CIT-MEM: COUNTERFACTUAL INTERVENTION TRAINING FOR EVIDENCE-GROUNDED MEMORY POLICIES IN LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon LLM agents rely on memory to interpret conversational history, yet most systems optimize semantic relevance rather than whether a record remains valid, directly supports the query, and provides sufficient evidence. A correct answer may still reflect parametric knowledge or a shortcut; inference-time interventions can reveal this dependence but do not teach a policy how to use evidence. We introduce CIT-MEM (**C**ounterfactual **I**ntervention **T**raining for Memory), a framework that jointly trains invariance when the role of decisive evidence is preserved, sensitivity when that role changes, and abstention when no valid and sufficient evidence remains. It constructs validated non-causal twins that preserve decisive evidence and causal twins that remove, replace, contradict, or update it. Validation checks structural consistency, evidence provenance, recoverability, and answer direction, discarding unreliable transformations before training. Aligned policy objectives directly supervise memory operations, retrieval, and answerability, while adaptive replay prioritizes unresolved constraints with the backend and reader fixed. Experiments on LoCoMo-10 and LongMemEval-S show that CIT-MEM improves intervention consistency and abstention reliability while attaining the best overall answer quality across three frozen readers. On LoCoMo-10 with Qwen3-8B, it improves all-query F1 by 14.3% and BLEU-1 by 17.2% relative to the strongest baseline while using the smallest retrieved-context budget. It also achieves the best overall judged accuracy on LongMemEval-S, demonstrating more effective evidence-grounded memory selection in long conversations. Our code can be found at https://anonymous.4open.science/r/cit-mem-2B43.

1 INTRODUCTION

Consciously retrieving past experiences is a hallmark of human intelligence. This "mental time travel", or episodic memory, organizes experiences along spatiotemporal axes to capture event order, duration and causal links. Unlike semantic memory for general world knowledge, episodic memory stores context-dependent personal encounters and enables continuous environmental adaptation. Building such episodic-like memory for long-horizon LLM agents remains challenging (Maharana et al., 2024). Existing external-memory frameworks maintain records for growing interaction contexts to support agent reasoning (Wang et al., 2023; Packer et al., 2023; Zhong et al., 2024; Gutiérrez et al., 2024; Pan et al., 2025; Xu et al., 2025; Tan et al., 2025). While these systems improve memory recall by semantic relevance, relevance alone cannot guarantee that queries obtain timely, direct, and sufficient supporting evidence (Wu et al., 2025; Zhang et al., 2026a). Rather than redesigning the underlying storage or retrieval backend, we focus on the memory-policy problem: how to select decisive evidence from candidate records returned by a fixed storage-and-access pipeline.

Retrieved memories can provide conflicting support for the same query: a stale record and its later temporal correction may both appear topically relevant, yet only the updated entry reflects current ground truth (Xu et al., 2024; Fatemi et al., 2025). More fundamentally, superficial answer correctness cannot reveal which piece of memory the policy actually relies on (Cohen-Wang et al., 2024). The reader model may produce correct outputs via parametric world knowledge, spurious correlated cues, or by simply ignoring contradictory or irrelevant memory entries (Zhou et al., 2023; Xie et al., 2024; Du et al., 2024; Yoran et al., 2024; Amiraz et al., 2025). Consequently, memory

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.

policies with incompatible evidence-dependence patterns can receive identical reward signals on observed examples but fail catastrophically once decisive evidence changes. This motivates us to study memory-policy behavior that properly resolves memory conflicts, maintains stable inference under non-critical context edits, and abstains when valid and sufficient evidence is absent.

Prior work assumes better answers imply faithful evidence use, but shortcut learning invalidates this. Figure 1 contrasts two dominant memory-policy learning paradigms. Outcome-supervised policies reward selections yielding correct answers (Yan et al., 2026; Yu et al., 2026; Zhou et al., 2026). They neither verify evidence reliance nor prevent shortcuts. Inference-time intervention perturbs context to diagnose evidence reliance (Chang et al., 2025; Liu et al., 2025). It only probes after training without updating the memory policy. To address these limitations, this work leverages validated counterfactual twins to encode the three target properties as trainable constraints. They are invariance under irrelevant perturbation, sensitivity to decisive-evidence modification, and evidence-driven abstention. Given a history, we build non-causal twins altering nuisance content while preserving decisive-evidence roles, and causal twins removing, replacing, contradicting or updating such evidence. Because supervision hinges on evidential manipulation, multi-dimensional validation filters invalid samples pre-training. This principle underpins CIT-MEM, translating offline interventions into direct supervision for evidence-grounded behavior with frozen storage backend and reader.

Building on this principle, we present CIT-MEM, a counterfactual training framework for evidence-grounded memory policies in long-horizon LLM agents. Unlike prior work optimizing only final-answer accuracy, it focuses on how memory policies utilize decisive evidence. It builds validated non-causal and causal twins and applies strict multi-dimensional filtering to discard unreliable counterfactual transformations before training. Leveraging shared-support twin alignment, our multi-head objective jointly supervises memory operations, evidence retrieval, and answer-aware abstention, enforcing expected policy behaviors (Wen et al., 2025). All counterfactual processing occurs exclusively at training time with zero deployment overhead. To further optimize training efficiency, this work adopts Adaptive Counterfactual Replay (ACR) as a lightweight training-time scheduler. It neither modifies the core objective function nor participates in inference (Schapiro et al., 2018; Mattar & Daw, 2018). As an auxiliary mechanism, it concentrates replay on unresolved twin constraints while retaining broad coverage. Our contributions are summarized as follows:

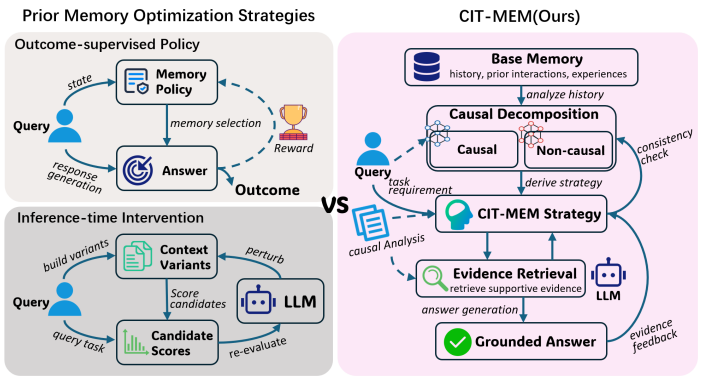


Figure 1: Left: outcome-supervised policy (top) and inference-time intervention (bottom). Right: CIT-MEM learns causal strategies for evidence retrieval and grounded answering.

- We formulate decisive-evidence selection among query-related memories as a memory-policy problem and identify a fundamental limitation of answer-only supervision: policies with incompatible evidence dependencies can obtain the same answer loss. We define reliability through invariance, sensitivity, and evidence-aware abstention.
- We introduce validated non-causal and causal twins that directly target controlled history changes. Structural, evidence-provenance, recoverability, and answer-direction checks reject unreliable transformations before training, including causal removals leaving sufficient alternative evidence.
- We develop an aligned multi-head objective that compares twin behaviors on shared support and jointly trains memory operations, evidence retrieval, and answerability. It converts counterfactual diagnostics into parameter updates without modifying the fixed storage backend or reader.
- We propose ACR to allocate a fixed update budget toward unresolved twin constraints while preserving uniform coverage. All counterfactual processing remains training-only; at deployment, the controller selects at most K records in one pass and invokes the reader at most once.

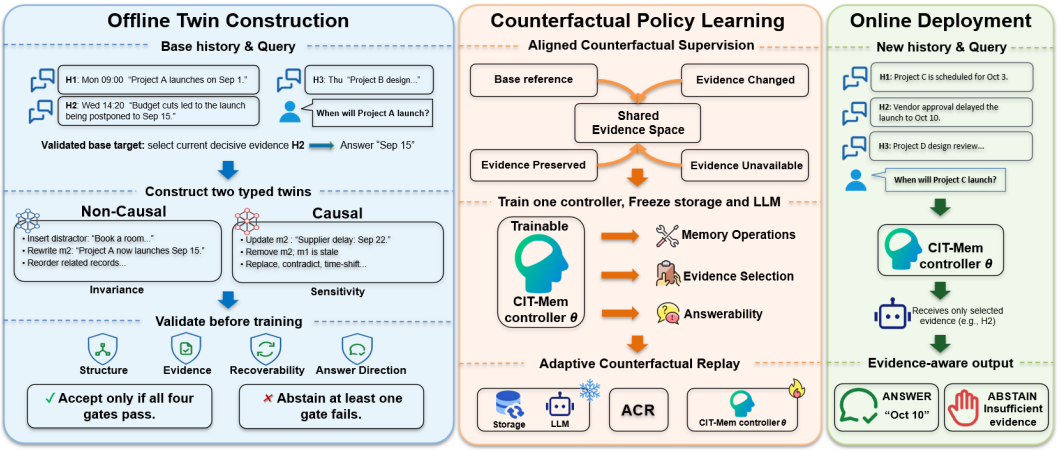


Figure 2: The figure illustrates how CIT-Mem learns evidence-grounded memory behavior from validated counterfactual twins and performs efficient one-pass inference with Top-K evidence selection and reliable abstention.

2 RELATED WORK

Memory systems determine what information is reachable. External memory controls recording, indexing, and cross-session reachability. SeCom and LightMem improve segmentation and maintenance (Pan et al., 2025; Zhang et al., 2026b); graph memories connect records missed by pairwise retrieval (Xu et al., 2025; Gutiérrez et al., 2024; Anokhin et al., 2025); and MemGuide/RMM condition access on intent or citations (Du et al., 2026; Tan et al., 2025). Long-context input remains position-sensitive (Liu et al., 2024). Overall, memory systems improve reachability and relevance but cannot ensure retrieved evidence remains current, directly supports queries, and is sufficient under distraction, temporal updates, or knowledge conflict (Amiraz et al., 2025; Fatemi et al., 2025; Xie et al., 2024).

Task outcomes provide indirect supervision. Memory-R1 learns operations from frozen-reader correctness; AgeMem propagates terminal reward through a tool policy (Yan et al., 2026; Yu et al., 2026). ReasoningBank uses trajectories; Memory-as-Action and MEM1 supervise context editing and recurrent state (Ouyang et al., 2026; Zhang et al., 2026c; Zhou et al., 2026). Yet answer reward cannot expose evidence dependence: accuracy may arise from redundant evidence, nuisance correlations, or parametric answering (Zhou et al., 2023; Du et al., 2024). Overall, outcome-supervised policies improve accuracy but cannot determine whether predictions depend on selected evidence.

Interventions identify evidence dependence. Attribution and false-premise probes diagnose context use without training policies (Cohen-Wang et al., 2024; Enouen et al., 2024; Chang et al., 2025; Ong et al., 2025). Counterfactual interventions can train policies efficiently (Kaushik et al., 2020; Geiger et al., 2022; Liu et al., 2025). CMI scores candidates at inference by comparing a frozen reader under no-, observed-, and perturbed-memory conditions (Srivastava, 2026); it trains neither writes nor abstention. In contrast, CIT-MEM validates typed twins offline and supervises invariance to evidence-preserving edits, directional response to causal edits, and abstention without sufficient evidence. Pairs update parameters with fixed backend and reader. Both directions matter: ignoring memory satisfies invariance, whereas sensitivity alone permits instability. Because reader confidence poorly indicates evidence sufficiency (Wen et al., 2025; Soudani et al., 2025), ACR prioritizes recent policy violations while retaining coverage (Schaul et al., 2016; Schapiro et al., 2018; Mattar & Daw, 2018). Overall, intervention-based approaches either remain post-hoc diagnostics or omit joint training of memory operations, evidence selection, and evidence-aware abstention.

3 METHODS

We train only the memory policy, freezing the storage executor, retrieval backend, and reader. It combines online operations and budgeted selection with validated evidence-preserving and evidence-

changing twins aligned on shared support. Twins supervise invariance and sensitivity alongside base loss; Adaptive Counterfactual Replay (ACR) schedules pairs without objective or deployment changes.

Policy formulation and supervision gap. Given $\mathcal{H}_T = (e_1, \ldots, e_T)$, the controller observes only the prefix $\mathcal{F}_t = \sigma(e_1, \ldots, e_t)$ and samples $a_t \in \{\textsc{Write}, \textsc{Update}, \textsc{Delete}, \textsc{NoOp}\}$ from its operation head. The fixed executor W_B applies it:

$$\mathcal{M}_t = W_B(\mathcal{M}_{t-1}, e_t, a_t). \tag{1}$$

A frozen scorer ρ_B exposes at most C live records, and the controller adds a trainable residual r_i. κ deterministically groups records only when canonical entity and relation match, validity intervals are compatible, and normalized texts bidirectionally entail; topical overlap is insufficient. Section 3.3 specifies class scoring, capped selection, and answerability. Only θ is trainable; the evidence budget is structural (Lewis et al., 2020).

Answer loss alone cannot identify evidence dependence. Two selectors may retrieve different evidence while the frozen reader produces the same correct answer:

$$f_\phi(q, Z_{\theta_1}) = f_\phi(q, Z_{\theta_2}) = y \;\not\Rightarrow\; Z_{\theta_1} = Z_{\theta_2}. \tag{2}$$

We therefore supervise the full policy signature g_θ, comprising operation, retrieval, and decision heads. A reliable policy remains stable after non-causal edits but moves toward validated targets after causal edits. Without sufficient evidence, the target is abstention rather than a confidence heuristic (Wen et al., 2025; Soudani et al., 2025).

3.1 Offline Twin Construction

Figure 2 (left) constructs evidence-preserving ($r = \mathrm{nc}$) and evidence-changing ($r = \mathrm{c}$) twins from each base $x = (\mathcal{H}, q, y, \mathcal{E}^{\mathrm{ann}}, b^\star)$:

$$x^r = \big(\mathcal{I}_r(\mathcal{H}), q, y^r, \mathcal{E}^{\mathrm{ann},r}, b^r_{\mathcal{I}}\big), \qquad r \in \{\mathrm{nc}, \mathrm{c}\}. \tag{3}$$

Non-causal twins transfer y^r, evidence annotations, and target behavior through $F_{\mathcal{I}}$; they may insert irrelevant events, paraphrase, or duplicate records without changing decisive evidence. Causal edits delete, replace, contradict, update, or temporally shift decisive evidence, requiring a new target. A causal twin is retained only if its validated target differs from the base on a shared head (Kaushik et al., 2020; Wu et al., 2025; Maharana et al., 2024; Hu et al., 2026).

Operational validation and audit. Targets are fixed before training, independent of θ. Let $D(x)$ denote decisive evidence roles and $\mathcal{S}_K(x)$ contain candidate subsets of at most K from which validator v produces the target but fails after matched removal of decisive members. This bounded test covers only the C exposed candidates, not the unexposed history. For transformation $\mathcal{I}_r$, define $C_{\mathrm{nc}} := [D(x^{\mathrm{nc}}) = F_{\mathcal{I}}(D(x))]$ and $C_{\mathrm{c}} := [D(x^{\mathrm{c}}) \neq F_{\mathcal{I}}(D(x))]$. The gates are

$$\begin{aligned} &G_{\mathrm{str}} = \mathbb{I}[\mathrm{Legal}(x^r)], && G_{\mathrm{evid}} = \mathbb{I}[\mathrm{Prov}(\mathcal{I}_r, x) \wedge \neg\,\mathrm{Leak}(\mathcal{I}_r, x)]\,\mathbb{I}[C_r], \\ &G_{\mathrm{rec}} = \mathbb{I}[b^r_{\mathcal{I}} \neq \textsc{Abstain} \;\vee\; \mathcal{S}_K(x^r) = \varnothing], && G^{(v)}_{\mathrm{ans}} = \mathbb{I}[\mathrm{Rel}_v(x, x^r) = R_r]\,\mathbb{I}[\mathrm{Ctrl}_v(x^r) = 1]. \end{aligned} \tag{4}$$

Legal checks schema, chronology, identifiers, and write/update/delete preconditions; Prov requires generated spans to map to a declared source; and Leak detects a target copied outside the declared edit. R_r specifies the same-answer relation for a non-causal twin and the changed-answer or abstention relation for a causal twin. Ctrl_v requires the answer to fail after decisive-evidence removal, preventing parametric knowledge from making irrelevant context appear sufficient (Zhou et al., 2023). Thus G_{rec} rejects residual sufficient evidence or duplicate recovery paths, while $G^{(v)}_{\mathrm{ans}}$ rejects unstable base answers and changes in the wrong direction.

A twin is retained only if every gate passes:

$$V^{(v)}(\mathcal{I}_r, x) = G_{\mathrm{str}} G_{\mathrm{evid}} G_{\mathrm{rec}} G^{(v)}_{\mathrm{ans}} = 1. \tag{5}$$

Gates are evaluated in order, logging the first failure. We report family-level acceptance, first-failure, and accepted answer-change/abstention distributions. Because $G^{(v)}_{\mathrm{ans}}$ queries v, validation is reader-conditioned. Reader-dependent validation is restricted to the training split; Section 4.5 tests leave-one-validator-out transfer to unseen validators.

Independent human audit. We conducted an independent stratified audit of 100 *accepted, held-out* twins covering every intervention family and both changed-answer and abstention targets (Appendix C.1). Two annotators blinded to the validation reader and policy output independently judged (i) structural legality, (ii) preservation or change of the declared decisive evidence, (iii) absence of target leakage, (iv) no residual sufficient evidence for abstention targets, and (v) correctness of the target answer relation; a third annotator resolved disagreements. The audit yielded a 92.0% *Human-Valid* rate, the fraction satisfying all five criteria, with eight label errors after adjudication and Cohen's κ of 0.78. These adjudicated labels provide independent evidence of twin validity for the audited subset; validator agreement alone is not treated as ground-truth twin validity.

3.2 Counterfactual Policy Learning

Figure 2 (center) aligns twins on union support $\mathcal{K}_{\mathcal{I}}$ with nuisance class $\perp$ and no-evidence class $\varnothing$. The map $P_{\mathcal{I}}$ permutes equivalent coordinates and zero-pads missing ones without changing normalization. A shared backbone, optionally LoRA-adapted (Hu et al., 2022), produces three policy heads; Appendix A.1 gives the adapter parameterization and trainable parameter count. For behavior b under input x', each shared head h supplies normalized log-likelihood s_h; their weighted mean is

$$s_\theta(b \mid x') = \frac{\sum_{h\in\Omega_{\mathcal{I}}} \widetilde{\mu}_h s_h(b \mid x')}{\sum_{h\in\Omega_{\mathcal{I}}} \widetilde{\mu}_h}. \tag{6}$$

$\Omega_{\mathcal{I}}$ contains only shared operation, retrieval, and decision heads, weighted by $\widetilde{\mu}_h > 0$. Operation scores average target-operation log-probabilities, retrieval scores use aligned evidence membership, and decision scores use the target decision. This removes cardinality and positional mismatches.

Base supervision jointly trains the three heads:

$$\mathcal{L}_{\text{base}} = \lambda_{\text{op}}\mathcal{L}_{\text{op}} + \lambda_{\text{set}}\mathcal{L}_{\text{set}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}. \tag{7}$$

For an aligned J-class universe, let $z_S \in \{0,1\}^J$ encode $S \in \mathcal{S}_K(x)$. The retrieval term is a soft minimum over separately validated supports,

$$\mathcal{L}_{\text{set}}(x) = -\frac{1}{\beta}\log\frac{1}{|\mathcal{S}_K(x)|}\sum_{S\in\mathcal{S}_K(x)} e^{-\beta\ell_{\text{BCE}}(p_\theta, z_S)}, \tag{8}$$

where ℓ_{BCE} is class-balanced binary cross-entropy and $\beta > 0$. This accepts alternative valid supports without merging incompatible sets into an unvalidated hybrid. If $\mathcal{S}_K(x) = \varnothing$, set and decision heads target $\varnothing$ and ABSTAIN, respectively.

For a non-causal twin, $\mathcal{L}_{\text{inv}}$ is the normalized ω_h-weighted Jensen–Shannon divergence after alignment; mask $m_h(\mathcal{I})$ excludes unalignable heads, so different sample spaces are never compared. For a causal twin, let $b^{\text{c}}_{\mathcal{I}}$ be the validated target and Δs_θ its score advantage over the hardest incompatible behavior. Sensitivity is trained by

$$\mathcal{L}_{\text{sens}} = -s_\theta(b^{\text{c}}_{\mathcal{I}} \mid x^{\text{c}}) + \lambda_{\text{mar}}[\delta_{\text{c}} - \Delta s_\theta]_+. \tag{9}$$

The first term raises the validated target's probability, while the hinge enforces a directional margin instead of merely maximizing an unlabeled distance. Appendix A.2 details the roles and interactions of the associated hyperparameters.

Training objective and adaptive replay. Let $\mathcal{R}_0$, $\mathcal{R}_{\text{inv}}$, and $\mathcal{R}_{\text{sens}}$ denote the expected base, invariance, and sensitivity losses. The latter two arise as penalties for the dataset-level reliability constraints, giving the complete objective

$$\mathcal{L}_{\text{CIT}} = \mathcal{R}_0 + \lambda_{\text{inv}}\mathcal{R}_{\text{inv}} + \lambda_{\text{sens}}\mathcal{R}_{\text{sens}}. \tag{10}$$

ACR assigns non-causal and causal pools separate budgets, prioritizing larger detached moving-average losses (Schaul et al., 2016). Uniform sampling preserves coverage; reciprocal-probability weights keep the estimator unbiased; ACR changes allocation, not Equation 10, in expectation. Appendix A.3 specifies moving-average updates, sampling distribution, and importance correction.

3.3 ONLINE DEPLOYMENT

The right panel of Figure 2 shows online inference. After κ groups records, class k uses maximum member score u_k and representative $i_k^\star$, tie-broken by latest valid timestamp then identifier. Classes with $u_k \geq u_\varnothing$ form $\mathcal{K}_\theta^+$; the controller selects at most K representatives:

$$Z_\theta(x) = \left\{ m_{i_k^\star} : k \in \mathrm{TopK}_K(\mathcal{K}_\theta^+; u_k) \right\}. \tag{11}$$

For $a_\theta(x) = p_\theta(\textsc{Answer} \mid x)$, output is

$$\widehat{y}_\theta(x) = \begin{cases} f_\phi(q, Z_\theta(x)), & a_\theta(x) \geq \tau_{\mathrm{abs}}, \\ \textsc{Abstain}, & a_\theta(x) < \tau_{\mathrm{abs}}. \end{cases} \tag{12}$$

Increasing τ_{abs} is more conservative and cannot increase coverage. Deployment uses no twins or replay: one controller pass calls the frozen reader once for ANSWER and never for ABSTAIN, rather than $O(C)$ times.

4 EXPERIMENTS

We conduct experiments to answer the following research questions (RQs), progressing from overall effectiveness to attribution, mechanism, operating conditions, and generalization:

RQ1: Does CIT-MEM boost answer quality across datasets and readers with compact evidence?

RQ2: Which components explain gains in RQ1 and are their contributions consistent across datasets?

RQ3: Do the components identified in RQ2 improve the invariance, causal-sensitivity, and optimization behaviors they are designed to govern?

RQ4: How do the evidence budget K and abstention threshold τ_{abs} trade off answer quality, selective reliability, coverage, and context cost?

RQ5: With the operating point fixed, do the learned behaviors transfer across validation readers rather than encode validator-specific preferences?

4.1 EXPERIMENTAL SETTINGS

Datasets. LoCoMo-10 contains ten long multi-session conversations; we evaluate its single-hop, multi-hop, temporal, and open-domain questions (Maharana et al., 2024). **LongMemEval-S** contains 500 questions over timestamped multi-session histories; we evaluate its knowledge-update, multi-session, temporal-reasoning, and three single-session categories (Wu et al., 2025).

Baselines. We compare CIT-MEM with the six memory systems in Tables 1 and 2. **A-MEM** builds linked structured notes and updates their attributes as the graph evolves (Xu et al., 2025). **Mem0** extracts compact records and retrieves them through a scalable long-term-memory pipeline (Chhikara et al., 2025). **MemoryOS** coordinates short-, mid-, and long-term personal-memory tiers (Kang et al., 2025). **MemOS** treats plaintext, activation, and parameter memory as unified system resources (Li et al., 2025). **TiMem** consolidates conversations in a temporal memory tree and performs complexity-aware recall (Li et al., 2026). **CMI-Mem** selects records by their estimated causal effect on the answer rather than by similarity alone (Srivastava, 2026).

Metrics. LoCoMo-10. We report token-level **F1/BLEU-1** between normalized predictions and references across four categories, plus **Token**, mean retrieved context per question (thousands). All-query scores retain every query and assign zero to abstentions before category-seed aggregation. Section 4.4 reports **answered-only F1**, **answer coverage**, and **abstention precision (Abst-P)**: answered-query quality, answered fraction, and appropriate-abstention fraction under the evidence-sufficiency criterion. Appendix D.1 details accounting. **LongMemEval-S.** Under the official protocol (Wu et al., 2025), a category-specific LLM judge assigns binary correctness using reference or preference rubrics. **LLJ accuracy** is per-category accuracy; **Avg.** is the category macro-average in Table 2. Appendix D.3 also reports per-question accuracy from identical outputs. Frozen submissions cover every question; controller abstentions receive the same rubric as answers, preserving baseline comparability. Higher scores and lower Token are better.

Table 1: LoCoMo-10 results. F1/BLEU-1 measure quality; Token averages retrieved context. Bold/underline mark reader-wise top-two. Methods share protocols/seeds; abstentions score zero. Default CIT-MEM uses $K = 5$ and $\tau_{\text{abs}} = 0.3$; Appendix D.1 reports GPT-4o-mini coverage and answered-only F1.

LoCoMo-10										
Model	Method	Category								Avg.
		Single Hop		Multi Hop		Temporal		Open Domain		
		F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	Token
GPT-4o-mini	A-MEM	48.01	45.76	26.35	20.67	40.10	37.26	14.00	12.71	3.46
	Mem0	53.24	48.47	29.09	25.78	39.31	33.61	25.56	23.19	8.92
	MemoryOS	33.10	29.61	20.03	18.44	27.84	23.18	9.77	8.47	4.15
	MemOS	57.42	51.65	36.86	32.66	49.85	42.62	36.95	32.30	5.24
	TiMem	54.68	49.75	31.24	27.70	**53.20**	**48.10**	28.90	26.20	3.88
	CMI-Mem	62.15	51.90	32.55	28.85	43.10	36.90	36.40	31.80	9.65
	CIT-MEM(ours)	**64.87**	**57.10**	**41.07**	**36.46**	51.68	47.75	**40.27**	**36.66**	2.96
Qwen3-8B	A-MEM	40.59	35.19	25.99	24.16	31.58	29.60	18.51	14.93	3.39
	Mem0	49.85	46.92	30.41	28.86	36.39	34.70	29.89	25.74	9.03
	MemoryOS	32.86	29.23	18.15	14.63	27.52	24.45	10.47	9.20	4.17
	MemOS	53.45	47.90	34.10	30.20	46.30	39.55	34.55	31.15	5.31
	TiMem	51.20	46.64	29.15	25.85	**52.15**	**48.90**	27.30	24.80	3.92
	CMI-Mem	59.91	51.30	30.40	26.95	40.25	34.45	36.40	31.10	9.72
	CIT-MEM(ours)	**63.17**	**55.03**	**38.79**	**37.05**	50.36	47.42	**40.21**	**34.89**	2.98
Llama3.2-3B	A-MEM	21.45	18.90	16.72	14.12	22.52	18.04	10.89	9.50	3.11
	Mem0	28.46	25.58	23.09	19.23	31.71	29.20	17.76	14.03	8.28
	MemoryOS	29.08	25.92	17.33	14.27	25.58	22.94	9.32	8.21	3.86
	MemOS	41.68	36.30	26.75	23.55	36.80	31.40	27.20	24.65	5.02
	TiMem	38.90	34.31	23.40	20.30	**45.05**	**40.68**	21.85	19.70	3.79
	CMI-Mem	44.10	36.45	24.15	20.85	31.05	26.55	27.40	22.15	9.48
	CIT-MEM(ours)	**47.28**	**39.07**	**31.14**	**25.60**	43.71	39.69	**28.15**	**26.09**	2.86

Implementation Details. All methods use GPT-4o-mini, Qwen3-8B, and Llama 3.2 3B readers with a fixed Qwen3-Embedding-8B encoder, answer prompt, and fact extractor. Dense retrieval exposes 50 candidates; budget-matched methods select at most five. Defaults are $K = 5$ and $\tau_{\text{abs}} = 0.3$; Section 4.4 evaluates their trade-offs. Decoding uses temperature 0, disabled thinking where supported, and a 64-token limit. The controller uses Adam for up to ten epochs (learning rate 10^{-4}, batch size 8; seeds 42, 43, and 44). Appendix A.3 gives the complete configuration. LoCoMo-10 uses conversation-level cross-validation; frozen checkpoints and thresholds transfer to LongMemEval-S without tuning. Training-disjoint twins are intervention-stratified; prompts, validation rules, and comparison thresholds are fixed on training folds before held-out scoring.

4.2 Main Results and Analysis (RQ1)

4.2.1 LoCoMo-10 Results and Analysis

Table 1 shows CIT-MEM delivers the best quality and most compact retrieval across three readers. It leads 18 of 24 comparisons—F1/BLEU-1 on single-hop, multi-hop, and open-domain questions—while TiMem alone leads temporal questions. All-query F1 beats the strongest baseline by $+4.2$ to $+6.0$ points. Zero-scored abstentions prevent gains from dropping difficult queries; Appendix D.1 reports coverage and answered-only quality. At 2.86–2.98K tokens, CIT-MEM is lowest in every reader block and at least 8% below the next-cheapest baseline. This answers RQ1: cross-reader gains come from concentrating decisive records, not more input.

4.2.2 LongMemEval-S Results and Analysis

Table 2 confirms transfer without target tuning. CIT-MEM leads every reader's Avg. LLJ and four categories (TR, KU, SSP, and reader-dependent SSU/SSA), beating TiMem by $+3.2$ to $+4.3$ points; TiMem leads MS. Both aggregations preserve the ranking despite category imbalance (Appendix D.3). This answers RQ1: counterfactual training yields compact, evidence-selective policies across datasets and readers. Fixed checkpoints and LoCoMo token results exclude tuning and evidence budgets.

Table 2: LongMemEval-S LLJ accuracy: TR/MS/KU; SSU/SSA/SSP (user/assistant/preference). Avg. is the category macro-average; Appendix D.3: same-submission per-question accuracy. Bold/underline: reader-column top-two.

LongMemEval-S								
		Category						Avg.
Model	Method	TR LLJ	MS LLJ	KU LLJ	SSU LLJ	SSA LLJ	SSP LLJ	LLJ
GPT-4o-mini	A-MEM	36.09	40.30	72.82	82.86	87.50	39.33	59.82
	Mem0	49.17	66.17	78.72	94.29	51.79	50.00	65.02
	MemoryOS	53.38	44.81	56.15	81.14	78.18	51.33	60.83
	MemOS	65.11	58.80	76.67	93.71	67.86	50.67	68.80
	TiMem	68.42	**70.83**	86.16	**95.71**	82.14	63.33	77.77
	CMI-Mem	61.20	55.40	79.60	90.25	89.20	58.45	72.35
	CIT-MEM(ours)	**72.59**	69.17	**90.56**	94.57	**92.79**	**70.07**	**81.63**
Qwen3-8B	A-MEM	34.83	41.40	72.36	79.77	83.47	38.04	58.31
	Mem0	47.30	62.23	78.41	88.51	50.68	50.52	62.94
	MemoryOS	49.22	44.01	53.13	82.73	74.59	52.13	59.30
	MemOS	62.11	57.48	76.55	86.10	62.15	46.40	65.13
	TiMem	67.21	**69.81**	84.37	93.90	80.63	61.75	76.28
	CMI-Mem	58.32	54.27	76.85	87.45	89.92	58.10	70.82
	CIT-MEM(ours)	**71.03**	67.11	**88.75**	**94.85**	89.03	**66.22**	**79.50**
Llama3.2-3B	A-MEM	31.65	33.45	61.36	64.80	69.84	25.19	47.72
	Mem0	37.61	50.25	57.08	69.18	33.89	43.16	48.53
	MemoryOS	40.27	34.65	48.69	61.71	57.29	44.59	47.87
	MemOS	48.90	44.62	60.75	70.23	41.35	38.47	50.72
	TiMem	52.86	**65.31**	68.46	**79.12**	74.75	51.92	65.40
	CMI-Mem	47.15	46.84	65.20	77.45	76.61	50.25	60.58
	CIT-MEM(ours)	**61.47**	63.95	**72.59**	77.88	**81.19**	**60.84**	**69.65**

4.3 Ablation Study (RQs 2–3)

4.3.1 Component Effectiveness (RQ2)

We test four variants: *no twins* removes twin supervision and ACR; single-loss variants drop λ_{inv} or λ_{sens}; uniform replay replaces ACR but retains both objectives. All share identical splits, backend, reader, evidence/update budgets, optimizer, and seeds. Figure 3 shows consistent LoCoMo-10 F1 and LongMemEval-S LLJ ordering. Versus no twins, CIT-MEM gains $+28.9$ and $+25.2$ points, respectively, leading every category. Single-loss variants improve yet trail their combination, confirming complementarity. Uniform replay is the strongest ablation: objectives drive gains, while ACR adds benefit. Thus, counterfactual training, not one question type, explains the RQ1 advantage and answers RQ2. Appendix B separates twin data, answer-direction validation, paired objectives, and replay under matched exposure and compute.

Table 3: Three-seed results on held-out LoCoMo-10 twins (GPT-4o-mini). IS/CDA are percentages; loss ablations test objectives; Epoch@95 isolates ACR optimization.

Method	IS ↑	CDA ↑	AURC ↓	Epoch@95 ↓
No twins	78.6	58.4	0.184	7.2
w/o invariance loss	82.1	70.9	0.145	7.5
w/o sensitivity loss	88.7	56.1	0.179	7.1
w/o ACR (uniform replay)	90.9	72.1	0.141	8.8
CIT-MEM	**93.6**	**75.8**	**0.129**	**6.4**

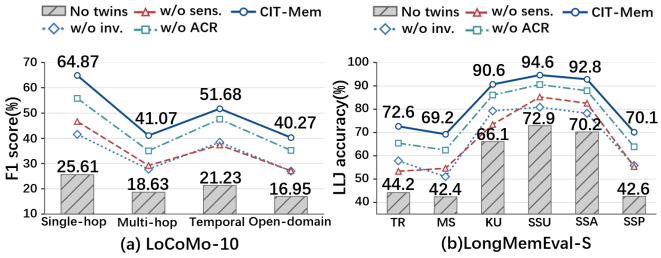


Figure 3: GPT-4o-mini component ablations: (a) LoCoMo-10 all-query F1; (b) LongMemEval-S LLJ accuracy. Panels share variant encodings/setup.

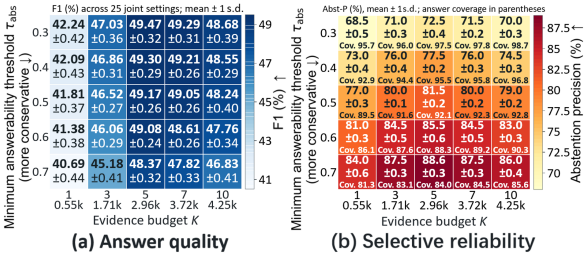


Figure 4: LoCoMo-10/GPT-4o-mini, K–τ_{abs}: (a) zero-scored all-query F1; (b) Abst-P (coverage). Three-seed means ± SD; context lengths below.

Table 4: LoCoMo-10 answered-only F1 (%; GPT-4o-mini). Rows: τ_{abs}; columns: K; parentheses: mean context. One frozen checkpoint separates quality/coverage; Figure 4(a): all-query F1.

τ_{abs}	K=1 (0.55k)	K=3 (1.71k)	K=5 (2.96k)	K=7 (3.72k)	K=10 (4.25k)
0.3	44.23	48.99	51.14	50.50	49.67
0.4	45.50	49.59	52.06	51.42	50.57
0.5	46.72	51.12	53.77	53.14	51.98
0.6	48.12	52.64	55.46	54.62	53.07
0.7	50.23	54.43	57.58	56.59	55.09

4.3.2 Behavioral Alignment and Replay Efficiency (RQ3)

Table 3 verifies the intended mechanisms: invariance most affects IS; sensitivity most affects CDA and calibration. Each objective targets its behavior, not generic regularization. The full model reaches 93.6 IS/75.8 CDA; uniform replay needs 8.8 versus 6.4 epochs. This identifies ACR as an optimization scheduler, not an inference-time evidence source. Thus, complementary losses and ACR improve training efficiency, answering RQ3 and motivating RQ4.

4.4 Hyperparameter Analysis (RQ4)

We vary K and τ_{abs} on LoCoMo-10 (GPT-4o-mini); using fixed components/outputs isolates deployment choices. Table 4 and Figure 4 report answered-only/all-query F1, Abst-P, coverage, and context cost. F1/Abst-P favor $K = 5$: smaller budgets omit support; larger ones add quality-reducing context. Higher τ_{abs} raises answered-only F1/Abst-P but lowers coverage/all-query F1. At $K = 5$, answered-only F1 rises 6.44 points as coverage falls 13.5 points. $(K, \tau_{\text{abs}}) = (5, 0.3)$ yields 49.47 all-query F1 and 97.5% coverage with zero-scored abstentions. K controls evidence efficiency; τ_{abs} trades reliability for coverage without retraining, answering RQ4.

4.5 Cross-Validator Generalization (RQ5)

$G_{\text{ans}}^{(v)}$ may overfit validator-specific patterns (Zhou et al., 2023). LOVO trains on one reader's twins and tests others' accepted twins. Metrics: invariance score (IS), causal-direction accuracy (CDA), AURC, and abstention precision (Abst-P); matched controls isolate validation/training. CDA: -3.3 versus matched; $+10.4$ versus unvalidated. Robustness persists with weaker Llama validation. Removing G_{ans} outweighs changing validators; answer-only remains weaker despite stable F1. Validation filters misleading twins; counterfactual training transfers four cross-validator behaviors, answering RQ5.

Table 5: Held-out LoCoMo-10 twins: LOVO nears matched validation, beating controls across validators.

Condition	IS ↑	CDA ↑	AURC ↓	Abst-P ↑
Answer-only (no twins)	78.2	57.9	0.186	60.4
Unvalidated (w/o G_{ans})	84.9	62.1	0.165	67.8
LOVO, val.: GPT-4o-mini	92.0	72.4	0.139	79.1
LOVO, val.: Qwen3-8B	92.7	73.4	0.135	80.1
LOVO, val.: Llama 3.2 3B	91.6	71.8	0.140	78.5
LOVO avg.	92.1	72.5	0.138	79.2
Matched val. (Table 3)	**93.6**	**75.8**	**0.129**	**81.4**

5 Conclusions

Validated non-causal and causal twins train evidence-grounded CIT-Mem policies to remain invariant to evidence-preserving edits, respond to decisive-evidence changes, and abstain when sufficient evidence is unavailable. The objectives jointly supervise memory operations, retrieval, and answerability with fixed backends and readers. Across the evaluated datasets and readers, CIT-Mem achieves strong overall answer quality with compact evidence, reliable abstention, and no added deployment overhead, while transferring effectively across diverse evaluation settings. It remains robust to stale, conflicting, distracting, or insufficient evidence. Future work will further reduce validator dependence by expanding human audits and strengthening validation for complex causal interventions.

AI USE STATEMENT

Generative-AI tools were adopted to assist parts of this research workflow. AI was used to support literature survey and preliminary literature-related organization; nevertheless, literature screening, relevance judgment and all research insights, core hypotheses, and paper arguments were conceived independently by the authors. The logical skeleton of the whole codebase was manually designed by authors, while generative AI provided auxiliary support for concrete code implementation. Every AI-generated code snippet was manually inspected, debugged, and verified through repeated experimental runs before integration. The manuscript draft was originally written by authors; generative-AI tools were only leveraged for language polishing, expression refinement and grammatical correction, without generating core claims, experimental interpretations or new arguments. All AI-suggested textual revisions were reviewed, filtered and adjusted by authors. The authors take full and final responsibility for all content, claims, experimental results, code artifacts and conclusions presented in this submission.

ETHICS STATEMENT

This work focuses on retrieval-augmented memory for large language models. No human subjects, private personal data, or sensitive information are involved in our experiments. We use only publicly released datasets, and all data comply with their original licensing terms. The research does not introduce discriminatory or biased modeling objectives, nor does it develop techniques intended for harmful applications. We identify no ethical risks, privacy concerns, or conflicts of interest associated with this submission. The authors confirm that all methodological choices and experimental evaluations adhere to standard research integrity norms.

REPRODUCIBILITY STATEMENT

To support reproducibility of the proposed framework, partial key source code has been deposited in an anonymous repository for double-blind review; full implementation materials covering model training, counterfactual sample construction, replay scheduling, inference pipelines and evaluation workflows will be officially released upon paper publication. Key hyperparameters, adaptation configurations, optimizer choices, random seeds and deployment-stage control parameters are comprehensively documented within the paper appendices. All experiments are performed on publicly available benchmarks following their official evaluation protocols, including metric-computation conventions, validation pipelines for counterfactual samples, and full prompt templates for model intervention and external judging. Throughout our experiments, we adopt fixed pre-trained backbone models without target-dataset fine-tuning. All prompt designs for sample generation, memory processing and performance evaluation are formally recorded in the appendix. Ablation studies, hyperparameter analysis, generalization and transfer experiments can be reproduced under the documented hardware environment. Reported numerical metrics are averaged across multiple repeated runs with distinct random seeds, and detailed raw experimental numbers are presented in appendix tables for reference.

REFERENCES

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18228–18258. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.892/.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. AriGraph: Learning knowledge graph world models with episodic memory for LLM agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 12–20, 2025. doi: 10.24963/ijcai.2025/2. URL https://www.ijcai.org/proceedings/2025/2.

Yurui Chang, Bochuan Cao, Yujia Wang, Jinghui Chen, and Lu Lin. JoPA: Explaining large language model's generation via joint prompt attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22106–22122. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1074. URL https://aclanthology.org/2025.acl-long.1074/.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/adbea136219b64db96a9941e4249a857-Abstract-Conference.html.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120, 2024. doi: 10.1145/3596490. URL https://dl.acm.org/doi/10.1145/3596490.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z. Pan, Ruifeng Xu, and Kam-Fai Wong. MemGuide: Intent-driven memory selection for goal-oriented multi-session LLM agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30584–30592, 2026. doi: 10.1609/aaai.v40i36.40313. URL https://ojs.aaai.org/index.php/AAAI/article/view/40313.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Ö. Arik, Yan Liu, and Tomas Pfister. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13984–14011. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.832. URL https://aclanthology.org/2024.findings-acl.832/.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb7295a8bc613b375726659c2ecd6f14-Abstract-Conference.html.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

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://arxiv.org/abs/2106.09685.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.05257.

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. 25961–25970. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.1318/.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.12434.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, Chengbao Liu, and Jie Tan. TiMem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. *arXiv preprint arXiv:2601.02845*, 2026. URL https://arxiv.org/abs/2601.02845.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025. URL https://arxiv.org/abs/2507.03724.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.15102.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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*, 2024. URL https://arxiv.org/abs/2402.17753.

Marcelo G. Mattar and Nathaniel D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11):1609–1617, 2018. doi: 10.1038/s41593-018-0232-z. URL https://www.nature.com/articles/s41593-018-0232-z.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 8631–8661. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.435. URL https://aclanthology.org/2025.naacl-long.435/.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jL7fwchScm.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

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. SeCom: On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e56f394bbd4f0ec81393d767caa5a31b-Abstract-Conference.html.

Anna C. Schapiro, Elizabeth A. McDevitt, Timothy T. Rogers, Sara C. Mednick, and Kenneth A. Norman. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, 9(1):3920, 2018. doi: 10.1038/s41467-018-06213-1. URL https://www.nature.com/articles/s41467-018-06213-1.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16596–16616. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.852/.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon LLM agents. *arXiv preprint arXiv:2605.17641*, 2026. URL https://arxiv.org/abs/2605.17641.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 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. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.413/.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.07174.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2502.12110.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.583/.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01558.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.acl-long.981/.

Dylan Zhang, Yanshan Lin, Zhengkun Wu, Yihang Sun, Bingxuan Li, Dianqi Li, and Hao Peng. Useful memories become faulty when continuously updated by LLMs. *arXiv preprint arXiv:2605.12978*, 2026a. URL https://arxiv.org/abs/2605.12978.

Jiaquan Zhang, Chaoning Zhang, Shuxu Chen, Zhenzhen Huang, Pengcheng Zheng, Zhicheng Wang, Ping Guo, Fan Mo, Sung-Ho Bae, Jie Zou, Jiwei Wei, and Yang Yang. Lightweight LLM agent memory with small language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12914–12929. Association for Computational Linguistics, 2026b. URL https://aclanthology.org/2026.acl-long.588.

Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. Memory as action: Autonomous context curation for long-horizon agentic tasks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19149–19164. Association for Computational Linguistics, 2026c. URL https://aclanthology.org/2026.findings-acl.956/.

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. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.15841.

Contents of Appendix

A CIT-Mem Training and Inference Procedure **16**

A.1 LoRA Parameterization 16

A.2 Hyperparameter Roles and Interactions 16

A.3 Adaptive Replay and Importance Correction 17

A.4 End-to-End Training and Deployment Algorithm 18

B Contribution Attribution with Matched Twins **18**

B.1 Matched Attribution Design 18

B.2 Attribution Results and Interpretation 19

C Independent Twin Audit Protocol and Results **20**

C.1 Independent Human Audit 20

C.2 Zero-Shot Transfer to Long-MT-Bench+ 20

D Scoring-Protocol Audits **21**

D.1 Abstention Coverage Accounting on LoCoMo-10 21

D.2 Threshold-Sweep Monotonicity 22

D.3 Average Results on LongMemEval-S 22

E Prompt Templates **23**

E.1 Intervention-Generation Prompts 23

E.2 Agent and Memory-Processing Prompts 24

E.3 External Evaluation Prompts 26

A CIT-MEM TRAINING AND INFERENCE PROCEDURE

This appendix provides the parameterization and sampling details underlying Algorithm 1, supplementing Sections 3.2 and 3.2.

A.1 LORA PARAMETERIZATION

For the optional LoRA adaptation in Section 3.2, a selected backbone projection $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is parameterized as

$$\begin{aligned} h_{\text{out}} &= W_0 h + \frac{\alpha_{\text{L}}}{r_{\text{L}}} B_{\text{L}} A_{\text{L}} \, \text{Dropout}_{p_{\text{L}}}(h), \\ A_{\text{L}} &\in \mathbb{R}^{r_{\text{L}} \times d_{\text{in}}}, \qquad B_{\text{L}} \in \mathbb{R}^{d_{\text{out}} \times r_{\text{L}}}. \end{aligned} \tag{13}$$

Here r_{L} is the adapter rank, $\alpha_{\text{L}}/r_{\text{L}}$ its scaling factor, and p_{L} the adapter-input dropout probability. Under LoRA, W_0 is frozen while the adapters and task heads receive gradients. For the set $\mathcal{T}_{\text{L}}$ of adapted projections, the trainable parameter count is

$$N_{\text{train}} = N_{\text{heads}} + \sum_{\ell \in \mathcal{T}_{\text{L}}} r_{\text{L}} \left(d_{\text{in}}^{(\ell)} + d_{\text{out}}^{(\ell)} \right), \tag{14}$$

where N_{heads} counts the operation, retrieval, and decision heads. This expression makes the memory cost of the selected projection modules and adapter rank explicit. Table 6 reports the concrete adapter placement and optimization configuration.

Parameter	Meaning	Value
b_{L}	Backbone adaptation mode	Cross-encoder LoRA
$\mathcal{T}_{\text{L}}$	Adapted projections	`q_proj`, `v_proj`; all attention blocks
r_{L}	LoRA rank	16
α_{L}	LoRA scaling coefficient	32
$\alpha_{\text{L}}/r_{\text{L}}$	Effective adapter scale	2.0
p_{L}	Adapter-input dropout	0.1
b_{bias}	Trainable LoRA bias	None
$(d_{\text{op}}, d_{\text{ret}}, d_{\text{dec}})$	Task-head hidden widths	$(256, 256, 128)$
ξ_{opt}	Optimization algorithm	Adam
η_{opt}	Initial learning rate	10^{-4}
B	Minibatch size	8
$E_{\max}$	Maximum training epochs	10
$\mathcal{S}_{\text{seed}}$	Repeated-run seeds	$\{42, 43, 44\}$

Table 6: LoRA configuration used for the controller backbone, including adapter placement, rank, scaling, dropout, trainable modules, and optimization settings.

Configuration rationale. The rank-16 adapters use $\alpha_{\text{L}} = 32$, giving an effective residual scale of 2.0. Restricting adaptation to the query and value projections in every self-attention block leaves the key/output projections, feed-forward layers, embeddings, and normalization parameters frozen. Only the low-rank matrices and the three policy heads are optimized, using the same optimizer, learning rate, batch size, epoch cap, and seeds as the main configuration in Table 9. This keeps the adaptation budget fixed across runs and attributes changes to the trained memory policy rather than to full-backbone fine-tuning.

A.2 HYPERPARAMETER ROLES AND INTERACTIONS

The losses and their coefficients are defined in Equations 7–10. Three interactions matter when specifying a configuration. First, λ_{sens} scales both the target likelihood and the margin penalty, so the latter has effective coefficient $\lambda_{\text{sens}}\lambda_{\text{mar}}$ in the total objective. In contrast, changing δ_{c} changes which score differences violate the margin. Its scale is tied to the normalized head score in Equation 6; replacing the head average with a sum would change the meaning of a fixed margin. Multiplying all $\widetilde{\mu}_h$ by the same positive constant leaves that score unchanged; the analogous property holds for the ω_h weights in the normalized invariance loss.

Second, β in Equation 8 controls concentration on the best-fitting validated support. As $\beta \to \infty$, the loss approaches the minimum support loss; as $\beta \to 0^+$, it approaches their arithmetic mean. This distinction matters when several sufficient supports exist. Third, λ_{inv} and λ_{sens} weight pool-level expectations. Each sampled pool must therefore be averaged separately: pooling all sampled pairs into one unweighted minibatch mean would additionally scale the two terms by their sampling proportions. The estimator below avoids that implicit change of objective. Table 7 reports the fixed values used in all runs.

Parameter	Meaning	Value
λ_{op}	Operation-loss weight	0.7
λ_{set}	Evidence-set-loss weight	1.2
λ_{dec}	Decision-loss weight	0.9
λ_{inv}	Invariance-loss weight	0.5
λ_{sens}	Sensitivity-loss weight	1.2
λ_{mar}	Margin-penalty weight	0.6
β	Soft-minimum inverse temperature	5.0
δ_{c}	Causal directional margin	0.2
$(\widetilde{\mu}_{\text{op}}, \widetilde{\mu}_{\text{set}}, \widetilde{\mu}_{\text{dec}})$	Aligned-behavior head weights	$(0.8, 1.2, 1.0)$
$(\omega_{\text{op}}, \omega_{\text{set}}, \omega_{\text{dec}})$	Invariance head weights	$(0.7, 1.3, 1.0)$

Table 7: Default loss, soft-minimum, margin, and policy-head weights for CIT-Mem training. These settings balance supervision, invariance, sensitivity, and evidence-selection behavior.

A.3 Adaptive Replay and Importance Correction

The ACR description in Section 3.2 admits the following explicit sampling rule. Let $N_r = |\mathcal{Q}_r| > 0$ for $r \in \{\text{nc}, \text{c}\}$, with pair losses $\ell_{\text{nc},i} = \mathcal{L}_{\text{inv},i}$ and $\ell_{\text{c},i} = \mathcal{L}_{\text{sens},i}$. Let $\bar{\ell}_{r,i}^{(t)}$ store the smoothed loss of pair i. Initialize these values to the same nonnegative constant. After observing a pair's loss, update its stored value by

$$\bar{\ell}_{r,i}^{(t+1)} = \begin{cases} \gamma \bar{\ell}_{r,i}^{(t)} + (1-\gamma)\,\text{sg}(\ell_{r,i}^{(t)}), & i \text{ is observed}, \\ \bar{\ell}_{r,i}^{(t)}, & \text{otherwise}, \end{cases} \tag{15}$$

where $0 \le \gamma < 1$ controls smoothing and sg stops gradients. Sampling within each pool can then use

$$v_{r,i}^{(t)} = (\bar{\ell}_{r,i}^{(t)} + \varepsilon)^{\eta}, \qquad q_{r,i}^{(t)} = \frac{\zeta}{N_r} + (1-\zeta)\frac{v_{r,i}^{(t)}}{\sum_{j=1}^{N_r} v_{r,j}^{(t)}}, \tag{16}$$

with numerical floor $\varepsilon > 0$, priority exponent $\eta \ge 0$, and uniform-mixture fraction $0 < \zeta \le 1$. Larger η concentrates updates on higher-loss pairs; ζ guarantees $q_{r,i}^{(t)} \ge \zeta / N_r$. Setting $\zeta = 1$ recovers the uniform-replay ablation. Table 8 gives the fixed replay settings.

Parameter	Meaning	Value
γ	Priority moving-average decay	0.95
ε	Numerical priority floor	10^{-6}
η	Priority exponent	0.7
ζ	Uniform-mixture fraction	0.2

Table 8: Adaptive Counterfactual Replay parameters used for priority updates and mixed prioritized–uniform sampling.

For m_r independent draws with replacement from pool r, use the sampling probability at the time of the draw and the full reciprocal-probability correction:

$$w_{r,i}^{(t)} = \frac{1}{N_r q_{r,i}^{(t)}}, \qquad \widehat{\mathcal{R}}_r = \frac{1}{m_r}\sum_{b=1}^{m_r} \text{sg}(w_{r,i_b}^{(t)})\ell_{r,i_b}, \tag{17}$$
$$\mathbb{E}[\widehat{\mathcal{R}}_r \mid \theta, q^{(t)}] = \frac{1}{N_r}\sum_{i=1}^{N_r} \ell_{r,i}.$$

Use $\widehat{\mathcal{R}}_{\mathrm{nc}}$ and $\widehat{\mathcal{R}}_{\mathrm{c}}$ to estimate $\mathcal{R}_{\mathrm{inv}}$ and $\mathcal{R}_{\mathrm{sens}}$ in Equation 10, respectively. Sampling probabilities and weights remain detached during each gradient update. The equality above requires the full weights, without clipping or minibatch self-normalization; the uniform mixture already bounds $w_{r,i}^{(t)} \leq 1/\zeta$. Duplicate draws count separately in the estimator; their observed detached losses may be averaged for a single stored-priority update. An empty pool contributes no counterfactual term and is not sampled.

For a base minibatch $\mathcal{B}_0^{(t)}$ of size m_0, the actual stochastic objective and controller update are

$$\widehat{\mathcal{L}}_{\mathrm{CIT}}^{(t)} = \frac{1}{m_0} \sum_{x \in \mathcal{B}_0^{(t)}} \mathcal{L}_{\mathrm{base}}(x) + \lambda_{\mathrm{inv}} \widehat{\mathcal{R}}_{\mathrm{nc}}^{(t)} + \lambda_{\mathrm{sens}} \widehat{\mathcal{R}}_{\mathrm{c}}^{(t)},$$
$$g_t = \nabla_\theta \widehat{\mathcal{L}}_{\mathrm{CIT}}^{(t)}, \qquad \theta_{t+1} = \mathrm{Adam}(\theta_t, g_t; \xi_{\mathrm{opt}}), \tag{18}$$

where ξ_{opt} denotes the optimizer settings summarized below. Equation 18 is the finite-minibatch estimator used to optimize the population objective in Equation 10; the two replay pools remain separately normalized.

Parameter	Meaning	Value
ξ_{opt}	Optimization algorithm	Adam
η_{opt}	Initial learning rate	10^{-4}
$E_{\max}$	Maximum training epochs	10
B	Minibatch size	8
$\mathcal{S}_{\mathrm{seed}}$	Repeated-run seeds	$\{42, 43, 44\}$
C	Retrieved candidate limit	50 records
K	Selected evidence limit	5 records
τ_{abs}	Minimum answerability score	0.3
T_{dec}	Reader decoding temperature	0
L_{ans}	Maximum answer length	64 tokens
I_{think}	Reader thinking-mode switch	Disabled
H_{train}	Training device	1× NVIDIA GeForce RTX 4090 (24 GB VRAM)

Table 9: Core training and inference parameters used by Algorithm 1. Each row reports one parameter, its operational meaning, and its default setting. Section 4.4 varies K and τ_{abs} to characterize the deployment trade-off.

Analysis and conclusion. Table 9 specifies a single reproducible operating point rather than a per-dataset search. The candidate limit $C = 50$ bounds backend exposure, while $K = 5$ restricts the reader to compact evidence. The threshold $\tau_{\mathrm{abs}} = 0.3$ favors broad coverage; later audits show how increasing it changes selectivity. Deterministic decoding and disabled thinking reduce variance attributable to the frozen reader, so repeated-run differences primarily reflect controller training. With three fixed seeds and a 10-epoch cap on one RTX 4090, the configuration also makes the reported compute budget explicit. Overall, the table defines a controlled basis for every subsequent appendix comparison.

A.4 End-to-End Training and Deployment Algorithm

Algorithm 1 summarizes the complete training and inference workflow.

B Contribution Attribution with Matched Twins

B.1 Matched Attribution Design

The no-twins comparison in Figure 3 changes the training data and the objective simultaneously, so the 28.9-point gain on LoCoMo-10 could in principle come from more and cleaner training samples, from stronger validation and labels, or from the paired invariance and sensitivity losses themselves. We separate these sources with the factorial design in Table 10: all arms share the same base examples, twin corpora, labels, optimizer, batch size, update budget, and seeds, and differ only in twin validation and the training objective. The plain-SFT arms supervise the identical twins and their target labels

Algorithm 1 CIT-Mem training and deployment

Require: Training set $\mathcal{D}$; interventions $\{\mathcal{I}_r\}_{r\in\{\mathrm{nc,c}\}}$; validator v; fixed scorer ρ_B and reader f_ϕ; budgets C, K
Ensure: Trained memory controller g_θ and evidence-grounded prediction

Initialize θ and replay pools $\mathcal{Q}_{\mathrm{nc}}, \mathcal{Q}_{\mathrm{c}}$
for all base examples $x = (\mathcal{H}, q, y, \mathcal{E}^{\mathrm{ann}}, b^\star) \in \mathcal{D}$ **do**
  Expose at most C candidate records with ρ_B and construct valid supports $\mathcal{S}_K(x)$
  for $r \in \{\mathrm{nc}, \mathrm{c}\}$ **do**
    Generate the typed twin $x^r \leftarrow \mathcal{I}_r(x)$ and its target behavior $b^r_{\mathcal{I}}$
    Evaluate $G_{\mathrm{str}}, G_{\mathrm{evid}}, G_{\mathrm{rec}}, G^{(v)}_{\mathrm{ans}}$ in order
    if $V^{(v)}(\mathcal{I}_r, x) = 1$ **then**
      Align equivalent records on $\mathcal{K}_{\mathcal{I}}$ using $P_{\mathcal{I}}$
      Add $(x, x^r, b^r_{\mathcal{I}}, P_{\mathcal{I}})$ to $\mathcal{Q}_r$
    end if
  end for
end for
for each training epoch **do**
  Sample base examples uniformly and twin pairs with Equation 16
  Compute $\mathcal{L}_{\mathrm{base}}$, $\mathcal{L}_{\mathrm{inv}}$, and $\mathcal{L}_{\mathrm{sens}}$
  Update θ using Equation 10 with the estimator in Equation 17
  Update detached moving-average losses with Equation 15 and refresh priorities
end for
Deployment:
Apply the predicted memory operation and expose at most C live candidates with ρ_B
Group candidates with κ, reject classes below $u_\varnothing$, and form $Z_\theta(x)$ from the top K
if $p_\theta(\textsc{Answer} \mid x) < \tau_{\mathrm{abs}}$ **then**
  return ABSTAIN
else
  return $f_\phi(q, Z_\theta(x))$
end if

Training condition	Twins	Objective	F1	IS	CDA	Abst-P
Answer-only	–	SFT	20.6	78.2	57.9	60.4
w/o G_{ans}, plain SFT	✓(no G_{ans})	SFT	26.1	79.4	58.8	67.5
w/o G_{ans}, counterfactual	✓(no G_{ans})	IS+CS	38.6	84.9	62.1	67.8
Validated, plain SFT	✓(validated)	SFT	32.5	83.0	64.9	70.1
Validated, counterfactual (uniform replay)	✓(validated)	IS+CS	43.4	90.9	72.1	72.3
CIT-MEM (full)	✓(validated)	IS+CS+ACR	**49.5**	**93.6**	**75.8**	**81.4**

Table 10: Contribution attribution on LoCoMo-10 with GPT-4o-mini under matched data, optimization, compute budgets, and seeds (42–44). All arms share identical base examples, twin exposure, batch size, and update budget. Plain-SFT comparisons isolate the effects of twin data and answer-direction validation; validated plain SFT versus counterfactual training isolates the paired losses; uniform replay versus the full method isolates ACR. F1 uses all-query accounting, while IS, CDA, and Abst-P follow Tables 3 and 5. Results are three-seed averages.

with standard cross-entropy, without paired contrastive objectives; counterfactual arms train with the full $\mathcal{L}_{\mathrm{inv}}$ and $\mathcal{L}_{\mathrm{sens}}$. Under this design, the answer-only to plain-SFT increment isolates the data effect; the validated-versus-w/o-G_{ans} contrast isolates the label effect of the answer-direction gate G_{ans} specifically—the only validation gate removed in the w/o-G_{ans} arms, while the structural, provenance, and recoverability gates remain active—so label-effect claims are scoped to this gate rather than to the full validation pipeline; the validated plain-SFT versus validated counterfactual (uniform replay) contrast isolates the contribution of the paired losses under identical twins, sampling, and compute; and the uniform-replay-to-ACR contrast on the same counterfactual objective isolates the end-task effect of adaptive scheduling.

B.2 Attribution Results and Interpretation

Table 10 separates improvements that were confounded in the coarse ablation. Merely adding unvalidated twins raises F1 from 20.6 to 26.1 under plain SFT, while answer-direction validation lifts

the matched plain-SFT score to 32.5. Switching from validated plain SFT to paired counterfactual losses with uniform replay adds 10.9 F1 points and improves all behavioral metrics. ACR then contributes another 6.1 F1 points, with the strongest gain in Abst-P. The consistent progression across F1, IS, CDA, and Abst-P indicates that clean labels, paired objectives, and adaptive scheduling are complementary; the full method's advantage cannot be attributed only to additional twin samples.

C Independent Twin Audit Protocol and Results

C.1 Independent Human Audit

Intervention family	Audited n	Intended target relation	Human-Valid (%)	Label errors
Paraphrase	13	Same answer; evidence preserved	92.3	1
Distractor insertion	13	Same answer; nuisance added	100.0	0
Equivalent duplication	12	Same answer; support duplicated	91.7	1
Evidence deletion	13	Abstain if no support remains	92.3	1
Evidence replacement	13	Changed answer	84.6	2
Contradiction	12	Changed answer or abstain	91.7	1
Temporal update	12	Changed answer to latest value	100.0	0
Temporal shift	12	Changed answer or abstain	83.3	2
Total	100	38 non-causal; 62 causal	92.0	8

Cohen's κ: 0.78 overall (structure 0.86; evidence 0.79; leakage 0.81; sufficiency 0.74; answer relation 0.72).
Human-confirmed subset after adjudication: 92 twins.

Table 11: Independent Human Audit Report. Label errors count accepted twins whose adjudicated validity disagrees with the automated acceptance.

We conducted an independent stratified audit of 100 accepted, held-out twins, with annotators blinded to the validation reader and policy output. Table 11 reports the audited sample counts and results for each intervention family. Two blinded annotators applied the five checks defined in Section 3.1, and a third annotator adjudicated disagreements. *Human-Valid* is the fraction of audited twins in each family that pass all five checks after adjudication; *label errors* count accepted twins whose adjudicated validity disagrees with the automated acceptance; Cohen's κ measures pre-adjudication inter-annotator agreement, pooled over the five binary checks; and the *human-confirmed subset* contains the 92 twins passing all five checks. The validity and label-error results were computed from adjudicated human labels, and Cohen's κ from the independent annotations before adjudication.

Analysis and conclusion. Table 11 shows that 92 of 100 automatically accepted twins pass every human check. Distractor insertion and temporal update achieve 100.0% Human-Valid, whereas temporal shift has the lowest rate at 83.3%. The eight label errors comprise two non-causal and six causal cases; evidence replacement and temporal shift contribute the largest counts, with two each, while distractor insertion and temporal update contribute none. Agreement remains substantial overall ($\kappa = 0.78$), although answer-relation and sufficiency checks are less reliable than structure or leakage checks. We conclude that the automated gates produce valid training pairs in 92.0% of the audited cases, but the remaining errors are weighted toward causal interventions, motivating larger audits and stronger causal-evidence validation.

C.2 Zero-Shot Transfer to Long-MT-Bench+

Analysis and conclusion. Table 12 demonstrates zero-shot transfer from the LoCoMo-selected checkpoint to Long-MT-Bench+ across all three frozen readers. CIT-MEM improves F1 over the strongest baseline by 11.86%, 15.79%, and 10.65%, and improves BLEU-1 by 17.44%, 13.51%, and 13.27% for GPT-4o-mini, Qwen3-8B, and Llama 3.2 3B, respectively. Its retrieved context remains near five units, far below MemGPT and Mem0, although slightly above the cheapest systems. The gain is therefore not explained by a larger context budget. The consistent ranking across reader families supports the conclusion that the learned evidence-selection policy transfers beyond the training benchmark and is not tied to one answer model. Beyond the headline ranking, we

Table 12: Long-MT-Bench+ transfer results for GPT-4o-mini, Qwen3-8B, and Llama 3.2 3B readers. F1 and BLEU-1 measure answer quality, and Token is the average retrieved context length per question. The LoCoMo-selected checkpoint is applied without target-dataset tuning. Improvement (%) is the relative gain of CIT-MEM over the strongest baseline in each answer-quality column; Token is omitted because it measures context cost.

Long-MT-Bench+				
		Answer Quality		Cost
Model	Method	F1	BLEU-1	Token (in thousands)
GPT-4o-mini	MemoryBank	28.08	24.93	0.84
	MemGPT	41.63	35.98	20.12
	A-MEM	43.90	35.12	5.87
	Mem0	48.15	40.78	12.58
	LightMem	23.28	18.67	4.71
	MemoryOS	33.10	27.58	4.81
	CIT-MEM(ours)	**53.86**	**47.89**	5.03
	Improvement (%)	11.86%↑	17.44%↑	–
Qwen3-8B	MemoryBank	27.21	24.68	0.86
	MemGPT	42.14	37.55	19.98
	A-MEM	41.01	30.88	5.91
	Mem0	44.39	39.07	12.63
	LightMem	24.12	20.77	4.70
	MemoryOS	32.97	27.16	4.97
	CIT-MEM(ours)	**51.40**	**44.35**	5.11
	Improvement (%)	15.79%↑	13.51%↑	–
Llama3.2-3B	MemoryBank	24.22	23.40	0.81
	MemGPT	37.91	33.16	20.05
	A-MEM	37.22	30.53	5.70
	Mem0	42.63	38.42	12.44
	LightMem	21.91	15.14	4.66
	MemoryOS	30.05	22.41	4.65
	CIT-MEM(ours)	**47.17**	**43.52**	5.08
	Improvement (%)	10.65%↑	13.27%↑	–

analyze F1 and BLEU-1 jointly to distinguish token-level answer coverage from unigram overlap, and examine retrieved-context length to test whether improvements merely reflect a larger evidence budget. We also repeat the comparison with three frozen readers of different capacities while transferring the same LoCoMo-selected checkpoint without Long-MT-Bench+ tuning. The qualitative pattern remains stable across these views: CIT-MEM selects compact supporting records, preserves answer quality under a new dialogue distribution, and does not rely on a particular reader family. These results indicate that counterfactual training learns reusable evidence-selection behavior rather than benchmark-specific surface correlations, providing complementary evidence of generalization across datasets, evaluation metrics, and reader architectures.

D SCORING-PROTOCOL AUDITS

Because CIT-MEM can abstain while every baseline answers all queries, and because the Overall in Table 2 is the average of the six category scores, we examine both accounting choices. Table 13 summarizes two operating points from the LoCoMo-10 hyperparameter analysis, separating all-query F1 from answered-only F1 and coverage. Table 14 compares the averages over categories and over individual questions on LongMemEval-S using the same judge outputs.

D.1 ABSTENTION COVERAGE ACCOUNTING ON LOCOMO-10

Analysis and conclusion. Table 13 makes the RQ4 trade-off explicit. At the default threshold of 0.3, CIT-MEM answers 97.5% of queries and obtains 49.47 all-query F1, exceeding the strongest listed baseline by 4.20 points. Raising the threshold to 0.5 improves answered-only F1 from 51.14 to 53.77 but reduces coverage to 92.0%, so all-query F1 declines slightly to 49.17. Because abstentions remain zero-scored and every retained LoCoMo-10 question has a reference answer, gains cannot

Method	Coverage (%)	Answered-only F1	All-query F1
MemOS	100.0	45.27	45.27
TiMem	100.0	42.01	42.01
CMI-Mem	100.0	43.55	43.55
CIT-MEM ($\tau_{\text{abs}} = 0.3$, default)	97.5	51.14	49.47
CIT-MEM ($\tau_{\text{abs}} = 0.5$)	92.0	53.77	49.17

Table 13: Abstention accounting on LoCoMo-10 with GPT-4o-mini. Baselines answer all queries. The CIT-MEM rows use $K = 5$: answered-only F1 is reproduced from Table 4, and coverage and all-query F1 from Figure 4. The default threshold prioritizes all-query F1; the higher threshold illustrates improved selective answer quality at reduced coverage.

come from excluding unanswered cases. We conclude that the default setting best serves overall utility, while the higher threshold suits applications that value retained-answer quality more than coverage.

The accounting identity must be applied before aggregation. For a single run and question category, let c be the answered fraction, F_{ans} the mean F1 over answered queries, and F_{all} the mean after assigning zero to abstentions. Then $F_{\text{all}} = cF_{\text{ans}}$. With category or seed averaging, the corresponding all-query score averages the within-group products; multiplying separately averaged coverage and answered-only F1 is not generally equivalent. The table therefore lists each metric directly, without reconstructing either F1 measure from rounded aggregate values.

D.2 THRESHOLD-SWEEP MONOTONICITY

For a fixed budget and evaluation group g (one seed and question category), let N_g be its fixed query count, a_i the answerability score, and $s_i \geq 0$ the F1 of the fixed reader prediction before abstention. For $\tau_2 \geq \tau_1$,

$$F_g(\tau) = \frac{1}{N_g} \sum_{i \in g} \mathbb{I}[a_i \geq \tau] s_i,$$
$$F_g(\tau_2) - F_g(\tau_1) = -\frac{1}{N_g} \sum_{\substack{i \in g: \\ \tau_1 \leq a_i < \tau_2}} s_i \leq 0. \tag{19}$$

Averaging these scores across categories or seeds with fixed nonnegative weights preserves this inequality. Consequently, taking these averages cannot explain an increase in all-query F1 when only the threshold changes. A reproducible threshold audit must apply every threshold to the same stored predictions, retain all queries in each denominator, and check the inequality within each group before aggregation. If predictions, selected evidence, or checkpoints change, the comparison evaluates a different protocol. Higher answered-only F1 remains compatible with Equation 19, because its answered subset and denominator change with the threshold.

D.3 AVERAGE RESULTS ON LONGMEMEVAL-S

	Reader		
Method	GPT-4o-mini	Qwen3-8B	Llama 3.2 3B
MemOS	68.80 / 68.68	65.13 / 65.55	50.72 / 51.12
TiMem	77.77 / 76.88	76.28 / 75.49	65.40 / 64.68
CMI-Mem	72.35 / 69.57	70.82 / 67.74	60.58 / 57.61
CIT-MEM(ours)	**81.63 / 79.67**	**79.50 / 77.81**	**69.65 / 68.33**

Each cell: average over categories / average over all questions.

Table 14: LongMemEval-S accuracy under category-macro and question-weighted averaging. CIT-MEM ranks first with all three readers under both aggregation conventions.

Analysis and conclusion. Table 14 shows that the ranking is invariant to aggregation choice. CIT-MEM leads under category-macro and question-weighted accuracy for each reader, despite unequal category sizes. Its question-weighted scores are 79.67, 77.81, and 68.33 for GPT-4o-mini, Qwen3-8B, and Llama 3.2 3B, respectively; the corresponding margins over TiMem are 2.79, 2.32, and 3.65 points. The macro averages are slightly higher because smaller categories receive equal weight, but no conclusion reverses. Thus, the reported advantage is not an artifact of averaging category percentages rather than individual questions, and it remains strongest for the smallest reader under both accounting conventions.

E PROMPT TEMPLATES

For reproducibility, we report below the ten core frozen prompt templates used by the reader, memory-processing calls, and answer-quality evaluators. Placeholders enclosed in braces are replaced at runtime. All compared methods share the same reader prompt; only the retrieved memory content changes. The controller emits an abstention before reader invocation, so the LoCoMo formatting rule that forbids the reader from returning "Unknown" does not override the controller's selective-answering decision. The eight intervention families used in the experiments are paraphrase, distractor insertion, equivalent duplication, evidence deletion, evidence replacement, contradiction, temporal update, and temporal shift, matching the taxonomy in Section 3.1 and Table 11. Distractor insertion, equivalent duplication, evidence deletion, evidence replacement, and temporal shift are deterministic programmatic transformations and therefore require no generation prompt. The remaining three families—paraphrase, contradiction, and temporal update—use the fixed templates reported consecutively below. For paraphrase, the timestamp prefix is removed before prompting and reattached unchanged afterward. Every model-assisted proposal must pass the validation gates in Section 3.1 before entering training. In the LoCoMo-formatting and LongMemEval-S boxes, bracketed headings identify mutually exclusive runtime branches and are not themselves sent to the model.

E.1 INTERVENTION-GENERATION PROMPTS

Prompt Template for the Paraphrase Intervention.

```
Rewrite the following utterance so that it conveys exactly the same facts but uses different wording
    and sentence structure. Keep all names, numbers, and dates identical. Reply with the rewritten
    utterance only, no quotes, no explanation.

{utterance_body}
```

Prompt Template for the Contradiction Intervention.

```
You are constructing a counterfactual memory record for training. Given a question, its current
    answer, and one decisive memory record, rewrite only the decisive factual claim so that the
    rewritten record explicitly contradicts the original claim. Preserve all entities, dates, and
    unrelated facts unless one of them is the contradicted value. Do not add explanations or
    unsupported details.

Return STRICT JSON: {"content": str, "new_answer": str|null, "target": "changed_answer"|"abstain
    "}

Question: {question}
Current answer: {answer}
Decisive memory: {decisive_memory}
```

Prompt Template for the Temporal-Update Intervention.

```
You are constructing a counterfactual memory record for training. Given a question, its current
    answer, one decisive memory record, and a later timestamp, write exactly one later record that
```

updates exactly one mutable value in the decisive fact. Preserve the entity, relation, and all unrelated facts. The updated value must differ from the original value and directly determine the new answer. Do not add explanations or unsupported details.

Return STRICT JSON: {"content": str, "new_answer": str}

Question: {question}
Current answer: {answer}
Decisive memory: {decisive_memory}
Later timestamp: {new_timestamp}

E.2 Agent and Memory-Processing Prompts

Prompt Template for Shared Answer Generation.

You are an intelligent Memory Assistant.
Please answer the following question based ONLY on the provided retrieved memories.
Question: {question}
Only output the result following the format instruction based on the retieved memories.
{format_instruction}

Retrieved Memories:
[1] {memory_1}
[2] {memory_2}
...
[K] {memory_K}

Prompt Templates for LoCoMo Answer Formatting.

[Categories 1, 2, and 4 branch]
The Final Result's Format Must Follow These Rules:
1. For questions requiring a date or time, strictly follow the format '15 July, 2023', 'July, 2023'.
2. Pay special attention to relative times like 'yesterday', 'last week', 'last Friday' in the text:
   + Only for last year/ last month/yesterday, calculate the absolute date, precise to year/month/day respectively, eg. 'July, 2023' or '19 July, 2023'.
   + For last week/weekend/Friday/Saturday, or few days ago etc, use 'the week/weekend/Friday before [certain absolute time]' to express MUST NOT calculate the exact date, just use the week /weekend/Friday before the certain absolute time, eg. 'the week/weekend/Friday before 15 July, 2023'/ few days before 15 July, 2023;
3. The answer should be the form of a short phrase (roughly a few words) for the following question, not a full sentence.
4. Use exact wording from the original conversation whenever possible.
5. For quantities, if English/Arabic numerals are used in the original text, use English/Arabic numerals in the answer respectively. If it is a quantity or frequency counted by yourself, default to using English word, eg. prefer two not 2.
6. When the answer has multiple phrases, connect them with commas don't use 'and'.
7. Ensure your response aligns directly with the question. For instance, start with 'Yes' or 'No' for binary questions, and do not name a province when asked for a country.
8. If the information is not enough, you MUST NOT answer 'Unknown' or 'I don't know.'

[Category 3 branch]
The Final Result's Format Must Follow These Rules:
1. Provide a short phrase answer, not a full sentence.
2. The question may require you to analyze and infer the answer from the retrieved information.
3. For quantities, if English/Arabic numerals are used in the original text, use English/Arabic numerals in the answer respectively. Numbers are represented by English words by default., eg. prefer two not 2.
4. This is an open−domain problem. When answering this type of question, you can ignore all other requirements that you must be completely confident before responding. NEVER answer 'I don't

know./None./Unknown'. You can perform reasoning based on the retrieved information and your model knowledge. Uncertain inferences can be expressed using 'likely'.
5. When the answer has multiple phrases, connect them with commas don't use 'and'.
6. Ensure your response aligns directly with the question. For instance, start with 'Yes' or 'No' for binary questions, and do not name a province when asked for a country.

Prompt Template for Typed-Store Retrieval.

You are an intelligent Memory Assistant. Decide which memory store to search next in order to answer the user's question.

[Available Tools]
{tools}
– finish: call this when the retrieved information is already sufficient.

[Rules]
1. Reply with EXACTLY one tool call in this format, nothing else:
<tool_call> {"name": "tool_name", "arguments": {"query": "..."}} </tool_call>
2. Do NOT call the same tool with the same query twice.
3. Prefer a different tool or a more specific query than what was already retrieved.
4. If the retrieved information already answers the question, call finish.

Prompt Template for Atomic Memory Formation.

Extract factual memories from the dialogue turns below.

CRITICAL RULES:
1. Full Entity Scan: Extract attributes and relationships between ALL entities mentioned (specific objects, places, third parties), not just the main speakers.
2. Pay special attention to time information, including relative times like 'yesterday' or 'the week before' in the text.
3. Each memory must be ATOMIC – one single piece of information per memory. Keep it brief and standalone (understandable without the dialogue).
4. Output STRICT JSON: {"memories": [{"content": str, "time_raw": str, "source_dia_ids": [str]}]}
Use "" for time_raw when no time is mentioned. Output at most {max_facts} memories.

Dialogue turns:
Session time: {session_time}
[{dia_id}] {speaker}: {utterance}
...

Prompt Template for Memory Evolution.

You are an Insights Database Administrator. Decide how to process a 'Potential Memory Item' against the 'Related Items'.

[Decision Logic]
Step 1 – Contradiction: does the new info explicitly say the old info is completely wrong? –> DELETE
Step 2 – Identity: is the new info exactly the same as an existing memory (content + time)? –> IGNORE
Step 3 – Atomic update: does the new info unambiguously overwrite a single mutable field (count, quantity, status, boolean, time) of the SAME entity? –> UPDATE
Never use UPDATE to merge two separate pieces of information – ADD them separately.
Step 4 – Default: prefer ADD (different time, different person, or new topic).

Output STRICT JSON: {"action": "ADD"|"UPDATE"|"DELETE"|"IGNORE", "target_index": int|null, "new_content": str|null, "reason": str}

E.3 External Evaluation Prompts

Prompt Template for the LoCoMo External Judge.

[System message]
Your task is to label an answer to a question as 'CORRECT' or 'WRONG'.
You will be given the following data:
(1) a question (posed by one user to another user),
(2) a 'gold' (ground truth) answer,
(3) a generated answer.

The point of the question is to ask about something one user should know about the other user based on their prior conversations.
The gold answer will usually be a concise and short answer that includes the referenced topic.

Grading Rules:
1. Be generous. As long as the generated answer touches on the same topic as the gold answer, it should be counted as CORRECT.
2. For time−related questions, if it refers to the same time period as the gold answer (e.g., "last Tuesday" vs "May 7th"), mark it CORRECT.
3. Ignore minor formatting differences.

Output Format:
Return the label in JSON format with the keys "reasoning" and "label".
Do NOT output anything else outside the JSON object.
Example: {"reasoning": "The answer identifies the correct item mentioned in the gold answer.", "label": "CORRECT"}

[User message]
Question: {question}
Gold answer: {gold_answer}
Generated answer: {generated_answer}

First, provide a short (one sentence) explanation of your reasoning, then finish with CORRECT or WRONG.

Prompt Templates for the LongMemEval-S External Judge.

[Common system message]
Your task is to evaluate whether a model's response correctly answers a given question.
You will be provided with:
(1) A question
(2) A correct answer or rubric
(3) A model's response

Based on the evaluation criteria provided in the user prompt, determine if the model's response is correct.

Output Format:
Return your judgment in JSON format with the keys "reasoning" and "label".
The "label" should be either "CORRECT" or "WRONG".
Example: {"reasoning": "The response contains the correct answer and addresses all key points.", "label": "CORRECT"}

[User message: single−session−user, single−session−assistant, or multi−session]
I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response is equivalent to the correct answer or contains all the intermediate steps to get the correct answer, you should also answer yes. If the response only contains a subset of the information required by the answer, answer no.

Question: {question}

Correct Answer: {correct_answer}

Model Response: {model_response}

Is the model response correct? Provide your reasoning and answer with CORRECT or WRONG in JSON format.

[User message: temporal−reasoning]
I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response is equivalent to the correct answer or contains all the intermediate steps to get the correct answer, you should also answer yes. If the response only contains a subset of the information required by the answer, answer no. In addition, do not penalize off−by−one errors for the number of days. If the question asks for the number of days/weeks/months, etc., and the model makes off−by−one errors (e.g., predicting 19 days when the answer is 18), the model's response is still correct.

Question: {question}

Correct Answer: {correct_answer}

Model Response: {model_response}

Is the model response correct? Provide your reasoning and answer with CORRECT or WRONG in JSON format.

[User message: knowledge−update]
I will give you a question, a correct answer, and a response from a model. Please answer yes if the response contains the correct answer. Otherwise, answer no. If the response contains some previous information along with an updated answer, the response should be considered as correct as long as the updated answer is the required answer.

Question: {question}

Correct Answer: {correct_answer}

Model Response: {model_response}

Is the model response correct? Provide your reasoning and answer with CORRECT or WRONG in JSON format.

[User message: single−session−preference]
I will give you a question, a rubric for desired personalized response, and a response from a model. Please answer yes if the response satisfies the desired response. Otherwise, answer no. The model does not need to reflect all the points in the rubric. The response is correct as long as it recalls and utilizes the user's personal information correctly.

Question: {question}

Rubric: {rubric}

Model Response: {model_response}

Is the model response correct? Provide your reasoning and answer with CORRECT or WRONG in JSON format.

[User message: unanswerable/abstention sample]
I will give you an unanswerable question, an explanation, and a response from a model. Please answer yes if the model correctly identifies the question as unanswerable. The model could say that the information is incomplete, or some other information is given but the asked information is not.

Question: {question}

Explanation: {explanation}

Model Response: {model_response}

Does the model correctly identify the question as unanswerable? Provide your reasoning and answer with CORRECT or WRONG in JSON format.