. The retrieval-budget analysis varies $K \in \{4, 8, 10, 16\}$.

Table 2: **Cross-model gains.** Eight-task mean capped Recall@10 and NDCG@10 (%).

	Recall@10		NDCG@10	
Model	Base	**NEAR**	Base	**NEAR**
Qwen3-Embedding-0.6B	65.34	**67.17**	50.93	**52.99**
Qwen3-Embedding-4B	62.67	**69.76**	47.88	**55.33**
Qwen3-Embedding-8B	68.65	**71.76**	53.95	**57.00**
Llama-Nemotron-Embed-1B-v2	74.61	**75.05**	61.25	**61.55**
Llama-NV-Embed-Reasoning-3B	77.50	**78.78**	60.96	**62.40**
Llama-Embed-Nemotron-8B	76.51	**77.88**	62.73	**63.71**
EmbeddingGemma-300M	72.66	**76.09**	59.36	**61.31**
KaLM-Embedding-Gemma3-12B	72.67	**75.19**	57.11	**59.21**
ModernBERT-Embed-Base	69.76	**70.88**	55.72	**57.08**

4.2 Main Retrieval Results

Overall retrieval performance. Table 1 shows that NEAR achieves the highest average Recall@10 and NDCG@10 among the evaluated same-backbone methods at all three Qwen3-Embedding sizes. For Qwen-4B, average Recall@10 increases from 62.67 to 69.76, while NDCG@10 rises from 47.88 to 55.33. On Qwen-8B, NEAR exceeds the strongest same-backbone baseline for each metric by 1.68 percentage points in average Recall@10 and 1.50 points in average NDCG@10.

Performance across tasks. On LoCoMo, NEAR achieves the best Recall@10 and NDCG@10 among same-backbone methods at all three Qwen sizes. On LongMemEval, it also improves both metrics over Base. For Qwen-4B, the two metrics increase by 14.76 and 12.34 percentage points on LoCoMo and by 30.70 and 31.39 points on LongMemEval, respectively. On LongMemEval, Qwen-8B with NEAR also outperforms ColBERTv2 and XTR on both metrics. For document retrieval, NEAR achieves the best same-backbone scores on both metrics for ESG-Reports with Qwen-4B and Qwen-8B, and the highest same-backbone NDCG@10 for PeerQA at all three Qwen sizes.

Results across embedding models. NEAR's improvements are not limited to the Qwen series. Table 2 shows that NEAR improves average Recall@10 and NDCG@10 over Base for all nine frozen embedding models, including both small and large models. For example, the two metrics increase by 3.43 and 1.95 percentage points on EmbeddingGemma-300M and by 2.52 and 2.11 points on KaLM-Embedding-Gemma3-12B, respectively. Complete comparisons with the other methods for each model appear in Appendix D.

4.3 Downstream Question Answering

We further examine whether retrieval improvements translate into better downstream answers. Table 3 compares Base and NEAR across nine backbones using the original conversation histories of LoCoMo and LongMemEval. All comparisons use the same answering model and prompts, with a fixed budget of ten retrieved memories.

On LoCoMo, NEAR improves both Accuracy and F1 across all nine backbones. For Qwen-4B, Accuracy increases from 62.7 to 72.3, while F1 rises from 30.9 to 35.5. The improvements extend across model families: Llama-3B gains 13.3 and 7.3 percentage points on the two metrics, respectively. These results show that NEAR's retrieval gains can translate into better answers to questions about conversation histories.

On LongMemEval, NEAR improves Accuracy on seven of the nine backbones, including Qwen-4B from 69.3 to 72.1 and KaLM-12B from 75.8 to 77.4; Llama-3B and EmbeddingGemma show small decreases. Supplementary lexical F1 does not always follow Accuracy: Qwen-8B gains 0.8 percentage points in Accuracy while losing 0.2 points in F1, reflecting the distinction between answer correctness and lexical overlap.

4.4 Ablation Study

We conduct ablations on all eight retrieval tasks using the three Qwen3-Embedding sizes and Llama-3B. Each variant retains the depth, residual normalization, and fusion weight selected for full NEAR

Table 3: **Downstream QA at Top-10 (%).** NEAR scores are followed by blue parenthesized changes (NEAR − Base, in percentage points). Category columns report accuracy.

(a) LoCoMo					**Overall**	
Model	S-hop	Multi	Temp.	Open	**Acc.**	**F1**
Qwen3-Embedding-0.6B	83.8 (+7.4)	64.9 (+6.0)	49.2 (+4.4)	57.3 (+6.2)	71.5 (+6.4)	34.6 (+3.7)
Qwen3-Embedding-4B	85.7 (+9.2)	66.3 (+14.2)	46.7 (+8.1)	58.3 (+5.2)	72.3 (+9.6)	35.5 (+4.6)
Qwen3-Embedding-8B	85.1 (+1.7)	63.5 (+1.1)	46.7 (+4.0)	60.4 (+12.5)	71.6 (+2.7)	34.2 (+0.8)
Llama-Nemotron-Embed-1B-v2	85.1 (+5.8)	55.0 (+3.5)	42.1 (+0.3)	57.3 (+7.3)	68.9 (+4.4)	34.0 (+2.3)
Llama-NV-Embed-Reasoning-3B	54.3 (+18.1)	35.5 (+9.9)	21.8 (+6.2)	40.6 (+5.2)	43.2 (+13.3)	21.5 (+7.3)
Llama-Embed-Nemotron-8B	89.2 (+2.5)	64.9 (+6.0)	47.7 (+2.8)	52.1 (−1.0)	73.8 (+3.0)	36.7 (+1.9)
EmbeddingGemma-300M	88.0 (+4.0)	62.4 (+6.4)	47.4 (+3.4)	55.2 (0.0)	72.8 (+4.1)	35.9 (+2.2)
KaLM-Embedding-Gemma3-12B	88.6 (−0.1)	65.6 (+2.8)	48.3 (+1.6)	53.1 (−1.0)	73.8 (+0.7)	36.5 (+0.1)
ModernBERT-Embed-Base	70.5 (+18.8)	47.5 (+11.0)	36.1 (+8.4)	51.0 (+12.5)	57.9 (+14.8)	29.8 (+7.9)

(b) LongMemEval							**Overall**	
Model	User	Asst.	Pref.	Multi	Temp.	Upd.	**Acc.**	**F1**
Qwen3-Embedding-0.6B	91.4 (+5.7)	98.2 (+1.8)	70.0 (+10.0)	47.4 (−0.8)	54.1 (+4.5)	80.8 (−3.8)	67.6 (+2.0)	11.5 (+0.2)
Qwen3-Embedding-4B	97.1 (+2.9)	98.2 (0.0)	70.0 (+16.7)	52.6 (+2.3)	60.2 (−0.8)	85.9 (+6.4)	72.1 (+2.8)	12.3 (+0.1)
Qwen3-Embedding-8B	95.7 (0.0)	98.2 (+1.8)	76.7 (+6.7)	60.9 (0.0)	58.6 (−1.5)	84.6 (+3.8)	74.0 (+0.8)	11.6 (−0.2)
Llama-Nemotron-Embed-1B-v2	88.6 (+4.3)	96.4 (0.0)	63.3 (+6.7)	42.9 (+3.0)	49.6 (+6.0)	85.9 (+1.3)	65.0 (+3.6)	11.8 (+0.2)
Llama-NV-Embed-Reasoning-3B	82.9 (+1.4)	100.0 (+1.8)	66.7 (−10.0)	36.1 (−0.8)	44.4 (−3.0)	79.5 (+2.6)	60.6 (−0.8)	10.9 (+0.4)
Llama-Embed-Nemotron-8B	95.7 (0.0)	98.2 (−1.8)	70.0 (−10.0)	51.9 (+3.0)	55.6 (+2.3)	85.9 (+3.8)	70.6 (+1.2)	12.0 (0.0)
EmbeddingGemma-300M	95.7 (+2.9)	98.2 (+1.8)	76.7 (+3.3)	48.1 (−2.3)	51.9 (−6.8)	85.9 (+6.4)	69.0 (−0.6)	11.7 (+0.3)
KaLM-Embedding-Gemma3-12B	95.7 (−1.4)	100.0 (+1.8)	76.7 (+6.7)	66.2 (+5.3)	66.2 (0.0)	83.3 (−1.3)	77.4 (+1.6)	12.7 (+0.3)
ModernBERT-Embed-Base	90.0 (+1.4)	94.6 (−3.6)	63.3 (0.0)	46.6 (+0.8)	52.6 (+3.8)	76.9 (+1.3)	65.4 (+1.2)	11.4 (+0.2)

on development data, without retuning, and preserves original embedding matching. Table 4 summarizes the results, and Appendix F details the controls.

Table 4: **Effect of neighborhood references (%).** Eight-task mean capped Recall@10 (R@10) and NDCG@10 (N@10). Bold marks column bests.

	Qwen3-Embedding-0.6B		**Qwen3-Embedding-4B**		**Qwen3-Embedding-8B**		**Llama-NV-Embed-Reasoning-3B**	
Variant	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10
Global reference	66.18	51.84	65.12	51.14	69.69	54.93	77.85	61.23
Shuffled references	65.80	51.80	68.14	53.96	70.22	55.72	77.69	61.79
Memory only	65.97	52.12	67.54	53.76	70.86	56.18	78.69	62.13
Query only	66.74	52.24	67.94	53.17	70.19	55.15	77.77	60.80
NEAR	**67.17**	**52.99**	**69.76**	**55.33**	**71.76**	**57.00**	**78.78**	**62.40**

Local references and their correspondence. We first replace per-memory local references with the memory bank's global mean direction. Full NEAR outperforms this variant on both metrics for all four backbones, with average gains of 2.16 and 2.15 percentage points in Recall@10 and NDCG@10, respectively. Keeping the local references but shuffling their assignments to memories reduces the two metrics by 1.40 and 1.12 points on average. These results indicate that the gains depend not just on introducing reference directions, but also on their locality and correspondence with individual memories.

Query- and memory-side references. We construct residuals on only the memory or query side, retaining the original embeddings on the other side. Full NEAR outperforms both one-sided variants on all four backbones: average Recall@10 and NDCG@10 increase by 1.10 and 0.88 points over memory-only matching, and by 1.21 and 1.59 points over query-only matching. These results support using both query- and memory-side references to construct complementary matching representations.

Case study. In the LoCoMo example in Figure 1, Base retrieves personal updates and endorsement experiences but misses Tim's advice to align endorsements with John's values and brand. Local references define the projection directions. Here, distractors' raw scores fall more than the correct evidence's after projection, indicating that the removed components contribute to their original rank-

ing advantage. The evidence outscores Base's top ten in the first residual layer. Fusing all residual scores with original matching moves it from rank 75 to rank 1.

4.5 Sensitivity Analysis

Fusion weight and residual construction. We study the effects of fusion weight, residual normalization, and depth using Qwen3-Embedding-4B on the complete query sets of LoCoMo and LongMemEval. We vary the total residual weight over $\alpha \in \{0.25, 0.5, 1, 2, 4\}$. Different depths use the same total residual weight, with all other settings held fixed.

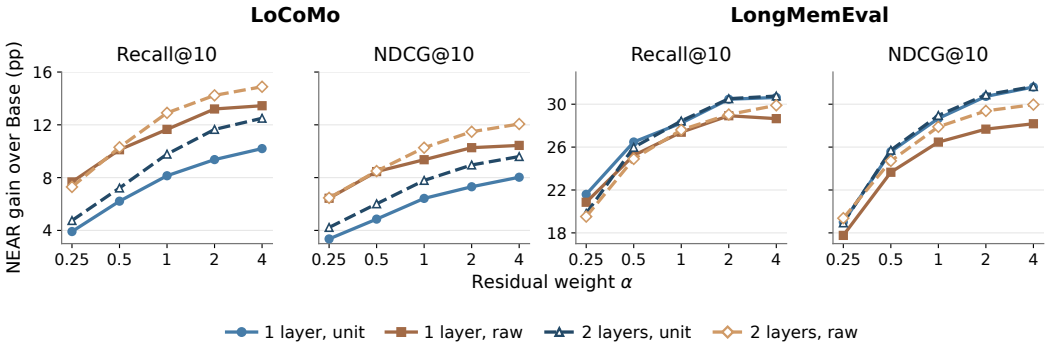


Figure 3: **Sensitivity to residual weight and construction.** Each curve reports the NEAR gain over Base for Qwen3-Embedding-4B. Unit and raw denote normalized and unnormalized scoring views; layer choices use the same total residual weight α. Vertical ranges differ by task.

As shown in Figure 3, all displayed configurations outperform Base in both Recall@10 and NDCG@10 on both tasks throughout the tested weight range. Performance generally improves as the residual weight increases and levels off at larger weights. The tasks favor different residual constructions: unnormalized residuals perform better on LoCoMo, whereas normalized residuals achieve higher NDCG@10 at moderate and large weights on LongMemEval. Thus, NEAR maintains gains over Base across multiple residual constructions and fusion weights, rather than only at a particular configuration.

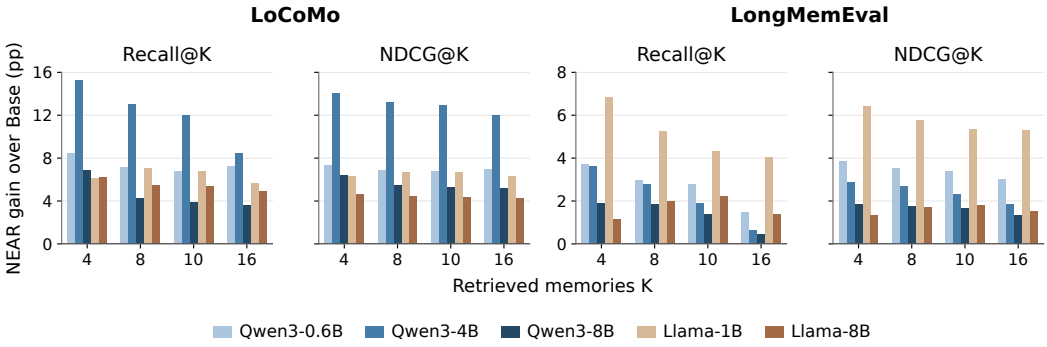


Figure 4: **Gains across retrieved-memory budgets.** NEAR minus Base across five embedding models; vertical ranges differ by task.

Retrieved memory budget. We further vary the number of returned memories over $K \in \{4, 8, 10, 16\}$ using the original conversational candidate banks of LoCoMo and LongMemEval, with the remaining retrieval configuration held fixed. Across the five embedding models shown in Figure 4, NEAR improves Recall@K and NDCG@K over Base at every tested budget on both datasets. The retrieval gains therefore extend beyond the Top-10 setting used in the main comparison. These curves measure evidence retrieval; the Top-10 QA results in Table 3 separately assess downstream answer quality.

4.6 Retrieval Cost Analysis

We compare retrieval quality, storage, and query latency using Qwen-0.6B/4B on LoCoMo and LongMemEval. Quality is evaluated on the complete tasks, while storage and latency are measured on four representative candidate banks. To examine the effect of index quantization, we compare Base and NEAR in FP32 and with four-bit scalar quantization (SQ4). All indexes use the same HNSW settings (Malkov & Yashunin, 2020; Johnson et al., 2021). Query latency includes vector preparation and ANN search, but excludes encoding. Appendix G provides additional details.

Table 5: **Effect of index quantization.** Unmarked rows use FP32. Storage is relative to FP32 Base, and query latency excludes encoding.

Method	Recall@10	NDCG@10	Storage (×Base)	Latency (ms)
Base	57.91	45.88	1.00	0.242
Base (SQ4)	57.81	45.82	0.17	0.239
NEAR	71.82	59.21	2.02	0.579
NEAR (SQ4)	71.97	59.16	0.36	0.590

Retrieval quality and latency. Table 5 shows that uncompressed NEAR reaches mean Recall@10 and NDCG@10 of 71.82 and 59.21, compared with 57.91 and 45.88 for Base. Median query latency increases from 0.242ms to 0.579ms, an additional 0.337ms, or approximately 2.39 times the Base latency. NEAR therefore trades additional retrieval computation for improved quality rather than reducing query latency.

Offline construction. On the same representative banks of 476 or 629 memories, building NEAR's references, scales, keys, and FP32 index takes 0.129–1.147 seconds, compared with 0.040–0.140 seconds for Base. NEAR with SQ4 takes 0.244–1.988 seconds, including quantizer training. These are per-bank build times from cached embeddings, separate from query latency; the index and reference table are reused across queries. Appendix G reports detailed construction times for all compared methods.

Storage and quantization. Uncompressed NEAR has a median total-storage ratio of 2.02 relative to FP32 Base. SQ4 reduces NEAR's total assets by a median 82.7%, while mean Recall@10 changes from 71.82 to 71.97 and NDCG@10 from 59.21 to 59.16. Its total assets are a median 36.0% of FP32 Base; quantizing Base itself reduces its assets to 16.7%, with mean NDCG@10 of 45.82. Quantization thus substantially reduces NEAR's storage requirements while retaining its retrieval quality, although its assets remain larger than those of quantized Base. NEAR's median query latency with SQ4 is 0.590ms, close to its uncompressed latency, so the measured benefit is primarily in storage. These timings characterize cached-vector retrieval on a shared CPU host, not end-to-end question answering.

5 Conclusion

We studied the difficulty of retrieving correct evidence from memory records with shared topics, entities, or events. Such shared contents would make records without the required evidence more likely to outrank the correct evidence when ranking relied only on query–memory similarity. Our key observation was that these similar records were not only competing candidates; they also provided a local neighborhood that could serve as a reference for each memory. In this paper, we proposed NEAR, a training-free framework that turned this potential source of confusion into useful matching information by constructing memory-specific local references. NEAR obtained query references from a precomputed table and combined residual representations with the original embeddings in a fixed index for a single approximate nearest-neighbor search. Experiments on eight long-context retrieval benchmarks showed improvements in mean Recall@10 and NDCG@10 across nine embedding models, with further gains in downstream question answering. In future work, we will extend NEAR to frequently updated memory banks and explore more compact reference representations.

Reproducibility Statement

We describe NEAR in Section 3 and the experimental setup in Section 4.1. Appendices A–C document dataset statistics, model and baseline configurations, hyperparameter selection, and implementation details. Complete retrieval results and additional efficiency measurements are provided in Appendices D and G, respectively.

AI Use Statement

We used generative AI tools to revise the manuscript and make the text in figures and qualitative examples clearer and more concise. The underlying data, evaluation labels, and reported results were unchanged by these editorial revisions. Responsibility for the final content of the paper rests with the authors.

References

Dhananjay Ashok, Suraj Nair, Mutasem Al-Darabsah, Choon Hui Teo, Tarun Agarwal, and Jonathan May. A representation sharpening framework for zero shot dense retrieval. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*, 2025.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. PeerQA: A scientific question answering dataset from peer reviews. 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)*, 2025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Max Conti, Manuel Faysse, Gautier Viaud, Antoine Bosselut, Celine Hudelot, and Pierre Colombo. Context is gold to find the gold passage: Evaluating and training contextual document embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu, and Miloš Radovanović. Localized centering: Reducing hubness in large-sample data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Xiangru Jian and Yimu Wang. InvGC: Robust cross-modal retrieval by inverse graph convolution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. Realtalk: A 21-day real-world dataset for long-term conversation. *arXiv preprint arXiv:2502.13270*, 2025.

Jinhyuk Lee, Zhuyun Dai, Sai Meher Karthik Duddu, Tao Lei, Iftekhar Naim, Ming-Wei Chang, and Vincent Zhao. Rethinking the role of token retrieval in multi-vector retrieval. In *Advances in Neural Information Processing Systems*, 2023.

Yunxiao Luo, Siyuan Wang, Ben Chen, and Chenyi Lei. Prq-kmeans: Projection residual quantization for semantic id tokenization. *arXiv preprint arXiv:2608.24207*, 2026.

Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Yury A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

John X. Morris and Alexander M. Rush. Contextual document embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chang Nie, Chaoyou Fu, Junlan Feng, and Caifeng Shan. Evoembedding: Evolvable representations for long-context retrieval and agentic memory. *arXiv preprint arXiv:2606.21649*, 2026.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic Embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*, 2025.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

Zhengxin Pan, Haishuai Wang, Fangyu Wu, Peng Zhang, and Jiajun Bu. Hubness reduction with dual bank sinkhorn normalization for cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia*, 2025.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022a.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022b.

Jan Luca Scheerer, Matei Zaharia, Christopher Potts, Gustavo Alonso, and Omar Khattab. WARP: An efficient engine for multi-vector retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

Shaoyan Sun, Ying Li, Wengang Zhou, Qi Tian, and Houqiang Li. Local residual similarity for image re-ranking. *Information Sciences*, 417:143–153, 2017.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025.

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.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. Search-adaptor: Embedding customization for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Kai Zhang, Xinyuan Zhang, Ejaz Ahmed, Hongda Jiang, Caleb Kumar, Kai Sun, Zhaojiang Lin, Sanat Sharma, Shereen Oraby, Aaron Colak, Ahmed Aly, Anuj Kumar, Xiaozhong Liu, and Xin Dong. AssoMem: Scalable memory QA with multi-signal associative retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

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. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Xinping Zhao, Xinshuo Hu, Jiaxin Xu, Danyu Tang, Xin Zhang, Mengjia Zhou, Yan Zhong, Yao Zhou, Zifei Shan, Meishan Zhang, et al. Lmeb: Long-horizon memory embedding benchmark. *arXiv preprint arXiv:2603.12572*, 2026b.

A Evaluation details

We evaluate 5,873 queries across eight benchmarks. Following LMEB, we report capped Recall@10 and NDCG@10 with linear relevance gains.

Dataset	Queries	Documents	Banks	Subsets
LoCoMo	1,976	5,882	10	5
LongMemEval	500	237,655	500	6
REALTALK	679	8,944	10	3
QASPER	1,335	20,508	413	1
ESG-Reports	36	2,407	18	1
PeerQA	136	18,593	70	1
CovidQA	1,111	3,351	112	1
MLDR	100	1,536	100	1

Model configurations. Each encoder follows its released query/document format and pooling strategy. Same-encoder comparisons use identical inputs and cached embeddings. The KaLM and ModernBERT citations follow their official model cards; the checkpoints used are KaLM-Embedding-Gemma3-12B-2511 and ModernBERT-Embed-Base. The two smaller Llama models are Llama-Nemotron-Embed-1B-v2 (Moreira et al., 2024) and Llama-NV-Embed-Reasoning-3B.

Baselines. LC, InvGC, and DBSN use the same frozen embeddings as NEAR, with configurations selected on aligned development splits. Search-Adaptor uses a shared residual MLP with hidden width 256, trained with Adam at learning rate 10^{-3} and batch size up to 128 for at most 2,000 steps (Yoon et al., 2024). ColBERTv2 and XTR retain their native encoders and input settings. We use PLAID (Santhanam et al., 2022a) for ColBERTv2 and WARP (Scheerer et al., 2025) for XTR. BM25 uses the default BM25Okapi parameters.

Fixed settings. We fix the neighborhood width to 50, the reference-table size to 16, the number of routing entries to four, and the inverse temperature to 10. Depth, residual normalization, and total residual weight are selected on development data.

Downstream question answering. Retrieved memories are restored to chronological order before answer generation, which uses temperature zero and a maximum of 1,024 output tokens. Correctness is evaluated with fixed task-specific judge prompts. LoCoMo F1 follows its category-specific procedure; LongMemEval F1 measures normalized, stemmed word overlap on answerable questions. Base and NEAR are compared on the same questions.

B Development selection

For NEAR, we search depth $L \in \{1, 2, 3\}$, raw or normalized residuals, and total residual weight in $\{0.25, 0.5, 1, 2, 4\}$. For seven benchmarks, we use grouped five-fold evaluation, assigning queries that share candidate banks or overlapping candidates to the same group. MLDR uses disjoint ConTEB queries for development. All methods use aligned splits and select their configurations by development NDCG@10. The selected configurations are reused for held-out Top-10 and Top-5 evaluation. Results for the seven grouped benchmarks are averaged over three repetitions.

C Implementation details

For each matching representation $\ell = 0, \ldots, L$, we compute a fixed score scale from the memory bank:

$$\tau_\ell = \max\left\{10^{-6}, \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}_{j=1,\ldots,N}\left[(\mathbf{z}_i^{(\ell)})^\top \mathbf{z}_j^{(\ell)}\right]}\right\},$$

where Var is the population variance over all N memories, including $j = i$. The scales are computed offline without queries or relevance labels and shared by all queries in the same bank. The implementation evaluates this expression using the covariance of the memory representations, without constructing all pairwise scores.

The query-reference table contains at most 16 entries and is constructed from the memory embeddings using spherical K-means. At retrieval time, the original and residual matching components are combined into a single inner-product representation, allowing NEAR to use one ANN search over a fixed index.

For ANN retrieval, we use FAISS HNSW with inner-product search. We set $M = 32$, `efConstruction` to 200, `efSearch` to 128, and the random seed to 42.

D COMPLETE RETRIEVAL RESULTS

Table A1: **Complete cross-model comparison (%).** Eight-task mean capped Recall@10 and NDCG@10 for all same-backbone methods. Bold marks row bests, including ties.

(a) Capped Recall@10

Model	Base	LC	InvGC	DBSN	Search-Adaptor	NEAR
Qwen3-Embedding-0.6B	65.34	65.74	65.67	65.42	66.42	**67.17**
Qwen3-Embedding-4B	62.67	62.91	68.44	69.17	67.97	**69.76**
Qwen3-Embedding-8B	68.65	69.32	69.35	70.08	69.73	**71.76**
EmbeddingGemma-300M	72.66	73.80	72.49	72.78	72.22	**76.09**
KaLM-Embedding-Gemma3-12B	72.67	73.44	72.29	73.33	73.51	**75.19**
ModernBERT-Embed-Base	69.76	70.00	69.26	69.42	69.53	**70.88**
Llama-Nemotron-Embed-1B-v2	74.61	74.80	74.66	74.32	74.62	**75.05**
Llama-NV-Embed-Reasoning-3B	77.50	76.85	76.95	76.71	75.65	**78.78**
Llama-Embed-Nemotron-8B	76.51	76.67	76.19	76.24	76.67	**77.88**

(b) NDCG@10

Model	Base	LC	InvGC	DBSN	Search-Adaptor	NEAR
Qwen3-Embedding-0.6B	50.93	51.69	51.67	51.15	52.23	**52.99**
Qwen3-Embedding-4B	47.88	48.50	54.36	54.31	54.40	**55.33**
Qwen3-Embedding-8B	53.95	54.23	54.64	55.37	55.50	**57.00**
EmbeddingGemma-300M	59.36	59.38	59.00	59.03	59.25	**61.31**
KaLM-Embedding-Gemma3-12B	57.11	57.52	57.75	58.21	58.80	**59.21**
ModernBERT-Embed-Base	55.72	55.78	55.71	55.41	55.45	**57.08**
Llama-Nemotron-Embed-1B-v2	61.25	61.01	61.09	61.01	60.91	**61.55**
Llama-NV-Embed-Reasoning-3B	60.96	60.50	61.00	60.60	59.49	**62.40**
Llama-Embed-Nemotron-8B	62.73	62.47	62.48	62.47	63.03	**63.71**

The tables report task-level Recall@10 and NDCG@10 for nine encoders and six same-encoder methods. The two smaller Llama-derived encoders are grouped for legibility.

Table A2: **Additional Llama model results (%).** Bold marks the best value within each model and column.

(a) Capped Recall@10 ↑

Method	LoCoMo	LongMemEval	REALTALK	ESG-Reports	QASPER	PeerQA	CovidQA	MLDR	Overall
Llama-Nemotron-Embed-1B-v2									
Base	68.33	93.60	56.37	60.60	77.48	45.84	95.68	**99.00**	74.61
LC	68.40	93.69	56.71	60.83	77.46	46.69	95.65	**99.00**	74.80 ±0.25
InvGC	68.86	**93.91**	**57.07**	59.10	77.47	45.76	96.07	**99.00**	74.66 ±0.43
DBSN	68.04	93.60	56.61	59.06	77.48	45.13	95.68	**99.00**	74.32 ±0.33
Search-Adaptor	69.00	93.75	56.89	55.77	**78.81**	**47.64**	**96.10**	**99.00**	74.62 ±0.04
NEAR	**69.53**	93.45	56.13	**61.88**	77.74	46.97	95.71	**99.00**	**75.05** ±0.36
Llama-NV-Embed-Reasoning-3B									
Base	72.09	89.39	61.20	78.98	80.78	44.97	93.61	**99.00**	77.50
LC	71.92	89.23	61.20	73.77	80.70	45.37	93.58	**99.00**	76.85 ±0.26
InvGC	72.02	88.83	61.61	74.24	80.71	45.26	**93.91**	**99.00**	76.95 ±0.45
DBSN	72.09	89.56	60.84	73.52	80.78	44.39	93.55	**99.00**	76.71 ±0.11
Search-Adaptor	71.87	90.08	61.22	64.65	**81.57**	43.66	93.19	**99.00**	75.65 ±0.81
NEAR	**74.42**	**90.52**	**61.73**	**83.66**	80.23	**47.02**	93.64	**99.00**	**78.78** ±0.10

(b) NDCG@10 ↑

Method	LoCoMo	LongMemEval	REALTALK	ESG-Reports	QASPER	PeerQA	CovidQA	MLDR	Overall
Llama-Nemotron-Embed-1B-v2									
Base	53.88	83.04	47.42	48.21	56.32	**36.31**	84.56	**80.26**	61.25
LC	54.31	83.19	47.28	47.05	56.06	35.58	84.36	**80.26**	61.01 ±0.03
InvGC	54.21	84.10	**48.27**	46.01	56.22	35.14	84.51	80.24	61.09 ±0.22
DBSN	53.58	83.97	47.38	47.13	56.32	34.91	84.56	**80.26**	61.01 ±0.16
Search-Adaptor	54.43	83.14	47.88	42.27	**58.73**	35.83	**84.76**	**80.26**	60.91 ±0.20
NEAR	**55.22**	**84.45**	47.06	**48.52**	56.32	36.00	84.56	**80.26**	**61.55** ±0.27
Llama-NV-Embed-Reasoning-3B									
Base	56.13	75.51	49.19	56.74	58.12	34.23	76.69	81.10	60.96
LC	55.98	75.34	49.02	53.63	58.01	34.29	76.66	81.10	60.50 ±0.23
InvGC	56.24	75.73	**49.63**	55.30	58.18	33.68	77.20	**82.07**	61.00 ±0.18
DBSN	56.13	76.43	48.70	54.50	58.12	33.28	76.52	81.10	60.60 ±0.46
Search-Adaptor	56.14	75.82	48.96	42.87	**60.13**	32.80	**77.38**	81.84	59.49 ±1.01
NEAR	**58.20**	**76.83**	49.60	**61.37**	58.33	**37.23**	76.54	81.11	**62.40** ±0.13

Table A3: **Complete capped Recall@10 results (%).** Table A2 adds the 1B and 3B Llama models. Overall reports the mean and sample SD across the three splits.

(a) Capped Recall@10 ↑

Method	LoCoMo	LongMemEval	REALTALK	ESG-Reports	QASPER	PeerQA	CovidQA	MLDR	Overall
Standalone retrieval systems									
ColBERTv2	62.95	87.53	54.34	56.94	75.46	45.25	95.23	96.00	71.71
XTR	43.55	89.42	42.65	62.87	71.19	38.82	93.88	98.00	67.55
Qwen3-Embedding-0.6B									
Base	46.44	84.84	46.39	**52.82**	69.71	36.17	91.36	**95.00**	65.34
LC	50.38	85.87	51.06	46.13	69.55	37.11	90.85	**95.00**	65.74 ±0.39
InvGC	48.10	**88.70**	49.27	45.62	69.79	36.69	92.23	**95.00**	65.67 ±0.27
DBSN	46.55	86.72	46.71	48.07	69.71	38.08	92.53	**95.00**	65.42 ±0.84
Search-Adaptor	51.88	88.45	49.31	41.40	**73.69**	38.57	**93.07**	**95.00**	66.42 ±0.27
NEAR	**55.75**	86.24	**51.24**	49.41	69.43	**39.75**	90.55	**95.00**	**67.17** ±0.21
Qwen3-Embedding-4B									
Base	41.39	58.96	42.47	60.69	73.11	35.19	93.52	96.00	62.67
LC	44.59	58.96	44.12	56.19	72.97	37.36	93.07	96.00	62.91 ±0.35
InvGC	49.31	91.33	46.77	59.06	73.25	38.34	93.49	96.00	68.44 ±0.77
DBSN	51.71	91.69	**49.51**	58.41	73.11	**39.58**	93.37	96.00	69.17 ±0.29
Search-Adaptor	50.07	**93.57**	43.71	52.90	**75.65**	37.88	**94.00**	96.00	67.97 ±1.17
NEAR	**56.15**	89.65	47.17	**62.73**	72.76	38.87	93.79	**97.00**	**69.76** ±0.28
Qwen3-Embedding-8B									
Base	49.15	87.93	49.78	62.22	70.38	40.76	93.97	95.00	68.65
LC	50.66	87.89	50.50	63.12	70.27	42.12	93.97	**96.00**	69.32 ±0.35
InvGC	51.97	**94.74**	51.93	56.20	71.62	38.18	94.15	**96.00**	69.35 ±0.23
DBSN	51.75	94.69	51.34	62.33	70.38	41.43	93.70	95.00	70.08 ±0.23
Search-Adaptor	54.65	93.07	50.57	52.85	**74.92**	41.18	**94.57**	**96.00**	69.73 ±0.50
NEAR	**59.88**	90.70	**52.56**	**68.01**	69.85	**43.15**	93.91	**96.00**	**71.76** ±0.16
EmbeddingGemma-300M									
Base	64.66	**91.88**	58.71	50.46	78.64	46.41	93.52	97.00	72.66
LC	65.72	91.73	59.91	56.57	78.39	45.32	93.73	99.00	73.80 ±0.35
InvGC	63.76	91.73	60.61	47.69	78.53	46.70	93.88	97.00	72.49 ±0.39
DBSN	64.66	91.23	58.84	51.88	78.64	46.63	93.40	97.00	72.78 ±0.38
Search-Adaptor	65.03	91.36	59.24	44.34	**79.73**	**47.06**	94.03	97.00	72.22 ±0.86
NEAR	**70.38**	91.76	**60.85**	**68.09**	78.43	45.09	**94.12**	**100.00**	**76.09** ±0.41
KaLM-Embedding-Gemma3-12B									
Base	58.50	93.07	54.99	63.61	73.46	44.78	92.98	**100.00**	72.67
LC	61.60	92.89	**57.47**	64.29	73.46	45.07	92.74	**100.00**	73.44 ±0.18
InvGC	60.34	92.85	55.68	58.73	73.84	44.24	93.67	99.00	72.29 ±0.21
DBSN	61.09	93.56	55.04	64.12	73.46	**47.22**	93.16	99.00	73.33 ±0.16
Search-Adaptor	64.76	**94.85**	56.76	55.12	**76.88**	45.36	**94.39**	**100.00**	73.51 ±0.94
NEAR	**66.59**	93.29	56.97	**72.65**	73.86	45.08	93.10	**100.00**	**75.19** ±0.17
ModernBERT-Embed-Base									
Base	58.93	87.40	51.44	**51.76**	76.57	42.40	92.62	**97.00**	69.76
LC	61.24	87.23	53.30	48.80	76.40	43.54	92.50	**97.00**	70.00 ±0.24
InvGC	58.30	88.47	52.43	46.90	76.35	42.01	92.59	**97.00**	69.26 ±0.41
DBSN	58.93	86.61	51.14	50.56	76.57	41.97	92.62	**97.00**	69.42 ±0.25
Search-Adaptor	60.36	**88.54**	52.14	46.16	**77.21**	42.15	**92.68**	**97.00**	69.53 ±0.87
NEAR	**66.06**	87.21	**53.82**	50.93	76.10	**44.38**	92.53	96.00	**70.88** ±0.40
Llama-Embed-Nemotron-8B									
Base	71.90	95.11	60.73	67.08	77.74	46.45	95.05	98.00	76.51
LC	71.63	94.62	61.46	68.35	77.74	47.11	94.45	98.00	76.67 ±0.21
InvGC	71.71	**95.62**	61.10	64.78	77.70	45.36	95.26	98.00	76.19 ±0.28
DBSN	71.90	94.80	60.51	65.20	77.74	45.04	**95.71**	**99.00**	76.24 ±0.58
Search-Adaptor	73.10	95.13	62.10	63.38	**78.89**	47.18	95.56	98.00	76.67 ±1.10
NEAR	**73.60**	94.82	**62.37**	**72.62**	77.91	**47.97**	94.78	**99.00**	**77.88** ±0.17

Table A4: **Complete NDCG@10 results (%).** The evaluation settings match Table 1.

(b) NDCG@10 ↑

Method	LoCoMo	LongMemEval	REALTALK	ESG-Reports	QASPER	PeerQA	CovidQA	MLDR	Overall
Standalone retrieval systems									
ColBERTv2	49.97	79.98	45.81	45.05	51.97	36.29	84.44	85.35	59.86
XTR	31.59	79.12	33.11	51.91	48.76	29.80	83.25	88.18	55.71
Qwen3-Embedding-0.6B									
Base	33.92	72.82	37.45	40.08	47.02	26.69	73.11	76.40	50.93
LC	37.35	73.85	**41.05**	36.39	47.21	27.33	73.49	**76.87**	51.69 ±0.22
InvGC	35.05	77.03	39.74	36.24	47.07	27.39	74.51	76.35	51.67 ±0.22
DBSN	34.09	75.89	37.66	35.33	47.02	28.43	74.41	76.40	51.15 ±0.38
Search-Adaptor	38.01	**77.29**	39.28	30.76	**51.55**	28.24	**76.05**	76.68	52.23 ±0.13
NEAR	**41.46**	75.14	40.82	**40.48**	47.37	**29.03**	73.59	76.07	**52.99** ±0.27
Qwen3-Embedding-4B									
Base	29.69	47.11	34.46	41.21	50.23	26.74	76.56	77.02	47.88
LC	32.21	47.11	34.75	41.47	50.08	28.03	76.68	77.70	48.50 ±0.34
InvGC	35.65	81.22	37.60	44.69	51.22	27.92	78.09	**78.49**	54.36 ±0.67
DBSN	38.14	80.46	**40.20**	42.64	50.23	29.10	76.68	77.02	54.31 ±0.31
Search-Adaptor	37.38	**83.06**	35.51	38.34	**54.76**	28.84	**79.06**	78.26	54.40 ±0.62
NEAR	**42.04**	78.50	37.70	**47.76**	50.90	**30.30**	77.63	77.83	**55.33** ±0.25
Qwen3-Embedding-8B									
Base	36.17	77.40	40.66	43.00	47.35	30.11	79.46	77.40	53.95
LC	37.13	77.28	40.22	44.22	47.26	30.22	79.32	78.15	54.23 ±0.17
InvGC	38.75	84.24	**42.39**	37.93	48.28	28.56	**79.95**	77.03	54.64 ±0.05
DBSN	39.21	**84.71**	42.30	42.38	47.35	30.66	78.94	77.40	55.37 ±0.06
Search-Adaptor	41.29	83.00	41.04	37.19	**52.06**	30.83	79.94	**78.69**	55.50 ±0.38
NEAR	**45.34**	81.64	42.33	**49.86**	47.13	**32.09**	79.15	78.48	**57.00** ±0.17
EmbeddingGemma-300M									
Base	50.29	79.81	48.47	42.33	56.85	37.49	78.79	80.88	59.36
LC	51.12	79.70	48.99	43.08	56.84	36.37	78.55	80.37	59.38 ±0.12
InvGC	49.94	80.14	**49.66**	38.49	56.70	37.34	78.84	80.88	59.00 ±0.41
DBSN	50.29	79.76	48.32	40.03	56.85	37.47	78.63	80.88	59.03 ±0.11
Search-Adaptor	51.05	79.87	48.85	35.27	**59.85**	**38.21**	**78.86**	**82.05**	59.25 ±0.62
NEAR	**55.01**	**80.45**	49.46	**52.18**	56.91	37.23	78.70	80.56	**61.31** ±0.26
KaLM-Embedding-Gemma3-12B									
Base	42.58	81.64	44.66	49.51	51.56	32.56	77.98	76.37	57.11
LC	45.00	81.53	45.55	50.22	51.34	32.47	77.68	76.37	57.52 ±0.04
InvGC	44.18	81.92	44.89	46.50	52.03	34.03	78.31	80.12	57.75 ±0.22
DBSN	44.93	82.78	44.34	48.05	51.56	**35.82**	77.59	**80.65**	58.21 ±0.28
Search-Adaptor	**49.60**	**83.37**	**46.03**	41.55	**55.69**	34.44	**80.02**	79.65	58.80 ±0.36
NEAR	49.43	81.74	45.53	**54.78**	51.69	34.67	77.87	78.00	**59.21** ±0.31
ModernBERT-Embed-Base									
Base	45.80	74.36	42.47	35.98	54.06	**34.61**	**76.93**	81.53	55.72
LC	47.15	74.45	43.41	35.45	53.75	34.24	76.58	81.19	55.78 ±0.14
InvGC	45.44	**76.59**	43.77	33.86	53.98	33.61	76.92	81.53	55.71 ±0.39
DBSN	45.80	74.60	42.28	34.44	54.06	33.65	**76.93**	81.53	55.41 ±0.22
Search-Adaptor	46.89	76.30	43.03	29.76	**55.34**	33.39	76.76	**82.13**	55.45 ±0.76
NEAR	**51.46**	74.92	**44.08**	**39.68**	53.87	34.46	76.91	81.28	**57.08** ±0.12
Llama-Embed-Nemotron-8B									
Base	57.69	85.84	51.13	50.94	56.55	35.55	80.98	83.16	62.73
LC	57.19	85.51	51.44	50.46	56.55	35.04	80.65	82.94	62.47 ±0.09
InvGC	57.52	**86.17**	51.13	48.86	56.41	34.67	81.56	**83.53**	62.48 ±0.02
DBSN	57.69	85.61	50.98	49.85	56.55	34.42	81.64	83.03	62.47 ±0.25
Search-Adaptor	59.13	85.71	51.93	46.88	**58.71**	36.25	**82.48**	83.16	63.03 ±0.72
NEAR	**60.05**	85.80	**52.03**	**55.11**	56.61	**36.46**	81.31	82.35	**63.71** ±0.14

E TOP-5 RESULTS

We report Top-5 results using the configurations selected for Top-10 evaluation.

Table A5: **Top-5 retrieval results (%).** Each cell reports capped Recall@5 above NDCG@5, averaged across eight benchmarks. Values are mean ± SD over three splits. Bold marks the best mean, including ties.

Model	Base	LC	InvGC	DBSN	Search-Adaptor	**NEAR**
	55.76	57.19 ±0.30	56.89 ±0.28	56.29 ±0.68	57.68 ±0.52	**57.80** ±0.21
Qwen3-Embedding-0.6B	47.40	48.51 ±0.17	48.40 ±0.31	47.78 ±0.27	48.95 ±0.30	**49.49** ±0.37
	52.18	53.00 ±0.40	59.34 ±0.35	59.66 ±0.23	58.35 ±0.42	**60.01** ±0.33
Qwen3-Embedding-4B	44.03	44.87 ±0.39	50.94 ±0.65	50.74 ±0.27	50.82 ±0.37	**51.73** ±0.31
	59.73	60.05 ±0.19	59.76 ±0.44	59.93 ±0.24	60.70 ±0.40	**62.84** ±0.30
Qwen3-Embedding-8B	50.57	50.76 ±0.12	51.09 ±0.03	51.56 ±0.05	52.09 ±0.34	**53.60** ±0.20
	63.74	63.98 ±0.32	63.78 ±0.23	64.04 ±0.44	63.58 ±0.72	**66.72** ±0.17
EmbeddingGemma-300M	56.02	55.70 ±0.06	55.74 ±0.31	55.75 ±0.19	55.96 ±0.54	**57.83** ±0.11
	63.34	64.20 ±0.43	63.48 ±0.51	64.03 ±0.27	64.86 ±0.73	**65.34** ±0.53
KaLM-Embedding-Gemma3-12B	53.56	53.98 ±0.18	54.40 ±0.38	54.69 ±0.37	**55.49** ±0.31	55.46 ±0.41
	59.32	60.48 ±0.91	60.10 ±0.43	59.41 ±0.14	60.30 ±0.89	**61.87** ±0.15
ModernBERT-Embed-Base	51.88	52.27 ±0.36	52.36 ±0.27	51.73 ±0.19	52.05 ±0.75	**53.71** ±0.07
	66.31	66.30 ±0.22	66.03 ±0.20	66.00 ±0.12	66.18 ±0.44	**67.14** ±0.12
Llama-Nemotron-Embed-1B-v2	58.13	57.81 ±0.15	57.87 ±0.35	57.89 ±0.03	57.76 ±0.23	**58.56** ±0.21
	67.52	67.03 ±0.56	67.18 ±0.48	66.74 ±0.47	65.31 ±0.85	**68.35** ±0.31
Llama-NV-Embed-Reasoning-3B	57.23	56.82 ±0.38	57.32 ±0.22	56.87 ±0.55	55.62 ±1.04	**58.49** ±0.20
	68.03	68.18 ±0.14	68.01 ±0.24	68.17 ±0.21	67.80 ±1.00	**69.82** ±0.11
Llama-Embed-Nemotron-8B	59.49	59.21 ±0.06	59.37 ±0.02	59.40 ±0.13	59.64 ±0.70	**60.61** ±0.11

F ADDITIONAL MECHANISM CONTROLS

All controls retain the depth, residual normalization, and fusion weight selected for NEAR. Shuffled-reference results are averaged over three fixed permutations.

G EFFICIENCY DETAILS

Storage and latency are measured on representative memory banks using one CPU core. Storage includes the index and query-side state; query latency includes vector preparation and ANN search, excluding encoding.

Offline construction. Offline construction starts from cached embeddings. We report the median of five runs.

Table A6: **Offline bank construction (seconds).** LoCoMo and LongMemEval banks contain 629 and 476 memories, respectively.

	Qwen3-Embedding-0.6B		**Qwen3-Embedding-4B**	
Method	LoCoMo	LongMemEval	LoCoMo	LongMemEval
Base (FP32)	0.063	0.040	0.140	0.087
Base (SQ4)	0.131	0.089	0.270	0.187
LC	0.072	0.065	0.229	0.102
InvGC	0.071	0.050	0.158	0.110
DBSN	0.073	0.050	0.157	0.101
Search-Adaptor	0.065	0.045	0.141	0.097
NEAR (FP32)	0.166	0.129	1.147	0.397
NEAR (SQ4)	0.274	0.244	1.988	0.648

Table A7: **Qwen3-Embedding-0.6B quality and retrieval costs.** Quality uses complete tasks; costs use representative banks. Aux. denotes query-side state.

Method	R@10	N@10	Index (MiB)	Aux. (KiB)	Total (MiB)	Query (ms)
LoCoMo; 629 candidates; 8 timed queries						
Base (FP32)	46.44	33.92	2.621	0.03	2.621	0.148
Base (SQ4)	45.96	33.57	0.479	0.03	0.479	0.156
LC	50.66	37.76	2.623	0.04	2.623	0.131
InvGC	48.78	35.62	2.621	0.04	2.621	0.109
DBSN	46.63	34.40	2.623	0.04	2.623	0.137
Search-Adaptor	52.23	38.08	2.621	2054.34	4.627	0.267
NEAR (FP32)	55.09	41.28	5.080	128.82	5.206	0.347
NEAR (SQ4)	55.23	41.19	0.799	128.82	0.925	0.441
LongMemEval; 476 candidates; 1 timed query						
Base (FP32)	84.84	72.82	1.983	0.03	1.983	0.139
Base (SQ4)	84.81	73.00	0.364	0.03	0.364	0.148
LC	86.08	73.95	1.985	0.04	1.985	0.129
InvGC	88.74	77.07	1.983	0.04	1.983	0.112
DBSN	86.67	76.17	1.985	0.04	1.985	0.130
Search-Adaptor	88.71	77.60	1.983	2054.34	3.990	0.255
NEAR (FP32)	86.45	75.53	3.843	128.82	3.969	0.315
NEAR (SQ4)	86.65	75.49	0.605	128.82	0.730	0.393

Table A8: **Qwen3-Embedding-4B quality and retrieval costs.** Same settings as the preceding table.

Method	R@10	N@10	Index (MiB)	Aux. (KiB)	Total (MiB)	Query (ms)
LoCoMo; 629 candidates; 8 timed queries						
Base (FP32)	41.39	29.69	6.306	0.03	6.307	0.396
Base (SQ4)	41.38	29.62	0.951	0.03	0.951	0.358
LC	45.12	32.23	6.309	0.04	6.309	0.281
InvGC	49.34	35.88	6.306	0.04	6.307	0.319
DBSN	51.93	38.24	6.309	0.04	6.309	0.351
Search-Adaptor	50.67	37.81	6.306	5132.34	11.319	0.679
NEAR (FP32)	56.09	41.88	24.737	962.04	25.676	1.817
NEAR (SQ4)	56.38	41.98	3.318	962.04	4.258	1.569
LongMemEval; 476 candidates; 1 timed query						
Base (FP32)	58.96	47.11	4.773	0.03	4.773	0.335
Base (SQ4)	59.07	47.09	0.725	0.03	0.725	0.322
LC	58.96	47.11	4.774	0.04	4.774	0.219
InvGC	91.52	81.05	4.773	0.04	4.773	0.240
DBSN	91.92	80.56	4.774	0.04	4.774	0.280
Search-Adaptor	93.28	82.86	4.773	5132.34	9.785	0.512
NEAR (FP32)	89.67	78.15	9.421	320.82	9.734	0.810
NEAR (SQ4)	89.63	77.97	1.325	320.82	1.639	0.740