MODEL REPORT / RETRIEVAL EVALUATION

Atlas V14 Turn Aware Session Retrieval on LongMemEval S

A LoCoMo trained BGE small encoder with turn maximum session scoring

Abstract. Atlas V14 retrieves the sessions needed to answer long-memory questions by embedding individual conversation turns and ranking each session by its strongest turn-level match. On the cleaned LongMemEval-S non-abstention set, the released session-level `recall_all@5` evaluator records **436/470** (92.8%) all-gold passes. A direct check against every labeled `answer_session_id` records **433/470** (92.1%). The two scores differ because the released evaluator has an empty gold set for 51 questions; restricting it to nonempty gold sets gives **385/419** (91.9%). With the V14 encoder held fixed, turn maximum scoring adds 30 net passes over one concatenated user-session embedding, moving the released result from 406/470 to 436/470. The rule was selected on held-out LoCoMo conversations before the LongMemEval-S evaluation. This report describes the training data, retrieval operator, evaluation definitions, ablation, category results, and remaining failure concentration.

1 Introduction

Long conversation sessions often contain only a few turns relevant to a later question. A single embedding for the whole session must compress relevant and irrelevant content together. Atlas V14 instead represents each user and assistant turn separately, then assigns the session the score of its best matching turn. The retrieved unit remains a session: five distinct session IDs are returned for each question.

The principal result is a per-question all-gold retrieval pass rate, not answer accuracy. On 470 cleaned LongMemEval-S questions that do not ask the system to abstain, the released session evaluator passes 436 questions. A direct test requiring all annotated answer-session IDs in the five returned sessions passes 433. These two definitions are reported together because the released evaluator treats 51 empty gold sets as automatic passes.

The result is most informative as a within-model comparison. Using the same V14 encoder and the released metric, one vector per user session passes 406 questions; turn maximum scoring passes 436. The paired outcomes contain 41 gains and 11 losses. This isolates the change in retrieval representation and aggregation more closely than a cross-model benchmark comparison.

2 Encoder and training

2.1 Model and examples

Atlas V14 begins with `BAAI/bge-small-en-v1.5`, a roughly 33.3-million-parameter encoder with 384-dimensional output vectors. The model has 12 encoder layers. Training updates only the final four layers and leaves the other eight frozen. The resulting weight file is approximately 128 MB.

The training set contains 1,623 mined query/evidence examples from eight LoCoMo conversations. A development set contains 354 examples from two disjoint held-out conversations. Each example includes positive evidence turns and up to 32 hard-negative turns. The query is prefixed with `Represent this sentence for searching relevant passages:`; document turns are not prefixed. Both sides use the final hidden state of the CLS token followed by ℓ_2 normalization.

2.2 Optimization and selection

The training script uses in-batch contrastive cross-entropy with cosine-similarity logits divided by 0.045. Each batch contains 12 questions, one sampled positive per question, and up to three sampled hard negatives per question. AdamW uses a learning rate of 1.5×10^{-5} and weight decay 0.02. Training runs for four epochs with gradient clipping at 1.0. The training tokenizer cap is 160 tokens; the retrieval evaluation uses a 512-token cap.

The saved checkpoint maximizes the sum of development `any@1` and `all@5` on the mined evidence task. Development `all@5` increases from 44.6% before fine-tuning to 64.1% after epoch four. These are within-example evidence-ranking numbers; they are distinct from the session retrieval evaluation below. No LongMemEval examples were used for training or checkpoint selection.

3 Turn maximum session retrieval

For question q, let $e(q)$ be the normalized query embedding and $e(t)$ the normalized embedding of turn t. The turn text includes its session date, speaker role, and content. The score of session S is

$$s(S, q) = \max_{t \in S} \; e(q)^{\top} e(t). \tag{1}$$

Sessions are sorted by $s(S, q)$, and the top five distinct session IDs are returned. Because both vectors are normalized, the

Table 1: Held-out LoCoMo session retrieval rule comparison. All entries use the V14 encoder and 354 development questions.

Session representation	Passes	All-gold@5
One user-session vector	217/354	61.3%
Turn maximum	**292/354**	**82.5%**
Top two turns	290/354	81.9%
Half-weight hybrid	287/354	81.1%

dot product is cosine similarity. Equation 1 lets one localized evidence turn surface the session without requiring the whole conversation to share a single representation.

The retrieval rule was chosen using only the two held-out LoCoMo conversations. Table 1 compares four ways to represent or aggregate sessions with the V14 encoder. Turn maximum yields the highest all-gold top-five rate, 292/354 (82.5%), compared with 217/354 (61.3%) for one concatenated user-session vector. The one empty-gold development question is retained in the reported denominator. The LongMemEval-S check followed this selection.

4 Evaluation protocol

The evaluation uses the official cleaned LongMemEval-S data, identified by the SHA-256 digest in Section 7. The full set has 500 questions. Thirty abstention questions are excluded, leaving $N = 470$ retrieval questions. For each question, five distinct sessions are selected without using benchmark labels during ranking.

Let $R_5(q)$ denote the returned session IDs, $G_{\text{rel}}(q)$ the released evaluator's gold session IDs, and $G_{\text{dir}}(q)$ the annotated `answer_session_ids`. For either gold definition G, the reported rate is

$$\text{all@5}(G) = \frac{1}{N}\sum_{q=1}^{N} \mathbf{1}[G(q) \subseteq R_5(q)]\,. \tag{2}$$

This is the fraction of questions for which *every* gold session appears in the top five. It is not mean fractional recall. The released evaluator constructs gold sessions from user turns marked `has_answer`; its gold set is empty for 51 of the 470 evaluated questions. An empty set satisfies Eq. 2 automatically. The direct check uses `answer_session_ids` instead. A third view restricts the released metric to its 419 nonempty-gold questions.

5 Results

5.1 Primary result and retrieval ablation

Table 2 separates the three interpretations of the turn maximum result. On the released metric, V14 passes 436/470 (92.8%). Direct answer-session coverage is 433/470 (92.1%); on nonempty released gold sets it is 385/419 (91.9%). The 0.7-point difference between the two full-set metrics is a consequence of their gold definitions, not a change in retrieval outputs.

Table 2: Cleaned LongMemEval-S session retrieval. All rates are per-question all-gold top-five pass rates. The Qwen run is a separate complete system.

System and gold definition	Passes	Rate
V14 turn maximum, released	**436/470**	**92.8%**
V14 turn maximum, direct IDs	433/470	92.1%
V14 turn maximum, nonempty released	385/419	91.9%
V14 one user-session vector, released	406/470	86.4%
Base BGE one user-session vector, released	405/470	86.2%
Qwen3-Embedding-8B Q4, released	413/470	87.9%
Qwen3-Embedding-8B Q4, direct IDs	343/470	73.0%

Table 3: Direct-gold all@5 by LongMemEval-S question type.

Question type	Passes	Rate
Single-session user	64/64	100.0%
Single-session preference	30/30	100.0%
Single-session assistant	56/56	100.0%
Knowledge update	71/72	98.6%
Multi-session	107/121	88.4%
Temporal reasoning	105/127	82.7%

Holding the V14 encoder fixed, the turn maximum layout improves the released metric by 30/470 questions, or 6.4 percentage points, relative to a single user-session vector. The paired contingency is 395 passes by both layouts, 41 passes unique to turn maximum, 11 passes unique to the single-vector layout, and 23 failures by both. The asymmetry indicates that the gain is not merely a reshuffling of equally many solved and unsolved questions.

The Qwen3-Embedding-8B Q4 run uses one user-only vector per session, its own query instruction, last-token pooling, and an 8,190-token server cap. Its 413/470 released score and 343/470 direct-gold score are useful system references, but they do not isolate model size or encoder quality: representation, query formatting, pooling, and context length also differ.

5.2 Performance by question type

Table 3 uses the direct answer-session IDs so that all six rows share a nonvacuous definition. V14 recovers every labeled session for all 150 single-session user, preference, and assistant questions. It passes 71/72 knowledge-update questions. Multi-session and temporal-reasoning questions account for almost all remaining errors.

The direct check has 37 misses. Twenty-two are temporal-reasoning questions, 14 are multi-session questions, and one is a knowledge-update question. Thus 36/37 misses (97.3%) fall in the two question types that most often require order-

ing events or assembling evidence across sessions. This concentration describes the observed errors; it does not establish whether the missing evidence was poorly encoded, outranked by distractors, or absent from the five selected sessions for another reason.

5.3 Turn ranking diagnostics

The stored turn-level ranking provides a separate view of how evidence accumulates with retrieval depth. Table 4 asks whether the top K individual turns contain any gold turn, every gold turn, or at least one turn from every labeled gold session. These are diagnostics over *turns*, not top-K distinct-session `recall_all` scores. Their retrieval budget therefore differs from the five-session result in Table 2.

Table 4: Turn-ranked evidence coverage on 470 non-abstention questions. Each column uses the top K individual turns.

K turns	Any gold turn	All gold turns	All gold sessions
1	60.4%	21.1%	31.3%
5	88.5%	64.0%	78.5%
10	95.1%	79.4%	91.1%
20	97.4%	87.7%	96.0%

At five turns, at least one gold turn appears for 416/470 questions, but all gold turns appear for only 301/470. At 20 turns those counts rise to 458 and 412. The gap is consistent with evidence being distributed across multiple turns; it also shows why a hit on one relevant turn should not be read as complete evidence coverage. Scoring sessions by their best turn and then returning distinct IDs targets a different budget: it preserves the localized match while giving each returned session one place in the final set.

6 Interpretation and limitations

The fixed-encoder comparison supports a narrow conclusion: turn-level indexing and maximum aggregation improve session coverage under this evaluation. Fine-tuning alone is less clearly implicated by the one-vector comparison: V14 passes 406/470 and the base BGE model passes 405/470 under that layout. The large improvement appears when V14 is paired with turn maximum retrieval, although this report does not include a base-BGE turn maximum run that would fully separate training and aggregation effects.

The LoCoMo train/development split is conversation-disjoint. A normalized exact-match audit found no identical LoCoMo and LongMemEval-S questions or gold turns. Exact matching cannot rule out semantic overlap. The public LongMemEval-S test and its failures have now been inspected; any later tuning against those failures should be treated as exploratory. An earlier V13 system used LongMemEval-S examples in training and is excluded from the clean comparison.

These experiments measure evidence-session retrieval only. They do not measure generated-answer correctness, abstention behavior, or latency. Turn-level top-K scores and gold-session coverage within the top-K *turns* use a different retrieval unit from the five-distinct-session metric reported here. The strongest next analysis would inspect temporal and multi-session misses at the question level, then validate any revised rule on a fresh held-out benchmark.

7 Reproducibility

The following SHA-256 digests identify the evaluated inputs. The exact checkpoint and cleaned benchmark JSON remain on the originating machine.

LongMemEval-S data
`d6f21ea9d60a0d56f34a05b609c79c88a451d2ae03597821ea3d5a9678c3a442`

Atlas V14 weights
`997869f33cbfc7c2a04916cd8957d4ce160540299a07653ea591d47a07445b89`

LoCoMo source
`79fa87e90f04081343b8c8debecb80a9a6842b76a7aa537dc9fdf651ea698ff4`

The accompanying packet includes the turn maximum summary and per-question outcomes, category and paired analyses, the held-out LoCoMo method comparison, and training and evaluation scripts under `methods/`.

Source repositories. LongMemEval: https://github.com/xiaowu0162/LongMemEval; LoCoMo: https://github.com/snap-research/locomo; BGE-small: https://huggingface.co/BAAI/bge-small-en-v1.5; Qwen3-Embedding-8B Q4: https://huggingface.co/Qwen/Qwen3-Embedding-8B-GGUF.