MACC: MULTI-ASPECT COMMITMENT CONTROL FOR DRIFT-AWARE MEMORY MANAGEMENT IN LONG-HORIZON LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon dialogue agents built on large language models (LLMs) must preserve still-active user commitments while operating under limited context budgets. We study *context drift*, unintended divergence from such commitments, and ask whether efficient memory editing also preserves them. Multi-Aspect Commitment Control (MACC) scores historical segments with several projection-based commitment alignments and learns to retain, summarize, or prune them, using a response-consistency proxy computed by an encoder separate from the generator. On the LoCoMo long-term conversational-memory benchmark and on Blended Skill Talk, with a shared Qwen2.5-1.5B-Instruct generator, MACC reaches 62.9% judge-rated acceptability on LoCoMo at 0.092 of the Full-History token cost. A No-History control reaches 60.4% at 0.087, so acceptability alone does not establish memory use. On a human-labelled diagnostic of 60 distance, distractor, and revision probes, three heuristic memory methods fail on 55–67% of probes, and a post-hoc evaluation of one MACC checkpoint fails on 78% with no dialogue history retained. These results expose a gap between task–cost performance and commitment fidelity and motivate explicit commitment testing alongside task–cost evaluation.

1 INTRODUCTION

Consider an assistant helping with meals, travel, or study. An early constraint, such as vegetarian options, no spoilers, or wheelchair access, may drop out of the current query yet still govern the correct response. We call such goals, preferences, and constraints *still-active commitments*, and we call unintended divergence from them *context drift*, distinct from legitimate user revisions. Long-term memory evaluation has separately emphasized knowledge updates and latent constraints (Wu et al., 2025; Li et al., 2026); we additionally distinguish omission from contradiction and from stale adherence to a revised commitment.

Standard memory strategies address this only indirectly. Models can fail to use information placed in the middle of long contexts (Liu et al., 2024); compression and pruning may discard delayed constraints; retrieval can miss obligations that remain active but are not salient to the current query. Multi-view retrieval work shows that a single document vector can underrepresent different potential query views (Zhang et al., 2022). This motivates testing multiple commitment-segment alignment features, without assuming that a retrieval result transfers directly to memory control.

We ask which historical segments an agent should retain under a token budget. MACC scores segments with several projection-based commitment-alignment features and computes its drift proxy with a sentence encoder separate from the generator, so the proxy is not produced by the backbone it evaluates.

Contributions.

- *Design.* We formulate memory editing around explicitly represented still-active commitments and introduce a controller that scores segments with several projection-based segment–commitment alignment features, with a drift proxy computed by a separate encoder (§3, §4).

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.

- *Task–cost results.* We report judge-rated acceptability and Full-History-normalized token cost on two dialogue datasets with one shared 1.5B generator, against ten baselines spanning heuristic, learned, compression, and external memory approaches. A No-History control and a separate head-count pilot delimit what these results establish about memory use and the multi-aspect design (§6.1, §6.4).
- *Commitment diagnostic.* We construct and human-label a fixed commitment diagnostic whose heuristic results and post-hoc MACC evaluation expose failures that task–cost metrics do not reveal. We also describe an adaptive stress-testing protocol and report a partial exploratory transfer that does not validate an adaptive-search advantage (§6.3).

2 Related Work

Retrieval vs. compression for memory. Memory strategies decide which context reaches the generator. Prior dialogue, compression, and retrieval work supplies useful context-management components (Roller et al., 2021; Chevalier et al., 2023; Pan et al., 2024; Karpukhin et al., 2020; Lewis et al., 2020), and external memory systems extract and retrieve persistent memories across sessions (Chhikara et al., 2025). Multi-view dense retrieval addresses the limits of a single document representation for different queries (Zhang et al., 2022). MACC edits a persistent buffer by retaining, summarizing, or pruning segments, whereas retrieval methods select historical segments for the current response.

Learned memory controllers. Recent systems learn to manage context or external memory using policy actions and reinforcement learning (Zhang et al., 2026; Zhou et al., 2026; Yan et al., 2026). Released implementations differ in backbone, retrieval stack and training budget, so rather than quoting published numbers we build one compact controller in this family at a matched scale (§5) and report a small joint-training pilot.

Drift and commitment evaluation. Multi-turn interaction can degrade reliability even when single-turn task ability remains high (Laban et al., 2026). Long-term memory benchmarks test knowledge updates and latent constraints (Wu et al., 2025; Li et al., 2026), while curiosity-driven red-teaming motivates novelty-aware test search (Hong et al., 2024). We propose Adaptive Counterfactual Commitment Stress Testing (ACCST) to evaluate active commitments and legitimate revisions with an explicit commitment ledger; this paper reports its fixed diagnostic and a partial exploratory transfer study.

3 Problem Formulation

At turn t the controller observes user input u_t, an editable memory buffer M_t, and a still-active commitment set $\mathcal{C}_t$, forming state $s_t = (u_t, M_t, \mathcal{C}_t)$. For each editable segment it chooses an action in $\{\texttt{RETAIN}, \texttt{SUMMARIZE}, \texttt{PRUNE}\}$; the edited memory is passed to a *fixed* generator that produces r_t. Memory is hierarchical: a read-only core M_{core} (task instructions, profile, persistent constraints) and an editable segment buffer M_{seg}. The objective is to preserve $\mathcal{C}_t$ (low drift, high task utility) at low memory/compute cost over long horizons.

4 Method: MACC

4.1 Multi-aspect commitment representation

Figure 1 summarizes the architecture. Let a commitment $c \in \mathcal{C}_t$ (or a segment $m \in M_{\text{seg}}$) be encoded not by one vector but by H aspect vectors $\{e^{(h)}\}_{h=1}^{H}$ obtained from H projection heads over a frozen encoder, motivated by the multi-view representation problem studied by Zhang et al. (2022); the heads stay fixed in the main runs and are updated only in the separate joint-training pilot. The controller scores an action on segment m against commitment c using per-aspect dot products $\{\langle e_m^{(h)}, e_c^{(h)} \rangle\}_{h=1}^{H}$ rather than a single one. These scores provide several projection-based alignment features; we do not assume that each head captures a separate semantic facet, and §6.4 reports a one-

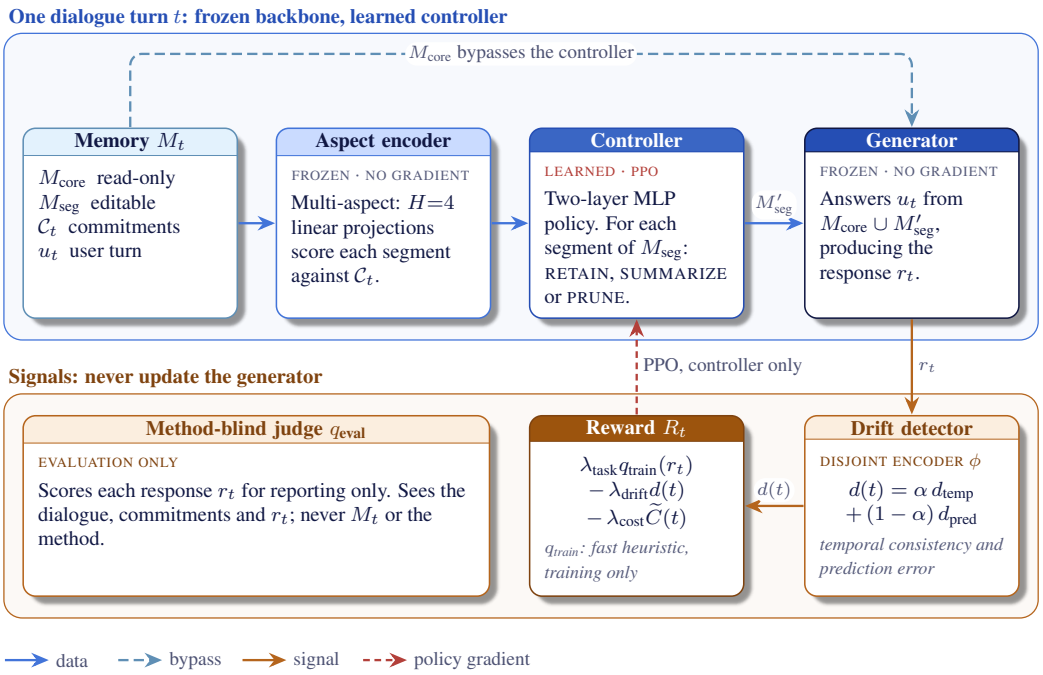

Figure 1: MACC architecture at one dialogue turn. The controller learns to retain, summarize, or prune editable memory segments using multi-aspect commitment alignments.

seed head-count pilot. The independent drift score below remains a scalar and is not decomposed into aspect-specific measurements.

Implementation. The aspect vectors are H linear projections of the frozen encoder's mean-pooled activations ($H = 4$). In the three-seed main evaluations, proximal policy optimization (PPO) (Schulman et al., 2017) updates only the controller. A separate joint-training pilot updates the aggregation and commitment modules jointly with the controller for 60 updates per arm; its one-seed, two-conversation evaluation is reported in §6.4. The controller is a two-layer multilayer perceptron (MLP) policy over the aspect alignments plus buffer statistics (segment length, age, summary status), trained with PPO (clip range 0.2, entropy bonus 0.01, generalized advantage estimation (GAE) (Schulman et al., 2016) with $\lambda_{\text{GAE}} = 0.95$; Appendix A.1) while the encoder and generator stay frozen.

Figure 1 illustrates the architecture. The controller reads per-aspect alignments between the editable buffer M_{seg} and the active commitments $\mathcal{C}_t$, then retains, summarizes, or prunes segments; it is the only module updated by PPO. The read-only core M_{core} reaches the frozen generator directly, bypassing the controller. Amber signals never update the generator: the drift detector scores r_t in a disjoint encoder space, and the method-blind evaluation judge q_{eval} scores responses without seeing the memory state or method identity. PPO rewards use a fast heuristic q_{train} instead. The separate joint-training pilot also updates the commitment and aggregation modules.

4.2 Separate-encoder drift proxy

The drift detector combines temporal consistency d_{temp} and prediction error d_{pred} with weight α. Both terms are computed with a sentence encoder ϕ (all-MiniLM-L6-v2) separate from the generator, so the proxy is not produced by the backbone it evaluates. With sim cosine similarity and f_θ a two-layer MLP head (hidden width 256, rectified linear activation) trained online by mean squared error on a first-in, first-out buffer of the 256 most recent (M_t, u_t, r_t) tuples,

$$d_{\text{temp}}(t) = 1 - \frac{1}{k}\sum_{i=1}^{k} \text{sim}\big(\phi(r_t), \phi(r_{t-i})\big), \qquad d_{\text{pred}}(t) = \big\|\phi(r_t) - f_\theta\big(\phi(M_t), \phi(u_t)\big)\big\|_2,$$

and d_{pred} is running-mean/variance normalized to $[0, 1]$; the combined score is

$$d(t) = \alpha\, d_{\text{temp}}(t) + (1 - \alpha)\, d_{\text{pred}}(t),$$

with $\alpha = 0.6$ and lookback $k = 8$ for LoCoMo (Maharana et al., 2024), $k = 5$ for Blended Skill Talk (BST) (Smith et al., 2020). We call d_{temp} *response dissimilarity*: it measures change between consecutive responses, not commitment error. We report the disjoint-encoder proxy as the headline drift; the external pilot diagnostic in §6.4 measures response self-consistency and has a different interpretation.

Design hypothesis. A single-embedding stability proxy may reward stylistic predictability even when commitment content erodes. Aspect alignments expose multiple directions in the commitment space, though they do not mathematically guarantee fidelity or eliminate proxy gaming. We test this design only in the one-seed pilot of §6.4 and discuss its limitations in §7.

4.3 Training objective and token-cost accounting

The per-turn reward is

$$R_t = \lambda_{\text{task}}\, q_{\text{train}}(r_t) \;-\; \lambda_{\text{drift}}\, d(t) \;-\; \lambda_{\text{cost}}\, \widetilde{C}(t),$$

where q_{train} is a fast task heuristic used only for PPO rewards, so training makes no judge calls. We set the reward weights $\lambda_{\text{task}} = 1.0$, $\lambda_{\text{drift}} = 0.3$, $\lambda_{\text{cost}} = 0.1$. Standard evaluation instead reports q_{eval}, the *judge acceptability* of a method-blind (i.e., identity-blinded) rubric judge that sees only the dialogue prefix, the current user message, the response r_t, and the commitment source $\mathcal{C}_t$—never the method name, the memory state M_t, or drift features—and scores commitment fidelity rather than generic fluency. q_{eval} never provides PPO rewards, and human commitment-failure labels (§6.3) are a separate outcome. Normalized cost divides token-equivalent cost by the Full-History reference cost, so *Full-History normalizes to* 1.0 *by construction*; other methods can exceed 1.0. The supplementary material also reports absolute input/output and summarization-token counts. The controller is trained with PPO; the generator is never updated.

5 Experimental Setup

Backbone (unified, small). All methods share a single `Qwen2.5-1.5B-Instruct` generator (Yang et al., 2024), isolating memory control from backbone capacity and enabling training on a single Apple-Silicon device (Metal Performance Shaders backend). With three seeds, we report descriptive statistics and 95% confidence intervals (CIs); establishing statistical significance would require additional runs.

Prompt limits. LoCoMo baselines and the learned comparator use an 8,192-token prompt limit (the longest Full-History prompt reaches about 6,600 tokens). MACC uses a 4,096-token limit; no MACC prompt exceeds 135 tokens in practice. BST uses 4,096 tokens for all methods. Appendix A.1 provides full details.

Decoding. Standard evaluations use temperature 0.7, top-p 0.9, and a 64-token response cap with the model's default generation settings. BST heuristics additionally apply repetition penalty 1.2 and no repeated 4-grams, which limits comparability of their acceptability, token F1, and response-dissimilarity scores (Appendix A.4).

Benchmarks. We focus the standard evaluation on LoCoMo (Maharana et al., 2024) (long-term conversational memory) and BST (Smith et al., 2020) (persona-grounded dialogue), scoring generated responses with our own judge protocol; the scores are therefore not interchangeable with published LoCoMo question-answering results. ACCST separately targets active commitments and legitimate revisions rather than factual QA recall.

Baselines. We evaluate ten baselines across four categories and a No-History control. *Heuristic memory management* (seven methods): Full-History retains all prior turns; Fixed-Window and Anchored-Window use sliding windows of different designs; Recency-Pruning removes the oldest

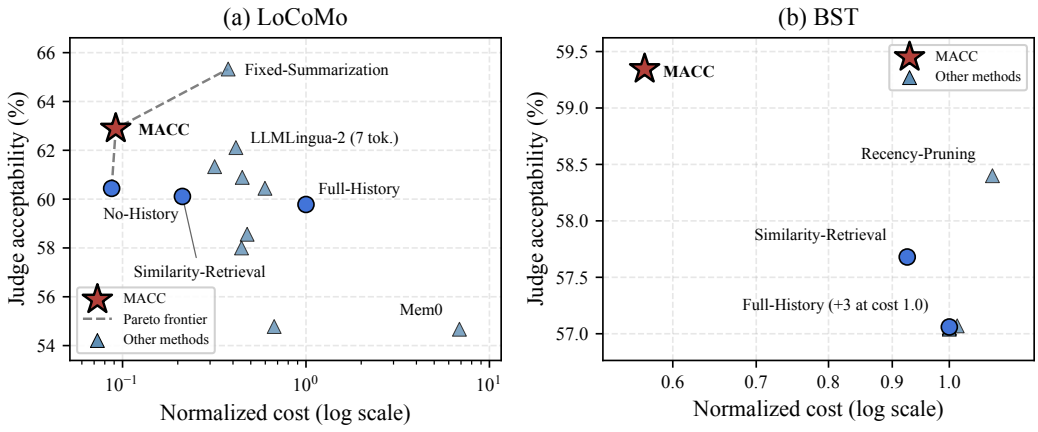

Figure 2: Judge acceptability versus normalized cost for all evaluated methods on LoCoMo and BST (seed means). MACC (red star) lies on the empirical LoCoMo frontier next to the No-History control. Full-History is the cost reference (1.0); on BST, three other heuristics overlap it near cost 1.0. Triangles are the remaining methods (Appendix A.1).

segments; Goal-Reinjection prepends the active commitment set; Fixed-Summarization compresses at fixed intervals; Similarity-Retrieval selects top-k segments by embedding similarity. *Learned controllers* (one method): MemAct-matched is our own compact implementation in the memory-as-action family (Zhang et al., 2026), trained with the same PPO budget but using a single embedding per segment and a task-only reward. *Compression* (one method): LLMLingua-2 (Pan et al., 2024) at two budgets (matched to MACC's mean history length, and one-third compression). The No-History control still receives the task instruction and answers the current message without dialogue history. *External memory systems* (one method): Mem0 (Chhikara et al., 2025) retrieves the 10 most relevant stored memories per query using hybrid dense-vector and BM25 lexical search and extracts new memories after each turn. Appendices A.1 and A.2 provide full details, and Appendix A.4 compares response dissimilarity.

Metrics. We report judge acceptability q_{eval} (`gpt-4o-mini`), lexical response F1 as a secondary diagnostic, response dissimilarity d_{temp} from the separate encoder, context length, Full-History-normalized cost, and absolute token costs. Standard results are the mean over seeds 42/123/456 $\pm$ the half-width of a Student-t 95% CI across the three seed means. The ACCST fixed diagnostic uses human outcome labels and base-scenario-clustered inference (§6.3).

6 Experiments

6.1 Main results

Judge acceptability and token F1 are percentages. The 60-update joint-training pilot is reported separately in §6.4 because its seed, budget, and evaluation subset differ from the three-seed standard evaluation. Table 1 reports standard comparisons for representative methods; Appendix A.1 (Tables 4 and 5) reports judge acceptability, token F1, and cost for every LoCoMo and BST method; the supplementary JSON provides all metrics and per-seed values.

On LoCoMo, Fixed-Summarization has the highest judge acceptability (65.3%) at 4$\times$ MACC's cost. MACC reaches 62.9% at normalized cost 0.092, against 59.8% at 1.000 for Full-History. The 91% cost reduction is descriptive under our accounting. The No-History control reaches 60.4% at 0.087, and MACC's mean prompt is only seven tokens longer (79 vs. 72). These results support a task–cost comparison, not a conclusion that MACC preserves active commitments, which §6.3 tests separately with human-labelled probes. Against Similarity-Retrieval, MACC costs 57% less (0.092 vs. 0.212) at similar acceptability (62.9 vs. 60.1%). LLMLingua-2 at the matched 7-token budget reaches 62.1% at 4.5$\times$ MACC's cost, but retains only 0.6 history tokens on average; Mem0 reaches

Table 1: Standard evaluation results: mean over three seeds ± Student-t 95% CI half-width. Full-History's cost is 1 by construction. Lower d_{temp} means more similar consecutive responses, not better commitment retention.

Dataset	Method	Judge accept. (%)↑	Resp. dissim. d_{temp}	Token F1 (%)↑	Norm. Cost↓
LoCoMo	Full-History	59.8 ± 4.1	$.687 \pm .015$	16.1 ± 1.1	$1.000 \pm .000$
	Similarity-Retrieval	60.1 ± 3.4	$.675 \pm .014$	15.5 ± 0.6	$.212 \pm .005$
	Fixed-Summarization	65.3 ± 7.6	$.684 \pm .025$	16.0 ± 1.4	$.377 \pm .008$
	No-History (control)	60.4 ± 7.5	$.670 \pm .018$	17.3 ± 0.5	$.087 \pm .005$
	LLMLingua-2 (7 tokens)	62.1 ± 4.2	$.671 \pm .012$	17.5 ± 0.3	$.415 \pm .003$
	Mem0	54.7 ± 5.0	$.654 \pm .013$	16.0 ± 1.5	$6.862 \pm .422$
	MemAct-matched	60.4 ± 6.6	$.675 \pm .014$	16.7 ± 2.2	$.597 \pm 1.131$
	MACC	62.9 ± 5.1	$.676 \pm .012$	17.3 ± 0.9	$.092 \pm .001$
BST	Full-History	57.1 ± 1.0	$.577 \pm .005$	6.8 ± 0.1	$1.000 \pm .000$
	Similarity-Retrieval	57.7 ± 0.3	$.566 \pm .007$	7.1 ± 0.1	$.925 \pm .001$
	MACC	59.3 ± 1.3	$.557 \pm .010$	10.2 ± 0.3	$.570 \pm .002$

54.7% at 75× MACC's cost because of per-turn memory extraction (Appendix A.3). Figure 2 shows the task–cost trade-off.

On BST, MACC has higher judge acceptability and token F1 than Similarity-Retrieval while using less context. The acceptability difference is 1.6 percentage points (59.3 vs. 57.7%) with cost falling from 0.925 to 0.570. For cross-method comparison, we report only the response-dissimilarity component d_{temp} of the drift proxy; MACC records 0.557 and Similarity-Retrieval 0.566. Lower values mean more similar consecutive responses, not higher commitment fidelity. The prediction-error component is method-specific and is excluded from comparative tables. Appendix A.4 provides the full temporal analysis.

6.2 Learned-baseline comparison at matched scale

The comparator is evaluated with an 8,192-token prompt limit (Appendix A.1). Table 1 compares MACC with the matched-scale learned memory controller on the same backbone, action space, and PPO budget. The point estimates of judge acceptability differ by 2.5 percentage points (62.9 vs. 60.4%) with substantial three-seed uncertainty, but the comparator retains far longer contexts on two of three seeds, giving it a normalized cost of 0.597 against 0.092 for MACC. This wide cost interval reflects strong seed-to-seed variation in context length. Because the comparator differs in both representation and reward, this comparison is descriptive rather than a component ablation.

6.3 Fixed commitment diagnostic and ACCST protocol

Standard dialogue scores do not test commitment fidelity directly. To address this, we propose the ACCST protocol (Figure 3), which combines method-blind adaptive search, matched controls, semantic validation, and frozen evaluation on held-out controllers. Here we report its completed *fixed diagnostic*: 20 discovery base scenarios, each with one distance, distractor, and revision probe, evaluated with Full-History, Fixed-Window, and Similarity-Retrieval. These 60 probes yield 180 responses from the shared Qwen2.5-1.5B-Instruct backbone using greedy decoding, seed 20260924, and a 256-token output cap. A post-hoc addendum applies the standard LoCoMo MACC seed-42 checkpoint to the same 60 probes.

Core-failure rates are percentages combining FORGOTTEN, CONTRADICTED, and STALE ADHERENCE labels from final adjudication. Percentile 95% CIs resample the 20 base scenarios jointly across methods (10,000 draws). Stale opportunities are the 20 revision probes per method. Cap hits are retained in all primary estimates.

Review and reliability. Both initial validity reviews rejected all 60 probes because they did not activate concrete commitments. After iterative repair, both reviewers accepted the revised set of 60. Annotator 1 and Annotator 2 submitted separate six-class outcome labels for each of the 180 responses, achieving 145/180 exact matches (80.6%; nominal Krippendorff's $\alpha = 0.724$). For FORGOTTEN, binary $\alpha = 0.588$ and positive agreement is 0.675, below the protocol's respective 0.67 and 0.70 thresholds. Annotator 1 served as both an initial reviewer and the final adjudicator for

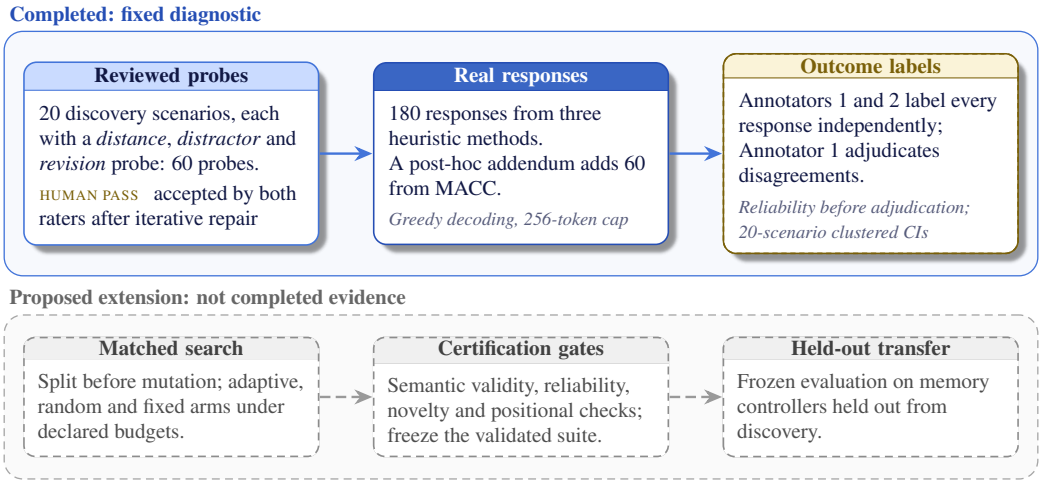

Figure 3: Completed fixed diagnostic (top) and the proposed adaptive benchmark extension (bottom, grey and dashed). The 180 main diagnostic responses come from three heuristic methods on reviewed discovery scenarios; a post-hoc MACC addendum adds 60 responses on the same probes.

Table 2: Fixed diagnostic core-failure rates (60 responses per method).

Method	Core failure (95% CI)	Stale / opportunity	Cap hits
Full-History	56.7 [45.0, 68.3]	4/20	18/60
Fixed-Window	66.7 [58.3, 76.7]	4/20	21/60
Similarity-Retrieval	55.0 [40.0, 68.3]	5/20	24/60

the 35 disagreements; original submissions and final decisions are retained. Only pre-adjudication labels enter reliability estimates.

Diagnostic findings. Table 2 gives final failure rates. By distance/distractor/revision family, failure counts are 13/13/8 for Full-History, 19/13/8 for Fixed-Window, and 11/12/10 for Similarity-Retrieval (20 responses per cell). Stale adherence occurs in 4/20, 4/20, and 5/20 revision opportunities, respectively, and never where the ledger has no superseded commitment. Full-History and Fixed-Window share identical contexts on the distractor and revision probes, so their equal results in those cells are not independent replications. The supplementary material reports all label counts, original-rater sensitivity, and paired contrasts.

Figure 4 visualizes failure patterns across probe types and methods. Distance probes consistently yield the highest failure rates, indicating that temporal separation from the commitment is a key challenge. The post-hoc MACC addendum shows elevated failures across all probe families, confirming that task-cost gains do not imply commitment retention.

Post-hoc MACC addendum. A separate run produced 60 MACC responses on the same reviewed probes. Annotator 1 and Annotator 2 independently labelled them, agreeing on 50/60 (83.3%; nominal $\alpha = 0.767$); FORGOTTEN binary $\alpha = 0.651$ remained below the protocol threshold. Annotator 1 adjudicated the 10 disagreements. Final core failures were 47/60 (78.3%; 95% scenario-cluster CI [66.7, 90.0]), with 16/20 distance, 16/20 distractor, and 15/20 revision failures; stale adherence occurred in 7/20 revision opportunities. The scored context was exactly the neutral core plus current probe in all 60 cases, verified by context hashes; no dialogue history survived MACC's memory actions at those turns. Paired contrasts relative to the heuristic arms are in the supplement. This post-hoc run uses one previously selected checkpoint on previously inspected scenarios; it shows that standard MACC training did not retain commitment history at these evaluation turns.

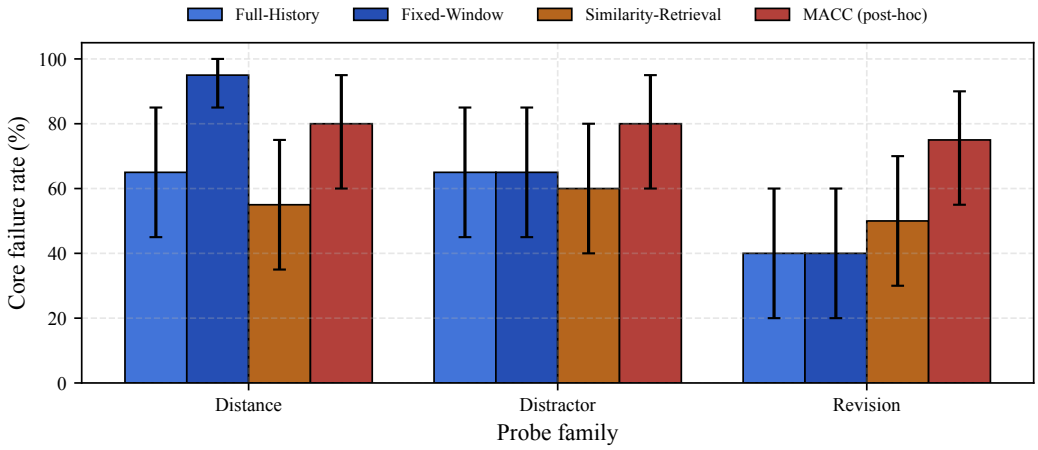

Figure 4: Core failure rates by probe type across four methods (with 95% CIs). Distance probes expose the most failures across all methods. MACC (post-hoc) shows high failure rates across all probe families, consistent with the finding that standard training did not retain commitment history at evaluation turns.

Exploratory held-out transfer. A planned held-out transfer applied the ACCST protocol to two held-out methods: Similarity-Retrieval and a multi-aspect pilot checkpoint trained separately from the standard MACC model ("multi-aspect pilot"). Memory pressure stopped generation at 197 of 280 planned responses; the 196 responses from the 14 complete base scenarios form the analysis set, and after removing 8 repeated dialogue–response pairs, Annotator 1 and Annotator 2 independently labelled the 188 unique responses with binary core-failure agreement of 86.7% (nominal $\alpha = 0.727$, positive agreement 0.886). The three search arms (fixed, adaptive, random) used method-blind candidate selection; the adaptive arm relied on an automated steering judge that failed prespecified calibration thresholds. For a stress test, a higher human-validated core-failure yield is the intended search outcome. Final binary core-failure rates were 52 to 71% for Similarity-Retrieval and 57 to 86% for the multi-aspect pilot across arms. Adaptive search found fewer core failures than random selection on both held-out methods: 53.6% versus 71.4% for Similarity-Retrieval and 60.7% versus 85.7% for the multi-aspect pilot. The paired adaptive-minus-random differences are -17.9 percentage points (95% base-bootstrap CI $[-42.9, 3.6]$) and -25.0 points ($[-46.4, -7.1]$), respectively. These exploratory intervals condition on the 14 complete bases and the final binary labels; the uncertified steering judge and incomplete transfer preclude an adaptive-yield claim. This exploratory analysis demonstrates protocol feasibility with human-labelled held-out transfer but does not support transfer-success claims, and the benchmark is not certified. Appendix A.6 provides full details.

Across the three heuristic methods, 63/180 responses (35%) hit the output cap; none has a truncated prompt. We retain these responses and report capped/uncapped strata in the supplement. Automated outcome-judge calibration on these 180 responses reveals a limitation: the rubric-based `gpt-4o-mini` judge returned 180 valid labels but detected none of 13 human-adjudicated stale cases (nominal Krippendorff's $\alpha = 0.486$ against final human labels). Its labels do not alter the human outcome table. The separate exploratory pilot used this uncertified judge for steering; its machine outcomes are excluded from the human result tables. Full ACCST benchmark certification remains future work because the fixed diagnostic's FORGOTTEN reliability is below threshold, the pilot's six-class transfer agreement is low, the transfer covers only 14/20 planned bases, and adaptive-yield analysis is incomplete (§7).

6.4 JOINT-TRAINING ASPECT PILOT

Both pilot arms completed 60 PPO updates with seed 42 and were evaluated on the same two LoCoMo test conversations (150 turns total). They differ in the number of aspect heads ($H = 4$ vs. $H = 1$); both include the trained commitment and aggregation modules. The task score uses a

Table 3: Joint-training aspect pilot point estimates (seed 42, 60 updates per arm, two LoCoMo conversations).

Arm	Task score (%)↑	Norm. cost↓	External diagnostic↓
Multi-aspect ($H = 4$)	86.8	0.139	0.661
Single-aspect ($H = 1$)	79.7	0.096	0.643

strict coherence judge. The external diagnostic is a frozen-encoder response self-consistency metric scoring 148 turns after excluding the first turn of each conversation; smaller values mean more local coherence, not necessarily better commitment retention. Table 3 reports the point estimates. The four-head arm has a 7.1-point higher task score than the one-head arm, but costs 44.8% more and has a 0.018 higher diagnostic; with one seed and two conversations, this is an exploratory trade-off rather than evidence of general multi-aspect superiority.

7 Limitations

Our evaluation uses a single 1.5B backbone and two dialogue benchmarks; generalization to other scales and domains requires further study. The fixed diagnostic uses 20 templated scenarios with one generation seed, limiting coverage of natural deployment variation. The joint-training pilot evaluates only two conversations, insufficient for population-level claims about the multi-aspect component. Commitment sets $\mathcal{C}_t$ are derived from controlled metadata rather than extracted from deployment. The fixed diagnostic's class-specific reliability and incomplete adaptive benchmark components do not support ACCST certification. The exploratory transfer includes only 14 of 20 planned base scenarios after a memory-pressure stop; its 188 unique annotated responses have low six-class pre-adjudication agreement. Binary core-failure decisions are complete after 25 documented adjudications, but 30 within-core class disagreements remain unresolved. The partial subset cannot establish full held-out transfer or generalize to the six unrun bases. Our separate-encoder drift proxy measures response dissimilarity, not commitment fidelity, as the post-hoc MACC diagnostic demonstrates. Multi-aspect representations expose multiple directions in the commitment space but do not mathematically guarantee fidelity or eliminate proxy gaming. The main evaluation lacks a replicated, reward- and capacity-controlled ablation of the number of aspect heads H; the joint-training pilot in §6.4 compares $H = 4$ with $H = 1$ on one seed only. LoCoMo scores come from two test conversations, so seed intervals do not capture variation across conversations. We hypothesize that MACC helps most beyond a crossover horizon, but testing this requires longer held-out trajectories. LoCoMo baselines use an 8,192-token prompt limit while MACC uses 4,096 tokens; although no MACC prompt exceeds 135 tokens in practice, this asymmetry means the comparison may understate what unconstrained baselines can achieve. BST heuristics use repetition penalties that differ from MACC, limiting direct comparability of their acceptability, token F1, and response-dissimilarity scores.

8 Conclusion

MACC combines multi-aspect commitment-alignment features, a response-consistency proxy computed by an encoder separate from the generator, and token-cost reporting normalized against a Full-History reference. On LoCoMo and BST with one 1.5B generator, and against ten baselines spanning heuristic, learned, compression, and external memory approaches, MACC attains low normalized cost with competitive judge acceptability, close to a No-History control. The fixed commitment diagnostic reveals substantial core failures across memory methods, and the post-hoc MACC evaluation shows that standard training did not retain commitment history at the evaluation turns. An exploratory transfer on held-out methods shows protocol feasibility but does not support transfer-success claims. Task–cost scores alone therefore do not capture commitment fidelity, which motivates explicit commitment testing alongside them.

AI USE STATEMENT

We used a large language model to assist with prose polishing of this manuscript for clarity and conciseness. Large language models also serve as experimental subjects in this work: the fixed generator and all baselines are LLM-based, and a rubric-based `gpt-4o-mini` judge provides judge-acceptability scores for the standard evaluation. The authors reviewed all AI-assisted prose, checked reported claims against the experiment artifacts and cited sources, and take full responsibility for the final content of this work.

ETHICS STATEMENT

Experiments use public dialogue benchmarks and constructed commitment scenarios. The fixed diagnostic includes human-labelled model responses. The two human annotators are referred to as Annotator 1 and Annotator 2 in the manuscript. Annotator 1 also adjudicated disagreements after providing an original independent annotation; this dual role is disclosed in the main text and supplementary material. Original keyed records remain unchanged for reproducibility.

REPRODUCIBILITY STATEMENT

Standard main results use seeds 42/123/456; the joint-training pilot uses seed 42 with a separate 60-update budget. The original fixed diagnostic uses generation seed 20260924 and three heuristic methods. A separate, post-hoc fixed addendum uses the standard LoCoMo MACC seed-42 checkpoint on the same reviewed probes. The supplementary material includes the 21 reported standard-result summaries, per-seed values, joint-training pilot metrics and source hashes, reviewed ACCST inputs, real responses, both original label sets, final decisions, failed automated-judge calibration, generation settings, rubric, and file hashes. A separate exploratory 14-base transfer snapshot reports final binary core-failure outcomes, original pre-adjudication agreement, and 10,000 paired base-bootstrap draws; 30 within-core six-class disagreements remain unresolved. Separate standalone scripts regenerate the original diagnostic and MACC addendum analyses using 10,000 paired base-scenario bootstrap draws (seed 20260924). Code and all released data are available at an anonymous repository, `https://anonymous.4open.science/r/ICRL-MACC-Drift-Aware-Memory-Control-Artifact-E271/`, which includes the source archive and a reproduction map; checkpoints, operational logs, and local filesystem metadata are excluded.

REFERENCES

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, 2023. doi: 10.18653/v1/2023.emnlp-main.232.

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.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026.

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 33*, 2020.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. LoCoMo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026. doi: 10.18653/v1/2026.acl-long.1150.

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.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, 2021. doi: 10.18653/v1/2021.eacl-main.24.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *The Fourth International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, 2020. doi: 10.18653/v1/2020.acl-main.183.

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

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, 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, 2026. doi: 10.18653/v1/2026.acl-long.583.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5990–6000, 2022. doi: 10.18653/v1/2022.acl-long.414.

Table 4: Complete LoCoMo standard comparison (mean ± 95% CI over three seeds). Judge acceptability and token F1 are percentages; normalized cost includes method-specific processing charges.

Method	Judge accept. (%)↑	Token F1 (%)↑	Norm. cost↓
MACC	62.9 ± 5.1	17.3 ± 0.9	0.092 ± 0.001
Full-History	59.8 ± 4.1	16.1 ± 1.1	1.000 ± 0.000
Fixed-Window	58.0 ± 3.8	15.4 ± 0.1	0.445 ± 0.005
Anchored-Window	58.6 ± 6.3	16.0 ± 1.0	0.478 ± 0.005
Recency-Pruning	61.3 ± 3.6	16.1 ± 0.8	0.318 ± 0.003
Goal-Reinjection	60.9 ± 1.9	15.8 ± 0.7	0.449 ± 0.004
Fixed-Summarization	65.3 ± 7.6	16.0 ± 1.4	0.377 ± 0.008
Similarity-Retrieval	60.1 ± 3.4	15.5 ± 0.6	0.212 ± 0.005
MemAct-matched	60.4 ± 6.6	16.7 ± 2.2	0.597 ± 1.131
No-History	60.4 ± 7.5	17.3 ± 0.5	0.087 ± 0.005
LLMLingua-2 (7 tokens)	62.1 ± 4.2	17.5 ± 0.3	0.415 ± 0.003
LLMLingua-2 (1/3)	54.8 ± 2.4	16.7 ± 1.3	0.671 ± 0.005
Mem0	54.7 ± 5.0	16.0 ± 1.5	6.862 ± 0.422

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, 2026. doi: 10.18653/v1/2026.findings-acl.956.

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

A Reproducibility and Evaluation Details

A.1 Implementation and experimental scope

All standard-evaluation methods use the same Qwen2.5-1.5B-Instruct generator and decoder. On LoCoMo, the baselines and the learned comparator use an 8,192-token prompt limit (the longest Full-History prompt has about 6,600 tokens); no MACC prompt exceeds 135 tokens. No-History answers from the task instruction and the current message only; LLMLingua-2 compresses the full history with its XLM-RoBERTa-large compressor to 7 tokens (MACC's mean per-turn history budget) or to one third of its length, and its compressor pass is charged as parameter-scaled token-equivalents. The reported LoCoMo and Blended Skill Talk (BST) summaries use seeds 42, 123, and 456. The LoCoMo split contains seven training, one validation, and two test conversations; BST uses its test split (Hugging Face `ParlAI/blended_skill_talk`). The decoder uses temperature 0.7, top-p 0.9, a 64-token response cap and the model's default generation settings; the BST heuristics additionally used repetition penalty 1.2 and no repeated 4-grams. MACC and the matched learned comparator share the action space and PPO implementation; their task rewards and segment representations differ as described in the main paper. The separate joint-training aspect pilot uses seed 42, 60 updates per arm, and the same two LoCoMo test conversations (150 turns); its results must not be pooled with the three-seed standard table.

The disjoint drift encoder is all-MiniLM-L6-v2. Normalized token cost uses the Full-History reference, so Full-History equals 1 by definition; absolute counts remain in the JSON summaries. Three-seed intervals are descriptive, while the one-seed joint-training pilot reports point estimates. The anonymous code archive omits model weights and judge credentials.

The following tables give the complete standard task, token F1, and cost comparisons. The machine-readable supplementary summaries additionally contain absolute token counts, per-seed metrics, and judge settings. Temporal consistency is reported separately in Appendix A.4.

Table 5: Complete BST standard comparison (mean ± 95% CI over three seeds). Judge acceptability and token F1 are percentages; normalized cost includes method-specific processing charges.

Method	Judge accept. (%)↑	Token F1 (%)↑	Norm. cost↓
MACC	59.3 ± 1.3	10.2 ± 0.3	0.570 ± 0.002
Full-History	57.1 ± 1.0	6.8 ± 0.1	1.000 ± 0.000
Fixed-Window	57.0 ± 1.0	6.8 ± 0.1	1.000 ± 0.000
Anchored-Window	57.1 ± 0.9	6.8 ± 0.1	1.000 ± 0.000
Recency-Pruning	58.4 ± 0.7	7.1 ± 0.2	1.083 ± 0.001
Goal-Reinjection	57.1 ± 0.5	6.9 ± 0.1	1.015 ± 0.000
Fixed-Summarization	57.1 ± 0.9	6.8 ± 0.1	1.000 ± 0.000
Similarity-Retrieval	57.7 ± 0.3	7.1 ± 0.1	0.925 ± 0.001

A.2 Additional LoCoMo baselines

LLMLingua-2 compresses the whole dialogue history before every turn, and the compressed text enters the system message. At the 7-token budget no tokens are forced; at one third of the history, newlines and question marks are always kept. The compressor reads the full history each turn, so its pass is charged as 0.36 token-equivalents per input token (compressor-to-generator parameter ratio, $0.56\text{B}/1.54\text{B}$); raw compressor token counts are in the supplementary results. Mem0 (v2.2.0, default pipeline) retrieves the 10 most relevant stored memories for each user message (MiniLM vectors with BM25), places them in the system message, and after each turn extracts and consolidates new memories with the same frozen Qwen backbone (greedy, 512-token cap); each extraction call is charged as its full prompt plus output tokens. Its extraction prompt is about 8,000 tokens, and 9 of 450 extraction outputs (2.0%) were invalid JSON and added no memory. Table 6 reports these baselines with MACC and No-History.

Table 6: Additional LoCoMo baselines (mean ± 95% CI over three seeds). Judge acceptability and token F1 are percentages; context is the mean prompt length in tokens; "kept" is the mean compressed-history length.

Method	Judge accept.	Token F1	Norm. cost	Context	Kept
MACC	62.9 ± 5.1	17.3 ± 0.9	$.092 \pm .001$	79	–
No-History	60.4 ± 7.5	17.3 ± 0.5	$.087 \pm .005$	72	–
LLMLingua-2 (7 tokens)	62.1 ± 4.2	17.5 ± 0.3	$.415 \pm .003$	73	0.6
LLMLingua-2 (1/3)	54.8 ± 2.4	16.7 ± 1.3	$.671 \pm .005$	1023	958
Mem0	54.7 ± 5.0	16.0 ± 1.5	$6.862 \pm .422$	219	–

A.3 Cost breakdown analysis

Figure 5 breaks down the absolute token count by component for selected LoCoMo methods. MACC's total is close to No-History's because almost no history survives its memory actions, whereas Mem0's 75× higher normalized cost is driven by per-turn memory extraction and consolidation (1.34M tokens).

A.4 Response dissimilarity

The combined drift score adds a prediction-error term whose predictor is trained only inside MACC's own PPO training loop; for every other method it is an untrained network, so that term is not comparable across methods. We therefore compare methods on its response-dissimilarity term alone, d_{temp}: one minus the mean cosine similarity between a response and the previous k responses of the same episode, in the all-MiniLM-L6-v2 space ($k = 8$ on LoCoMo, 5 on BST), averaged over turns, episodes, and then seeds. It depends only on each method's own responses. Lower is more consistent, and a method that repeats itself scores low, so d_{temp} is descriptive rather than a quality score. The BST heuristics' repetition penalties discourage reusing earlier words, which raises their d_{temp}. Values are in `standard/temporal_drift.json`.

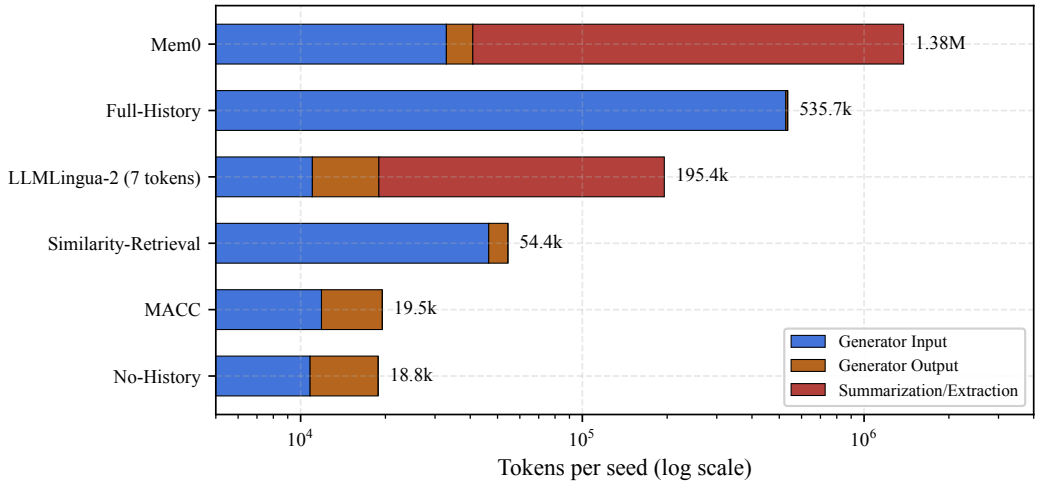

Figure 5: Absolute tokens per seed by component for selected LoCoMo methods (mean over three seeds; log scale). Summarization/extraction covers LLMLingua-2 compression and Mem0 memory extraction.

Table 7: Response dissimilarity d_{temp} (mean $\pm$ 95% CI over three seeds).

Method	LoCoMo	BST
MACC	$.676 \pm .012$	$.557 \pm .010$
MemAct-matched	$.675 \pm .014$	–
No-History	$.670 \pm .018$	–
LLMLingua-2 (7 tokens)	$.671 \pm .012$	–
LLMLingua-2 (1/3)	$.664 \pm .034$	–
Mem0	$.654 \pm .013$	–
Full-History	$.687 \pm .015$	$.577 \pm .005$
Fixed-Window	$.695 \pm .019$	$.577 \pm .005$
Anchored-Window	$.679 \pm .009$	$.577 \pm .005$
Recency-Pruning	$.681 \pm .024$	$.588 \pm .002$
Goal-Reinjection	$.677 \pm .026$	$.577 \pm .005$
Fixed-Summarization	$.684 \pm .025$	$.577 \pm .005$
Similarity-Retrieval	$.675 \pm .014$	$.566 \pm .007$

Table 7 lists all methods, and Figure 6 compares response dissimilarity d_{temp}, computed with the separate encoder, across LoCoMo methods. MACC (0.676) lies between Similarity-Retrieval (0.675) and Full-History (0.687), with overlapping seed intervals; this measures change between consecutive responses, not commitment retention.

A.5 Fixed commitment diagnostic

The reported diagnostic has 20 base scenarios. Each base produces one distance, one distractor, and one revision probe, giving 60 reviewed trajectories. Full-History, Fixed-Window, and Similarity-Retrieval answer each probe once, giving 180 responses under greedy decoding and a 256-token output cap. A separate post-hoc addendum applies one standard LoCoMo MACC checkpoint to the same 60 probes; it is not part of the three-method diagnostic or a held-out transfer evaluation.

Pass V asks whether a mutation preserves the intended commitment semantics, whether the final request activates the commitment, whether the dialogue is coherent, and whether the prompt leaks the expected answer. The final reviewed probes were accepted by both raters after repair. Pass O applies six mutually exclusive response labels in this order: NOT APPLICABLE if no commitment is activated; STALE ADHERENCE if an explicitly superseded commitment is followed instead of

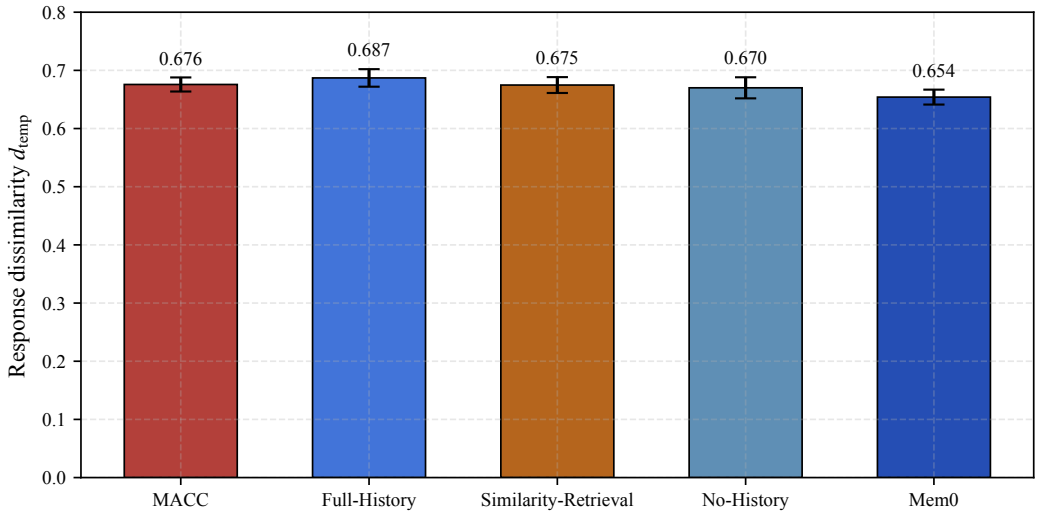

Figure 6: Response dissimilarity d_{temp} on LoCoMo (mean over three seeds with 95% CI; lower means more similar consecutive responses, not better commitment retention). All five methods lie within 0.654–0.687.

its replacement; CONTRADICTED if the active commitment is violated; FORGOTTEN if a relevant active commitment is omitted without explicit contradiction; UNJUSTIFIED REFUSAL if an otherwise permitted answer is refused; and PRESERVED otherwise. A stale label is impossible when the superseded ledger is empty. Annotators 1 and 2 saw blinded responses and commitment ledgers, not method identity or memory state. Annotator 1 had a recorded seed-curation assignment spanning work, technology, academic, health, and finance scenarios; Annotator 2 covered food, travel, lifestyle, hobbies, and shopping. These are recorded task assignments, not verified professional qualifications. Their original independent labels, reasons, evidence spans, and confidence are retained in the supplementary observations. Annotator 1 also adjudicated the disagreements; this dual role is disclosed rather than treated as a third independent rating.

For each method, core failure is the union of FORGOTTEN, CONTRADICTED, and STALE ADHERENCE. Stale rates use only probes with a superseded commitment as their denominator. The 63 cap-hit heuristic responses and 39 cap-hit MACC responses remain in their respective primary estimates. Cap-stratified results are descriptive because cap status is observed after generation. The 95% intervals for fixed-diagnostic rates resample the 20 base scenarios with replacement using 10,000 percentile draws and preserve the matching of methods and families within each draw. They condition on the templates and final labels; they do not measure annotation or construction uncertainty. The original three-method diagnostic and the post-hoc MACC addendum have separate protocols, analyses, and reproduction scripts in the supplementary files.

A.6 SCOPE OF THE ADAPTIVE PROTOCOL

ACCST's adaptive extension compares method-blind search with budget-matched random proposals. Its acceptance conditions include independent semantic review, duplicate and position checks, human-labelled per-family valid-failure yield, and transfer to methods excluded from search. The fixed diagnostic and post-hoc addendum establish none of these adaptive conditions. A separate 20-base discovery pilot completed 60 fixed/adaptive/random jobs and 1,020 method responses, but used an uncertified automated judge for search steering. The planned transfer generated 197 of 280 responses before a memory-pressure stop. The separately archived complete-case prefix contains 196 responses from 14 base scenarios: per held-out method, 42 fixed, 28 adaptive, and 28 random responses. The two held-out methods are Similarity-Retrieval and a multi-aspect pilot checkpoint trained for 60 updates ("multi-aspect pilot"), which is distinct from the MACC checkpoint in the standard tables. Eight exact dialogue–response pairs recur across methods, so the independent blinded two-annotator packet had 188 unique units. Six-class agreement was 133/188 (nominal

Table 8: Exploratory partial transfer: core-failure rate in 14 complete base scenarios (count/denominator; 95% base-cluster bootstrap CI). Fixed has 42 responses per method, adaptive and random 28 each. The multi-aspect pilot is a 60-update checkpoint, distinct from the MACC model in the standard tables.

Held-out method	Fixed	Adaptive	Random
Similarity-Retrieval	22/42 (52.4%; [33.3, 71.4])	15/28 (53.6%; [28.6, 75.0])	20/28 (71.4%; [50.0, 89.3])
Multi-aspect pilot	24/42 (57.1%; [40.5, 73.8])	17/28 (60.7%; [39.3, 82.1])	24/28 (85.7%; [75.0, 96.4])

Krippendorff's $\alpha = 0.583$); binary core-failure agreement was 163/188 ($\alpha = 0.727$; positive agreement 0.886). Twenty-five of 55 six-class disagreements cross the core-failure boundary. Annotator 1 documented final decisions for all 25 core-boundary disagreements; the other 30 within-core class disagreements remain unresolved. The 71 cap hits remain in the denominator. Table 8 gives descriptive final binary core-failure rates. The original 20-base plan, stopped-run provenance, original labels, decisions, and separate annotator sensitivities are in `accst/transfer_partial14/` of the supplementary material. Intervals resample the 14 complete base scenarios jointly across arms and methods (10,000 unadjusted percentile draws) and condition on this post-hoc subset and final binary labels. Adaptive has a lower point failure rate than random for both held-out methods; this is not evidence for a certified adaptive gain. The partial results do not establish the planned full transfer, adaptive yield, or benchmark certification.