A PAST CRISIS, A PRESENT REPLY: LEARNING TO REGULATE MEMORY INFLUENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Conversational memory helps language models maintain continuity in supportive dialogue, but salient past experiences can also draw responses away from a user's current needs. Historical information may remain valuable for factual access even when it is no longer relevant to the present request, creating a need to regulate memory influence without removing the memory itself. We introduce Memory Influence Budget (MIB), a fine-tuning framework that separates memory access from memory influence by placing a calibrated budget on changes in response distributions under matched memory interventions. Experiments demonstrate that MIB substantially reduces historical influence without a corresponding loss of memory dependent capability. Across model backbones and public memory QA tasks, MIB preserves factual access while maintaining current evidence recognition. Comparisons across control strengths further demonstrate stronger factual access than reverse context-aware decoding at comparable influence levels. These results demonstrate separate control of memory access and influence, enabling reduced historical influence without sacrificing access to retained information.

1 INTRODUCTION

As Large Language Models (LLMs) become increasingly integrated into repeated and longer-term user interactions, maintaining continuity across conversations has become an important capability (Pan et al., 2025). Conversational memory enables LLM-based assistants to retain information from earlier interactions, such as user preferences and goals, to inform subsequent responses. Recent work has advanced LLM memory through mechanisms for organising and updating interaction histories (Zhong et al., 2024; Xu et al., 2026), alongside benchmarks for reliable retrieval and reasoning over relevant information across extended histories (Maharana et al., 2024; Wu et al., 2025b).

However, access to historical information does not ensure appropriate usage in a given interaction. Irrelevant information can reduce topical relevance and impair constraint awareness, thereby diverting LLMs from the current interaction (Wang et al., 2026; Shi et al., 2023). Therefore, a central challenge lies in regulating the extent of memory influence on a response as its relevance to the current request changes (Huang et al., 2026). Figure 1 illustrates the pertinence of such regulation in risk-sensitive supportive dialogue settings. Although a salient past-crisis memory of an abusive relationship is peripheral to the user's current request, it substantially shapes the supervised fine-tuning (SFT) response. Overemphasising such memories may displace attention from immediate needs and unnecessarily reintroduce distressing material, which motivates development of explicit control mechanisms over how retrieved memories contribute to response generation.

Existing approaches regulate contextual influence at different stages of the memory pipeline through robustness training (Yoran et al., 2024; Shen et al., 2024), selective context compression (Xu et al., 2024), preference based optimisation of memory dependence (Huang et al., 2026; Bi et al., 2025), and generation time control of contextual use (Shi et al., 2024; Wang et al., 2025; Wu et al., 2025a). Although these approaches improve reliability and controllability of context use, they do not explicitly constrain the influence of retained memory while preserving access to its factual content. We study this problem in the context of high-stakes supportive dialogue, where historical information may support conversational continuity and subsequent factual use, yet inappropriate carryover can disproportionately shape responses to a user's immediate needs. Factual recall alone therefore does not establish appropriate use of history, while reduced historical sensitivity does not establish that

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.

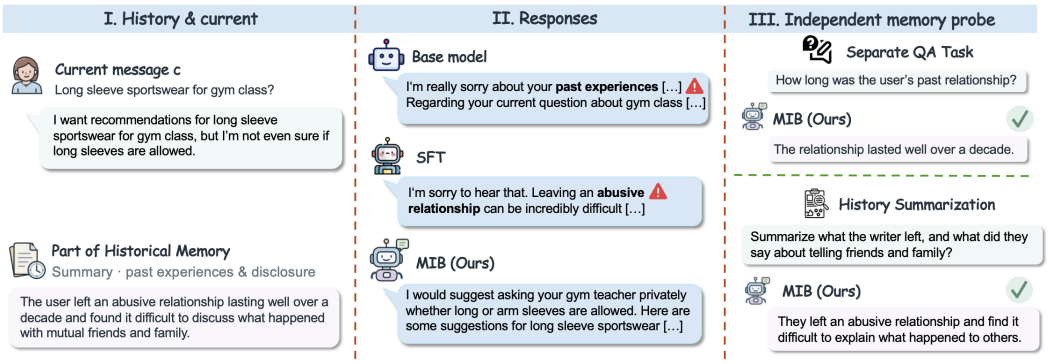

Figure 1: Illustration of MIB. Prior work utilising SFT approaches unnecessarily focuses on a past abusive relationship, whereas MIB correctly addresses the current sportswear concern. MIB also retrieves accurate historical information in a separate factual QA and summarisation task.

useful information remains accessible. Addressing this gap requires jointly considering historical influence, factual access, and recognition of evidence in the current message.

To address these challenges, we introduce **Memory Influence Budget (MIB)**, a fine-tuning framework that separates *memory access* and *memory influence* to control the influence of retained memory on a model response. MIB constrains distributional divergence under matched memory conditions within a validation calibrated budget and casts learning as constrained optimization. Within this budget, it improves supportive responses while preserving capabilities needed to utilise historical information and recognize current evidence. At inference, the model generates directly from the original memory and current message, without the comparison branch used during training.

Empirically, MIB regulates historical influence while preserving the model's ability to utilise memory when it remains relevant. Compared with competing memory control approaches, MIB achieves stronger current-evidence recognition and factual access, while avoiding the loss of memory-dependent capabilities associated with more aggressive memory interventions. These findings generalise across model backbones and public memory benchmarks.

Our contributions are summarized as follows:

- We introduce MIB, a fine-tuning framework that uses a calibrated budget on matched-memory prediction divergence, directly regulating historical influence while preserving factual recall, recognition of current evidence, and supportive response utility.
- We formulate a controlled evaluation setting for retained-memory use that disentangles historical influence from memory dependent capability.
- Experiments demonstrate that historical influence can be substantially reduced without a corresponding loss of memory dependent capability.

2 Related Work

2.1 Conversational History and Memory

Long-term memory systems have extended dialogue models beyond immediate context by storing and retrieving information from prior interactions. Existing memory architectures have explored different mechanisms for maintaining such histories. MemoryBank combines memory retrieval and updating with selective forgetting (Zhong et al., 2024); A-Mem constructs interconnected memory notes (Xu et al., 2026); MemoryOS organizes personal memories along short-, mid-, and long-term timescales (Kang et al., 2025); and Reflective Memory Management combines multi-granularity summarization with evidence-guided refinement to improve retrieval (Tan et al., 2025).

Beyond memory organization, benchmarks and dialogue tasks evaluate whether models can effectively utilize stored histories. The MapDia task studies when to reintroduce historical topics through

memory-aware proactive dialogue (Wu et al., 2025a). LoCoMo evaluates long conversations through question answering, event summarization, and multimodal dialogue generation (Maharana et al., 2024). LongMemEval targets information extraction, multi-session and temporal reasoning, knowledge updates, and abstention (Wu et al., 2025b). Collectively, these studies advance how conversational memories are accessed and reintroduced, but do not directly control how strongly retained memory shapes the current response once it becomes available. MIB addresses this gap by regulating memory influence while preserving access to the information it contains.

2.2 Robustness to Contextual Distractions

Irrelevant historical information can impair model performance even when task-relevant evidence remains accessible (Shi et al., 2023). Prior work addresses this problem at different stages of the model pipeline. At training, exposing retrieval-augmented models to mixtures of relevant and distracting passages improves robustness while retaining their ability to use relevant evidence (Yoran et al., 2024; Shen et al., 2024). During context processing, RECOMP uses learned context compression and selective augmentation to omit unhelpful retrieved content (Xu et al., 2024).

These methods primarily promote appropriate context use through selection and controllable reliance, but do not explicitly constrain how much retained memory may alter the model's predictive distribution. MIB instead imposes a calibrated budget on distributional changes under matched memory interventions, while preserving factual recall and sensitivity to current evidence.

2.3 Controlling Memory Influence

Memory influence can be regulated by how models use contextual information through prompting, decoding, internal representations, and training. Context-faithful prompting and Context-DPO encourage adherence to supplied evidence (Zhou et al., 2023; Bi et al., 2025). CAD(Shi et al., 2024) strengthens contextual influence by contrasting predictions with and without context, and Ada-CAD(Wang et al., 2025) adapts this contrast using Jensen–Shannon divergence. CAA(Rimsky et al., 2024) steers residual activations using contrastive directions, while PASTA(Zhang et al., 2024a) adjusts attention to user-specified input segments. SteeM aligns response-level memory dependence with user preferences through supervised fine-tuning and GRPO(Huang et al., 2026).

Existing methods steer contextual reliance without explicitly isolating the effect of retained history. MIB does so through matched historical-memory substitutions, holding the current message and reference prefixes fixed. It uses a differentiable Jensen–Shannon influence objective and a validation-calibrated training budget, together with auxiliary objectives that preserve recall, current-evidence, and response-utility. This formulation connects to constrained alignment methods such as Safe RLHF(Dai et al., 2024), but constraints the influence of retained conversational history.

3 Methodology

We introduce **Memory Influence Budget (MIB)**, a fine-tuning framework that regulates the influence of retained histories on current responses. As illustrated in Figure 2, MIB compares predictions under matched memory conditions and combines a calibrated influence budget with constraints on factual recall, current-evidence recognition, and response utility. At inference, the model generates directly from the original memory and current message.

3.1 Task Formulation and Matched Memory Interventions

Let c denote the current user message, m a historical memory, and $p_\theta(y \mid c, m)$ the response model. We seek to preserve access to historical information while limiting its tendency to redirect the response away from the needs expressed in c. The model should remain responsive to relevant evidence in the current message even as the influence of past events is reduced.

We construct matched intervention triples (c, m, m_0), where m carries a past-crisis annotation and m_0 is a comparison memory matched by topic and length selected within each data split. The influence-training population uses current messages without annotated current-crisis evidence. Each

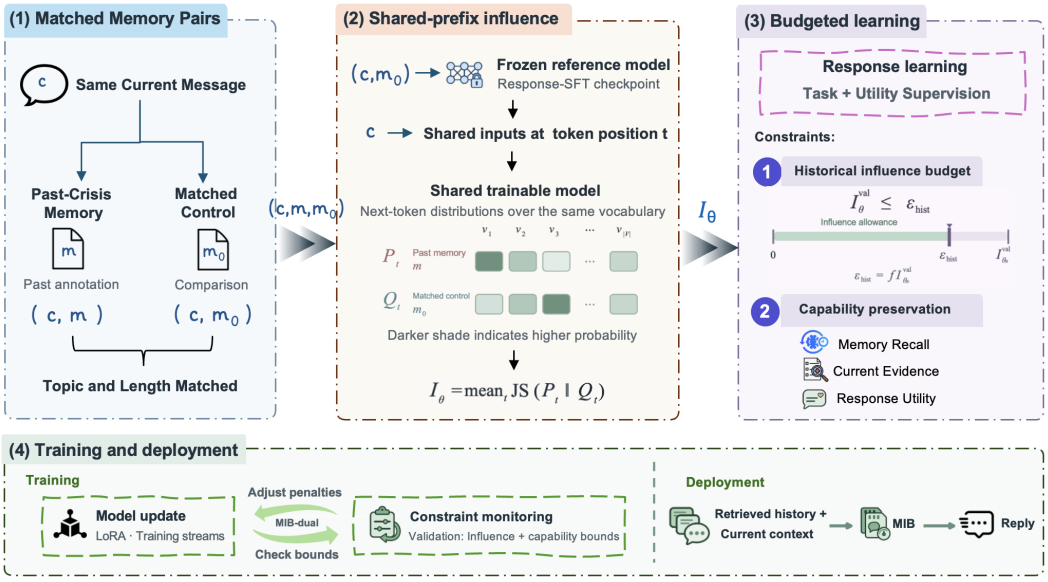

Figure 2: Overview of Memory Influence Budget (MIB).

intervention compares (c, m) with (c, m_0) using identical model and current text, isolating historical variation constrained by the influence budget while retaining history in context.

3.2 Influence Along Shared Reference Prefixes

To obtain a differentiable influence signal, we compare next-token distributions under the two memory conditions along a common response trajectory. Let θ_0 be the response-SFT initialization. We keep a frozen copy of this model and greedily generate a reference response $y^\star$ conditioned on (c, m_0). The resulting reference prefixes remain fixed during fine-tuning. At each scored position t, the trainable model produces

$$\begin{aligned} P_t &= p_\theta(\cdot \mid c, m, y^\star_{<t}), \\ Q_t &= p_\theta(\cdot \mid c, m_0, y^\star_{<t}). \end{aligned} \tag{1}$$

Both distributions are produced by the same trainable model, with gradients through both. Sharing prefixes fix preceding response tokens, excluding divergence from independently generated continuations. Dropout is disabled during the paired forward passes.

We define the influence for one matched input as

$$I_\theta(c, m, m_0; y^\star) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{JS}(P_t \| Q_t), \tag{2}$$

where T is the number of scored reference positions and JS is Jensen–Shannon divergence over the full next-token distributions (Lin, 1991). Its validation aggregate is denoted as I^{val}_θ.

3.3 Budgeted Fine-Tuning with Auxiliary Loss Bounds

The primary objective, L_{resp}, is the supervised negative log-likelihood (NLL) of supportive responses on the response training stream. The resulting objective is

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & L^{\mathrm{train}}_{\mathrm{resp}}(\theta) \\ \text{subject to} \quad & I^{\mathrm{val}}_\theta \le \varepsilon_{\mathrm{hist}}, \qquad \boldsymbol{\ell}^{\mathrm{val}}(\theta) \preceq \mathbf{b}. \end{aligned} \tag{3}$$

Here, $\boldsymbol{\ell}^{\mathrm{val}}(\theta) = \big(L^{\mathrm{val}}_{\mathrm{ret}}(\theta), L^{\mathrm{val}}_{\mathrm{cur}}(\theta), L^{\mathrm{val}}_{\mathrm{util}}(\theta)\big)^\top$ collects the auxiliary losses, $\mathbf{b} = (b_r, b_c, b_u)^\top$ contains their bounds, and $\preceq$ denotes componentwise inequality.

The utility loss L_{util} is the NLL of memory-free supportive responses on a separate utility stream. Together with L_{resp}, it provides supervision for the response-learning component: the primary stream trains task responses, while the utility bound anchors allowable changes in general supportive-response likelihood to the starting model. L_{ret} is the NLL on auxiliary memory-question-answering examples, providing recall supervision under the supplied memory. L_{cur} is the NLL for predicting the released labels of the current user turn from a user-only dialogue prefix, with historical event-state context supplied as an additional input. Listener and assistant turns are excluded from this classification input. Each supervised loss uses its own task-specific inputs and targets. The auxiliary recall task is distinct from the narrative factual QA used in evaluation.

We calibrate the bounds using the frozen response-SFT model on validation data:

$$\begin{aligned} \varepsilon_{\text{hist}} &= f I^{\text{val}}_{\theta_0}, & b_u &= (1+\rho_u) L^{\text{val}}_{\text{util}}(\theta_0), \\ b_r &= L^{\text{val}}_{\text{ret}}(\theta_0) + \delta_r, & b_c &= L^{\text{val}}_{\text{cur}}(\theta_0) + \delta_c. \end{aligned} \tag{4}$$

The budget factor $0 < f \leq 1$ specifies the allowed influence relative to the starting model; smaller values request stronger control. The nonnegative tolerances δ_r and δ_c specify additive NLL allowances, while ρ_u specifies a fractional utility-loss allowance. This calibration expresses the desired influence reduction relative to each backbone's initial behavior while anchoring the auxiliary bounds to its initial capabilities. The bounds remain fixed during training and apply to aggregate validation quantities. Parameter updates use the corresponding training streams.

3.4 Optimization and Inference

We initialize the model from response-SFT and optimize LoRA parameters (Hu et al., 2022). The response-learning component $L^{\text{train}}_{\text{gen}}(\theta; \lambda_u)$ augments the primary response NLL with an auxiliary NLL on memory-free supportive responses, scaled by λ_u, to maintain general response capability during task adaptation. Both MIB variants then use the four-component training objective

$$\begin{aligned} \mathcal{L}_{\text{train}}(\theta; \boldsymbol{\lambda}) = L^{\text{train}}_{\text{gen}}(\theta; \lambda_u) + \lambda_h \widehat{I}_{\text{train}}(\theta) \\ + \lambda_r L^{\text{train}}_{\text{ret}}(\theta) + \lambda_c L^{\text{train}}_{\text{cur}}(\theta), \end{aligned} \tag{5}$$

where $\widehat{I}_{\text{train}}$ is the influence estimate from training pairs and the penalty weights are nonnegative.

MIB-dual. The adaptive implementation alternates model updates on training streams with periodic penalty updates based on validation constraint values. For $k \in \{h, r, c, u\}$, the quantities g^{val}_k correspond to I^{val}_θ, $L^{\text{val}}_{\text{ret}}$, $L^{\text{val}}_{\text{cur}}$, and $L^{\text{val}}_{\text{util}}$, respectively, with $b_h = \varepsilon_{\text{hist}}$. We update

$$\lambda_k \leftarrow \text{clip}\left(\lambda_k + \eta \frac{g^{\text{val}}_k - b_k}{b_k},\, 0,\, \lambda_{\max}\right), \tag{6}$$

where η is the multiplier step size and $\lambda_{\max}$ is its upper bound. A violated constraint increases its penalty, whereas slack allows the penalty to decrease. Normalizing violations by their bounds accounts for differences in scale across the four quantities. The controller uses complete deterministic validation streams; gradients for model updates are computed on training data. The multiplier λ_u within the response-learning component follows the same update rule, using the utility constraint.

MIB-fixed. The fixed-penalty implementation optimizes Equation 5 with validation-selected constant weights. MIB-fixed shares the same four training components with MIB-dual while using a fixed trade-off throughout training, whereas MIB-dual adjusts these weights according to constraint violations. Additional training implementation details are provided in Appendix B.

Inference. After fine-tuning, the model generates from the original historical memory and current message using standard autoregressive decoding, $p_\theta(y \mid c, m)$. The comparison memory, reference trajectory, and paired forward passes are used only for training and analysis.

4 Experiment Setup

Our experiments examine whether MIB can control the influence of past-crisis memories while preserving access to historical facts and recognition of current evidence. We evaluate these aspects

Table 1: Task performance and public memory QA on Qwen3-8B. Scores are mean F1 over three seeds; best scores are bold.

	Task-specific evaluation				Public memory QA			
	Narrative QA	Current evidence			LoCoMo		LongMemEval	
Method	F1 ↑	Stage ↑	Type ↑	Status ↑	TF-IDF ↑	BGE ↑	TF-IDF ↑	BGE ↑
Original + full memory	0.3622	0.1004	0.0379	0.0692	0.1308	0.0979	0.0698	0.0696
SFT + full memory	0.5138	0.1721	0.2516	0.3170	0.2163	0.1681	0.0608	0.0630
Memory-level methods								
Attributed-past	0.4926	0.1870	0.2671	0.2900	0.2202	0.1663	0.0601	0.0602
MemoryBank-style‡	0.3028	0.1150	0.1663	0.1974	–	–	–	–
Inference-time model interventions								
CAD-R	0.0163	0.0021	0.0027	0.0067	0.0339	0.0339	0.0557	0.0557
AdaCAD-R	0.2203	0.0942	0.1277	0.1566	0.1144	0.0951	0.0562	0.0537
CAA-R	0.5019	0.1670	0.2340	0.2980	0.2085	0.1623	0.0588	0.0612
PASTA-R ($\alpha = 0.05$)	0.2857	0.0636	0.0873	0.0915	0.1584	0.1268	0.0550	0.0545
Training-based methods								
Context-DPO-R	0.4923	0.1585	0.2310	0.3046	0.2160	0.1659	0.0647	0.0649
NPO-R	0.5180	0.1642	0.2365	0.3077	0.2176	0.1655	**0.0706**	0.0718
Sham-target SFT + retain	0.5700	0.1999	0.2913	0.3238	0.2331	0.1772	0.0611	0.0644
SteeM-style SFT+GRPO	0.3300	0.1907	0.3098	0.2749	0.1501	0.1133	0.0599	0.0601
Ours								
MIB-dual	0.5878	0.1894	0.2804	0.2967	**0.2398**	**0.1815**	0.0669	**0.0735**
MIB-fixed	**0.5925**	**0.3949**	**0.4654**	**0.6062**	0.2393	0.1800	0.0654	0.0731

‡ MemoryBank-style is omitted from the full-set public QA comparison: LoCoMo lacks query timestamps, and some retrieved LongMemEval sessions postdate their queries.

against inference-time and training-based alternatives, then examine factual access across model families and public memory QA tasks. Controlled training comparisons assess the contribution of the historical-influence constraint under matched response supervision.

4.1 Datasets and Tasks

Historical memories and supportive responses. We use the annotated text data released with CRADLE Bench (Byun et al., 2026a) as the source of historical memories. Its crisis-type and temporal annotations support the construction of matched past-crisis and control-memory conditions. ESConv (Liu et al., 2021) and PsyDial-D4 (Qiu & Lan, 2025) provide supportive responses for response training and utility evaluation, while PsyDial-D101 serves as an additional held-out utility set. ESConv (Liu et al., 2021) and PsyDial-D4 (Qiu & Lan, 2025) provide 68,618 response-objective records and 17,552 utility-stream records. PsyDial-D101 provides 1,278 additional test-only utility records. For current-evidence classification, we use 25,662/3,529/4,527 training/validation/test user turns from CRADLE-Dialogue (Byun et al., 2026b).

Memory interventions and factual access. Matched interventions hold the current message fixed while substituting past-crisis and control memories to measure historical influence on the current response. To evaluate whether historical facts remain accessible, we construct a narrative QA test set derived from the source memories, distinct from the auxiliary recall task in training. Each method answers using the memory representation produced by its policy. We assess recognition of current evidence on 4,527 CRADLE-Dialogue test turns (Byun et al., 2026b), using stage, type, and status labels. Data and task construction are detailed in Appendix A.

Public memory QA. To assess whether memory QA remains available beyond the constructed crisis-memory task, we evaluate on 1,540 factual questions from ten LoCoMo conversations (Maharana et al., 2024) and the 500-question LongMemEval cleaned-small set (Wu et al., 2025b).

4.2 Model and Baselines

Backbones and reference models. The main baseline comparison uses Qwen3-8B (Yang et al., 2025). We extend factual QA evaluation to Qwen2.5-1.5B-Instruct (Qwen et al., 2025), Qwen3-1.7B/4B/14B, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Olmo-3-7B-Instruct (Olmo et al., 2026), using the same narrative QA test set across backbones. We include two reference models. Original + full memory uses the released Qwen3-8B checkpoint without project-specific fine-tuning. SFT + full memory applies response supervised fine-tuning while retaining the historical text in the input, and serves as the initialization for both MIB variants. The original and fine-tuned models are evaluated using the same task prompts, input construction, and decoding settings.

Baselines. We organize the baselines into three groups according to where they intervene in memory use. (1) *Memory-level methods* modify the presentation or retention of historical information. **Attributed-past**, inspired by context-faithful prompting (Zhou et al., 2023), explicitly frames the memory as a past account; **MemoryBank-style Ebbinghaus** (Zhong et al., 2024) applies an exponential retention rule based on the session gap. (2) *Inference-time model interventions* alter decoding or internal representations during generation. **CAD-R** (Shi et al., 2024) uses reverse context-aware decoding with a fixed mixing coefficient, while **AdaCAD-R** (Wang et al., 2025) adapts this coefficient to the divergence between output distributions. **CAA-R** (Rimsky et al., 2024) subtracts an activation direction estimated from paired past- and control-memory inputs, and **PASTA-R** (Zhang et al., 2024a) adapts attention scaling to the memory span. (3) *Training-based methods* change model parameters through alternative learning objectives. **Context-DPO-R** (Bi et al., 2025) favors control-memory responses over past-crisis-conditioned responses; **NPO-R** (Zhang et al., 2024b) uses negative preference optimization to suppress the latter; and **Sham-target SFT + retain** distills control-memory responses with auxiliary recall supervision. We also include **SteeM-style SFT+GRPO** (Huang et al., 2026), which adapts memory-dependence preference alignment to our task through supervised fine-tuning followed by GRPO. The suffix *-R* denotes our task adaptation, and dependence-control instructions are listed in Appendix E.2. Cross-backbone comparisons include no memory, SFT, MIB-dual, and MIB-fixed in Table 2. Baseline adaptations are described in Appendix B.

4.3 Evaluation Metrics

Factual access and current evidence. Narrative QA uses token-level F1 averaged by source memory to measure factual access after applying each memory policy. Current-evidence recognition uses macro-F1 for stage, type, and status, with the main comparison conducted under the past-memory condition. Supportive response utility is measured by negative log likelihood.

Public memory QA. For LoCoMo (Maharana et al., 2024), we report F1 on factual questions (categories 1–4). For LongMemEval (Wu et al., 2025b), we report token-level F1.

Influence analysis. D_{null} measures sensitivity to memory as the mean token-level Jensen–Shannon divergence(Lin, 1991) between predictions with and without memory along shared reference prefixes. Together with factual QA, it characterizes the relationship between memory influence and factual access. The training quantity I_θ measures divergence between matched past- and control-memory conditions and supports the analysis of budget control.

4.4 Experiment Settings

We initialize MIB from response-SFT and fine-tune with LoRA (Hu et al., 2022). The main Qwen3-8B experiments use matched replay. The default budget factor is $f = 0.75$, with constraint bounds calibrated separately for each backbone. Penalty settings and baseline parameters are determined using validation data. Baseline control strengths are adapted to the historical-memory task. Scores are averaged over three seeds. Public memory QA uses TF–IDF and BGE retrieval under the same token input budget and a fixed generation protocol, with results reported separately by dataset and retrieval setting. The main comparison reports mean scores over three random seeds.

We conduct additional analyses on Qwen3-8B to examine the contribution of the historical constraint and the trade-off between influence control and capability preservation. The historical-constraint

(a) Historical-constraint ablation
Qwen3-8B

Backbone	Constraint	D_{null}	I_θ	QA F1 ↑
Qwen3-8B	Disabled	5.804	6.260	**0.5984**
	Enabled	**4.705**	**5.000**	0.5878

(b) Budget sensitivity on Qwen2.5-1.5B
MIB-dual

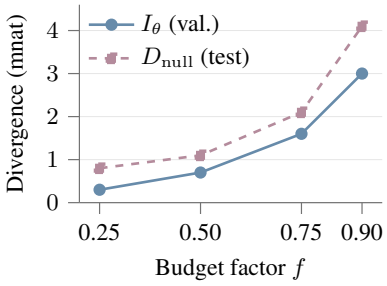

Figure 3: Historical influence control through constraint ablation (a) and budget adjustment (b). Divergences are in mnat; smaller f specifies a tighter budget.

Table 2: Historical QA performance across model sizes and families.

	Qwen2.5	Qwen3				Llama-3.1	Olmo-3
Method	1.5B	1.7B	4B	8B	14B	8B	7B
No memory	0.0495	0.0606	0.0731	0.0145	0.0398	0.0348	0.0368
SFT	0.4500	0.4348	0.5035	0.5138	0.3613	0.4914	0.3911
MIB-dual	0.4774	**0.4563**	0.4955	0.5878	0.3816	0.4955	0.4048
MIB-fixed	**0.4844**	0.4381	**0.5405**	**0.5925**	**0.5679**	**0.5235**	**0.4515**

ablation holds initialization, replay supervision, optimization settings, and the other three constraints fixed, and uses the full current-evidence test set. For the control-strength sweep, we report three settings each for MIB-dual and MIB-fixed alongside selected baseline configurations. All sweep settings use the same Narrative QA test set and a fixed, label-stratified subset of 600 current-evidence test turns. Influence is evaluated using shared frozen reference prefixes across configurations.

5 Results

5.1 Task Performance and Public Memory QA

Table 1 summarizes the main results across task-specific evaluation and public memory QA. Relative to the original checkpoint, response SFT improves Narrative QA and all three current-evidence tasks, establishing a stronger task-adapted reference. Both MIB variants further improve Narrative QA and outperform all comparison methods. MIB-fixed additionally achieves the strongest performance across the three current-evidence tasks. These joint gains show that stronger access to historical facts can coexist with more accurate recognition of the current message. The contribution of the historical constraint is examined separately in the ablation analysis.

Both MIB variants improve over SFT + full memory across all public memory QA settings. On LoCoMo, they also outperform the original checkpoint under both lexical and dense retrieval, extending the factual-access gains beyond the constructed historical-memory task. On LongMemEval, both variants exceed the original checkpoint with BGE retrieval while remaining slightly lower with TF–IDF retrieval. MIB-dual achieves the best LoCoMo and BGE-based LongMemEval scores, with MIB-fixed performing similarly. Overall, MIB consistently outperforms compared to response SFT, with gains over the original model depending on the evaluation setting.

5.2 Factual Access across Model Sizes and Families

Table 2 evaluates factual access across models from the Olmo, Llama, and Qwen families. MIB-fixed improves over the corresponding SFT reference on every evaluated backbone, while MIB-dual improves on most. Within Qwen3, the improvements of MIB-fixed persist from 1.7B through 14B,

Table 3: Influence and factual access on Qwen3-8B.

Method	SFT	MIB-dual			MIB-fixed			CAD-R	PASTA-R	Memory Bank-style	Context DPO-R
Setting	M	W	M	S	W	M	S	0.5	0.2	M	M
I_θ (mnat)	13.884	5.635	5.000	2.768	4.784	3.689	2.919	4.052	6.149	5.330	7.010
Narrative F1	0.514	0.589	0.588	**0.609**	0.594	0.593	0.593	0.250	0.444	0.303	0.492
Stage F1	0.222	0.216	0.243	0.192	0.409	0.419	**0.430**	0.093	0.168	0.130	0.190
Type F1	0.297	0.310	0.332	0.312	0.504	0.504	**0.512**	0.144	0.288	0.183	0.287
Status F1	0.373	0.339	0.339	0.327	**0.638**	0.632	0.630	0.133	0.280	0.220	0.363

Settings: W/M/S denote weak/main/strong influence control within each method. **Stage:** `alert` (an ambiguous warning signal) or `confirm` (a crisis type becomes explicitly identifiable). **Type:** crisis category: active/passive suicidal ideation, self-harm, domestic violence, rape, sexual harassment, or child abuse/endangerment. **Status:** event temporality, `ongoing` or `past`. CAD-R ($\alpha = 0.5$) and PASTA-R ($\alpha = 0.2$) are task-adapted exploratory sweep settings.

with pronounced gains on the 8B and 14B models. This shows that the factual-access benefits remain relevant as backbone capacity increases. The improvements on Llama and Olmo further extend the evidence across different model families. Meanwhile, the consistently weak no-memory reference confirms that this evaluation requires information from the supplied history, making these results evidence of retained factual access after influence-control training. Additional results on Qwen2.5-1.5B are reported in Appendix C, with paired bootstrap comparisons against SFT provided in Appendix D.

5.3 Influence Control and Capability Preservation

Historical-constraint ablation. Figure 3(a) isolates the historical constraint on Qwen3-8B, with response supervision, initialization, training schedule, and other constraints fixed. This separates its contribution to influence control from the effects of the shared training. Enabling the constraint lowers both matched-memory influence and D_{null}, while factual QA F1 decreases from 0.5984 to 0.5878. These results indicate additional influence control with an observed factual-access trade-off. Figure 3(b) further shows that tighter MIB-dual budgets reduce both validation influence and test D_{null}, enabling adjustable control over historical sensitivity.

Capabilities across control strengths. Table 3 compares historical influence, factual access, and current-evidence recognition across control settings. All six MIB configurations reduce historical influence and improve Narrative QA F1 over SFT. At similar influence levels, the main MIB-fixed configuration achieves higher Narrative QA F1 than CAD-R, together with higher Stage, Type, and Status F1. The main MIB-dual configuration likewise outperforms MemoryBank-style on these four measures at closely matched influence levels. Compared with the displayed PASTA-R and Context-DPO-R configurations, all MIB configurations achieve higher factual QA at lower influence. Trade-offs differ across MIB variants, where strong-control MIB-dual achieves the highest Narrative QA F1, while MIB-fixed performs better on current-evidence recognition, with its strong setting leading Stage and Type and its weak setting leading Status. Overall, capability preservation depends on how influence is controlled, rather than suppression strength alone.

6 Conclusion

We introduced Memory Influence Budget (MIB), a fine-tuning framework for regulating the influence of retained memories on current responses while preserving access to useful historical information. MIB separates memory access from memory influence by constraining prediction divergence under matched memory interventions within a calibrated budget, while jointly preserving factual recall, recognition of current evidence, and supportive response utility. Across controlled ablations and control strength analyses, MIB demonstrates response sensitivity regulation to history alongside response learning and capability preservation. Experiments show that historical influence can be substantially reduced while maintaining factual recall and current-evidence recognition. These findings support differentiating memory access and influence as distinct aspects of adaptation, enabling more selective use of history in memory-conditioned assistants.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature discovery, method implementation and polish writing. We also used AI to generate narrative factual question–answer pairs, with answers checked programmatically against source-memory text. For the adapted baselines, we used model-generated response targets, rule-constructed preference pairs, and model scores for SteeM-style training. All AI-assisted work was reviewed by the authors. AI-generated code was checked and tested by 2 authors, and the manuscript was proofread by all authors. We take responsibility for the final content of this work, including its text, mathematical claims, experimental results, and AI-assisted artifacts.

ETHICS STATEMENT

This work studies controlled language-model behavior in the presence of historical crisis memories. We use existing datasets containing sensitive personal narratives and synthetic dialogues; constructed history–current-message pairings should not be interpreted as real longitudinal interactions. Model-generated QA and model-based training rewards are distinguished from human annotations. The supplementary code does not redistribute the underlying dataset records.

REPRODUCIBILITY STATEMENT

The main paper and appendix describe the MIB objective, fixed-reference influence measurement, auxiliary constraints, validation-based budget calibration, and dual and fixed-penalty optimization. We also document the datasets and splits, traning details, baseline adaptations, evaluation protocols. The supplementary material provides source-code excerpts covering MIB training, loss computation, prompt construction, baseline inference interventions, and evaluation metrics. These materials support inspection of the implementation and its correspondence with the method described in the paper. We will release the full implementation and experiment configurations to facilitate reproduction and further research. These efforts are intended to enable the community to replicate our experiments and build upon our findings.

REFERENCES

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, 2025.

Grace Byun, Rebecca Lipschutz, Sean T Minton, Abigail Powers, and Jinho D Choi. Cradle bench: A clinician-annotated benchmark for multi-faceted mental health crisis and safety risk detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1572–1590, 2026a.

Grace Byun, Abigail Lott, Rebecca Lipschutz, Sean T Minton, Elizabeth A Stinson, and Jinho D Choi. Expert-level crisis detection in mental health conversations. *arXiv preprint arXiv:2606.10380*, 2026b.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zisu Huang, Muzhao Tian, Xiaohua Wang, Jingwen Xu, Zhengkang Guo, Qi Qian, Kaitao Song, Jiakang Yuan, Changze Lv, and Xiaoqing Zheng. Controllable memory usage: Balancing anchoring and innovation in long-term human–agent interaction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14699–14719, 2026.

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*, pp. 25972–25981, 2025.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

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)*, pp. 13851–13870, 2024.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. On memory construction and retrieval for personalized conversational agents, 2025. URL https://arxiv.org/abs/2502.05589.

Huachuan Qiu and Zhenzhong Lan. Psydial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21624–21655, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8988–9003, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. 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)*, pp. 11636–11652, 2025.

Yuxin Wang, Paul Thomas, Zhiwei Yu, Yuan Gao, Saeed Hassanpour, Soroush Vosoughi, Robert Sim, and Nick Craswell. Memory makes the difference: Evaluating how different memory roles shape conversational agents, 2026. URL https://arxiv.org/abs/2606.25361.

Bowen Wu, Wenqing Wang, Haoran Li, Yunhan Deng, Ying Li, Jingsong Yu, and Baoxun Wang. Interpersonal memory matters: A new task for proactive dialogue utilizing conversational history. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 47–67, 2025a.

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*, 2025b. URL https://openreview.net/forum?id=pZiyCaVuti.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024a.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024b.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, 2023.

A Data and Task Construction

Matched memory interventions. We use the annotated text released with CRADLE Bench (Byun et al., 2026a) to construct past-crisis and comparison-memory conditions. Within each split, candidate memories are matched by topic similarity and length, while the current message is held fixed. The resulting input pairs allow the effect of memory content to be measured independently of changes to the current message. A source-memory cluster groups evaluations associated with the same historical source. Table 4 summarizes statistics for the response and utility streams, current-evidence classification, and factual QA datasets.

Response learning and utility. ESConv (Liu et al., 2021) and PsyDial-D4 (Qiu & Lan, 2025) provide supportive-response supervision and a separate utility stream. PsyDial-D101 supplies held-out utility evaluation. These streams complement the matched interventions: the intervention set defines the memory-control problem, whereas the response and utility records supply targets for learning and evaluating supportive language generation.

Narrative factual QA. We generate candidate questions with Qwen3-8B from source-memory windows, requesting three questions per memory. Answers are checked against substrings of the source text, and questions that fail the construction filters are removed. Questions concern narrative facts without exposing structured annotation headers. Each evaluated policy is applied to the memory before the question is answered. This evaluation task is separate from the auxiliary recall task used in MIB training.

Current-evidence recognition. CRADLE-Dialogue (Byun et al., 2026b) supplies 4,527 test turns with stage, type, and status labels. We evaluate the same turns under past-memory, comparison-memory, and no-memory conditions, applying each policy to the structured-state input. This task tests whether the model recognizes evidence relevant to the current turn under different memory conditions.

B Training and Baseline Implementation

Training Implementation The reference denoted SFT + full memory is a response-SFT checkpoint with the historical text retained in its input. MIB-dual and MIB-fixed initialize from the corresponding response-SFT model. The Original + full memory reference denotes the official instruction-tuned checkpoint without a project adapter. We use these names consistently to distinguish the model initialization from the memory policy.

The original configuration uses AdamW and rank-16 LoRA (Hu et al., 2022), with LoRA scaling 32, dropout 0.05, a microbatch size of 2, and gradient accumulation over 8 steps. Paired forwards for the influence loss disable dropout.

MIB-dual updates nonnegative penalty weights using validation violations normalized by their respective bounds. In the original configuration, updates occur at initialization and every 500 optimizer steps, with update rate $\eta = 5$ and projection to $[0, 20]$. MIB-fixed uses a historical-influence weight of 4.0 and weights of 0.5 for each of the three auxiliary losses, selected on validation data and transferred across backbones.

The matched-replay configuration adds supportive-response supervision rendered in the intervention input format. Its training inputs and additional validation inputs are constructed examples; the latter reuse the original validation histories. They are therefore not new independent historical sources.

The historical-constraint ablation holds this replay supervision, initialization, schedule, and seed fixed, and disables only the historical term while retaining the recall, current-evidence, and utility constraints.

Baseline Implementation The suffix -R identifies an adaptation to the present memory-control task. CAD-R (Shi et al., 2024) reverses the direction of context-aware decoding using a fixed coefficient; AdaCAD-R (Wang et al., 2025) adapts the coefficient to the divergence between output distributions. CAA-R (Rimsky et al., 2024) subtracts an activation direction estimated from paired memory conditions. PASTA-R (Zhang et al., 2024a) scales attention associated with the memory span. CAD-R and AdaCAD-R retain their validation-selected strengths in the completed unified evaluation. For Qwen3-8B, the main PASTA-R setting uses $\alpha = 0.05$, with $\alpha = 0.2$ additionally reported in the control-strength comparison. For Qwen2.5-1.5B, the selected setting is $\alpha = 1$, which reproduces its eager-attention SFT reference.

Attributed-past uses an explicit past-account framing inspired by context-faithful prompting (Zhou et al., 2023). MemoryBank-style Ebbinghaus uses an exponential retention rule based on session gap (Zhong et al., 2024). Context-DPO-R (Bi et al., 2025) favors comparison-memory responses over past-crisis-conditioned responses, while NPO-R (Zhang et al., 2024b) applies negative preference optimization to the latter. Sham-target SFT + retain distills comparison-memory responses with auxiliary recall supervision.

CAD-R control strengths on Qwen3-8B. CAD-R combines memory-conditioned and memory-free logits from a fixed response-SFT checkpoint θ_0:

$$\widetilde{z}_{\alpha,t} = (1-\alpha) z_{\theta_0}(c, m, y_{<t}) + \alpha z_{\theta_0}(c, \emptyset, y_{<t}), \tag{7}$$

where c contains the current input and task instructions, m is the historical memory, and both branches share the prefix $y_{<t}$. Softmax is applied after logit combination. We evaluate $\alpha \in \{0.5, 0.75, 1.0\}$ without additional training. Larger α decreases the weight on memory-conditioned logits; $\alpha = 1.0$ is the memory-free endpoint.

For Qwen3-8B, $\alpha = 1.0$ remains the validation-selected main configuration, while $\alpha = 0.5$ and 0.75 are supplementary exploratory settings. Table 5 reports the complete test sweep. Table 3 displays $\alpha = 0.5$ for an exploratory comparison with MIB-fixed main at similar measured influence (4.052 versus 3.689 mnat), without replacing the main CAD-R configuration. Increasing α reduces measured influence but also lowers Narrative QA and current-evidence F1. Thus, low influence alone does not establish capability preservation.

C Additional Results on Qwen2.5-1.5B

We additionally evaluate Qwen2.5-1.5B-Instruct to examine influence control in a smaller model. In Table 6, the comparison covers factual access, current-evidence recognition, and public memory QA using the inference-time and training-based alternatives described in the experimental setup. In the original training configuration, both MIB variants achieve higher mean Narrative QA F1 than SFT, and MIB-fixed achieves the highest current-evidence F1. On public memory QA, MIB-fixed remains close to SFT on LoCoMo and improves LongMemEval token-F1 under the evaluated retrieval settings. These experiments provide complementary evidence about the joint behavior of influence control and task performance at a smaller scale.

D Statistical Reliability

Table 7 reports paired differences against SFT using bootstrap resampling over source histories. This preserves the dependence among questions derived from the same history. Both MIB variants have positive point estimates, with confidence intervals that include zero. The sham-target training control yields a positive QA difference with an interval above zero, reinforcing the distinction between improving factual QA and controlling historical influence.

Table 4: Statistics of response, classification, and factual QA data. Slash-separated counts indicate training, validation, and test splits; other counts are totals for the indicated stream or set.

Data or task	Unit	Count
Response, utility, and current-evidence data		
ESConv + PsyDial-D4	Response records	68,618
ESConv + PsyDial-D4	Utility records	17,552
PsyDial-D101	Test utility records	1,278
CRADLE-Dialogue	Current-evidence turns	25,662 / 3,529 / 4,527
Factual QA datasets		
Narrative QA	Questions	1,315 / 144 / 187
LoCoMo (categories 1–4)	Factual questions	1,540
LongMemEval cleaned-small	Questions	500

Table 5: Complete CAD-R strength sweep on Qwen3-8B. All settings use the same response-SFT checkpoint. Bold marks the highest F1 within this sweep.

α	I_θ (mnat)	Narrative F1	Stage F1	Type F1	Status F1
0.50	4.052	**0.250**	**0.093**	**0.144**	**0.133**
0.75	1.131	0.133	0.013	0.022	0.014
1.00*	≈ 0	0.016	0.000	0.000	0.000

*Validation-selected main configuration. Narrative QA is token-F1 averaged within each source memory and then across sources. Stage, Type, and Status are macro-F1 on the same fixed turn synthetic-dialogue test subset under the past-memory condition; they measure evidence stage, crisis category, and event temporality, respectively. I_θ uses the same frozen reference paths as Table 3; 1 mnat equals 10^{-3} nat. At $\alpha = 1.0$, the measured value is 5.78×10^{-6} mnat, consistent with the memory-free endpoint up to numerical precision.

E Label Definitions and Dependence Instructions

E.1 Current-Evidence Labels

Current-evidence recognition is evaluated along three dimensions: stage, crisis type, and temporal status. These dimensions are derived from the dataset's canonical labels and are summarized in Table 8. Multiple labels may be assigned to the same example.

E.2 SteeM-Style Dependence Levels

The trained, task-adapted SteeM-style baseline uses five levels of memory dependence, ranging from answering independently of history to treating history as the main source of evidence. Table 9 lists the instruction associated with each level. The selected instruction is prepended to the user-role input.

Table 6: Task performance and public memory QA on Qwen2.5-1.5B. Scores are mean F1 over three seeds; best scores are bold.

	Task-specific evaluation				Public memory QA			
	Narrative QA	Current evidence			LoCoMo		LongMemEval	
Method	F1 ↑	Stage ↑	Type ↑	Status ↑	TF-IDF ↑	BGE ↑	TF-IDF ↑	BGE ↑
Original + full memory	0.3721	0.0934	0.0240	0.0573	0.1926	0.1567	0.0848	0.0830
SFT + full memory	0.4500	0.2464	0.2605	0.5188	0.2240	0.1817	0.0928	0.0910
Memory-level methods								
Attributed-past	0.4328	0.2749	0.3016	0.4767	0.2106	0.1672	0.0917	0.0946
MemoryBank-style‡	0.3060	0.2050	0.2054	0.3899	–	–	–	–
Inference-time model interventions								
CAD-R	0.3030	0.2411	0.2678	0.4780	0.2095	0.1718	0.0988	0.0993
AdaCAD-R	0.1688	0.2437	0.2618	0.5004	0.1408	0.1215	0.0959	0.0954
CAA-R	0.4510	0.2481	0.2650	0.5240	0.2227	0.1826	0.0947	0.0912
PASTA-R ($\alpha = 1$)†	0.4335	0.2543	0.2763	0.5091	0.2250	0.1826	0.0961	0.0925
Training-based methods								
Context-DPO-R	0.4673	0.2360	0.2494	0.5154	0.2275	0.1875	0.0949	0.0962
NPO-R	0.4581	0.2496	0.2749	0.5146	**0.2330**	0.1847	0.1011	0.0959
Sham-target SFT + retain	**0.5055**	0.2713	0.3035	0.5269	0.2230	0.1830	0.0930	0.0932
Ours								
MIB-dual	0.4774	0.2672	0.2854	0.5380	0.2179	0.1795	0.0922	0.0875
MIB-fixed	0.4844	**0.4071**	**0.4580**	**0.5762**	0.2237	**0.1928**	**0.1074**	**0.1002**

‡ MemoryBank-style is omitted from the full-set public QA comparison: LoCoMo lacks query timestamps, and some retrieved LongMemEval sessions postdate their queries.

Method	Narrative F1 ↑	ΔF1 ↑	Paired 95% CI
SFT + full memory	0.4500	0.0000	[0.0000, 0.0000]
CAD-R	0.3030	−0.1470	[−0.1906, −0.1062]
AdaCAD-R	0.1688	−0.2812	[−0.3312, −0.2305]
CAA-R	0.4510	0.0010	[−0.0093, 0.0117]
Attributed-past	0.4328	−0.0172	[−0.0418, 0.0060]
MemoryBank-style	0.3060	−0.1440	[−0.1985, −0.0926]
Context-DPO-R	0.4673	0.0173	[−0.0021, 0.0409]
NPO-R	0.4581	0.0081	[−0.0187, 0.0357]
PASTA-R ($\alpha = 1$; eager)	0.4335	0.0000	[0.0000, 0.0000]
Sham-target SFT + retain	**0.5055**	**0.0555**	[0.0212, 0.0924]
MIB-dual	0.4774	0.0274	[−0.0007, 0.0562]
MIB-fixed	0.4844	0.0344	[−0.0056, 0.0749]

Table 7: ifferences are relative to SFT evaluated with the same attention implementation: PASTA-R at $\alpha = 1$ is identical to its eager SFT reference, explaining its zero difference despite a different absolute score from the first row. Bold marks the highest point estimate, not statistical significance.

Table 8: Label dimensions for current-evidence recognition.

Dimension	Values and interpretation
Stage	`alert`: possible evidence; `confirm`: explicitly identifiable evidence.
Type	Active suicidal ideation, passive suicidal ideation, self-harm, domestic violence, rape, sexual harassment, and child abuse or endangerment.
Status	`ongoing` or `past`, following the released temporal annotations.

Table 9: Dependence instructions for the trained SteeM-style baseline.

Level	Instruction
1	Answer the current request independently of historical records.
2	Let the historical record provide light context, while answering the current request mainly on its own terms.
3	Combine relevant historical details with the current request, letting both meaningfully shape your answer.
4	Let the relevant historical details guide the structure and substance of your answer to the current request.
5	Use the historical record as the main evidence for your answer.