Research Article

Coupled Spiking Memory for Budgeted Long-Dialogue Question Answering with Small Language Models

Yiwen Wang[1], **Jiahao Qin**[2]

1. Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, China; 2. School of Psychology, Shanghai Jiao Tong University, Shanghai, China

Longitudinal conversational assistance requires access to information that is scattered across sessions and changes over time. A small language model can retrieve this information from an external log, but independently ranking each record ignores the sequence in which the record was produced. We present Coupled Spiking Memory (CSM), an evidence-supervised log controller that augments each record with a causal temporal trace before query-dependent retrieval. CSM adapts two-compartment spiking dynamics, reads the post-reset somatic voltage, and learns a residual correction to lexical relevance while leaving the language model frozen. Whole-record selection restricts the retrieved context to a fixed token budget and preserves chronological presentation. With a frozen 3B reader, CSM obtains 30.33 mean QA F1 on 192 EvoEmo questions across five training seeds, exceeding BM25, BGE-M3, and a matched GRU controller by 0.31, 1.28, and 0.76 points, respectively. It uses 123,907 trainable controller parameters. Frozen transfer to LoCoMo and a three-budget analysis reveal that the benefit depends on the readout, evidence budget, and balance between answering and abstention. These results establish a compact spiking-controller approach to conversational evidence retrieval and identify the conditions that limit its transfer.

Corresponding author: Jiahao Qin, jiahao.qin19@gmail.com

1. Introduction

Persistent conversational assistance depends on remembering what has happened to a user, not merely recognizing the topic of the current utterance. A person may describe the same relationship, work situation, or source of distress over several sessions while changing their interpretation of it. Answering

a later question requires access to these earlier disclosures and their temporal context. This requirement is particularly relevant to personalized emotional support, where dispersed information and evolving circumstances are explicit components of the ES-MemEval benchmark [1]. It also appears in more general long-conversation question answering, as studied by LoCoMo [2]. For a locally deployed small language model, external memory offers a way to retain interaction-specific information without repeatedly updating the language model itself.

Multimodal affective computing addresses a complementary part of this problem: interpreting evidence within an interaction. Bi-stream graph learning separates within-modality context from cross-modal interaction [3], while transformer-based self-distillation learns multimodal conversational representations [4]. BC-PMJRS organizes cross-modal representations through information-based objectives [5], and Möbius Interaction Filtering regulates the contribution of different modality-interaction orders to humor and sentiment predictions [6]. Persistent assistance additionally requires access to earlier events and disclosures. This motivates studying historical evidence access as a computational component alongside affective perception, with the present work focusing on retrieval from textual conversation records.

Retrieval-augmented generation separates information storage from language generation [7]. This separation is attractive when the generator must remain fixed, but it introduces a second problem: deciding which records should occupy its limited reading context. Lexical retrieval favors explicit term overlap, while dense retrieval favors embedding similarity. Neither score, by itself, explicitly contextualizes a record using the sequence of preceding records. Two statements can be similarly relevant to a question even though one introduces a situation and the other follows a later revision. A temporal representation can make that distinction available to the retrieval module before the selected text reaches the language model.

We investigate a small recurrent controller attached to an external conversational log. Its role is to make stored evidence easier to use, rather than to replace the reader's language capabilities or rewrite the log into generated summaries. Each record retains its original text and receives a compact, causally computed temporal trace. A query-conditioned scorer combines that trace with the record representation, lexical relevance, and semantic similarity. The resulting score determines which complete records fit into the reading budget. The language model then answers from the selected records, presented in chronological order.

Spiking dynamics provide a useful design space for this controller because accumulation, thresholding, and reset separate persistent state from discrete events. In particular, two coupled compartments allow an input-driven dendritic state and a somatic state to influence one another. We adapt the established TC-LIF dynamics [8] to conversational-log processing and use the continuous post-reset somatic voltage for retrieval. This choice retains graded state information between firing events. The resulting framework, Coupled Spiking Memory (CSM), combines a recurrent temporal representation with an explicitly preserved lexical retrieval path. It is trained from supporting-record annotations, while both the record encoder and the answer-generating model remain frozen.

The central research question is whether this compact temporal controller improves evidence-conditioned answers under a fixed reading budget. We address it with user-disjoint EvoEmo partitions, matched recurrent controls, frozen cross-dataset transfer, and explicit changes to the context budget. The complete five-seed EvoEmo comparison shows a positive average advantage over BM25, BGE-M3, and GRU. The transfer experiment is more selective: a simpler voltage-readout LIF is stronger than CSM on LoCoMo, and larger evidence budgets improve answerable questions without necessarily improving the aggregate score. This variation motivates a distinction between retrieving relevant records, providing usable evidence, and eliciting an answer from that evidence.

The contributions are threefold:

1. We formulate a coupled-spiking log controller for a frozen small reader, integrating causal record traces, residual relevance scoring, and chronological token-budgeted retrieval.
2. We specify how the adopted two-compartment recurrence and its continuous readout interact with the retrieval objective, and compare them with matched GRU, spike-readout LIF, and voltage-readout LIF controllers.
3. We provide a multi-seed evaluation on longitudinal emotional-support logs, frozen transfer to general long conversations, and budget-dependent evidence analysis that separates overall QA, answerable questions, abstention, and supporting-record recall.

2. Related Work

2.1. Multimodal affective representation and evidence reliability

Multimodal representation learning studies how heterogeneous observations contribute to a common prediction. The Alternative Telescopic Displacement approach in *Zoom and Shift are All You Need* aligns

modality representations through learned transformations [9]. BC-PMJRS uses mutual-information objectives to balance cross-modal redundancy and sentiment relevance [5]. HAEMSA organizes modality-specific and shared experts through hierarchical learning and cross-modal knowledge transfer [10]. MoIF instead regulates the prediction contribution of interaction orders through a fixed operator over modality-subset estimates [6]. These designs organize information across modalities; CSM organizes access to records across a user's history.

Reliability is also studied through uncertainty modeling and missing-data recovery. DUAL models feature uncertainty and uncertain cross-modal relationships [11]. MTLP-MDG addresses missing-data generation within a multi-task learning framework [12], while DCER combines within- and cross-modality compression with energy-based reconstruction [13]. These approaches operate on the representations available to a predictor. The conversational-memory setting poses a related selection problem: relevant information may already exist in the log but remain outside the reader's context budget. CSM retains the original records and learns a temporal representation for choosing which of them to expose to a frozen reader.

Graph-based conversational fusion further distinguishes useful interactions from redundant context. GraphCFC uses directed contextual relations and pairwise cross-modal complementation [14]. Structure Aware Multi-Graph Networks construct modality-specific graphs and learn which utterance connections to retain [15]. These approaches make context structure part of representation learning. CSM applies a related information-selection principle at the interface to the reader: the learned representation determines which original records occupy the available text budget.

2.2. Conversational context and emotional dynamics

Context modeling in affective dialogue involves both dependencies between utterances and changes in the speaker's state. Quan et al. study multi-party dependencies through dialog-oriented memory and directed acyclic graph representations [16]. CFN-ESA explicitly models emotion shifts alongside cross-modal context [17]. Speaker identity is another modeling choice: progressive contrastive deep supervision jointly addresses speaker diarization and emotion recognition when speaker labels are unavailable at inference [18]. Personality-aware ERC combines inter-speaker differences with within-speaker changes [19], whereas contrastive adversarial tuning targets the discriminability and robustness of LLM-

based conversational emotion representations [20]. Together, these studies motivate preserving relational and evolving context rather than treating conversational observations as independent items.

Response-generation systems also organize context beyond isolated utterances. ChatMDG fuses discourse structure and role interactions with semantic states for multi-party generation [21]. Multi-level semantic encoding and gated extraction of context-relevant knowledge have also been studied for emotional-support conversation [22]. CSM addresses an earlier access decision in this computation: which historical records should be supplied to an unchanged reader. Its supervision concerns supporting records, and its temporal trace is used to rank evidence rather than to predict emotion labels or generate a support strategy.

2.3. External memory for conversational language models

Conversational memory systems differ in how they construct and revisit stored information. MemoryBank maintains and updates long-term memories with a time- and importance-sensitive mechanism [23]. A-MEM organizes generated notes through contextual attributes and links, allowing newly stored material to affect existing memories [24]. ReadAgent compresses document episodes into gist memories and revisits original passages when needed [25]. These approaches demonstrate several ways to separate long-term storage from the immediate context presented to a language model.

Evidence-grounded generation also benefits from organizing retrieval around the reader's task. OpenScholar combines a scientific literature store, learned retrieval and reranking, and iterative feedback to generate citation-supported answers [26]. Its scientific-synthesis setting illustrates how external evidence access can be developed as a dedicated computational component. CSM studies that component for chronological interaction logs, with a fixed reader and a compact learned controller.

CSM makes a different implementation choice: it retains original records and learns their retrieval scores through a small recurrent module. It does not require a language-model call to write a summary or generate a link for each record. This distinction concerns the computation performed by the memory system, rather than a general claim that learned record ranking is preferable to generated memory organization.

Selective use of history also appears in embodied tasks. Memory-Adaptive Vision-and-Language Navigation filters detrimental components of stored histories using current visual and textual information [27]. The shared design question is how to retain useful history without passing all historical

content to the decision module. Here, selection operates over complete textual records under a reader-token budget, with chronological state available as a retrieval feature.

2.4. Lexical, dense, and hybrid retrieval

BM25 models term relevance with document-length normalization [28]. BGE-M3 supports dense, sparse, and multi-vector retrieval within a multilingual embedding framework [29]; our comparison uses its normalized dense embeddings. Reciprocal rank fusion combines rankings without requiring comparable raw scores [30]. We use BM25 and BGE-M3 as individual retrieval comparators and report their fixed-RRF combination as an additional control.

The recurrent controller does not replace these retrieval principles with a new similarity metric. Instead, it learns how a record's temporal representation should adjust an explicit lexical score. The distinction matters for interpreting results: a gain over a lexical retriever can arise from additional semantic features, learned relevance supervision, temporal processing, or their interaction. Matched GRU and LIF implementations keep the surrounding feature and scoring pipeline fixed to make the recurrent design comparison more informative.

Contextualized retrieval is also explored beyond text logs. Context-Enhanced Video Moment Retrieval uses LLM-derived contextual semantics and language-conditioned decoding to locate moments within a video [31]. This neighboring setting illustrates the distinction between representing a candidate in context and matching it to a query. CSM implements that distinction through a causal record trace and a separate query-conditioned scorer, while retrieving original conversation text.

2.5. Recurrent and spiking sequence representations

Gated recurrent units provide a conventional mechanism for maintaining a learned sequence state [32]. Spiking networks instead combine state accumulation with threshold-triggered events, and surrogate gradients allow their discontinuous firing operations to participate in gradient-based learning [33]. TC-LIF introduces coupled dendritic and somatic states for long sequential processing [8]. We use that neuron design as an existing computational component, rather than presenting the two-compartment recurrence itself as a new invention.

Recent spiking models provide several ways to control temporal information. Nonlinear dendritic computation introduces spatial gating and temporal input adjustment [34]; delay learning makes spike

timing itself trainable [35]; and MSFI combines short- and long-timescale spatiotemporal features [36]. Training and decoding are separate design axes: masked surrogate gradients with temporally weighted outputs address gradient sparsity and timestep contributions [37], while temporal-consistency training constrains disagreement between outputs at different timesteps [38]. These studies motivate examining the state dynamics, readout, and supervision separately when adapting a spiking component to a new sequence task.

Within-neuron organization offers another axis of model design. Chavlis and Poirazi use structured dendritic connectivity and restricted input sampling to improve parameter efficiency and robustness to overfitting in image classification [39]. Sun et al. combine fast spiking activity with a compact, explicit slow-memory pathway for long-sequence processing [40]. These approaches organize information within computational units and across timescales, providing a broader context for using compartmental states as sequence representations.

The numerical realization of those states also matters. Baronig et al. analyze adaptive LIF dynamics and show that symplectic Euler discretization addresses stability and parameterization problems associated with forward Euler updates [41]. This connection between state equations and their discrete implementation motivates specifying the update order, reset, and readout explicitly when introducing spiking dynamics into a retrieval pipeline.

Our adaptation concerns its placement in a log-retrieval pipeline and the information exposed to the query scorer. The spiking state contextualizes each record; the original record representation remains available through a separate path. Continuous somatic readout additionally preserves subthreshold information that a binary spike vector would discard. The experiments compare these readout choices under the same task supervision and distinguish a useful representation from a demonstrated advantage on temporal reasoning questions.

3. Method

Figure 1 summarizes the computation. CSM first contextualizes records in chronological order, then uses the question to score the resulting record keys. The retained raw log supplies the selected evidence to the frozen reader; the recurrent trace guides access to that evidence rather than replacing its text.

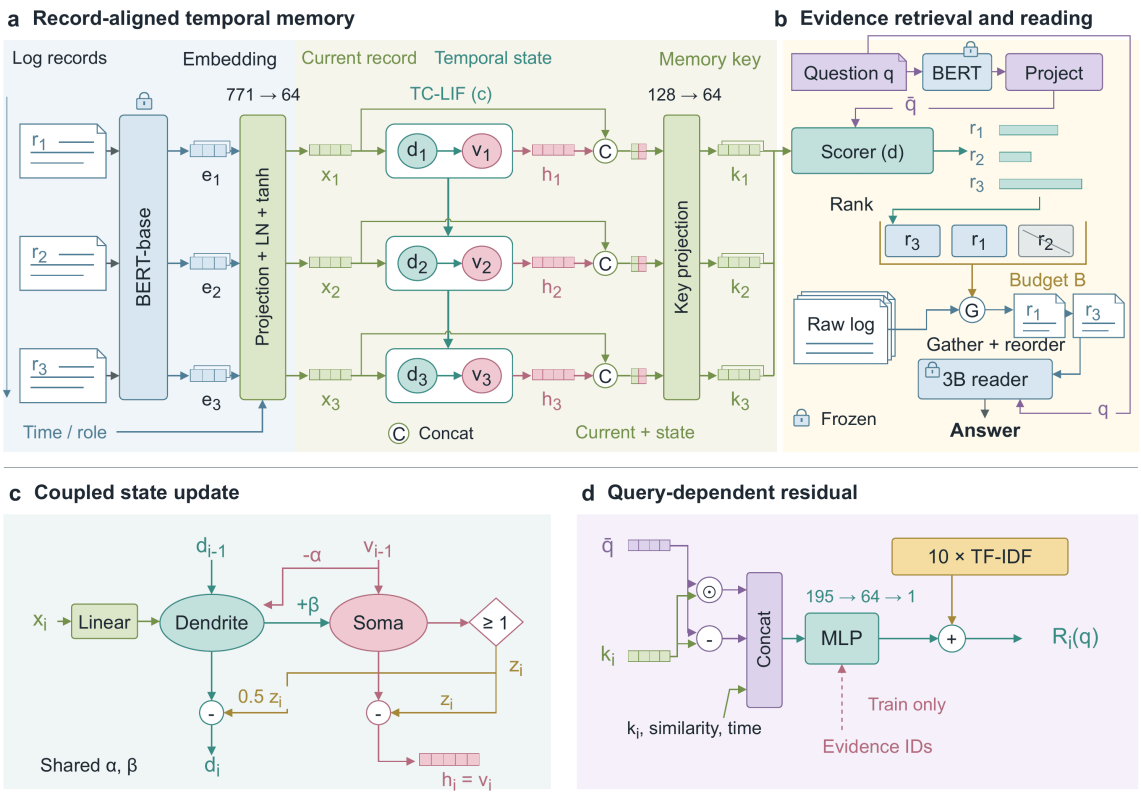


Figure 1. Architecture of Coupled Spiking Memory (CSM). (a) Each row traces one record from its frozen embedding to the immediate feature x_i, recurrent voltage h_i, and retrieval key k_i. Vertical state propagation follows the log chronology; the bypass retains the current record alongside its temporal context. (b) The question, encoded with the same frozen BERT, scores these keys. Whole-record admission under budget B selects identifiers used to gather original text from the same raw log shown in (a). Admitted text is reordered chronologically and passed with the question to the frozen 3B reader. (c) The adopted TC-LIF cell couples the previous soma to the dendrite and the updated dendrite to the soma, with distinct spike-triggered resets and post-reset voltage readout. (d) Query–key interactions, similarity and time features produce a learned residual added to lexical relevance. Locks mark frozen modules; the dashed arrow denotes training-only evidence supervision. The three-record selection and feature glyphs are schematic, not measured activations or an experimental example.

3.1. Problem formulation

Let a user's available history be an ordered log $\mathscr{H} = \{r_i\}_{i=1}^{N}$. Each record contains text c_i, a date t_i, a speaker role ρ_i, and a unique evidence identifier. For a question q, the memory system selects a subset $S_q \subseteq \{1, \ldots, N\}$ and serializes its records in chronological order. The retrieved text C_q must satisfy

$$|\mathrm{Tok}(C_q)| \leq B, \qquad \hat{a} = G_\phi(q, C_q), \tag{1}$$

where B is the log-context token budget and G_ϕ is the frozen reader. The question and instruction tokens are additional to this log budget. Controller parameters θ are learned using training questions and their supporting-record identifiers. Test answers and evidence annotations are used only for scoring.

The method separates three objects: the persistent raw log, the temporal states attached to its records, and the limited evidence context supplied for a particular question. A fixed recurrent-state dimension therefore does not imply fixed total storage. The query is answered from the raw selected text, not directly from a decoded neural state.

3.2. Record representation and temporal features

A frozen BERT-base encoder [42] maps record contents and questions into normalized 768-dimensional vectors, denoted e_i and e_q. We use attention-mask-weighted mean pooling over encoder outputs, with a maximum of 256 input tokens per record or question. The full record text remains in the external log even when its encoder input is truncated.

We add three scalar features to each record: elapsed days from the start of the history divided by 365, the gap from the preceding record divided by 30, and a binary indicator of the seeker role. The first feature is elapsed time since the beginning, not distance from the end of the history. Writing these features as η_i, the projected representation is

$$x_i = \tanh(\mathrm{LN}\left(W_p[e_i;\eta_i] + b_p\right)), \qquad x_i \in \mathbb{R}^{64}. \tag{2}$$

The same representation and the same surrounding retrieval network are used by the recurrent controls. The recurrent index corresponds to chronological records; calendar information enters through η_i.

3.3. Coupled spiking temporal traces

CSM maintains a dendritic state d_i and a somatic state v_i, each of width 64. Both start at zero for a new history. Following TC-LIF [8], two shared learnable coefficients are constrained through sigmoid functions, $\alpha = \sigma(a)$ and $\beta = \sigma(b)$. They are initialized at 0.5 and shared across the neurons. With $u_i = W_d x_i + b_d$, the update is

$$\tilde{d}_i = d_{i-1} - \alpha v_{i-1} + u_i, \tag{3}$$
$$\tilde{v}_i = v_{i-1} + \beta \tilde{d}_i, \tag{4}$$
$$z_i = H(\tilde{v}_i - 1), \tag{5}$$
$$d_i = \tilde{d}_i - 0.5 z_i, \qquad v_i = \tilde{v}_i - z_i. \tag{6}$$

All operations are elementwise except the input projection. The dendritic state receives the current record and somatic feedback; the soma integrates dendritic input and determines firing. Both compartments undergo a soft reset, with different reset magnitudes.

The retrieval trace is the post-reset somatic voltage, $h_i = v_i$. Thus, the controller uses spiking dynamics but exposes graded state information to the scorer. During backpropagation, the derivative of the step function is replaced by $1/(1+\pi^2 x^2)$, where $x = \tilde{v}_i - 1$. Surrogate gradients pass through both reset terms. A single chronological pass produces the record traces before individual questions are scored.

Subthreshold dynamics.

The coupling gives a concrete alternative to independently decaying state coordinates. With zero input and no firing, the transition of one neuron is

$$\begin{bmatrix} d_i \\ v_i \end{bmatrix} = M \begin{bmatrix} d_{i-1} \\ v_{i-1} \end{bmatrix}, \qquad M = \begin{bmatrix} 1 & -\alpha \\ \beta & 1-\alpha\beta \end{bmatrix}. \tag{7}$$

Its determinant is one and its characteristic polynomial is

$$\lambda^2 - (2 - \alpha\beta)\lambda + 1 = 0. \tag{8}$$

For $0 < \alpha, \beta < 1$, the two eigenvalues are complex conjugates of unit modulus. The no-input, non-firing transition therefore exchanges state between compartments rather than multiplying each coordinate by a contractive leak. This local calculation explains a representational distinction from leaky integration. It is not a stability guarantee for the input-driven, firing, surrogate-gradient system, whose reset events alter the transition.

3.4. Residual evidence scoring

The scorer combines a record's immediate representation with its causal temporal trace:

$$k_i = \mathrm{Normalize}\left(\mathrm{LN}(W_k[x_i; h_i] + b_k)\right), \qquad \bar{q} = \mathrm{Normalize}(W_q e_q + b_q). \tag{9}$$

Let $\ell_i(q)$ be TF-IDF cosine similarity between the question and the serialized record. The vectorizer uses unigrams and bigrams with sublinear term frequency, fitted independently on each history. We additionally compute the frozen-encoder similarity $s_i = e_q^\top e_i$. The scoring feature vector is

$$f_i(q) = [\bar{q} \odot k_i; \bar{q} - k_i; k_i; \ell_i(q); s_i; \eta_{i,1}]. \tag{10}$$

A two-layer perceptron with a 64-dimensional hidden layer supplies a residual:

$$R_i(q) = 10\ell_i(q) + w_2^\top \tanh\left(W_1 f_i(q) + b_1\right) + b_2. \tag{11}$$

The final layer is initialized to zero, so the initial ordering is exactly lexical retrieval. Training then learns a correction from semantic and temporal features. The lexical term remains a direct part of the final decision rather than being absorbed into the recurrent state.

3.5. Evidence supervision and budgeted reading

For a training question with nonempty supporting-record set E_q, define the target distribution $y_{qi} = 1[i \in E_q]/|E_q|$. The loss for that question is

$$\mathscr{L}_q = -\sum_i y_{qi} \log \frac{\exp R_i(q)}{\sum_j \exp R_j(q)}. \tag{12}$$

Questions without annotated supporting records are omitted from this loss. An update averages the loss over eligible questions from one training user. No reference-answer generation loss or language-model fine-tuning is used in the main CSM experiment. The checkpoint is selected by development-set user-macro evidence Recall@10.

At inference, records are visited in descending score order, with chronological index breaking ties. A record is admitted only if its estimated cost fits the remaining allowance and the re-tokenized chronological concatenation also fits B. Oversized records are skipped rather than shortened. The accepted records are then placed back in chronological order. This is a greedy feasibility procedure, not an exact solution to a knapsack objective. It preserves original evidence and separates relevance-based admission from temporal presentation.

3.6. Matched controllers and computational scope

The GRU controller replaces only the temporal module with a width-64 GRU. The spike-readout LIF uses half its coordinates with time constant 2 and half with time constant 8. Its decay depends on the inter-record gap, and its output is the firing vector. The voltage-readout LIF retains the same dynamics and initialized parameters but returns its post-reset voltage. Both single-compartment versions detach the spike in the reset operation, unlike the adopted TC-LIF implementation. Consequently, CSM versus spike LIF changes coupling, readout, and reset-gradient handling, whereas voltage versus spike LIF isolates the readout choice.

For fixed feature dimensions, temporal processing is linear in the number of records, with dense width-dependent projection costs. The current implementation retains a 768-dimensional embedding and a 64-dimensional trace for each record, in addition to raw text. Query-dependent scoring still visits stored

records. The architecture can reuse causal prefix states when a log is extended, but the experiments evaluate complete histories and do not benchmark an online append service.

4. Experimental Setup

4.1. Datasets and partitions

EvoEmo

We use the public EvoEmo release associated with ES-MemEval [1]. It represents longitudinal emotional-support conversations through simulated user trajectories rather than a prospective study of real users. We partition its 18 users into eight training users, four development users, and six held-out users. The controller-training set contains 551 questions, development contains 128, and the primary evaluation contains 192, with 32 questions per held-out user.

LoCoMo

The transfer experiment uses all 1,986 questions from the ten-conversation public release [2]. Controllers trained on EvoEmo are applied without LoCoMo training or checkpoint selection. We use text turns and retain dates and speaker names in the serialized evidence, without adding image captions or generated observations. Since LoCoMo does not use the seeker/supporter role convention, both speakers receive the seeker indicator; their distinct names remain visible in the retrieved text. This mapping is shared by the neural controls.

4.2. Comparators and reader configuration

All main comparisons use the same frozen Qwen2.5-Omni-3B Thinker and tokenizer [43]. Only text is supplied. Decoding is greedy with at most 80 generated tokens, BF16 parameters, SDPA attention, and TF32 disabled. The default memory budget is 2,048 tokens, with 512- and 1,024-token evaluations using the same checkpoints. The EvoEmo instruction asks for a concise answer from the supplied history and the literal `unknown` when evidence is insufficient. LoCoMo uses the same instruction with `not mentioned` as the refusal literal to match its scoring convention.

The individual retrieval methods are TF-IDF, BM25, and BGE-M3. BM25 uses lowercased word tokens and the BM25Okapi defaults $k_1 = 1.5$, $b = 0.75$, and $\epsilon = 0.25$. BGE-M3 uses normalized dense CLS

embeddings and a maximum encoder length of 8,192 tokens. The trained controller comparisons share frozen BERT features, evidence annotations, projection width, residual scorer, record packing, and reader. These controls compare temporal modules within our framework, not reproductions of separately published end-to-end conversational systems.

4.3. Training and reporting

Controllers are optimized on CPU using cached frozen features for 24 epochs with AdamW, learning rate 10^{-3}, weight decay 0.01, and gradient-norm clipping at 1. Reader inference uses an RTX 5090. Each epoch visits the eight training users in a seeded random order and performs one update per user. Models are selected using development evidence Recall@10, with earlier checkpoints retained on ties. The primary CSM, GRU, and spike-LIF comparison includes every completed seed: 7, 22, 42, 93, and 99.

The later budget, continuous-readout, and transfer experiments use seeds 7, 22, and 99. This subset was selected after inspecting EvoEmo outcomes and fixed before the LoCoMo transfer; it is not treated as a prospectively selected three-seed confirmation of EvoEmo performance. The full five-seed primary results and all paired differences are retained. These are exploratory model-development experiments, with held-out users excluded from parameter fitting, rather than a preregistered final evaluation. Reported SDs use the sample denominator $n - 1$; deterministic retrievers are not assigned an artificial seed variance.

4.4. Metrics and output verification

EvoEmo QA uses lowercase word-multiset overlap F1, averaged equally over users. This local reproducible metric is not the repository's alternative set-numerator implementation. On the balanced primary subset, user-macro and question-micro scores coincide. We also report question-micro F1 for answerable questions and for temporal/conflict questions, evidence Recall@10, and the fraction of annotated evidence included in the selected context.

LoCoMo uses the release's category-aware scoring functions unchanged, including its abstention rule for category 5. We report the conversation-macro aggregate, question-micro scores within categories, and answerable categories 1–4 separately. EvoEmo and LoCoMo scores are never pooled into one overall accuracy. Raw predictions, prompt-token hashes, record selections, and checkpoint hashes are retained. Exact prompt matches can reuse saved answers from the same frozen reader configuration; reused predictions are not independent repetitions. Outputs reaching the token cap remain in the evaluation.

5. Results and Analysis

5.1. Primary evidence on longitudinal support logs

Method	QA F1	Answerable	Temp./conflict	Context recall
TF-IDF	28.33	22.04	12.57	45.02
BM25	30.02	22.84	12.20	48.80
BGE-M3	29.04	**24.37**	**12.77**	56.49
GRU controller	29.56 $\pm$ 0.18	23.08 $\pm$ 0.37	12.30 $\pm$ 0.51	**57.86 $\pm$ 1.07**
Spike-readout LIF	29.19 $\pm$ 0.38	22.50 $\pm$ 0.16	11.71 $\pm$ 0.48	56.96 $\pm$ 1.91
CSM (coupled voltage)	**30.33 $\pm$ 0.50**	23.77 $\pm$ 0.43	12.52 $\pm$ 0.31	57.82 $\pm$ 0.51

Table 1. EvoEmo primary evaluation: 192 questions from six held-out users, 2,048 log tokens. Learned controllers use all five seeds (7, 22, 42, 93, 99); values are mean $\pm$ sample SD. Deterministic retrievers have no training-seed SD. All values are percentages; bold marks the largest displayed mean in each column.

Table 1 shows that the coupled controller improves the aggregate EvoEmo QA score relative to each individual retrieval method and both matched recurrent controls. Its five-seed mean exceeds GRU by 0.76 points and spike LIF by 1.13 points. The gain over BM25 is smaller, at 0.31 points, while the difference from BGE-M3 is 1.28 points. Thus, the strongest evidence is a positive mean improvement within the specified small-reader setting, rather than a large separation from every retrieval alternative.

The component metrics help locate this improvement. Relative to GRU, CSM increases answerable QA while producing almost the same selected-context evidence recall. More annotated evidence in the context is therefore not sufficient to explain the full answer-quality difference. Relative to BGE-M3, CSM has higher overall QA but lower answerable F1. The aggregate advantage does not imply that CSM is better whenever the history contains an answer. These comparisons motivate examining the reader's use of evidence and its decision to abstain alongside retrieval recall.

Seed	GRU	Spike LIF	CSM	Δ vs. GRU
7	29.43	29.19	31.00	+1.57
22	29.65	28.59	30.41	+0.75
42	29.68	29.61	30.43	+0.75
93	29.74	29.39	29.62	−0.11
99	29.31	29.20	30.16	+0.85

Table 2. All five paired EvoEmo seeds at 2,048 tokens. The final column is the signed CSM–GRU difference in percentage points.

The paired results in Table 2 retain the one negative CSM–GRU comparison. CSM improves on GRU in four of five seeds and on the spike-readout LIF in all five. This consistency supports the observed average effect but does not establish statistical significance across users or new datasets. A no-memory diagnostic gives 18.23 aggregate F1, with zero scores on the answerable capability groups and a high abstention score. The difference from that diagnostic primarily demonstrates the value of external evidence; the recurrent comparisons are the more discriminating tests of the controller.

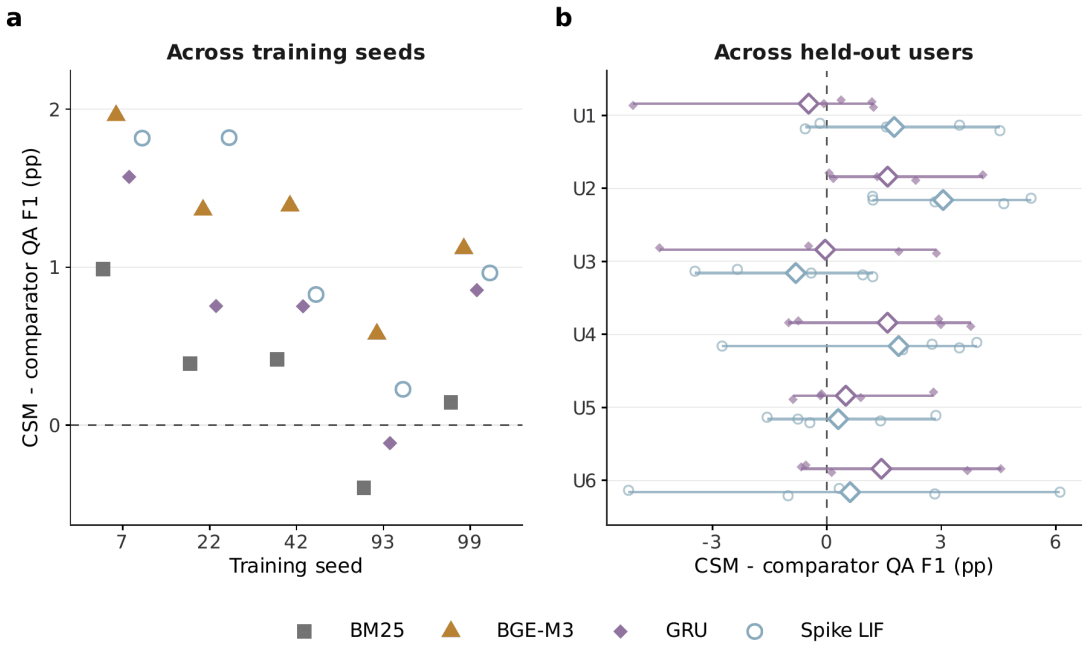


Figure 2. Paired variation in the primary EvoEmo comparison at a 2,048-token log budget. Positive differences favor CSM. (a) User-macro QA F1 differences for all five training seeds against each comparator. (b) User-level differences against the matched recurrent controls: small points retain all five seeds, open diamonds indicate their means, and horizontal segments span the minimum and maximum, not a confidence interval. Each user contributes the same 32 questions. U1–U6 map to p8, p9, p10, p11, p13, and p16, respectively. Deterministic BM25 and BGE-M3 predictions are reused across the five CSM comparisons.

Figure 2 separates variation across training initializations from variation across users. The five-seed mean improvement over GRU occurs for four of the six held-out users, while the comparison with spike-readout LIF is positive for five. The user-level ranges show where initialization changes the size or sign of the advantage. This view locates the aggregate improvement in individual histories and reveals which user comparisons account for the remaining variability.

5.2. Evidence coverage and answer changes

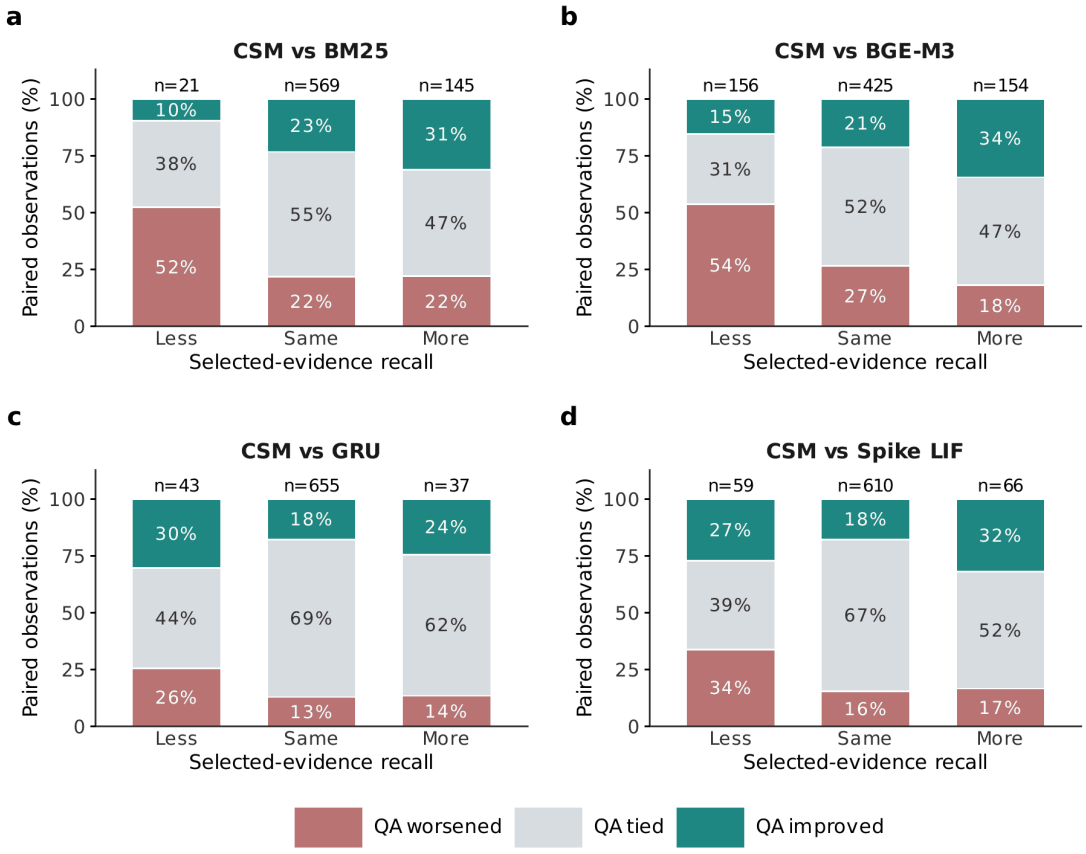


Figure 3. Relationship between evidence access and answer quality on EvoEmo. For each comparator, paired CSM predictions are grouped by lower, unchanged, or higher selected-context evidence recall. Stacked bars show the fractions with lower, tied, or higher answer F1; labels inside bars are rounded percentages. The analysis includes all 147 answerable questions with nonempty evidence annotations at each of five seeds, giving 735 seed–question pairs per comparator. The other nine answerable questions lack an annotated evidence set, and abstention questions are excluded from this analysis; all remain in the primary QA comparison. Counts above bars describe repeated paired observations, not independent questions.

Figure 3 examines the relationship at the level of individual questions. Against both lexical and dense retrieval, higher evidence recall is accompanied by a larger share of improved answers than lower recall. The matched recurrent comparisons are dominated by unchanged coverage, yet this group still contains both improved and degraded answers. Equal coverage can accompany different selected records, competing details, and surrounding context. CSM's evidence access should therefore be understood as

selecting a readable collection of records, whose value is only partly captured by the fraction of annotated evidence retrieved.

5.3. Frozen transfer to general long conversations

Method	QA score	Answerable	Temporal	Abstention
TF-IDF	32.12	37.23	12.99	15.47
BM25	32.44	37.86	13.73	14.35
BGE-M3	**33.96**	**40.43**	**15.17**	12.11
GRU controller	32.70 ± 0.42	37.38 ± 0.89	13.90 ± 1.16	17.86 ± 1.91
Spike-readout LIF	32.85 ± 1.60	37.53 ± 2.32	13.61 ± 2.41	17.49 ± 0.90
Voltage-readout LIF	32.85 ± 0.33	38.00 ± 0.40	14.06 ± 0.37	16.44 ± 0.26
CSM (coupled voltage)	32.12 ± 0.13	36.62 ± 0.31	12.53 ± 0.07	$\mathbf{18.01 \pm 0.13}$

Table 3. Frozen LoCoMo transfer at 2,048 tokens: all 1,986 questions in ten conversations, using the official category-aware scorer. Learned controllers use seeds 7, 22, 99, selected after EvoEmo inspection and fixed before LoCoMo. QA is conversation-macro; category and answerable scores are question-micro. Values are mean $\pm$ sample SD where applicable.

LoCoMo transfer changes the ordering of the methods. At the default budget, BGE-M3 provides the strongest aggregate and answerable performance in Table 3. Among the recurrent controllers, CSM falls below GRU and both single-compartment variants. This is a meaningful boundary of the current training setup: a controller learned from one set of support-oriented histories does not consistently improve general conversational retrieval without adaptation.

The category decomposition identifies a different balance between answering and abstaining. CSM has the highest abstention score among the displayed methods, but this is accompanied by lower answerable and temporal scores. Its transfer deficit is therefore not well described as a single loss of aggregate relevance. The controller changes which evidence the reader sees and how often the reader finds enough

information to answer. Differences in role structure, question distribution, and evidence organization are plausible contributors; the current transfer experiment does not isolate their individual effects.

(a) EvoEmo					
Method	Informationextraction	Temporalreasoning	Conflictdetection	Usermodeling	Abstention
BM25	47.7	23.2	4.0	19.7	**61.1**
BGE-M3	49.4	**24.5**	4.1	**23.0**	49.3
GRU controller	48.6	23.2	4.3	19.5	57.6
Spike-readout LIF	46.9	21.4	**4.5**	20.1	58.2
CSM (coupled voltage)	**49.5**	23.7	4.3	20.9	58.8
(b) LoCoMo					
Method	Multi-hop	Temporal	Open-domain	Single-hop	Adversarial(abstention)
BM25	26.9	13.7	13.5	53.5	14.3
BGE-M3	29.1	**15.2**	**19.3**	**56.3**	12.1
GRU controller	**29.7**	13.9	14.6	51.5	17.9
Spike-readout LIF	29.0	13.6	15.5	52.0	17.5
CSM (coupled voltage)	28.2	12.5	14.8	51.1	**18.0**

Table 4. Capability profiles at a 2,048-token log budget. Values are absolute mean category scores in percent, with the largest unrounded mean in each dataset/category shown in bold. (a) EvoEmo: 192 held-out questions, with all five training seeds for learned controllers. (b) LoCoMo: 1,986 frozen-transfer questions, with seeds 7, 22, and 99. Category scores are question-micro averages under each dataset's scorer; the two panels are not

pooled. BM25 and BGE-M3 are deterministic. Full-precision values, sample SDs, and EvoEmo category counts are retained in the accompanying source data.

Table 4 shows how these effects are distributed across tasks. Relative to GRU, CSM's EvoEmo gains extend beyond abstention to information extraction, temporal reasoning, and user modeling. Against BGE-M3, however, stronger abstention accompanies weaker temporal and user-modeling scores. The frozen LoCoMo comparison exhibits a different profile: CSM's increased abstention is paired with reduced single-hop performance relative to BM25, while BGE-M3 improves the answerable categories. The benefit of a memory controller therefore depends on both the evidence it selects and the question being answered; one aggregate score cannot identify which of these behaviors changed.

5.4. Readout and context budget interact

Dataset	Tokens	BM25	BGE-M3	GRU	Voltage LIF	CSM
EvoEmo	512	**30.91**	29.04	30.39	29.99	30.61
EvoEmo	1024	29.54	28.78	**29.84**	28.90	29.39
EvoEmo	2048	30.02	29.04	29.46	29.39	**30.52**
LoCoMo	512	33.39	32.25	33.29	**34.30**	33.63
LoCoMo	1024	33.63	33.96	33.53	**33.96**	33.38
LoCoMo	2048	32.44	**33.96**	32.70	32.85	32.12

Table 5. Context-budget comparison. Learned-controller means use the same three seeds (7, 22, 99) at every budget; all five EvoEmo seeds remain in Table 1. Bold marks the largest displayed mean within each row. EvoEmo and LoCoMo use different scorers and are not pooled.

doi.org/10.32388/89XYAO

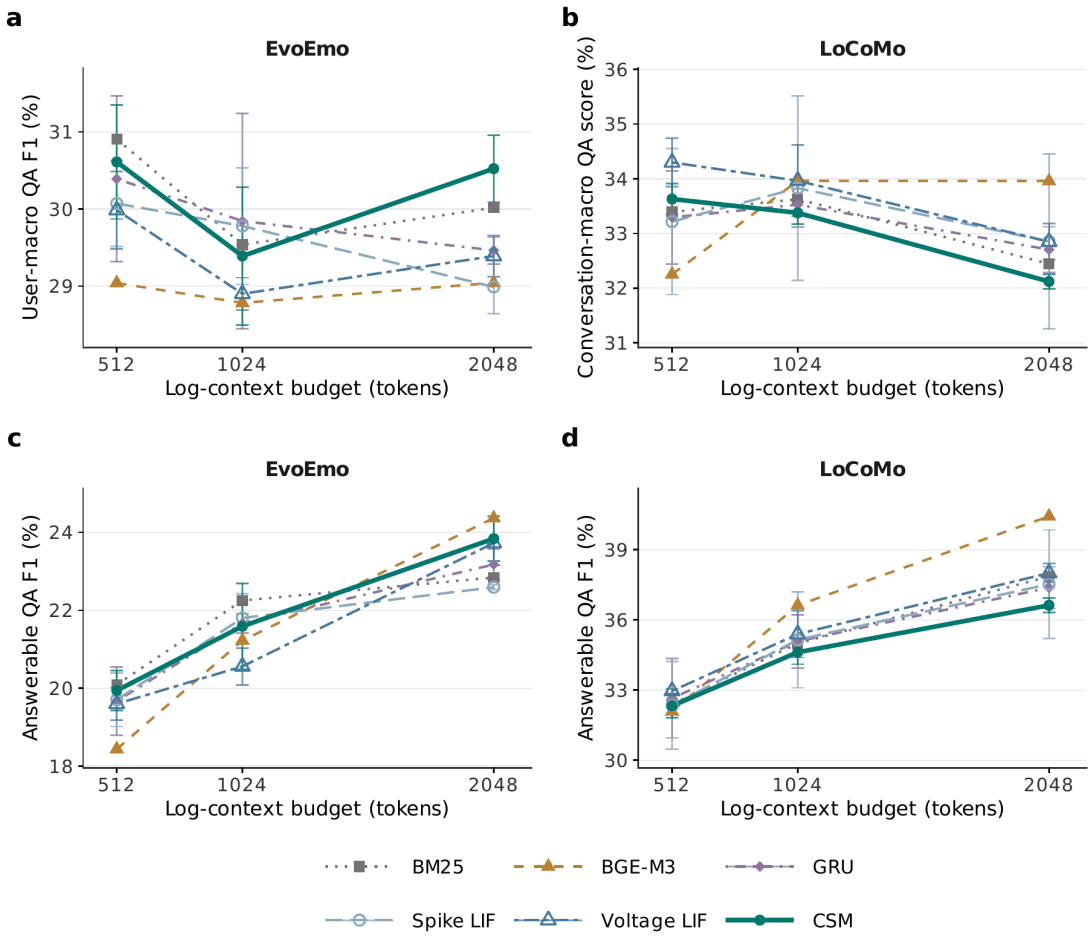


Figure 4. Interaction between readout and log-context budget. Panels (a,b) show the overall QA metric for EvoEmo and LoCoMo; panels (c,d) restrict evaluation to answerable questions. Points are means over the same three completed controller seeds (7, 22, 99), and error bars denote sample SD, not confidence intervals. Deterministic retrievers have no seed error bars. The EvoEmo budget cohort is separate from the full five-seed primary comparison. Lines connect the three measured budgets without smoothing. Each dataset retains its own scorer and aggregate definition.

The budget comparison in Figure 4 and Table 5 does not show a monotonic relation between available context and aggregate QA. On EvoEmo, CSM's clearest advantage over GRU occurs at 2,048 tokens, while its mean is lower at 1,024. On LoCoMo, voltage-readout LIF exceeds CSM at all three budgets. Continuous readout is therefore a useful control, not a guarantee that additional compartments will transfer better.

Reducing the budget also changes the mixture of successful answers and refusals. For CSM on LoCoMo, moving from 2,048 to 512 tokens raises the aggregate score while reducing answerable F1 by 4.30 points

and increasing abstention accuracy by 21.08 points. Each answerable category declines. The smaller context is not preserving the same factual capability more efficiently; it encourages a more abstention-heavy outcome. This is why the aggregate budget curve must be read together with the component scores.

5.5. Size and evidence access

Controller	Parameters	Temporal parameters	Recurrent state
GRU controller	144,705	24,960	64
Spike-readout LIF	123,905	4,160	64
Voltage-readout LIF	123,905	4,160	64
CSM (coupled voltage)	123,907	4,162	128

Table 6. Controller-only size. Recurrent state counts are scalar values, not whole-system storage. All methods additionally retain raw records, record embeddings, and per-record temporal traces.

CSM has 123,907 controller parameters, approximately 14.4% fewer than the matched GRU controller (Table 6). Its temporal module contains 4,162 parameters, while the majority of its controller parameters belong to the shared projections and scorer. The two-compartment recurrence doubles the active recurrent state from 64 to 128 scalars, but does not double the saved trace width because only somatic voltage is used for retrieval.

These counts describe a compact learned addition to the reader, not the size of the complete deployed system. Frozen BERT, the 3B reader, raw logs, and stored embeddings remain necessary resources. The completed five-seed validation pipeline took approximately 36.8 minutes and recorded 9.50 GiB peak PyTorch allocation during reader inference; it combined multiple controller runs and cached predictions. This is a reproducibility observation, not an isolated latency or energy comparison. The present implementation uses ordinary dense computation on an RTX 5090 and does not establish a neuromorphic efficiency advantage.

5.6. What stronger retrieval changes

An additional check combining BM25 and BGE-M3 by reciprocal rank fusion with fixed $k = 60$ yields stronger QA performance than the standalone coupled controller at the default budget on both datasets, using the same record budget, chronological packing, reader weights, and instruction. This result narrows the interpretation of the primary improvement: CSM improves on several individual retrieval and recurrent alternatives, but it does not yet dominate a stronger lexical–dense combination. The controller's contribution is therefore most clearly examined through matched recurrent substitutions and explicit evidence analysis, rather than through a claim of universal retrieval superiority. More effective integration with strong retrieval remains an open method-design question.

6. Discussion

6.1. A temporal trace is a retrieval representation

The proposed separation between stored text and recurrent state is important for interpreting what the method learns. The trace is not a replacement for the log and does not have to reconstruct every detail of a user's history. It contextualizes a record for a subsequent retrieval decision. The original content and lexical relevance remain available through direct paths. This arrangement allows a small trained module to affect evidence access without altering the frozen reader's parameters.

The two-compartment calculation provides a computational motivation for this representation, while the experiments test its usefulness in a specific retrieval pipeline. The local unit-modulus dynamics do not prove that the model remembers the right events or resolves contradictions. Those capabilities require task-level evidence. The observed improvement over the matched GRU on primary EvoEmo questions is compatible with a useful temporal representation, but the lower transfer scores and mixed temporal subsets limit stronger causal interpretations.

6.2. Evidence recall and answer utility are distinct

An evidence annotation marks relevant records, while the downstream metric measures the answer generated from a selected collection of records. These objectives are related but not identical. Record packing may omit a short but essential detail, include competing statements, or change the balance between a direct answer and a refusal. A controller can consequently improve evidence recall without

producing a better response. The primary and budget results expose this distinction rather than resolving it through a single retrieval metric.

The evaluation incentive is part of this distinction. Kalai et al. show that accuracy-based evaluations can reward guessing over admitting uncertainty [44]. Separately reporting answerable performance and abstention makes the reader's decision behavior visible alongside evidence access: an aggregate improvement can then be traced to better answers, more appropriate refusals, or a change in their balance.

This distinction suggests a more specific next step than simply increasing recurrent capacity: learning to select evidence sets whose content the fixed reader can use reliably. Such supervision must preserve both helpful and harmful examples and be evaluated on unseen users. The present manuscript does not merge later reader-instruction or utility-training variants into the CSM results, because they change different parts of the system. The reported CSM method remains the fixed evidence-supervised controller described in Section 3.

6.3. Limitations and deployment implications

The current evidence is limited to one small reader, a small number of EvoEmo users, and a frozen transfer setting. EvoEmo represents simulated longitudinal support histories; neither it nor the text-only LoCoMo experiment measures the quality of an actual emotional-support interaction. The evaluation establishes question-answering behavior under supplied histories, not clinical benefit, emotion-recognition accuracy, or empathetic response quality. Additional independent users and a fresh confirmatory evaluation are needed before estimating broad generalization.

The external log also creates practical storage, integrity, and privacy requirements. An edge assistant would need explicit retention rules, user-controlled deletion, and appropriate protection of personal records and derived embeddings. Knowledge-poisoning research on graph-based RAG shows how corrupted external evidence can redirect downstream generation [45]; provenance and integrity checks therefore belong alongside relevance selection in a deployed memory system. Future implementation work should additionally measure append-time processing, isolated end-to-end latency, and energy under the intended hardware. These are separate requirements from the parameter count of the neural controller.

The interaction setting should also guide future evaluation. Chien et al. identify differences in emotion annotation and speech-model behavior between conversational and monologue data [46]. Extending CSM

to affective interaction therefore calls for collecting and evaluating the intended conversational setting, including natural state changes and multimodal observations, rather than treating isolated affective samples as interchangeable with longitudinal histories.

7. Conclusion

We presented CSM, a compact coupled-spiking controller that contextualizes conversational records before budgeted retrieval for a frozen small language model. The complete five-seed EvoEmo evaluation supports a positive average QA improvement over BM25, BGE-M3, and a matched GRU, while LoCoMo transfer and budget analysis reveal substantial dependence on the readout and the answering–abstention balance. The practical direction is to improve how a small model accesses interaction-specific evidence without retraining the reader. Future work should connect temporal traces to reader-usable evidence selection and test that connection on independent users and continuously updated logs.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Data and Code Availability

EvoEmo/ES-MemEval is available at https://github.com/slptongji/ES-MemEval, and LoCoMo at https://github.com/snap-research/locomo. Dataset and pretrained-model use is subject to the respective source terms. A local evidence manifest, source hashes, table-generation code, and experiment artifacts are preserved by the authors. A public code-release location has not yet been assigned.

References

1. [a], [b] *Chen T, Lu J, Shen Y, Zhang L (2026). "ES-MemEval: Benchmarking Conversational Agents on Personalized Long-Term Emotional Support." arXiv:2602.01885. doi:10.1145/3774904.3792143. https://arxiv.org/abs/2602.01885.*

2. ^ a, b *Maharana A, Lee DH, Tulyakov S, Bansal M, Barbieri F, Fang Y (2024). "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). p. 13851–13870. doi:10.18653/v1/2024.acl-long.747. https://aclanthology.org/2024.acl-long.747/.*
3. ^ *Lu N, Han Z, Han M, Qian J (2024). "Bi-stream Graph Learning Based Multimodal Fusion for Emotion Recognition in Conversation." Inf Fusion. **106**:102272. https://doi.org/10.1016/j.inffus.2024.102272.*
4. ^ *Ma H, Wang J, Lin H, Zhang B, Zhang Y, Xu B (2024). "A Transformer-Based Model With Self-Distillation for Multimodal Emotion Recognition in Conversations." IEEE Trans Multimed. **26**:776–788. https://doi.org/10.1109/TMM.2023.3271019.*
5. ^ a, b *Qin J, Liu F, Zong L (2025). "BC-PMJRS: A Brain Computing-inspired Predefined Multimodal Joint Representation Spaces for Enhanced Cross-Modal Learning." Neural Netw. **188**:107449. https://doi.org/10.1016/j.neunet.2025.107449.*
6. ^ a, b *Qin J, Wang Y (2026). "Möbius Interaction Filtering: Order-Selective Spectral Regularization for Multimodal Humor Detection." arXiv preprint arXiv:2611.12913. doi:10.48550/arXiv.2611.12913. https://arxiv.org/abs/2611.12913.*
7. ^ *Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih WT, Rocktäschel T, Riedel S, Kiela D (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In: Advances in Neural Information Processing Systems. Vol. 33. https://arxiv.org/abs/2005.11401.*
8. ^ a, b, c *Zhang S, Yang Q, Ma C, Wu J, Li H, Tan KC (2024). "TC-LIF: A Two-Compartment Spiking Neuron Model for Long-Term Sequential Modelling." Proc AAAI Conf Artif Intell. **38**(15):16838–16847. doi:10.1609/aaai.v38i15.29625. https://ojs.aaai.org/index.php/AAAI/article/view/29625.*
9. ^ *Qin J (2024). "Zoom and Shift are All You Need." arXiv preprint arXiv:2406.08866. doi:10.48550/arXiv.2406.08866. https://arxiv.org/abs/2406.08866.*
10. ^ *Qin J, Liu F, Zong L (2025). "Hierarchical Adaptive Expert for Multimodal Sentiment Analysis." arXiv preprint arXiv:2503.22715. doi:10.48550/arXiv.2503.22715. https://arxiv.org/abs/2503.22715.*
11. ^ *Qin J, Peng B, Liu F, Cheng G, Zong L (2025). "DUAL: Dynamic Uncertainty-Aware Learning." arXiv preprint arXiv:2506.03158. doi:10.48550/arXiv.2506.03158. https://arxiv.org/abs/2506.03158.*
12. ^ *Qin J, Liu K, Cai Y, Ji T, Liu F (2025). "MTLP-MDG: Multi-Task Learning Framework using Probabilistic Distribution Perception for Missing Data Generation." In: 2025 International Joint Conference on Neural Networks (IJCNN). IEEE. p. 1–8. doi:10.1109/IJCNN64981.2025.11229220.*

13. ^ Wang Y, Qin J (2026). "DCER: Dual-Stage Compression and Energy-Based Reconstruction." arXiv preprint arXiv:2602.04904. doi:10.48550/arXiv.2602.04904. https://arxiv.org/abs/2602.04904.

14. ^ Li J, Wang X, Lv G, Zeng Z (2024). "GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition." IEEE Trans Multimed. **26**:77–89. https://doi.org/10.1109/TMM.2023.3260635.

15. ^ Zhang D, Chen F, Chang J, Chen X, Tian Q (2024). "Structure Aware Multi-Graph Network for Multi-Modal Emotion Recognition in Conversations." IEEE Trans Multimed. **26**:3987–3997. https://doi.org/10.1109/TMM.2023.3238314.

16. ^ Quan X, Wu S, Chen J, Shen W, Yu J (2024). "Multi-Party Conversation Modeling for Emotion Recognition." IEEE Trans Affect Comput. **15**(3):751–768. https://doi.org/10.1109/TAFFC.2023.3273589.

17. ^ Li J, Wang X, Liu Y, Zeng Z (2024). "CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition." IEEE Trans Affect Comput. **15**(4):1919–1933. https://doi.org/10.1109/TAFFC.2024.3389453.

18. ^ Shen S, Liu F, Wang H, Zhou A (2025). "Towards Speaker-Unknown Emotion Recognition in Conversation via Progressive Contrastive Deep Supervision." IEEE Trans Affect Comput. **16**(3):2261–2273. https://doi.org/10.1109/TAFFC.2025.3558222.

19. ^ Cheng Y, Chen H, Guo X, Ding B, Pei Z (2026). "Personality-aware Emotion Recognition in Conversation with Large Language Models." Pattern Recognit. **178**:113460. https://doi.org/10.1016/j.patcog.2026.113460.

20. ^ Lan K, Chen CLP, Zhang Z, Zhang T (2026). "Contrastive Adversarial Tuning: Enhancing Discriminability and Robustness of LLMs for Emotion Recognition in Conversation." Pattern Recognit. **174**:113025. https://doi.org/10.1016/j.patcog.2025.113025.

21. ^ Li J, Song S, Li Y, Zhang H, Hu G (2024). "ChatMDG: A Discourse Parsing Graph Fusion Based Approach for Multi-Party Dialogue Generation." Inf Fusion. **110**:102469. https://doi.org/10.1016/j.inffus.2024.102469.

22. ^ Zheng X, Du Y, Qin X (2026). "Multi-level Semantics-Aware and Multi-Granularity Knowledge-Infused Model for Emotional Support Conversation." Inf Fusion. **128**:103968. https://doi.org/10.1016/j.inffus.2025.103968.

23. ^ Zhong W, Guo L, Gao Q, Ye H, Wang Y (2024). "MemoryBank: Enhancing Large Language Models with Long-Term Memory." Proc AAAI Conf Artif Intell. **38**(17):19724–19731. doi:10.1609/aaai.v38i17.29946. https://ojs.aaai.org/index.php/AAAI/article/view/29946.

24. ^ Xu W, Liang Z, Mei K, Gao H, Tan J, Zhang Y (2025). "A-MEM: Agentic Memory for LLM Agents." In: Advances in Neural Information Processing Systems. https://arxiv.org/abs/2502.12110.

25. ^Lee KH, Chen X, Furuta H, Canny J, Fischer I (2024). "A Human-Inspired Reading Agent with Gist Memory of Very Long Contexts." In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research; vol. 235. p. 26396–26415. https://proceedings.mlr.press/v235/lee24c.html.

26. ^Asai A, He J, Shao R, Shi W, Singh A, Chang JC, Lo K, Soldaini L, Feldman S, D'Arcy M, Wadden D, Latzke M, Sparks J, Hwang JD, Kishore V, Tian M, Ji P, Liu S, Tong H, Wu B, Xiong Y, Zettlemoyer L, Neubig G, Weld DS, Downey D, Yih WT, Koh PW, Hajishirzi H (2026). "Synthesizing Scientific Literature with Retrieval-Augmented Language Models." Nature. **650**(8103):857–863. https://doi.org/10.1038/s41586-025-10072-4.

27. ^He K, Jing Y, Huang Y, Lu Z, An D, Wang L (2024). "Memory-Adaptive Vision-and-Language Navigation." Pattern Recognit. **153**:110511. https://doi.org/10.1016/j.patcog.2024.110511.

28. ^Robertson S, Zaragoza H (2009). "The Probabilistic Relevance Framework: BM25 and Beyond." Found Trends Inf Retr. **3**(4):333–389. doi:10.1561/1500000019. https://www.nowpublishers.com/article/Details/INR-019.

29. ^Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z (2024). "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation." In: Findings of the Association for Computational Linguistics: ACL 2024. p. 2318–2335. doi:10.18653/v1/2024.findings-acl.137. https://aclanthology.org/2024.findings-acl.137/.

30. ^Cormack GV, Clarke CLA, Büttcher S (2009). "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods." In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. https://plg.uwaterloo.ca/~gvcormac/cormacksigir09-rrf.pdf.

31. ^Liu W, Miao B, Cao J, Zhu X, Ge J, Liu B, Nasim M, Mian A (2025). "Context-Enhanced Video Moment Retrieval With Large Language Models." IEEE Trans Multimed. **27**:6296–6306. https://doi.org/10.1109/TMM.2025.3581797.

32. ^Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014). "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation." In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. p. 1724–1734. doi:10.3115/v1/D14-1179. https://aclanthology.org/D14-1179/.

33. ^Neftci EO, Mostafa H, Zenke F (2019). "Surrogate Gradient Learning in Spiking Neural Networks." arXiv:1901.09948. https://arxiv.org/abs/1901.09948.

34. ^Shen G, Zhao D, Zeng Y (2024). "Exploiting Nonlinear Dendritic Adaptive Computation in Training Deep Spiking Neural Networks." Neural Netw. **170**:190–201. https://doi.org/10.1016/j.neunet.2023.10.056.

35. ^Sun P, Wu J, Zhang M, Devos P, Botteldooren D (2024). "Delay Learning Based on Temporal Coding in Spiking Neural Networks." Neural Netw. **180**:106678. https://doi.org/10.1016/j.neunet.2024.106678.

36. ^Xue D, Li W, Yuan C, Liu H, Yao M, Liu W, Yang L, Li B, Hu W, Sun H, Li Z (2026). "MSFI: Multi-Timescale Spatio-Temporal Features Integration in Spiking Neural Networks." Neural Netw. **193**:108024. https://doi.org/10.1016/j.neunet.2025.108024.

37. ^Li Y, Zhao F, Zhao D, Zeng Y (2024). "Directly Training Temporal Spiking Neural Network With Sparse Surrogate Gradient." Neural Netw. **179**:106499. https://doi.org/10.1016/j.neunet.2024.106499.

38. ^Zhao D, Shen G, Dong Y, Li Y, Zeng Y (2025). "Improving Stability and Performance of Spiking Neural Networks Through Enhancing Temporal Consistency." Pattern Recognit. **159**:111094. https://doi.org/10.1016/j.patcog.2024.111094.

39. ^Chavlis S, Poirazi P (2025). "Dendrites Endow Artificial Neural Networks With Accurate, Robust and Parameter-Efficient Learning." Nat Commun. **16**(1):943. https://doi.org/10.1038/s41467-025-56297-9.

40. ^Sun P, Su Z, Achterberg J, Indiveri G, Goodman DFM, Akarca D (2026). "Algorithm–Hardware Co-Design of Neuromorphic Networks With Dual Memory Pathways." Nat Mach Intell. **8**(6):901–912. https://doi.org/10.1038/s42256-026-01255-3.

41. ^Baronig M, Ferrand R, Sabathiel S, Legenstein R (2025). "Advancing Spatio-Temporal Processing Through Adaptation in Spiking Neural Networks." Nat Commun. **16**(1):5776. https://doi.org/10.1038/s41467-025-60878-z.

42. ^Devlin J, Chang MW, Lee K, Toutanova K (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. p. 4171–4186. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423/.

43. ^Xu J, Guo Z, He J, Hu H, He T, Bai S, Chen K, Wang J, Fan Y, Dang K, Zhang B, Wang X, Chu Y, Lin J (2025). "Qwen2.5-Omni Technical Report." arXiv:2503.20215. https://arxiv.org/abs/2503.20215.

44. ^Kalai AT, Nachum O, Vempala SS, Zhang E (2026). "Evaluating Large Language Models for Accuracy Incentivizes Hallucinations." Nature. **653**(8116):1047–1051. https://doi.org/10.1038/s41586-026-10549-w.

45. ^Zhao T, Chen J, Ru Y, Zhu H, Hu N, Liu J, Lin Q (2026). "Exploring Knowledge Poisoning Attacks to Retrieval-Augmented Generation." Inf Fusion. **127**:103900. https://doi.org/10.1016/j.inffus.2025.103900.

46. ^Chien WS, Upadhyay SG, Lin WC, Busso C, Lee CC (2025). "Differential Impacts of Monologue and Conversation on Speech Emotion Recognition." IEEE Trans Affect Comput. **16**(2):485–498. https://doi.org/10.1109/TAFFC.2024.3509138.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.