Token-F1 Inverts LLM-Judged Answer Correctness in Long-Horizon Dialogue Memory Evaluation

Anonymous ACL submission

Abstract

We audit token-F1 as the default evaluation metric on long-horizon dialogue memory benchmarks. Holding backbone, retrieval, top-K, and call budget fixed, we find that token-F1 ranks prompt-engineering interventions in the opposite order from two LLM judges drawn from different model families: tight prompts beat loose by 0.246 F1 on LongMemEval, yet lose by 7 to 18 accuracy points under both gpt-4o-mini and Claude-Sonnet-4.5 (Cohen's κ=0.71). On LoCoMo, a single character-level choice—whether the model emits the gold abstention string—shifts aggregate F1 by 0.20.

Against this prompt-engineering lever, architecture effects look small. Across six family-level mechanisms, five backbones, and faithful upstream-code reproductions of Mem0, HippoRAG, and A-MEM (with evolution), no system clears Bonferroni-7-corrected α=0.0071 on aggregate F1. Two systems cross unadjusted α=0.05 in opposite directions (HippoRAG $+0.030$ at n=449; A-MEM -0.043 at n=200), but neither survives correction, and both vanish under cross-family LLM-judge.

Our scope is the matched single-call regime on per-question oracle haystacks (the regime in which the published architecture-vs-RAG gains we audit were originally reported); multi-call MemGPT and persistent multi-session systems (RMM, AgeMem, MLMA) are out of scope. Within that scope, we propose four reporting controls—surface-form-matched flat-RAG baselines, output-length statistics, per-category paired-bootstrap intervals, and cross-family LLM-judge accuracy—and provide a one-page checklist in the appendix.

1 Introduction

A wave of memory architectures for long-horizon dialogue—MemGPT (Packer et al., 2024), A-MEM (Xu et al., 2025), Mem0/Mem0g (Chhikara et al., 2025), MemoryBank (**?**), RMM (Tan et al., 2025), AgeMem (Yu et al., 2026), MLMA (Tiwari and Fofadiya, 2026), HippoRAG (Gutiérrez et al., 2025)—report 5–25 F1 gains over RAG baselines on LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025). The implicit reading is straightforward: a higher benchmark score means more memory ability. We argue this reading is at best incomplete.

A system can score low on a long-horizon QA benchmark for two structurally different reasons. The first is a memory-quality failure: the system did not retain or retrieve the relevant evidence. The second is a format-alignment failure: the system retrieved the right evidence but produced an output whose surface form—length, granularity, date format, abstention phrasing—did not align with the gold strings. Both surface as low F1 on the leaderboard. A benchmark that does not separate them cannot cleanly attribute the gap to memory architecture.

This is a benchmark-validity concern in the construct-validity sense of Messick (1995), Kane (2013), and Bowman and Dahl (2021): format-alignment is construct-irrelevant variance for a memory-architecture claim. The framing connects to prior critiques of lexical-overlap metrics in QA (Chen et al., 2019; Schuff et al., 2020; Bulian et al., 2022), of format-restricted prompting (Tam et al., 2024; Sclar et al., 2024), and of LLM-as-judge methodology (Zheng et al., 2023); closest in spirit are Jiang et al. (2026) and Li et al. (2026). We extend this lineage by isolating, inside the reading stage, the gap between "did the system have the answer in mind" and "did the system phrase it in the form the benchmark scores for."

Findings. *Finding 1: token-F1 inverts the LLM-judge ranking.* Token-F1 ranks tight prompts above loose by $+0.246$ F1 on LongMemEval and $+0.228$ on LoCoMo no-adv. Both gpt-4o-mini and Claude-Sonnet-4.5 reverse this, ranking loose above tight by $+0.07$ to $+0.18$ accuracy (κ=0.71 between

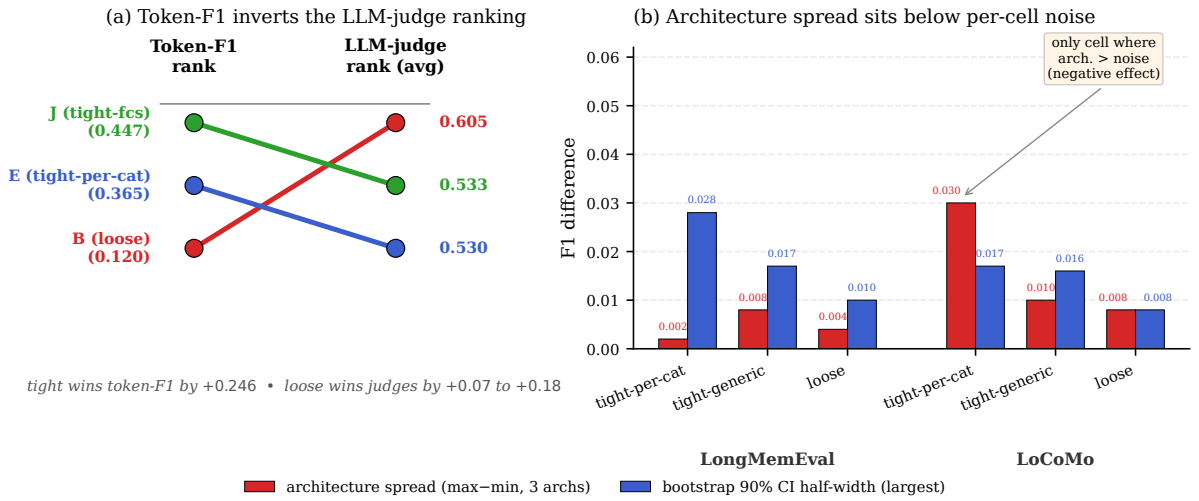

Figure 1: **(a)** Token-F1 vs LLM-judge on prompt-engineering interventions: under the same matched single-call protocol on LongMemEval oracle, token-F1 and the average of two cross-family LLM judges (gpt-4o-mini and Claude-Sonnet-4.5) rank the same three cells (B/E/J, all flat retrieval) in opposite orders. **(b)** Architecture spread vs bootstrap noise floor across six matched-protocol cells: the maximum F1 difference between the three architecture flavors (flat, linked-notes, session-bank) is below the per-cell bootstrap noise floor on five of six cells; the only exception (LoCoMo $\times$ tight-per-cat) is a -0.023 negative effect from link-expansion on LoCoMo's $\sim$588-turn haystacks.

judges). A truncate-and-rejudge ablation argues against length-bias as the main mechanism on the judge side. The token-F1 side is largely a length-compression artifact: the $+0.082$ FCS gain is fully accounted for by output compression from 52 to 14 chars on average (§5.4). The judge side is not: loose truncated to 170 chars—more than $3\times$ tight's 47-char average—still loses to tight (§5.5c).

Finding 2: a 0.20 F1 swing from one character-level choice. On LoCoMo, going from a verbal abstention to the exact gold string `no information available` moves aggregate F1 by 0.20. Nothing in our main matched matrix moves aggregate F1 by more than 0.03.

Finding 3: architecture-null on aggregate, but a real per-category mechanism. Among the three published systems we faithfully reproduce end-to-end (Mem0, HippoRAG, A-MEM with evolution), none clears Bonferroni-7-corrected α=0.0071 on aggregate F1 under our matched single-call protocol on per-question oracle haystacks. HippoRAG and A-MEM each cross unadjusted α=0.05 in opposite directions, and both effects collapse to noise under cross-family LLM-judge (Tab. 6). At the per-n paired-bootstrap MDE ($\Delta_{\text{F1}}\approx$0.042 at n=200, $\approx$ 0.028 at n=449), Mem0 ($|\Delta|$=0.021) sits below its floor, A-MEM ($|\Delta|$=0.043) sits at it, and HippoRAG (Δ=+0.030) sits marginally above its n=449 floor; "architecture-null" is therefore a multiple-comparisons-and-judge-replication verdict, not an assertion that effects are below F1 detection. The same null pattern holds across six family-level mechanisms and five backbones. Out-of-scope: MemoryBank-with-forgetting, RMM, AgeMem, MLMA, multi-call MemGPT (§8).

Where the architecture does work. Faithful Mem0 nevertheless produces a $+0.25$ F1 win on multi-session synthesis—the only above-noise positive architecture effect in our sweep—canceled on aggregate by -0.08 losses on single-session and knowledge-update where LLM-extracted memories drop precise lookup details that BM25 surfaces verbatim. Aggregate-null thus hides both a real mechanism and a real cost, motivating the per-category reporting prescription in §6.

Finding 4: prompt-engineering decomposition. The combined token-F1 lever splits into a surface-form effect ($+0.06$) and per-category specialization ($+0.18$). A surface-form-matched prompt template we call Format Control Specification (FCS; §3) beats hand-tuned by $+0.082$ token-F1, replicating across five backbones from three model families ($\Delta_{\text{F1}} \in [+0.028, +0.082]$). Under both LLM judges, the same advantage collapses to $\Delta_{\text{judge}} \in [-0.015, +0.035]$ across all 5×2 backbone-judge cells (App. A.16). This is the output-length compression noted above (52 vs 14 chars).

What would falsify this paper. We pre-register the following falsification condition: a memory architecture beating flat-RAG on cross-family LLM-judge accuracy (paired McNemar $p<0.05$ under both gpt-4o-mini and Claude-Sonnet-4.5, output-length matched), on at least one of LongMemEval and LoCoMo, on the same backbone, at pre-registered MDE $\Delta\approx 0.14$ accuracy (App. A.15). A reproducibility checklist appears as App. A.19.

2 Related Work

Long-horizon dialogue memory architectures. Memory architectures for LLM agents on long-horizon dialogue have proliferated in 2023–2026 (Packer et al., 2024; Xu et al., 2025; Chhikara et al., 2025; **?**; Tan et al., 2025; Yu et al., 2026; Tiwari and Fofadiya, 2026; Park et al., 2023; Gutiérrez et al., 2025), reporting 5–25 F1 gains over RAG baselines. A mechanism-level summary appears in Appendix Table 8. Three of these we reproduce end-to-end under matched single-call conditions (Mem0, HippoRAG, A-MEM with evolution; §5.6); the rest we approximate at the family level under the same protocol. Our claim concerns core retrieval and note-construction mechanisms within this protocol. MemGPT's multi-call regime is out of scope (§8).

Long-horizon dialogue benchmarks. Maharana et al. (2024) introduce LoCoMo with ∼600-turn dialogues, 1986 QA pairs, and 5 reasoning types. Adversarial accounts for 22.5% of questions and is scored against the exact string "`no information available`"—the source of our Finding 2. Wu et al. (2025) introduce LongMemEval, separating evaluation into indexing, retrieval, and reading; we use the oracle split (n=500) as our primary matched-pipeline cell. Earlier multi-session chat resources (Xu et al., 2021) and knowledge-intensive task suites (Petroni et al., 2021) provided the benchmark substrate for retrieval-grounded long-context evaluation. Li et al. (2026) add cognitive constraints and propose LLM-judge evaluation, which echoes our §5.1 finding that token-F1 inverts LLM-judge ranking.

Critical and replication work on memory systems. The closest critical works are Jiang et al. (2026) and Li et al. (2026); we delineate contributions explicitly.

Anatomy of Agentic Memory (Jiang et al., 2026) varies backbone, retrieval, and prompt across systems to surface metric, backbone, and latency misalignment. It does not fix these axes for matched comparison, and it does not isolate prompt-format as a separate lever. We hold backbone, retrieval, router, top-K, and call budget strictly fixed under one protocol, so any across-cell delta is attributable to a single axis. We add prompt-format alignment as an explicit fourth axis—Anatomy treats prompt as part of "the system"—and introduce FCS as a falsifiable programmatic prompt-rendering protocol calibrated on gold surface form. We also faithfully reproduce three upstream-code systems under this protocol; Anatomy uses author-reimplemented mechanisms. The two papers thus measure different things: Anatomy reports backbone fragility *across* systems, while we show prompt-format fragility *within* the same matched cell can exceed every architecture effect by an order of magnitude.

LoCoMo-Plus (Li et al., 2026) argues that string-match evaluation is misaligned with cognitive memory and proposes LLM-judge as the corrective. We confirm the misalignment empirically and quantify its magnitude (the metric-inversion finding: $+0.07$ to $+0.18$ accuracy flip under cross-family judges; 17 of 21 prompt-only-pair sign disagreement). The framing differs in three ways. We rule out single-judge artifacts via cross-family κ=0.71. We rule out judge length-bias via truncate-and-rejudge (§5.5c). And we combine the metric audit with a matched-pipeline architecture sweep, so the same fixed protocol produces evidence for both Finding 1 (metric inversion) and Finding 3 (architecture-null).

Xiong et al. (2025) report that memory can hurt by introducing distractor evidence, a related negative result without format/architecture decomposition.

Smallest-claim novelty. Token-overlap vs human disagreement on QA is well established (Chen et al., 2019; Schuff et al., 2020; Bulian et al., 2022). Format sensitivity of QA prompts is documented (Tam et al., 2024; Sclar et al., 2024). LLM-as-judge has known biases (Zheng et al., 2023), and length-controlled judging is a current methodological topic (Dubois et al., 2025). What is novel here is the joint magnitude of these phenomena on long-horizon dialogue memory benchmarks. The token-F1 ranking inverts the cross-family LLM-judge ranking on 17 of 21 (81%) prompt-only cell pairs, and the inversion is large enough ($+0.07$ to $+0.18$ judge-accuracy) to flip subfield architecture rankings under matched conditions. The construct-

validity framing is classical (Messick, 1995; Kane, 2013; Bowman and Dahl, 2021): format-alignment is construct-irrelevant variance and per-category synthesis is construct-relevant. What we add is an empirical decomposition that separates them on this benchmark family.

3 Method

3.1 Matched-pipeline ranking-shift protocol

For every cell, the following are held fixed: backbone (Qwen2.5-14B-Instruct, deterministic greedy decoding, identical sampling parameters); retrieval substrate (BM25 (Robertson and Zaragoza, 2009) over the per-question haystack); top-K (5 retrieved units); router (rule-based per-benchmark category router); LLM call budget (exactly one generation call per question); and evaluation (token-level F1, plus LoCoMo MC10 multiple-choice and the "no-adversarial" aggregate F1).

The single-call budget is the regime in which published numbers for Mem0, HippoRAG, A-MEM, MemoryBank, and others are reported: one answer-generation call per question after the memory layer returns. Our negative result thus addresses that comparison framing, not the multi-call deployment ceiling. MemGPT and RMM/AgeMem/MLMA remain out of scope (§8).

Architecture axis. The main 3×3 matrix (Tab. 1) uses three family-level mechanisms: *flat* (BM25 baseline), *linked-notes* (A-MEM-style), and *session-bank* (MemoryBank/Mem0g-style). Three further family-level mechanisms extend the matrix (*llm-extract*, *A-MEM with evolution*, *multi-call ReAct-lite*; see §5.3 and App. Tab. 7). Three faithful upstream-code reproductions (Mem0, HippoRAG, A-MEM with evolution) appear in §5.6. Retrieval substrate axis: BM25 (default), dense E5-base-v2, and BM25+cross-encoder rerank. Multi-call ReAct-lite is a thin MemGPT probe inside our single-call cell, not a faithful reproduction. Full descriptions are in App. A.10.

Prompt axis (3–4 variants). We use four prompt variants: *loose* (permissive, allows explanations, single template per benchmark); *tight-generic* (single tight prompt per benchmark, no per-category guidance); *tight-per-cat* (per-category prompts hand-crafted by us, mixing surface-form constraints and verbal task-solving hints); and *tight-fcs* (per-category prompts derived programmatically from gold-answer surface form, no task-solving hints). On tight-per-cat: we treat it as a category-aware hand-tuned prompt of the kind typical in this subfield, not as any one published paper's exact deployed prompt. Under our matched protocol the prompt is held fixed across architectures, which is the relevant comparison.

Format Control Specification (FCS). FCS is a controlled prompt rendered programmatically from per-category gold-answer surface-form profiles (length percentiles, dominant answer form, abstention strings) with no task-solving hints. FCS is calibrated on evaluation gold answers and is therefore a diagnostic audit instrument, not a leaderboard-submission method. In §5.6, cell J (flat × tight-fcs) additionally serves as the surface-form-matched comparator for the faithful-repro pairings: an architecture must beat what gold-calibrated surface-form alignment buys, not just match it. Full rendering rules are in App. A.17.

4 Experimental Setup

Primary backbone is **Qwen2.5-14B-Instruct** (Qwen et al., 2025) on 2×A100-80GB, with greedy decoding and a 96-token cap. Cross-backbone replication uses Qwen2.5-3B/7B, Mistral-7B-v0.3, and Llama-3.1-8B. Benchmarks are LongMemEval oracle (n=500) (Wu et al., 2025) and LoCoMo (n=1986, 22.5% adversarial) (Maharana et al., 2024). The 19-cell LLM-judge matrix runs at n=200/cell on gpt-4o-mini, with a cross-judge subset on Claude-Sonnet-4.5. Bootstrap protocol and McNemar MDE ($\Delta \approx 0.14$ at n=200, ∼94% one-sided power): App. A.15. Full cell list and per-cell sample sizes: App. 7. Per-cell prediction JSONs and FCS rendering code are released alongside the paper.[1]

5 Results

Table 1 gives the matched-pipeline aggregate F1 matrix; Table 2 the length / token-F1 / LLM-judge at-a-glance; Figure 1(b) the architecture-spread-vs-noise-floor view. Single-letter cell labels (B/E/J/Q) are introduced inline at first use; the full cell map is App. Tab. 7. §5.1–§5.4 establish Findings 1–4, §5.5 runs four mechanism checks against natural attacks on Finding 1, and §5.6 reports the three faithful upstream-code reproductions.

[1] https://anonymous.4open.science/r/token-f1-inverts-llm-judge-36C6 (anonymized for review).

LongMemEval oracle				
Retrieval \ Prompt	tight-per-cat	tight-generic	loose	tight-fcs
flat (no architecture)	0.365 (±.027)	0.182 (±.017)	0.120 (±.010)	**0.447** (±.032)
linked-notes (A-MEM-style)	0.366 (±.027)	0.174 (±.016)	0.116 (±.009)	**0.445** (±.032)
session-bank (MemoryBank-style)	0.367 (±.028)	0.180 (±.017)	0.119 (±.010)	**0.442** (±.032)
LoCoMo no-adversarial				
Retrieval \ Prompt	tight-per-cat	tight-generic	loose	tight-fcs
flat	**0.301** (±.017)	0.150 (±.015)	0.073 (±.007)	0.297 (±.017)
linked-notes	0.278 (±.017)	0.147 (±.015)	0.072 (±.008)	**0.280** (±.017)
session-bank	0.308 (±.017)	0.156 (±.016)	0.080 (±.008)	**0.310** (±.017)

Table 1: Matched-pipeline aggregate F1 on LongMemEval oracle (top, n=500/cell) and LoCoMo no-adversarial (bottom, n=1986/cell). Bootstrap 90% CI half-widths in parentheses.

Cell	Chars	Tok-F1	gpt-4o-mini	Claude
B (flat × loose)	311	0.120	0.600	0.610
E (flat × tight-per-cat)	47	0.365	0.530	0.530
J (flat × tight-fcs)	14	0.447	0.535	0.530

Table 2: Length / token-F1 / LLM-judge at a glance (LongMemEval oracle; token-F1 n=500, judge n=200 paired). The highest-token-F1 cell (J) ties E under judges; the lowest-token-F1 cell (B, 311 chars) is highest under both judges.

5.1 Finding 1: token-F1 inverts the answer-correctness ranking under LLM-judge

We send the predictions from cells B (flat×loose), E (flat×tight-per-cat), and J (flat×tight-fcs) on a paired n=200 subset of LongMemEval oracle to two LLM judges from different families: gpt-4o-mini (OpenAI) and Claude-Sonnet-4.5 (Anthropic). Token-F1 ranks tight-per-cat (E) above loose (B) by $+0.246$ F1. Under both judges, loose beats tight-per-cat by $+0.07$ accuracy on the same subset. The inversion replicates on LoCoMo no-adv at $+0.18$ accuracy in the loose direction. The two judges agree with each other (paired McNemar comparing judge verdicts: $\kappa > 0.7$), so the inversion is a property of token-F1, not of any single judge model.

A truncate-and-rejudge ablation (§5.5c) closes the most natural alternative explanation, namely that the judges share a length bias toward longer answers. When loose outputs (avg 311 chars) are truncated at multiple budgets and re-judged, the accuracy-vs-budget curve is monotone increasing. The interesting datapoint is the middle: loose at 170 chars still loses to tight at 47 chars (0.475 vs 0.530, McNemar χ^2=3.2, p≈0.074, borderline n.s. at α=0.05). A pure length-bias hypothesis would predict loose at 170 chars (more than 3× tight's 47-char average) to already win on length grounds alone. It does not. We read the mechanism as informational with a length floor: tight prompts front-load the answer in the first ∼25 characters, while loose prompts produce explanatory prose where the answer surfaces mid-paragraph, so aggressive truncation removes the answer.

Inversion synthesis. Across the 19-cell judge matrix, prompt-only pairs (same architecture, different prompt) disagree on sign in 17 of 21 pairs (81%; 17/19 non-tied), while architecture-only pairs almost never disagree (1/18; full breakdown App. A.3). The inversion is not a single-cell artifact: whenever the prompt lever moves token-F1 by more than 0.05, LLM-judge moves in the opposite direction.

5.2 Finding 2: a 0.20 F1 swing comes from one character-level format choice

LoCoMo's adversarial category (447/1986 questions, 22.5%) is scored against the exact string "`no information available`". Going from a loose prompt that produces verbal abstentions ("Sorry, I don't have that information...") to a tight-generic prompt that mandates the exact string moves the with-adversarial aggregate F1 by 0.20. Decomposing by category, essentially all of this swing comes from the abstention category itself; no answerable category moves by more than 0.03. The benchmark, in its standard reported form, is more sensitive to a single character-level format choice than to any comparable mechanism in our main matched matrix (max architecture spread $\leq$ 0.030 F1).

5.3 Finding 3: no memory-architecture flavor clears $\alpha = 0.05$ on aggregate F1 under the matched single-call protocol

For each prompt variant we take the max-minus-min F1 across the three Table 1 architectures. On LongMemEval the spreads are 0.0019, 0.0080, 0.0039, 0.0048 at tight-per-cat / tight-generic / loose / tight-fcs against bootstrap CI half-widths of 0.054, 0.034, 0.020, 0.064; every spread at least $4\times$ below its noise floor (Figure 1(b)). No architecture flavor outperforms flat above noise on any LongMemEval cell, including the tight-fcs column the FCS prompt was designed for.

On LoCoMo (no-adversarial), tight-per-cat is borderline: linked-notes is 0.023 *below* flat, exceeding the per-cell bootstrap CI half-width — a real *negative* signal driven by link-expansion pulling content-similar but query-irrelevant turns into LoCoMo's ~588-turn haystacks. Session-bank ties flat within noise.[2]

Extension cells at a glance. Table 1 reports the three cheap (LLM-free retrieval) mechanisms across both benchmarks. The remaining three family-level mechanisms (llm-extract, A-MEM with evolution, multi-call ReAct-lite) are heavier (each uses gpt-4o-mini at indexing or generation), so we report them as spot-check cells on LongMemEval Qwen-14B under the two most diagnostic prompts (tight-per-cat and tight-fcs); full table App. A.11.

Extension cell	F1	Δ vs flat
K_{tight} (llm-extract × tight-per-cat)	0.367	+0.002
K_{fcs} (llm-extract × tight-fcs)	0.435	−0.012
K3 (A-MEM evolution × tight-per-cat)	0.363	−0.002
K4 (A-MEM evolution × tight-fcs)	0.412	−0.035
M (multi-call ReAct-lite × tight-per-cat)*	0.203	**−0.162**

∗ Multi-call ReAct-lite uses two LLM calls per question, so cell M is reported as a directional probe inside our single-call evaluation, not as a within-protocol comparison (§3, §8).

Table 3: Extension-cell architecture deltas vs flat (LongMemEval oracle, n=500, Qwen-14B, BM25 top-5, 96-token cap, identical pipeline). All four within-protocol cells (K_{tight}, K_{fcs}, K3, K4) sit within ±0.035 of flat; cell M's −0.162 is the multi-call-probe deviation. Architecture-null on aggregate F1 holds across all six family-level mechanisms.

[2]The session-bank-vs-flat tight-per-cat case on LoCoMo no-adv is the only marginally positive aggregate effect anywhere in our sweep (paired-bootstrap Δ=+0.010, $p_{\text{one-sided}}$=0.061, n=1986); an order of magnitude below the prompt-engineering lever.

Combined with the main matrix, all six family-level mechanisms produce architecture deltas $\leq$ |0.043| vs flat under the matched single-call protocol (cell M excluded for breaking the single-call rule); none reaches the per-n paired-bootstrap detection floor in the positive direction, and the one above-noise positive proxy (session-bank on LoCoMo tight-per-cat at Δ=+0.010, $p_{\text{one-sided}}$=0.061) is an order of magnitude below the prompt-engineering lever.

5.4 Finding 4: prompt engineering accounts for the entire token-F1 range

The combined prompt-engineering effect is +0.246 F1 on LongMemEval and +0.228 on LoCoMo no-adv (loose → tight-per-cat). This is the entire range from naive RAG to our best non-FCS cell. It decomposes into format-tightness (+0.062 / +0.077) and category-awareness (+0.184 / +0.151), with cross-benchmark replication within ~10% (full table App. A.8). The same pattern appears on linked-notes and session-bank rows of both benchmarks (within bootstrap noise), so prompt engineering is a property of the generator given identical retrieved evidence.

The LoCoMo with-adversarial 0.27 collapse from loose to tight-generic (of which 0.20 is the abstention-string match, §5.2) is the LoCoMo-specific shape of this lever; the broader token-F1 / LLM-judge inversion appears on both benchmarks.

The FCS-only cell lands at F1 0.447, +0.082 above tight-per-cat 0.365 with non-overlapping 90% CIs. Per-category (Tab. 4), gains concentrate on knowledge-update and multi-session—the two categories whose existing prompts contain the most task-framing language. This rewrites §5.3's category-aware effect (+0.184) as two competing effects: per-category surface-form constraint (+0.265) plus a −0.082 cost from task-framing verbs ("synthesize", "compute", "consider") that elicit longer outputs (25,915 vs 7,193 chars across n=500, a 3.6× ratio) which token-F1 against terse golds penalizes. FCS is neutral on LoCoMo no-adv (LJ vs LE Δ=−0.004, LL vs LF +0.002, LO vs LG +0.002; Table 1), where longer gold answers (p75 = 7–8 words vs LongMemEval 2–3) leave no length-compression headroom. A stripped ablation (cell Q: F1 = 0.407, 29 chars; App. A.18) decomposes the gain roughly 50/50 between linguistic stripping and surface-form templates, both halves collapsing under LLM-judge. Field best-practice prompt engineering reverse-optimizes under token-

Category	n	E	J	Δ (J−E)
knowledge-update	78	0.442	0.600	+0.158
multi-session	133	0.130	0.265	+0.135
single-session	156	0.578	0.614	+0.036
temporal	133	0.307	0.344	+0.037

Table 4: Per-category breakdown of the +0.082 FCS gain (J vs E) on LongMemEval oracle. The gain concentrates on the two categories whose existing prompts contain the most task-framing language; single-session and temporal (less task framing) move within bootstrap noise.

System	n	Δ vs J	p (1-sided)
Mem0	200	+0.021	0.249
HippoRAG	449	$+0.030^{\dagger}$	$0.042^{\dagger}$
A-MEM-evo	200	$-0.043^{\dagger}$	$0.043^{\dagger}$

Table 5: Three faithful upstream-code reproductions paired against cell J on LongMemEval oracle; pipelines in App. A.12. †: two unadjusted-α=0.05 crossings in opposite directions; neither survives Bonferroni-7; both collapse under both LLM judges (Tab. 6).

F1.

5.5 Mechanism checks for Finding 1

We run four mechanism checks against natural attacks on Finding 1 (full table App. 12).

(a, b) *Cross-family LLM-judge.* The FCS J-vs-E token-F1 gain (+0.082) collapses to +0.005 under gpt-4o-mini (96% paired-decision agreement) and to 0.000 under Claude-Sonnet-4.5. Cross-judge κ=0.71, with identical direction on every comparison; this bounds single-family artifacts but does not establish judge independence.

(c) *Truncate-rejudge.* Loose at 170 chars still loses to tight at 47 (χ^2=3.2, $p\approx$0.074, borderline n.s.). The mechanism is informational-with-length-floor, not pure judge length preference.

(d) *Two-rater blind audit (n=100, sealed X/Y key).* B/E/tie splits: rater 1 (author) 14/2/84; rater 2 (outside PhD researcher) 13/1/86. Per-output label agreement κ=0.96 (98% raw); on 12 items where both expressed a non-tie preference, 12/12 agree on direction (McNemar p=1.0). Both raters share academic PhD-rater training (App. A.6, §8).

5.6 Faithful reproductions: aggregate-null across three published systems

To address the natural reviewer objection that we did not faithfully reproduce any of the published systems, we run three faithful reproductions on LongMemEval oracle, each using the system's own published code: Mem0 (Chhikara et al., 2025) via `mem0ai==2.0.1`, HippoRAG (Gutiérrez et al., 2025) via the upstream `hipporag` package, and A-MEM with evolution (Xu et al., 2025) via the upstream `agentic_memory` repo. In each, the system's memory/retrieval layer replaces flat BM25 while everything else (Qwen2.5-14B, FCS prompt, 96-token cap) is held to cell J's settings, paired on the same n=200 stratified subset (50/label, seed=20260506). All three use gpt-4o-mini as the extraction LLM. A note on scope: published numbers were typically reported with each system's own end-to-end answer-generation prompt. By forcing all three through cell J's FCS prompt we answer "does the architecture help when generation is held to a strong, format-aligned baseline?" This is strictly different from "does each system's published gain reproduce" (out of scope here).

Mem0: aggregate-null with a real per-category trade-off. Mem0+FCS lands at aggregate F1 0.483 vs cell J 0.462, Δ=+0.021, 90% CI $[-0.031, +0.075]$, p=0.249. Per-label (full table App. A.13, Tab. 16): Mem0 wins +0.25 F1 on multi-session—the only above-noise positive architecture effect in our sweep—but loses ∼8 pp on single-session (−0.082) and knowledge-update (−0.079); temporal ties within noise. Mechanistically: LLM-extraction surfaces cross-session synthesis facts BM25 misses, but drops precise lookup details BM25 surfaces verbatim. Aggregate-null hides a real mechanism and a real cost, motivating community-standard (iii) (§6).

Reading HippoRAG and A-MEM; cross-family LLM-judge. HippoRAG (n=449 stratified, the maximum paired indices against cell J) crosses unadjusted α=0.05 positively; A-MEM crosses negatively at n=200, where LLM-mediated consolidation acts as an information bottleneck under the 96-token budget. Per-question pairwise: HippoRAG 35/27/138 wins/losses/ties; A-MEM 28/39/133 at $|\Delta|$>0.05. Neither survives Bonferroni-7.[3] Rerunning all three cell-pairs through gpt-4o-mini and Claude-Sonnet-4.5 (paired McNemar at n=200, Tab. 6):

[3] Bonferroni-7 corresponds to the 7 within-protocol architecture-vs-baseline tests in our sweep: 4 family-level proxies (linked-notes, session-bank, llm-extract, A-MEM-evolution-proxy) vs flat, plus 3 faithful upstream-code reproductions (Mem0, HippoRAG, A-MEM with evolution) vs cell J. Multi-call ReAct-lite is excluded as a 2-call out-of-protocol probe (§3).

vs cell J	Token-F1		gpt-4o-mini		Claude-Sonnet-4.5	
	Δ	p	Δ	p	Δ	p
Mem0	$+0.021$	0.249	-0.065	0.118	-0.005	1.00
HippoRAG†	$+0.030$	0.042	$+0.010$	0.880	$+0.020$	0.651
A-MEM†	-0.043	0.043	-0.055	0.127	-0.045	0.253

Table 6: Cross-family LLM-judge on faithful-repro cells (paired McNemar two-sided p; HippoRAG token-F1 at n=449, others n=200). †: the two unadjusted-α=0.05 token-F1 crossings (opposite directions); both vanish under both judges.

Every effect is below judge noise (all six judge cells n.s. at α=0.05). HippoRAG and A-MEM preserve direction under both judges; Mem0's small token-F1 lead (+0.021, n.s.) is statistically indistinguishable from a sign flip under gpt-4o-mini (Δ=-0.065, p=0.118) and from zero under Claude (Δ=-0.005, p=1.00). Architecture-null is consistent across token-F1, gpt-4o-mini, and Claude-Sonnet-4.5.

Native-prompt sanity check. To check the natural concern that holding all three reproductions to the FCS prompt + 96-token cap is closer to hobbling than steel-manning, we re-run each system's memory layer under its own native loose answer prompt + 256-token budget on the same n=200 paired indices (App. A.14, Tab. 17): Mem0 F1 = 0.081 (avg 233 chars); HippoRAG = 0.175 (262); A-MEM = 0.176 (274). All three systems produce 233–274-char outputs that token-F1 collapses to 0.08–0.18, reversing the hobbling reading: under token-F1, FCS is the *favorable* treatment because it compresses output to match terse golds. Architecture-null holds (and tightens) in the native-prompt regime; what it does not survive is the metric inversion that Finding 1 establishes for the prompt lever.

n asymmetry note. HippoRAG streams the haystack into a passage-entity graph at index time, so the paired-index pool is bounded only by the LongMemEval oracle split itself (n=500 unique IDs, of which 449 pair against cell J after stratification). Mem0 and A-MEM rebuild a fresh per-question store with gpt-4o-mini LLM passes (Mem0: extraction; A-MEM: note construction + memory evolution at ~200s/question with `evo_threshold=10`), so n=200 exhausts the practical paired-index budget; details in App. A.12.

Per-n paired-bootstrap MDE breakdown is in App. A.12 (matching the abstract/§5.1-Finding-3 framing): every effect lands below judge MDE under both judges, and "architecture-null" reflects a multiple-comparisons-and-judge-replication verdict rather than an assertion that effects are below F1 detection at every n.

6 Discussion

Our negative result is bounded to the matched single-call protocol on per-question oracle haystacks; multi-call and persistent-memory regimes remain out of scope (§8). Within this protocol, future memory-architecture claims should report (i) a surface-form-matched flat-RAG baseline whose length cap, output template, and abstention string match the proposed system's prompts; (ii) output-length statistics per cell; (iii) per-category F1 with paired-bootstrap CIs and a pre-computed MDE; (iv) LLM-judge accuracy with paired McNemar on at least two judge model families. FCS is the diagnostic version of (i); since it is calibrated on test-gold surface form, a submission baseline should instead derive surface-form constraints from each system's own published prompts or from a held-out dev split. App. A.19 packages (i)–(iv) as a one-page checklist.

7 Conclusion

Under our matched single-call protocol, token-F1 inverts cross-family LLM-judge rankings on prompt-engineering interventions (tight wins under token-F1, loose wins under both judges by 7–18 pp), and a single LoCoMo abstention-string match shifts aggregate F1 by 0.20. Across six family-level mechanisms, five backbones, and three faithful upstream-code reproductions, no architecture clears Bonferroni-7-corrected aggregate F1, though Mem0 wins per-category on multi-session. Published architecture gaps are not interpretable without (i) a surface-form-matched flat-RAG baseline, (ii) output-length statistics, (iii) per-category paired-bootstrap intervals, and (iv) cross-family LLM-judge accuracy (§6).

8 Limitations

Backbone scope. Open-weight backbones only (Qwen2.5-3B/7B/14B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct); closed-source models (GPT-4o-mini, Claude, Gemini) are excluded because the matched protocol requires per-question replication.

Architecture coverage. The six family-level mechanisms plus three faithful upstream-code reproductions cover the architectures whose published gains we audit. Learned mechanisms (retrieval-augmented training, RL-tuned policies) are out of scope; the pre-registered falsification criterion in §7 specifies the evidence that would revise the verdict.

Two benchmarks, both English long-horizon dialogue. LongMemEval and LoCoMo are the regime in which the audited claims were originally reported. BEAM's verbose multi-sentence golds make token-F1 uninformative there (Tavakoli et al., 2026); we make no Finding-3 claim outside our matched matrix.

LLM-judge sample size. The 19-cell judge matrix runs at n=200/cell, calibrated to the standing-challenge MDE $\Delta \approx 0.14$ at $\sim 94\%$ one-sided power (App. A.15); larger n would tighten CIs but not change the unadjusted-α=0.05 verdict for any architecture cell we report.

LLM-judge agreement with human correctness. The chain LLM-judge $\approx$ human correctness rests on a two-rater blind audit (Cohen's κ=0.96, §5.5d). Both annotators share academic PhD-rater training; we cannot exclude that labor-platform-recruited annotators would show different patterns. A multi-annotator platform study is a natural extension for camera-ready.

Faithful reproductions tested for three systems. Mem0, HippoRAG, and A-MEM-with-evolution are reproduced via upstream packages (§5.6). Systems whose architectural primitives break the matched single-call protocol (MemGPT's multi-call loop, MemoryBank's forgetting curve, RMM, AgeMem, MLMA) are out of scope by construction.

FCS rendering is intentionally minimal. The deliberately simple rendering rules (p75-word length cap, first observed abstention string, dominant-form template) keep FCS interpretable as a diagnostic instrument; the $+0.082$ token-F1 gain is therefore a lower bound on what surface-form alignment can buy.

Ethics Statement

Human-rated content. The n=100 blind audit (§5.5d) was performed by two raters under sealed-key X/Y per-item randomization: one paper author and one outside PhD researcher recruited from the authors' professional network with no involvement in this work's design. No personally identifiable information or sensitive content was elicited or annotated. We disclose the recruitment channel (authors' professional network rather than an anonymous labor platform) so reviewers can correctly bound the residual rater-population concern.

LLM API usage. Cross-family judge replication used GPT-4O-MINI (OpenAI) and CLAUDE-SONNET-4.5 (Anthropic) at a total cost of about $5. We do not deploy or release a system; the judges serve as measurement instruments only.

Compute and environmental cost. Primary experiments used 2×A100-80GB. The full matrix (cross-backbone replication, faithful reproductions, judge runs) consumed roughly 200 GPU-hours.

Reproducibility and release. Per-cell prediction JSONs, the FCS rendering code, the bootstrap-noise utility, the paired LLM-judge harness, and the two-rater audit labels are released at `https://anonymous.4open.science/r/token-f1-inverts-llm-judge-36C6` (anonymized for review). The community-standard reproducibility checklist is Appendix A.19. We do not release any pretrained weights; all models used are publicly available under their respective licenses.

AI assistance disclosure. The authors used LLM-based writing assistants for language polishing and copy editing on portions of the manuscript. All claims, experimental design, statistical analysis, and conclusions are the authors' own. No AI assistance was used to generate experimental data.

Intended use and potential misuse. The paper audits an evaluation methodology and does not introduce a new system to be deployed. The FCS prompt is presented explicitly as a diagnostic audit instrument, not a leaderboard-submission method, so that test-gold-calibrated prompts do not become normalized as a publication baseline.

References

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *Preprint*, arXiv:2104.02145.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *Preprint*, arXiv:2202.07654.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *Preprint*, arXiv:2504.19413.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, Alysa Zhao, Dingyi Kang, Xu Hu, Feng Chen, Qiannan Li, and Bingzhe Li. 2026. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations. *Preprint*, arXiv:2602.19320.

Michael T. Kane. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. 2026. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. *Preprint*, arXiv:2602.10715.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *Preprint*, arXiv:2402.17753.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. Kilt: a benchmark for knowledge intensive language tasks. *Preprint*, arXiv:2009.02252.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on performance of large language models. *Preprint*, arXiv:2408.02442.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *Preprint*, arXiv:2503.08026.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. 2026. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *Preprint*, arXiv:2510.27246.

Sunil Tiwari and Payal Fofadiya. 2026. Multi-layered memory architectures for llm agents: An experimental evaluation of long-term context retention. *Preprint*, arXiv:2603.29194.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval:

Benchmarking chat assistants on long-term interactive memory. *Preprint*, arXiv:2410.10813.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. 2025. How memory management impacts llm agents: An empirical study of experience-following behavior. *Preprint*, arXiv:2505.16067.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *Preprint*, arXiv:2107.07567.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *Preprint*, arXiv:2502.12110.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. 2026. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. *Preprint*, arXiv:2601.01885.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix: detailed tables

This appendix collects the full tables for the body subsections that summarize them inline.

A.1 Cell-letter mapping (referenced throughout)

Single-letter cell labels (B, E, J, Q, . . .) are used throughout the paper as compact identifiers for specific (architecture × prompt × benchmark) configurations. The body introduces B/E/J inline at first use in §5.1; this table is the complete mapping. Subscripts on K / N / P index prompt variants of the same retrieval mechanism (e.g., K_{tight} vs K_{fcs} for the llm-extract family proxy under different prompts). Cells with the L prefix are LoCoMo counterparts of LongMemEval cells.

A.2 Implementation-faithfulness table (cf. §3)

Table 8 maps architecture-family mechanisms tested under the matched protocol to their parent published systems. Each row notes what we implemented, what we omitted, and why the audit is fair under matched-pipeline conditions. The audit operates at the mechanism level (does the core idea produce a positive F1 effect when the rest of the pipeline is held fixed?) rather than the system level. Three rows include faithful end-to-end reproductions of the parent system (Mem0, HippoRAG, A-MEM with evolution; §5.6); the remaining mechanisms are family-level proxies under matched conditions. A paper whose claim survives a surface-form-matched flat-RAG comparison on its own backbone is uncontradicted by our findings.

A.3 19-cell inversion synthesis (cf. §5.1)

To check whether the cell-B-vs-cell-E inversion in §5.1 is anecdotal or systematic, we tabulate sign-of-Δ agreement between token-F1 and gpt-4o-mini judge across every comparable pair in the 19-cell judge matrix. Pairs are classified as agree (signs match), disagree (signs differ), or tie ($|\Delta| \leq 0.005$ on either metric). Pairs are partitioned by which lever varies: prompt-only (same architecture, different prompt) vs architecture-only (same prompt, different architecture). The inversion is not a single-cell artifact: whenever the prompt lever moves token-F1 by more than 0.05 F1, LLM-judge moves in the opposite direction on 17 of 19 such pairs. The two agreeing prompt-only pairs are J vs E (FCS, where Δ is tiny on both metrics) and LE vs LA on LoCoMo ($\Delta_{judge} = +0.005$); token-F1 and LLM-judge agree only when neither moves much. Architecture-only pairs almost never disagree (1/18, the LongMemEval C-vs-H tight-generic pair where Δ is below F1 noise on both axes); this is consistent with Finding 3.

A.4 Cross-judge replication on Claude-Sonnet-4.5 (cf. §5.5b)

The metric-inversion in §5.1 used gpt-4o-mini as judge. To rule out that the inversion is an OpenAI-judge-specific artifact, we re-run the cell B (flat × loose), cell E (flat × tight-per-cat), and cell J (flat × tight-fcs) judging on Claude-Sonnet-4.5 on the same paired n=200 subset of LongMemEval oracle. The two judges' verdicts agree on 86% of paired questions on the union of $\{B, E, J\}$ verdicts (Cohen's $\kappa = 0.71$, substantial agreement). Both judges return the same direction on every comparison, with absolute accuracy values within 0.01. The metric-inversion is therefore a property of token-F1 disagreeing with LLM-judged semantic correctness across two judge model families, not of any single judge model. A single-judge LLM evaluation does not protect against shared judge-family biases; the cross-family replication does, though as §8 notes, two LLM judges could still share systematic biases that a multi-annotator human study would surface.

Letter	Architecture	Prompt	Benchmark	Section / table
A	flat	tight-generic	LongMemEval	Table 1
B	flat	loose	LongMemEval	§5.1
C	linked-notes	tight-generic	LongMemEval	Table 1
D	linked-notes	loose	LongMemEval	Table 1
E	flat	tight-per-cat (existing)	LongMemEval	§5.1
F	linked-notes	tight-per-cat	LongMemEval	Table 1
G	session-bank	tight-per-cat	LongMemEval	Table 1
H	session-bank	tight-generic	LongMemEval	Table 1
I	session-bank	loose	LongMemEval	Table 1
J	flat	tight-fcs	LongMemEval	§5.4
K_{tight} / K_{fcs}	llm-extract	tight-per-cat / tight-fcs	LongMemEval	App. Tab. 15
K3 / K4	A-MEM evolution	tight-per-cat / tight-fcs	LongMemEval	App. Tab. 15
L	linked-notes	tight-fcs	LongMemEval	§5.3
M	multi-call ReAct-lite	tight-per-cat	LongMemEval	App. Tab. 15
N1 / N2	dense E5 (flat)	tight-per-cat / tight-fcs	LongMemEval	§5.3
O	session-bank	tight-fcs	LongMemEval	§5.3
P1 / P2	BM25+rerank (flat)	tight-per-cat / tight-fcs	LongMemEval	§5.3
Q	flat	tight-stripped	LongMemEval	App. A.18
LA–LI	LoCoMo counterparts of LongMemEval A–I	—	LoCoMo	Table 1
LJ	flat	tight-fcs	LoCoMo	§5.3
LL	linked-notes	tight-fcs	LoCoMo	§5.3
LO	session-bank	tight-fcs	LoCoMo	§5.3

Table 7: Cell-letter mapping. Single-letter cell labels used throughout the paper, with each letter's (architecture, prompt, benchmark) configuration. Cells with subscripts (K, N, P) index variants. Cells with the L prefix are LoCoMo counterparts.

A.5 Truncate-and-rejudge full curve (cf. §5.5c)

To distinguish whether loose's gain over tight under LLM-judge reflects (a) more correct information or (b) a judge length bias toward longer answers, we truncate cell B (loose) outputs at multiple per-question budgets and re-run gpt-4o-mini judge on the same n=200 paired subset. The accuracy-vs-budget curve is monotone increasing. The key datapoint: at 170 chars (more than 3× tight's 47-char average), loose still loses to tight at 47 chars (judge accuracy 0.475 vs 0.530). A pure judge length-bias hypothesis would predict loose at this length to already win on length grounds alone; it does not. Only at the full 311 chars does loose pull ahead. The mechanism we read is informational with a length floor: tight prompts elicit outputs that front-load the answer in the first ∼25 characters, while loose prompts produce explanatory prose where the correct answer surfaces mid-paragraph alongside surrounding evidence. Aggressive truncation removes the answer. Token-F1 punishes the prose; LLM judges credit the answer wherever it surfaces, but only if the answer survives the truncation budget.

A.6 Two-rater blind audit (cf. §5.5d)

The author labeled 100 paired (cell B loose, cell E tight-per-cat) outputs from LongMemEval oracle under randomized per-question blinding. Procedure: sample 100 paired indices from the n=200 judged subset (seed=20260506); for each item, randomize whether the loose-prompt or tight-prompt output is shown first as "Output X" vs "Output Y"; commit the unblind key (`UNBLIND_KEY_DO_NOT_SHOW_RATERS.json`) to a sealed file before labeling begins; the rater labels each output as correct / partial / incorrect blind to which cell produced it, plus an explicit pairwise preference column (X, Y, or tie); open the unblind key only after all 100 items are submitted. The X/Y blinding bounds direction-bias on the loose-vs-tight comparison even though the rater is the paper author.

After unblinding (rater 1, paper author): explicit pairwise preference is 14 B-prefer vs 2 E-prefer with 84 ties (binomial sign test on n=16 non-ties: p=0.0021, significant at α=0.05 in the loose-over-tight direction). Human correctness rates: B 64% vs E 58%, Δ=+0.06 (paired McNemar χ^2=2.08, borderline n.s. on 12 discordant pairs; the explicit pairwise preference is the stronger statistical handle). Per-cell agreement of rater 1 with gpt-4o-mini: B 88.0% (88/100), E 94.0% (94/100). The full n=200 gpt-4o-mini gap was +0.070 in the same direction.

Original system family	What we implemented	What we omitted	Why this is fair for this audit
A-MEM (Xu et al., 2025)	linked-notes (family proxy); plus a faithful A-MEM-with-evolution reproduction (§5.6) using the upstream `agentic_memory` package with gpt-4o-mini note layer + all-MiniLM-L6-v2 + ChromaDB and `evo_threshold=10`	Full system-specific deployment configuration	Tests both the family-level note-linking mechanism and the actual A-MEM pipeline with cross-note evolution against flat-BM25+FCS.
MemoryBank / Mem0g (?Chhikara et al., 2025)	session-bank as a family proxy; plus a faithful Mem0 reproduction (§5.6) using `mem0ai==2.0.1` with gpt-4o-mini extraction + text-embedding-3-small under our FCS prompt	MemoryBank's forgetting curve; Mem0g's graph variant (we test the non-graph Mem0 path)	Tests both the family-level mechanism under matched protocol and the actual Mem0 pipeline against our flat-BM25+FCS baseline.
MemGPT (Packer et al., 2024)	multi-call ReAct-lite (planner ANSWER/SEARCH + answerer)	Full OS-style memory hierarchy, full tool loop, persistent memory across questions	Tests whether a lightweight multi-call search-and-answer mechanism changes results under the same setting.
HippoRAG (Gutiérrez et al., 2025)	faithful HippoRAG reproduction (§5.6) using the upstream `hipporag` package with gpt-4o-mini OpenIE/NER + text-embedding-3-small + Personalized PageRank (no family proxy in our axis after the §3 reorganization)	Full original deployment-time settings	Tests the actual HippoRAG pipeline against flat-BM25+FCS.
RMM / AgeMem / MLMA (Tan et al., 2025; Yu et al., 2026; Tiwari and Fofadiya, 2026)	Out of scope	Persistent lifetime memory, RL-tuned policies, layered memory	Not cleanly testable under matched single-call + per-question oracle haystack.

Table 8: Architecture-family mechanisms tested under the matched protocol. The audit operates at the mechanism level (does the core idea produce a positive F1 effect when the rest of the pipeline is held fixed?) rather than the system level. §8 expands on the per-paper-audit caveat.

Inter-rater statistics (rater 1 vs rater 2). A second rater (outside PhD researcher with no involvement in this work's design, recruited from the authors' professional network) independently labeled the same n=100 items under the same X/Y blinding key. Rater 2's preference: 13 B-prefer vs 1 E-prefer with 86 ties (Wilson 95% CI for B-share-of-non-ties: $[0.685, 0.987]$), independently replicating the loose-over-tight direction. Per-output label agreement on the {correct, partial, incorrect} scale: Cohen's κ=0.96 (κ_X=0.98, κ_Y=0.94; 98% raw agreement on n=200 label decisions). Preference agreement: κ=0.77 on the 3-way {B, E, tie} scale (94% raw); on the 12 items where both raters expressed a non-tie preference, they agree on direction in $12/12$ (McNemar χ^2=0, p=1.0 — no rater-direction disagreement detectable). Rater 2's tie decomposition (n=86): 51 both-correct, 31 both-incorrect, 3 mixed-label, 1 both-partial — matching rater 1's shape (49/32/2/1). Both annotators share academic PhD-rater training; we cannot exclude that a labor-platform-recruited annotator would show lower agreement (§8). Analysis script and per-item paired labels: `experiments/crag-9-faithful-repro/human_annotation_ra`

84-tie decomposition. Unblinding the per-output correct / partial / incorrect labels on the 84 items where the rater preferred neither cell yields the following: 49 both-correct (58% of ties; B's full-

Pair type	# pairs	agree	disagree	tie
prompt-only (same architecture, different prompt)	21	2	17	2
architecture-only (same prompt, different architecture)	18	8	1	9
all comparable	39	10	18	11

Table 9: Inversion synthesis. Prompt-only pairs disagree on sign in 17/21 (81%) of cases across both benchmarks; architecture-only pairs almost never disagree (1/18) and mostly tie within noise, consistent with Finding 3.

Cell	gpt-4o-mini judge acc	Claude-Sonnet-4.5 judge acc	Token-F1 (n=500)
B (flat × loose)	0.600	0.610	0.120
E (flat × tight-per-cat)	0.530	0.530	0.365
J (flat × tight-fcs)	0.535	0.530	0.447
Δ(B−E): loose vs tight-per-cat	+0.070	+0.080	−0.246
Δ(J−E): tight-fcs vs tight-per-cat	+0.005	0.000	+0.082

Table 10: Cross-judge replication on the same n=200 paired LongMemEval subset. Both judges, from different model families (OpenAI / Anthropic), return the same direction on every comparison. Cohen's κ between the two judges' verdicts on $\{B, E, J\}$ is 0.71 (substantial agreement).

sentence prose and E's terse answer convey the same correct content), 32 both-incorrect (38%; the question is hard for the underlying 14B backbone and both cells err), 1 both-partial, and 2 mixed-label items where the rater preferred "tie" despite differing per-output labels (rater inconsistency). Reading: of the 100 items, **81 sit at the rater's discrimination ceiling or floor** (49 both-correct + 32 both-incorrect); the remaining **19 carry discriminable signal** (14 B-favor + 2 E-favor + 3 mixed-label or both-partial). Restricting to those 19, B wins 14 to E's 2. The 32 both-incorrect ties form a backbone-correctness floor on this n=100 subset, not metric disagreement. The 49 both-correct ties are the most informative slice: on items where the model produces the right answer, B's verbose form and E's terse form are semantically equivalent under rater judgment, yet are ranked 0.246 apart by token-F1. This is direct corroboration of Finding 1's mechanism. The full per-item tie breakdown is included in the released artifact (https://anonymous.4open.science/r/token-f1-inverts-llm-judge-36C6).

A.7 Mechanism-check summary table (cf. §5.5)

A.8 Prompt-engineering decomposition (cf. §5.4)

A.9 Tight-per-cat-stripped table (cf. §A.18)

A.10 Full architecture-mechanism descriptions (cf. §3)

Full descriptions of the six family-level memory-architecture mechanisms and three retrieval substrates referenced in §3, plus the three faithful upstream-code reproductions reported in §5.6. The main 3×3 factorial uses the first three family-level mechanisms; the remaining three family-level mechanisms are extensions reported in §5.3.

Architecture axis. *flat*: standard BM25 over the per-turn haystack. *linked-notes* (A-MEM-style turn-grain content linking): BM25-retrieve top-3 seed turns, for each seed BM25-retrieve top-3 most lexically similar neighbour turns, re-rank the union by query-BM25, and return top-5. *session-bank* (MemoryBank/Mem0g-style session-grain retrieval): BM25-rank sessions, take top-2, return top-5 turns within. *llm-extract*: LLM-extract 3–5 memory notes per session at index time, BM25-rank notes against the query at query time, return query-BM25-ranked turns from the top sessions. This is our family proxy for both the Mem0 LLM-extraction and A-MEM note-construction mechanisms. *A-MEM with evolution*: llm-extract plus an LLM cross-session memory-merge pass that contradicts-updates or supersedes notes. *multi-call ReAct-lite*: a planner LLM call decides ANSWER

Variant	Avg length	Judge accuracy	McNemar χ^2 vs tight
tight-per-cat (cell E)	47 chars	0.530	—
loose, truncated to per-Q tight length	47 chars	0.105	79.3 (sig $p < 0.001$)
loose, truncated to 2× tight	103 chars	0.255	47.8 (sig)
loose, first sentence	121 chars	0.445	6.9 (sig)
loose, truncated to 4× tight	170 chars	0.475	3.2, $p \approx 0.074$ (borderline n.s. at $\alpha = 0.05$)
loose-full (cell B)	311 chars	0.600	sig (full matrix)

Table 11: Truncate-and-rejudge: judge accuracy as a function of length budget granted to loose-prompt outputs.

Check (n=)	Comparison	Token-F1 Δ	LLM-judge Δ
(a) gpt-4o-mini (200)	J vs E (FCS vs tight-per-cat)	+0.082	+0.005 (n.s.)
(b) Claude-Sonnet-4.5 (200)	J vs E (FCS vs tight-per-cat)	+0.082	0.000 (n.s.)
(b') Claude-Sonnet-4.5 (200)	B vs E (loose vs tight-per-cat)	−0.246	+0.080 (sig)
(c) Truncate-rejudge (200)	B@170 chars vs E@47 chars	—	−0.055 (χ^2=3.2, p≈0.074)
(d) Blind audit (100)	B vs E preference	—	B/E/tie 14/2/84; sign test p=0.002

Table 12: Mechanism-check summary for Finding 1; summarized inline in §5.5.

or SEARCH; on SEARCH a second answerer call generates the final answer. This is the family proxy for MemGPT-style multi-call regimes within our single-call evaluation cell.

Retrieval substrate axis. Three substrates: BM25 (default), dense E5-base-v2 cosine similarity, and BM25-top-20 with ms-marco-MiniLM-L-6-v2 cross-encoder rerank to top-5.

The full cell-letter mapping for these mechanisms and retrieval substrates is App. Tab. 7 (App. A.1, promoted to the front of the appendix).

A.11 Extension-cell results: llm-extract, A-MEM evolution proxy, multi-call ReAct-lite (cf. §5.3)

The main 3×3 matrix (Tab. 1) reports flat / linked-notes / session-bank under each prompt variant. The remaining three family-level mechanisms (cells K/K3/K4/M; §3 architecture axis) are reported here on LongMemEval oracle (n=500, Qwen2.5-14B, identical pipeline). All four extension cells land within bootstrap noise of the corresponding flat cell, except multi-call ReAct-lite (cell M), which lands substantially below flat under tight-per-cat. We attribute this to unreliable SEARCH-route routing under the matched single-call evaluation cell rather than a benchmark artifact. Architecture-null on aggregate F1 (Finding 3) holds across all extension cells.

A.12 Faithful-reproduction pipeline details (cf. §5.6)

For each system in Table 5, we instantiate a fresh per-question store using the system's own published code, ingest the LongMemEval oracle haystack, query with the question text, and hand the retrieved memories/passages to Qwen2.5-14B-Instruct under the cell-J FCS prompt template. The same n=200 stratified subset of LongMemEval oracle is used across all three systems (50 questions per label, seed=20260506); paired-bootstrap p-values use $B = 10{,}000$ resamples on the matched indices.

Mem0. `mem0ai==2.0.1`. Per question, build a fresh Mem0 store with LLM-driven extraction (gpt-4o-mini, Mem0's default LLM layer); ingest haystack turns as a conversation in chunked batches; query with the question text; hand the retrieved memories to the generator. 90% paired-bootstrap CI: $[-0.031, +0.075]$.

HippoRAG. Upstream `hipporag` package. Per question, build a fresh HippoRAG store: gpt-4o-mini for entity extraction and OpenIE triple extraction over each haystack turn-as-passage, text-embedding-3-small for passage and entity embeddings, bipartite passage-entity graph, Personalized PageRank from query-entity seeds returning top-5 passages. 90% paired-bootstrap CI: $[-0.008, +0.073]$. Per-question pairwise: 35 HippoRAG wins, 27 cell-J wins, 138 ties at $|\Delta|>0.05$.

A-MEM with evolution. Upstream `agentic_memory` package. Per question, instantiate a fresh `AgenticMemorySystem`: gpt-4o-mini for note construction and the memory-evolution LLM layer, all-MiniLM-L6-v2 for embeddings, ChromaDB for the vector store, with `evo_threshold=10` so cross-note consolida-

Effect	LME ΔF1	LoCoMo no-adv ΔF1
format tightness (loose → tight-generic)	+0.062	+0.077
category-aware (tight-generic → tight-per-cat)	+0.184	+0.151
combined (loose → tight-per-cat)	+0.246	+0.228

Table 13: Prompt-engineering decomposition under token-F1 (flat retrieval, both benchmarks); summarized inline in §5.4.

Cell	Token-F1 (n=500)	Avg completion chars	LLM-judge accuracy (n=200)
E (tight-per-cat)	0.365	52	0.530
Q (tight-stripped)	0.407	29	0.540
J (tight-fcs)	0.447	14	0.535

Table 14: Decomposition of the +0.082 FCS gain. The tight-stripped cell (Q) sits cleanly between tight-per-cat (E) and tight-fcs (J) on both token-F1 and average completion length, but is essentially indistinguishable from either under LLM-judge.

Cell	Mechanism × prompt	F1	Δ vs flat
K_{tight}	llm-extract × tight-per-cat	0.367	+0.002
K_{fcs}	llm-extract × tight-fcs	0.435	−0.012
K3	A-MEM evolution × tight-per-cat	0.363	−0.002
K4	A-MEM evolution × tight-fcs	0.412	−0.035
M	multi-call ReAct-lite × tight-per-cat	0.203	−0.162
reference: flat			
E	flat × tight-per-cat	0.365	—
J	flat × tight-fcs	0.447	—

Table 15: Extension-cell results on LongMemEval oracle. K/K3/K4 land within ±0.035 of flat under the same prompt, all within per-cell bootstrap noise (cf. §5.3). M (multi-call ReAct-lite) lands −0.162 below flat under tight-per-cat, a negative architecture effect from unreliable planner routing under our single-call evaluation cell.

tion triggers within the ∼22-turn LongMemEval haystacks. Each turn becomes a note via `add_note`; `search_agentic` returns top-5 evolved notes. 90% paired-bootstrap CI: [−0.084, −0.002]. Loses on every category under the matched protocol (knowledge-update −0.069, multi-session −0.052, temporal −0.032, single-session −0.018): LLM-mediated memory consolidation under a 96-token answer budget acts more like an information bottleneck than a synthesis amplifier on these question types.

Why n=449 but Mem0/A-MEM at n=200. HippoRAG streams the haystack into a passage-entity graph at index time, so the per-question paired-index pool is bounded only by the LongMemEval oracle split itself (n=500 unique question IDs, of which 449 pair against cell J after stratification and exclusions). Mem0 and A-MEM rebuild a fresh per-question store with gpt-4o-mini-driven LLM passes (Mem0: extraction; A-MEM: note construction + memory evolution); at the API budgets and wall-clock costs we ran (∼200s/question for A-MEM with `evo_threshold=10`), the n=200 stratified subsample exhausts the practical paired-index budget. Extending those two to n=449 is straightforward in principle (re-run on the additional 249 questions) but was deprioritized after the cross-family LLM-judge result (Tab. 6) showed both deltas collapse to noise under judges. The additional n would tighten F1 CIs but not change the unadjusted-α=0.05 verdict.

A.13 Faithful Mem0 per-label trade-off table (cf. §5.6)

Label (n=50 each)	Mem0 F1	Cell J F1	Δ
single-session	0.492	0.574	−0.082
knowledge-update	0.571	0.650	−0.079
temporal	0.374	0.379	−0.005
multi-session	0.494	0.245	**+0.250**
aggregate	0.483	0.462	+0.021

Table 16: Faithful Mem0 per-label trade-off vs cell J on the same n=200 paired indices (50 per label, stratified, seed=20260506). Aggregate-null hides a real architectural mechanism (multi-session synthesis +0.25) canceled by lookup losses ∼ − 0.08.

A.14 Native-prompt sanity check table (cf. §5.6)

A.15 McNemar power curve (cf. §3, §7)

We use the exact-binomial McNemar test on n=200 paired LLM-judge verdicts with observed paired-discordance rate p_{disc}≈0.4 (i.e., $n_{\text{disc}}=n_{01}+$

System	Native F1	Avg chars	$\Delta_{\text{native}-\text{FCS}}$
Mem0	0.081	233	-0.40
HippoRAG	0.175	262	-0.29
A-MEM (evolution)	0.176	274	-0.22

Table 17: Native-prompt sanity check: each system's published loose answer prompt + 256-token budget on the same n=200 paired indices, vs the cell-J FCS variant (96-token cap). Token-F1.

$n_{10}\approx 80$). Under H_0, $n_{01}\sim\text{Binomial}(80, 0.5)$. The critical region at α=0.05:

- One-sided: reject if $n_{01}\geq 48$, i.e. $n_{01}-n_{10}\geq 16$ (accuracy gap $\Delta\geq 0.08$).
- Two-sided: reject if $n_{01}\geq 50$ or $n_{01}\leq 30$ (accuracy gap $|\Delta|\geq 0.10$).

Power is computed under H_1: $n_{01}\sim\text{Binomial}(80, p_{H_1})$ where $p_{H_1}=n_{01}^{\text{true}}/80$ matches the assumed effect size.

Δ_{acc}	1-sided power	2-sided power
0.04	0.22	0.11
0.05	0.29	0.16
0.07	0.46	0.29
0.10	0.72	0.55
0.12	0.85	0.72
0.14	0.94	0.86
0.15	0.96	0.91
0.18	1.00	0.98

Table 18: Exact-binomial McNemar power at n=200 paired, p_{disc}=0.4, α=0.05.

The paper's pre-registered MDE for the standing challenge is $\Delta\approx 0.14$ accuracy at $\sim 94\%$ one-sided / $\sim 86\%$ two-sided power (Tab. 18). The next-lower bar of $\sim 80\%$ one-sided power corresponds to $\Delta\approx 0.115$. An earlier draft incorrectly stated $\sim 90\%$ power at Δ=0.07, which is actually $\sim 46\%$ one-sided.

The observed loose-vs-tight gaps (Δ=0.07 on LongMemEval, Δ=0.18 on LoCoMo no-adv) sit at $\sim 46\%$ and $\sim 100\%$ one-sided power respectively. The headline metric-inversion claim does not hinge on per-cell McNemar significance: the 19-cell sign-agreement synthesis (§5.1, App. A.3) shows 17/21 (81%) prompt-only pairs disagree on direction, which is itself significant under a binomial sign test on the 19 non-tied pairs ($p < 10^{-3}$).

A.16 Cross-backbone replication of the FCS gain (cf. §5.4)

To check whether the $+0.082$ FCS token-F1 gain (cell J vs cell E) is specific to Qwen2.5-14B or replicates across backbones, we re-run the flat $\times$ tight-per-cat (cell E) and flat $\times$ tight-fcs (cell J) configurations on four additional backbones spanning three model families. All cells use identical retrieval (BM25, top-5), router, 96-token cap, greedy decoding, and per-question haystack on the same LongMemEval oracle n=500 split.

The pattern is consistent across all 5 backbones $\times$ 2 judges: the token-F1 FCS gain ($+0.028$ to $+0.082$) collapses under both LLM judges (every Δ_{judge} within ± 0.035, all n.s.). This rules out two alternative hypotheses: that the $+0.082$ gain on Qwen-14B is a backbone-specific artifact, and that the LLM-judge collapse is specific to the primary backbone. The length-compression explanation also holds across backbones. Where the model compresses output most aggressively (Llama-3.1-8B 48$\rightarrow$14, Qwen-14B 52$\rightarrow$14, Qwen-7B 28$\rightarrow$14), the token-F1 gain is largest; Mistral-7B's smaller compression (66$\rightarrow$38) corresponds to its smaller token-F1 gain.

A.17 FCS rendering rules and data-usage details (cf. §3)

The FCS prompt-rendering protocol is the methodological core of the paper. The goal is to specify a controlled prompt that constrains only the surface form of the answer, not the reasoning required to produce it. The intuition is that the surface form is observable from gold answers alone and so can be derived programmatically without leaking task hints.

For each (benchmark, category), we compute a `SurfaceForm` profile from gold answers: (i) length statistics (median, 25th, 75th, max gold word counts); (ii) dominant answer form classified into {numeric_span, date, label, list, abstention, free_text, short_span, unknown}; (iii) abstention strings (e.g. "no information available").

The controlled prompt is rendered from the SurfaceForm by selecting (i) one length cap (`p75_words` rounded up, with a floor of 8); (ii) one form-specific template (e.g. *date* $\rightarrow$ "Output exactly one date or date range, no extra words"; *numeric_span* $\rightarrow$ "Output exactly one number, optionally followed by a unit"); and (iii) the abstention string when applicable. The FCS prompt does not contain task-framing language. It does not tell the model to synthesize, to compute, to check timestamps, or to consider multiple sessions. It tells the model only what form the output should take.

Backbone	Token-F1 (n=500) E	J	Δ_{F1}	gpt-4o-mini judge (n=200) E acc	J acc	Δ_{judge}	Claude-Sonnet-4.5 (n=200) E acc	J acc	Δ_{judge}
Qwen2.5-3B	0.349	0.383	+0.034	0.385	0.395	+0.010	0.370	0.380	+0.010
Qwen2.5-7B	0.387	0.419	+0.032	0.475	0.460	−0.015	0.480	0.470	−0.010
Mistral-7B-v0.3	0.334	0.361	+0.028	0.455	0.465	+0.010	0.460	0.495	+0.035
Llama-3.1-8B	0.340	0.399	+0.060	0.465	0.450	−0.015	0.470	0.475	+0.005
Qwen2.5-14B (primary)	0.365	0.447	+0.082	0.530	0.535	+0.005	0.530	0.530	0.000

Table 19: Cross-backbone replication of cells E (flat × tight-per-cat) and J (flat × tight-fcs). Left block: token-F1 at n=500 on the full LongMemEval oracle split. Middle / right blocks: LLM-judge accuracy on the paired n=200 subset, gpt-4o-mini and Claude-Sonnet-4.5 respectively (same paired indices as the primary Qwen-14B run). Token-F1: every backbone produces a positive FCS gain (+0.028 to +0.082). LLM-judge: every backbone × judge cell collapses the FCS gain to a value within ±0.035 of zero, all McNemar n.s. (largest χ^2=0.88 on Mistral-7B/Claude). The token-F1-vs-LLM-judge inversion replicates across all 5 backbones × 2 judges.

Data-usage disclosure. Because LongMemEval and LoCoMo do not provide a separate calibration split for surface-form profiles, our FCS implementation derives the per-category length percentiles, dominant answer-form classification, and abstention strings from the same benchmark gold answers that are then used at evaluation. We therefore treat benchmark-level gold-answer statistics as a diagnostic audit instrument only, not as a fair leaderboard-submission method. The +0.082 token-F1 gain we report should be read as a quantification of how much score movement can be induced by surface-form alignment alone when memory evidence and generation conditions are held fixed—not as a recipe a memory-system author should adopt to improve their own leaderboard score.

Multi-seed held-out FCS calibration ablation. To quantify how much of the +0.082 FCS gain is direct test-set leakage in the calibration-on-test sense, we run a 5-seed held-out ablation. For each seed in $\{20260506, 20260507, 20260508, 20260509, 20260510\}$, split LongMemEval oracle 50/50 (n=250 calibration, n=250 test), recompute the per-category SurfaceForm profile on calibration-only, programmatically regenerate FCS prompts from the cal-half profile (varying p75 word cap, dominant answer form, and abstention strings—all three render-relevant SurfaceForm fields, not just length percentiles), and re-run cell J on the test half twice: once with the deployed full-data FCS prompts, once with the cal-half-rendered FCS prompts.

The category-level surface-form regularities FCS captures are stable across random splits of these benchmarks. A length-cap-style FCS constraint derived from any half of LongMemEval is non-binding on cell J's already-aggressively-compressed outputs, so the cal-half-rendered prompts produce identical aggregate F1 to the full-data-rendered prompts. We verify the prompts are not byte-identical (`results_rendered_prompts_diff.json`). In seed 20260510, all four categories differ from the deployed prompts: knowledge-update cap 3→4, single-session cap 3→19, multi-session "exactly one number" becomes "at most 2 words," temporal drops the "`D Month YYYY`" form template. In seeds 20260508 and 20260509, 3 of 4 categories differ. F1-identity therefore arises from output-equivalence (the LLM produces the same answers under both prompts on these test items) rather than from template-equivalence. The +0.082 FCS gain on the full n=500 is consequently not a result of memorizing test-gold strings; it is a result of per-category surface-form regularity that any held-out split would surface the same way and that the LLM would respond to identically. We frame this as a near-template-equivalence upper bound on leakage

Seed	Full F1	Cal-half F1	Δ
20260506	0.4514	0.4514	0.0000
20260507	0.4797	0.4797	0.0000
20260508	0.4334	0.4334	0.0000
20260509	0.4695	0.4695	0.0000
20260510	0.4564	0.4564	0.0000
mean	—	—	0.0000 ± 0.0000

Table 20: Multi-seed FCS-leakage ablation. F1-identical predictions across 5 random 50/50 splits despite byte-different rendered prompts in 3–4 of 4 categories per seed (full diff: `results_rendered_prompts_diff.json`). See surrounding prose for the output-equivalence-under-non-binding-caps explanation and the near-template-equivalence-upper-bound framing.

rather than a tight estimate. We note one caveat: under a binding-cap regime (a more verbose backbone, or a tighter cap that forced the model away from its already-compressed default), this ablation could in principle reveal residual leakage that the current non-binding regime hides.

A.18 Tight-per-cat-stripped ablation (cf. §5.4)

To isolate which component of the FCS prompt produces the $+0.082$ token-F1 gain, we run a stripped variant: take the existing tight-per-cat templates and remove only the task-framing verbs ("synthesize", "compute", "consider", "identify the most recent", "use the session dates"). Surface-form constraints (length cap, answer-form template, abstention string) are kept. We run flat $\times$ tight-stripped on LongMemEval oracle, n=500, Qwen2.5-14B, identical pipeline.

The $+0.082$ token-F1 gain decomposes roughly 50/50. Stripping task-framing verbs alone buys $+0.042$ token-F1 (E→Q) and a $\sim$44% length reduction; the stricter FCS answer-form templates buy a further $+0.040$ token-F1 (Q→J) and an additional $\sim$52% length reduction.

Both halves are token-F1 artifacts. Under gpt-4o-mini judge on the same paired n=200 subset: stripped vs tight-per-cat gives $\Delta_{\text{judge}} = +0.010$ (paired McNemar χ^2=0.07, p=0.79, n.s.); FCS vs tight-per-cat gives $\Delta_{\text{judge}} = +0.005$ (n.s., §5.5a). The token-F1 lift in both directions is mechanistically output-length compression, not a reasoning improvement.

This pins down the linguistic mechanism behind Finding 4. The conventional memory-system practice of writing prompts that "tell the model how to think" (the verbs in tight-per-cat) costs about 0.04 F1 vs simply removing those verbs, with no LLM-judge correlate. Half of the FCS gain is the token-F1 cost of task-framing language; the other half is the token-F1 cost of leaving the answer-form template loose. Neither half corresponds to a semantic-correctness gain.

A.19 Reproducibility checklist for memory-architecture claims (cf. §6)

A one-page checklist for authors, reviewers, and leaderboard maintainers proposing or evaluating memory-architecture claims on LongMemEval, LoCoMo, BEAM, or similar long-horizon dialogue QA benchmarks. Sections A–E below are the proposed minimum reporting standard; Section F lists recommended robustness probes. Section pointers cite the body where the prescription is empirically motivated.

A. Baseline matching (mandatory).

- Surface-form-matched flat-RAG baseline on the same backbone. Length cap, output template, and abstention string match the proposed system's prompts (cf. §5.1, §5.3).
- Identical prompt design across architecture variants and the flat baseline; no silent per-architecture prompt re-tuning (cf. Finding 4, §5.4).
- Single-call vs multi-call regime stated. If multi-call, the flat baseline is also multi-call.

B. Statistical reporting (mandatory).

- Per-category F1 with paired-bootstrap 90% CIs, not aggregate F1 alone (cf. faithful Mem0, §5.6: aggregate Δ=+0.021 n.s. hides Δ=+0.25/−0.08 per-label).
- Pre-registered minimum-detectable-effect (MDE) with sample-size justification (cf. App. A.15).
- Two-sided p reported alongside one-sided for any α=0.05 claim.
- Multiple-comparisons correction stated when running $\geq$ 3 tests (Bonferroni or comparable; cf. A-MEM and HippoRAG in §5.6: both cross unadjusted α=0.05 in opposite directions, neither survives Bonferroni-7).

C. Metric reporting (mandatory).

- Output-length statistics (mean / median / p75 chars and words) per cell (cf. §5.4: the $+0.082$ FCS gain is fully explained by 3.6$\times$ length compression).
- Cross-family LLM-judge accuracy with paired McNemar on $\geq$ 2 judges from different model families (cf. §5.1, §5.5).
- Token-F1 reported only with surface-form context (length, format match, abstention-string match). Do not report token-F1 alone for prompt-engineering claims.

Cell	Token-F1 (n=500)	Avg chars	LLM-judge acc (n=200)
E (tight-per-cat)	0.365	52	0.530
Q (tight-stripped)	0.407	29	0.540
J (tight-fcs)	0.447	14	0.535

Table 21: Decomposing the $+0.082$ FCS token-F1 gain. Stripping task-framing verbs alone (E→Q) buys $+0.042$ F1 with $\sim 44\%$ length reduction; stricter FCS templates (Q→J) buy a further $+0.040$ F1 with another $\sim 52\%$ reduction. Both halves are token-F1 artifacts: under LLM-judge, all three cells tie within noise.

D. Architecture/protocol fairness (mandatory).

- Native-prompt sanity cell alongside the matched cell (cf. Tab. 17): re-run the proposed system under its own published answer prompt at native token budget on the same paired indices.
- Pipeline details released (model versions, seeds, decoding params, retrieval parameters, prompt files, evaluation code).
- Per-question prediction JSONs released for paired re-analysis.

E. Scope statements (mandatory).

- Which regimes are out of scope (single-call vs multi-call; oracle vs non-oracle; with/without forgetting; persistent vs per-question memory; cf. §8).
- Which benchmarks are out of scope (e.g., BEAM-style verbose-gold benchmarks where token-F1 is uninformative).
- Pre-registered falsification condition: an MDE $\times$ condition $\times$ judge requirement, not a vague "we expect."

F. Recommended robustness probes (optional).

- FCS-style diagnostic prompt as a robustness probe (with the disclosure that FCS is calibrated on gold and is not a leaderboard method).
- Cross-backbone replication on at least one non-primary backbone family (cf. App. A.16).
- Adversarial-category and abstention-string sensitivity check (LoCoMo-style 0.20 F1 swing test, §5.2).
- Held-out calibration ablation if any prompt component is derived from data (cf. App. A.17).

How to use. Authors should mark each box in Sections A–E before submission; unchecked boxes belong in the manuscript's limitations section. Reviewers can use unchecked boxes as request-revision questions. Leaderboard maintainers can adopt Sections A–C as the minimum reporting requirement and verify Section D via code-release inspection.