WorldConsistMem: Evaluating Representation-Sensitive State-Transition Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Long-term memory systems for evolving worlds must reconstruct state changes, not merely retrieve related records. WorldConsistMem separates evidence availability from state-transition reconstruction on gold evolving histories: a transition query asks for the previous state, the transition event, and the resulting state.

Under a fixed GPT-4.1 snapshot and byte-matched BM25 evidence packs with gold-support recall = 1, a generic response interface reconstructs only 13/150 transitions (8.7%). A preregistered bundled structured transition prompt/representation intervention—specifying the required prev/event/next representation and a canonical subtype-specific output schema—raises accuracy to 109/150 on the same evidence and snapshot. On the original explicitly/operationally supported slice (S1+S2; 93 packs), accuracy rises from 13/93 to 52/93 (paired Δ+41.9 percentage points; bundle-bootstrap 95% CI $[31.2, 52.7]$; McNemar $p \approx 3.8 \times 10^{-11}$). A prospective post-result factorization shows the gain is not reducible to formatting alone: schema-only reaches 90/150 and representation-only 39/150, both increments remain positive after the other, and their interaction is additive-compatible. A forensic audit finds leakage PASS, 0 format-only (M5) wins, and 0 scorer-weakness (M8) wins among 97 genuine non-format improvements. Document transitions and original-label S2 cases remain 0/34 in every arm; CEO transitions reach 23/30 under the bundled interface. We conclude that measured transition reconstruction is highly representation-sensitive under this benchmark, prompt bundle, and snapshot—with schema the larger measured lever and explicit prev/event/next independently contributing—not that either factor alone, intrinsic model incapacity, or complete resolution explains the residual floors.

1 Introduction

Memory systems for evolving worlds must retrieve relevant events and reconstruct how state changed: what held before, which event applied, and what holds after. Standard long-term memory evaluations emphasize per-query Acc and retrieval fitness (Wu et al., 2025; 2026; Hu et al., 2025; Tao et al., 2026; Xie et al., 2026; Wang et al., 2026; Hu et al., 2026; Liu et al., 2026). That design can leave open a different gap: when supporting evidence is already available, how much does measured transition Acc depend on how the update task and answer interface are represented to the reader?

Central thesis (scoped). State-transition reconstruction performance is highly sensitive to how retrieved evidence and the required state update are represented to the reader. Under fixed BM25 evidence packs and the same GPT-4.1 snapshot, a preregistered bundled structured transition intervention raises transition Acc from 13/150 to 109/150. A prospective post-result factorization (F0–F3) shows the bundled gain is not reducible to output formatting alone: canonical schema is the larger contributor (90/150), while explicit prev/event/next representation independently improves Acc (39/150); both increments remain positive when added after the other, and their interaction is compatible with additivity.

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.

The intervention does not eliminate the task: document transitions and original-label S2 cases remain 0/34 in every arm; CEO transitions remain imperfect at 23/30 under the bundle. We do not claim synergy beyond additivity, causal isolation beyond these arms, intrinsic GPT-4.1 incapacity, retrieval irrelevance, or generalization beyond this synthetic benchmark and snapshot.

Approach. WorldConsistMem generates gold evolving histories, emits dependent query bundles, and scores readers under controlled evidence conditions. Gold-support recall = 1 establishes evidence availability; it does not imply ideal presentation. The primary confirmatory experiment holds evidence packs and the reader snapshot fixed and compares the original generic response contract (F0) to a bundled structured transition prompt/representation intervention (F3) specifying prev/event/next and a canonical subtype-specific output schema. A prospective factorized ablation then isolates schema-only (F1) and representation-only (F2) on the same 150 queries after observing F3. We score with transition scorer v2 (explicit field-alias normalization; Appendix N). Retrieval comparisons, oracle/joint cells, BCR/null analyses, and complete reconstruction remain secondary (Appendix).

Primary evidence. Original B4 Acc = 13/150; structured intervention Acc = 109/150 (independent double-check agreement; leakage PASS). On the original S1+S2 confirmatory slice: $13/93 \rightarrow 52/93$ (Δ+41.9 pp; CI $[31.2, 52.7]$; McNemar $p \approx 3.8 \times 10^{-11}$; Outcome A). Factorized Acc: F1 90/150, F2 39/150; all-150 schema Δ+51.3 pp and representation Δ+17.3 pp; interaction -4.7 pp with CI including 0 (additive-compatible). Subtype heterogeneity: conflict/employment are schema-dominated; CEO benefits more from temporal representation; document stays at floor. A post-hoc forensic re-audit of sufficiency labels is reported separately and does not replace the original S1–S5 primary slice (Appendix P).

Contributions.

1. WorldConsistMem: a controlled benchmark for reconstruction over evolving histories with explicit transition and provenance families.
2. An evidence-conditioned evaluation that separates missing support from failure to reconstruct a resulting state after a transition event.
3. A confirmatory finding that, with identical evidence and snapshot, a preregistered bundled structured transition intervention raises Acc from 13/150 to 109/150, while document (0/34) and residual CEO errors remain.
4. A prospective post-result factorization showing schema-dominant, representation-contributing, approximately additive increments—without retroactively making F3 factorially preregistered.
5. Transparent scoring (scorer v2), runner-parser correction, and forensic leakage/mechanism audits that bound overclaiming.

2 Related Work

We do not claim to invent the general observation that retrieval does not guarantee correct answering. Retrieval–utilization / faithfulness work already separates evidence presence from correct use. Prompting and output-schema studies show that task formatting can change measured accuracy; we treat F3 as a bundled confirmatory intervention and report a prospective post-result factorization (schema vs. prev/event/next) without claiming a new prompting method. Multi-hop and temporal QA stress chaining and time, typically without an evolving multi-entity gold history plus an explicit prev/event/next transition object under controlled packs. Temporal knowledge graphs and event-sourced stores implement validity and state updates as systems; we evaluate LM readers on a fixed benchmark and do not claim a new TKG architecture. Long-term conversational memory suites emphasize per-query Acc under long contexts (Wu et al., 2025; 2026; Hu et al., 2025; 2026; Wang et al., 2026). Conflict / temporal memory shows Acc can diverge from retrieval fitness (Tao et al.,

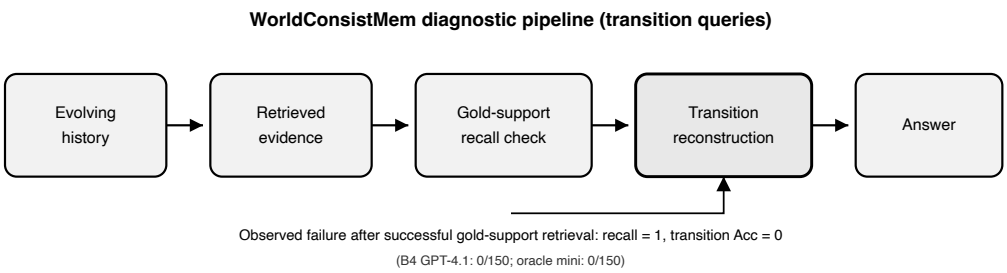


Figure 1: Diagnostic pipeline for transition queries. Primary experiment: fixed B4 evidence with original Acc = 13/150 vs. structured intervention Acc = 109/150 (scorer v2).

2026). Cross-query logical consistency separates Acc from coherence (Salla et al., 2026; Chaudhry, 2026); that separation is not claimed as generally new here.

Residual novelty (narrow). WorldConsistMem contributes (i) evolving gold histories with an explicit previous-state / transition-event / resulting-state object, (ii) evidence-conditioned analysis separating missing support from reconstruction failures, (iii) a confirmatory bundled finding (13/150 → 109/150) with a prospective schema-dominant factorization, while document/S2 floors remain, and (iv) matched retrieval/packaging cells retained as secondary context. Appendix F (Table 6) expands closest benchmarks.

3 Task: State Transitions over Evolving Histories

Let $W_0, \ldots, W_T$ be structured world states with events $E_{1:T}$ and $W_t = \text{Apply}(W_{t-1}, E_t)$. The gold history is retained for scoring only. The system receives observations $O_{1:T}$ and holds memory M_T. A dependent bundle $Q_b = \{q_1, \ldots, q_{n_b}\}$ asks interrelated questions (current/historical/transition/provenance/conflict/…) about one focal lifecycle. A reader maps evidence packs drawn from M_T to predictions $\hat{Y}_b$.

Transition query (primary object). A transition query asks the reader to reconstruct the state update

$$(\text{previous state},\ \text{transition event}) \longrightarrow \text{resulting state}.$$

Operationally, the gold answer is a structured tuple (e.g., role replacement: from, to, time, entity keys). Exact-match Acc on that tuple is transition Acc. We distinguish:

- Evidence retrieval: selecting candidate records from M_T (or oracle supports).
- Evidence packaging: serialization/layout of those records in the prompt.
- State-transition reconstruction: emitting the correct prev/event/next tuple.
- Provenance identification: naming supporting evidence identifiers when asked.
- Abstention / format failure: null/malformed outputs under the frozen parser.
- Complete bundle reconstruction: secondary joint metrics (BCR/CRR; appendix).

Evidence availability (evaluation only). Gold-support recall uses evaluation-time supporting_event_ids to check whether gold-supporting observations appear in the evidence pack. Recall = 1 means those supports are available in the pack. It does not imply optimal presentation, that the reader attended to the right span, or that retrieval can never matter for other queries/systems. supporting_event_ids are not a production routing signal; oracle packs that inject them are diagnostic only.

Pipeline. Figure 1 summarizes the diagnostic flow. Evidence availability (gold-support recall = 1) is necessary but not identical to a usable transition representation for the reader.

Table 1: Dataset statistics (scaled freeze, seed 100).

Quantity	Value
Worlds (S/M/L)	102 (34/34/34)
Bundles / queries	1,088 / 7,140
H4 LLM subset	18 / 192 / 1,260

Secondary joint metrics (not the mechanism). For completeness we also score item Acc, provenance Acc, and bundle-level Φ/BCR (Eqs. in Appendix A). BCR = 0 under LLM cells is treated as a consequence-level metric after family floors—not the diagnosed mechanism of this paper.

4 WorldConsistMem Benchmark

WorldConsistMem instantiates Section 3: typed entities, lifecycle events with validity/provenance, observation streams without gold W_t, and dependent bundles with history-derived constraints (Table 1).

Why synthetic? Causal transition scoring and gold-support recall need a known history. Ecological validity is not claimed. A confirmatory naturalistic set (Appendix J) tests bounded NL surface transfer only.

5 Evaluation Metrics

Primary. Transition Acc (structured exact match on prev/event/next tuples after scorer v2 alias normalization; Appendix N), including the paired original-vs-structured interface contrast on fixed B4 packs; provenance Acc; overall item Acc; gold-support recall (evaluation only); abstention/malformed rates under the corrected parser.

Scorer versions. Scorer v1 required gold key names on predictions (from/to/t). Scorer v2 permits only three explicit aliases (old_person_id→from, new_person_id→to, time→t), renames keys only, and retains strict value comparison. Conflicting canonical/alias values are rejected. v1 and v2 scores are both retained in the release; no raw predictions were altered.

Evidence-conditioned stages. With evaluation recall r: stage A if $r<1$ (partial r marked ambiguous); stage B if $r=1$ and a non-abstaining incorrect answer under the active scorer; stage C if $r=1$ with abstention/malformed output. Pack-level sufficiency labels S1–S5 further separate explicitly supported cases from underspecified packs (Appendix O).

Secondary joint metrics. Bundle constraint consistency BCR/Φ, CSR, complete reconstruction (CRR/all_correct), and $\text{Gap}_{\text{query}}$ are defined in Appendix A. They support robustness analyses and are not the diagnosed transition mechanism. Metrics are computed from frozen predictions by the released evaluators.

6 Diagnostic Systems and Setup

Systems are diagnostic references, not a SOTA leaderboard: B1 compact flat, B2 latest-only, B3 recency-k, B4 BM25-k, H0 hierarchical (cap150, ret8).

Readers and interventions. Confirmatory reader: OpenAI gpt-4.1-2025-04-14 (mini cells secondary) under temperature 0 on the 989/150 expansion. Evidence packs for the primary pair: matched B4/BM25 with gold-support recall = 1 on all 150 transitions. We define four arms on the same 150 query IDs and byte-identical B4 packs (snapshot fixed; scorer v2):

Table 2: Primary paired transition Acc under scorer v2 (identical B4 evidence; GPT-4.1 snapshot). S1/S2 use the original sufficiency labels. S1+S2 paired Δ+41.9 pp with bundle-bootstrap 95% CI [31.2, 52.7] and McNemar $p \approx 3.8 \times 10^{-11}$.

Slice	Original	Structured	Delta
All transitions	13/150	109/150	+64.0 pp
Original S1+S2	13/93	52/93	+41.9 pp
S1	13/59	52/59	+66.1 pp
S2	0/34	0/34	0
CEO	13/30	23/30	+33.3 pp
Document	0/34	0/34	0
Conflict	0/29	29/29	+100 pp
Employment	0/28	28/28	+100 pp
Ownership	0/29	29/29	+100 pp

- F0 (generic): original generic response interface; reuse frozen B4 predictions (13/150).
- F1 (schema-only): canonical subtype state schema with the original generic reasoning wrapper; no prev/event/next (90/150). This constrains semantic fields, not mere JSON syntax.
- F2 (representation-only): explicit prev/event/next representation with the original generic output wrapper; no subtype canonical schema (39/150).
- F3 (bundled): full structured prev/event/next+canonical schema intervention; reuse frozen structured outputs (109/150).

Primary confirmatory contrast: F0 vs. F3 (preregistered bundled intervention; aggregate F3 prompt SHA in artefacts). Prospective factorization: F1/F2 were designed and executed after observing F3, to decompose the bundled gain; they are not a factorial preregistration of F3. Secondary evidence conditions (generic interface): standard H0 packs; oracle gold-supporting observations; joint bundle reconstruction. Symbolic exact-match and Qwen H4 remain Option A / secondary (Appendix C). Fairness caveats: Appendix G.

7 Results

Primary analyses use the confirmatory expansion (989 queries / 150 transition queries; not Option A). Unless noted, transition Acc uses scorer v2 (Appendix N); no model was re-run for scoring or for the structured-intervention analysis. The structured condition reuses frozen raw outputs under offline bare-JSON parsing (Appendix Q).

7.1 Primary: structured transition representation intervention

Setup. Matched GPT-4.1 BM25 (B4) packs provide gold-support recall = 1 for all 150 transition queries. The original condition uses the generic response interface on those packs (paired baseline; Acc = 13/150 under scorer v2). The structured condition is a preregistered bundled structured transition prompt/representation intervention specifying the required prev/event/next state representation and a canonical subtype-specific output schema, with byte-matched evidence and query text relative to B4. Subtype schema selection uses only the public query-id slot, not gold answers. A forensic leakage audit reports PASS (Appendix R).

Headline paired result. Structured Acc = 109/150 vs. original 13/150 (Δ+64.0 pp). On the original S1+S2 confirmatory slice: 13/93 $\rightarrow$ 52/93 (Δ+41.9 pp; bundle-bootstrap 95% CI [31.2, 52.7] pp; McNemar exact $p \approx 3.8 \times 10^{-11}$). Independent strict semantic checker agreement: 109/109 structured corrects; 0 disagreements. Mechanism taxonomy (wins only): 97 non-format improvements; 0 true format-only (M5) wins; 0 leakage/scorer-weakness (M8) wins. Table 2 and Figure 2 summarize slices and subtypes.

Why conflict/employment/ownership improve sharply. Required fields are typically explicit in the query and/or evidence payload; the structured schema elicits the full tuple

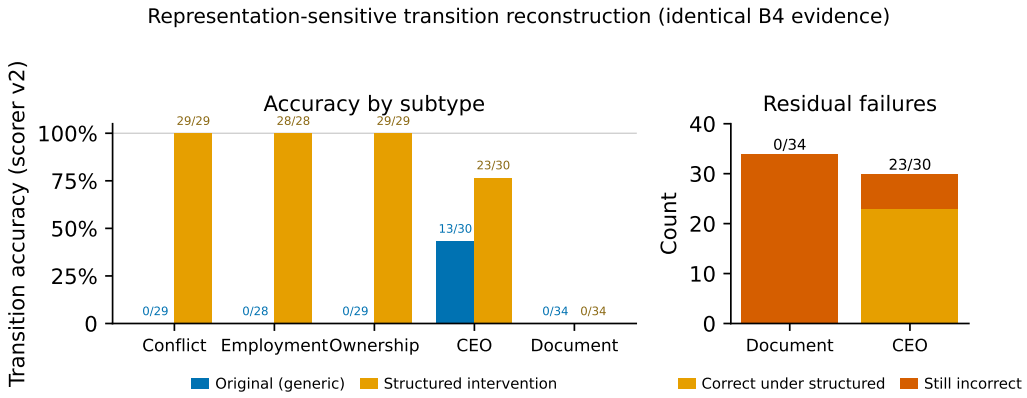


Figure 2: Original vs. structured transition Acc by subtype (denominators on bars; full y-axis). Right: residual failures under the structured intervention (document floor; imperfect CEO).

rather than a partial scalar. A forensic re-audit shows that many employment/ownership packs were previously labelled S3/S5 due to over-strict event-type matching (e.g., person_leaves_org vs. leave); under written global rules they reclassify as S1/S2 (Appendix P). Primary confirmatory S1+S2 Acc uses the original labels only; the re-audit is post-hoc.

Residual failures. Document transitions remain 0/34: slim packs often omit version identifiers required by the gold schema, so schema elicitation cannot invent missing values. CEO improves to 23/30 but retains distractor and previous-holder selection difficulty when multiple role replacements appear in the pack. S2 remains 0/34 under original labels: operational entailment without an explicit resulting-state statement is not rescued by the intervention.

Causal-claim discipline for the bundled arm. F3 bundles representation, task decomposition (prev/event/next), and canonical response structure. We therefore treat F3 as the confirmatory bundled intervention—not chain-of-thought alone—and use the prospective F0–F3 factorization below to measure component contributions without claiming isolation beyond these arms. Gains are not formatting-only under the forensic M5 definition (0 recoverable key-remap wins).

7.2 Prospective factorized ablation (post-result)

After observing F3, we executed a preregistered-for-execution factorization on the same 150 IDs (GPT-4.1 snapshot; T=0; max tokens 256; joint spend ceiling $1.50; F1/F2 150+150 new calls; F0/F3 reused). F1/F2 were not designed before F3 and do not retroactively make the bundled intervention factorially preregistered. Table 3 reports Acc and paired contrasts (bundle-cluster bootstrap 10^4, seed 20260917; McNemar exact).

Mechanism reading. Canonical state schema (not mere JSON syntax) is the larger aggregate lever; explicit prev/event/next independently contributes; both increments remain positive after introducing the other; interaction is compatible with approximate additivity (no synergy claim). Subtype counts (F0/F1/F2/F3): CEO 13/10/26/23 of 30; conflict 0/29/0/29; document 0/0/0/0 of 34; employment 0/28/0/28; ownership 0/23/13/29 of 29 (Table 16). Aggregate schema advantage is not universal: CEO benefits more from temporal representation; conflict/employment are schema-dominated; ownership benefits from both; document and original-label S2 remain 0/34 in every arm. Deterministic qualitative extracts (conflict schema recovery; CEO representation recovery; document/S2 floor): Appendix T.

Representative paired example (deterministic; F0 vs F3). Selection rule (documented; not cherry-picked): among CEO rows that are original-incorrect, structured-correct, original-S1, and mechanism M2 (previous-state wrong transition), take the lexicographically first

Table 3: Prospective F0–F3 factorization under scorer v2 (identical B4 evidence; GPT-4.1). Primary confirmatory result remains F3. Contrasts use bundle-bootstrap 95% CIs; interaction CI including 0 ⇒ additive-compatible.

Arm / contrast	All 150	Original S1+S2 (93)
F0 generic	13/150 (8.7%)	13/93 (14.0%)
F1 schema-only	90/150 (60.0%)	39/93 (41.9%)
F2 prev/event/next-only	39/150 (26.0%)	26/93 (28.0%)
F3 bundled	109/150 (72.7%)	52/93 (55.9%)
F1−F0 (schema)	+51.33 pp [42.67, 60.00]	+27.96 pp [17.20, 38.71]
F2−F0 (representation)	+17.33 pp [11.33, 24.00]	+13.98 pp [6.45, 21.51]
F3−F2 (schema after repr.)	+46.67 pp [38.00, 56.00]	+27.96 pp [16.13, 39.78]
F3−F1 (repr. after schema)	+12.67 pp [7.33, 18.00]	+13.98 pp [7.53, 21.51]
Interaction $(F3-F2)-(F1-F0)$	−4.67 pp [−10.67, +1.33]	0.00 pp

Table 4: Secondary confirmatory cells under the original generic interface (scorer v2). The structured intervention (Table 2) is the primary experiment.

Condition	Reader	Overall Acc	Support recall	Trans. Acc	Prov. Acc	Abstain
Standard H0	GPT-4.1 mini	0.397	—†	0/150	0/150	0.357
Standard H0	GPT-4.1	0.392	—†	0/150	0/150	0.372
Matched B4	GPT-4.1	0.423	0.707‡	13/150	0.687	0.293
Matched H0	GPT-4.1	0.392	0.488†	0/150	0/150	0.372
Oracle retrieval	GPT-4.1 mini	0.499	1.0	0/150	0.733	0.264
Joint reconstruction	GPT-4.1 mini	0.498	n/a	0/150	0.040	0.000

†H0 event-id recall proxy can understate structured transition records. ‡B4 mean post-truncation recall; transition subset recall = 1 for 150/150.

query ID—yielding sl0168_000:b:ceo:0:transition. Identical query and evidence ordering; gold is CEO person:4 → person:1 at t=32. Original output reports an earlier replacement (person:0 → person:1 at t=16). Structured output emits the gold tuple. The structured representation repaired previous-holder/time selection among distractors, not missing evidence. Full extract: Appendix B.

7.3 Secondary: retrieval packaging and other cells

Table 4 retains evidence-conditioned comparisons under the original generic interface (scorer v2). Transition Acc is 0/150 for mini standard H0, GPT-4.1 standard H0, mini oracle, mini joint, and budget $k \in \{1, 8\}$; matched B4 rises to 13/150. Oracle/joint raise overall Acc (≈ 0.50) and provenance Acc while transition Acc stays 0 under the generic interface—consistent with representation sensitivity rather than "retrieval irrelevance." Matched H0 vs. B4: no reliable overall H0 Acc advantage. Provenance Acc rises under B4 (≈ 0.69) and oracle (≈ 0.73) relative to standard H0 (0), supporting provenance as a distinct capability.

7.4 Secondary robustness (summary)

Bundle-complete BCR/CRR remain 0/150 under these LLM cells; post-hoc family-conditioned nulls show that zero completion is largely consistent with compounding given transition/provenance floors (Appendix L). Evaluable cross-query contradiction is not elevated vs. permutation nulls (Appendix M). Symbolic Option A Acc≠BCR is retained as a secondary evaluation-object demonstration (Appendix C), not as the central claim.

8 Failure Taxonomy and Qualitative Example

Original B4 under the generic interface (scorer v2).

- A — insufficient evidence (original labels): packs labelled S3–S5 (57/150).
- B — reconstruction error: S1/S2 packs incorrect under v2 (80/150).
- C — scorer alias false negatives under v1: 13/150 CEO rows corrected by v2 (not reconstruction failures under v2).
- D — abstention: B4 transition abstentions = 0.

Gold-support recall = 1 for all 150 does not imply every pack is evidence-sufficient under the original schema labels (S1+S2 = 93).

After the structured intervention and factorization. Bundled F3 Acc rises to 109/150; residual errors concentrate in document (0/34) and remaining CEO distractor cases (7/30 still wrong). Original-label S2 stays 0/34 under F0, F1, F2, and F3 alike—the factorization does not resolve operational entailment without an explicit resulting-state statement. Document transitions likewise remain 0/34 in every arm: slim packs often omit version identifiers required by the gold schema. Mechanism taxonomy among original→structured wins: predominantly incomplete scalar/non-object originals and previous-state CEO errors—not M5 format-only remaps (Appendix R).

Paired qualitative example. Appendix B expands the deterministic CEO example: identical evidence; original answers an earlier replacement; structured emits the gold prev/event/next tuple.

9 Discussion and Limitations

9.1 What the structured-intervention result does and does not mean

With identical evidence and the same GPT-4.1 snapshot, changing the prompt/representation interface raises transition Acc from 13/150 to 109/150. Preferred reading: measured state-transition reconstruction is representation-sensitive on this benchmark, and a prospective factorization shows schema as the larger lever with independent representation contribution under approximately additive increments. Forbidden readings: “decomposition alone caused the gain”; “schema formatting alone caused the gain”; “synergy beyond additivity”; “GPT-4.1 cannot reason about transitions”; “retrieval is irrelevant”; “109/150 is complete world reconstruction”; “F1/F2 were preregistered before F3”; “the revised S1–S5 taxonomy was preregistered.”

F1/F2 were designed and authorized after observing the bundled F3 result; they prospectively decompose that intervention and do not retroactively make F3 a factorially preregistered experiment. Residual document (0/34) and original-label S2 (0/34) floors across all arms, and imperfect CEO Acc under F3 (23/30), show the interface does not eliminate hard cases. Near-ceiling conflict/employment Acc under schema-bearing arms reflects packs where required fields are already explicit—not a general solver claim. Subtype heterogeneity warns against treating the aggregate schema advantage as universal.

9.2 Scorer correction (transparent)

The original v1 scorer false-negatived 13 CEO transitions that used event-payload field names. v2 accepts only those three declared aliases; values remain strict; no model was re-run for scoring. We retain v1 numbers for lineage and report v2 throughout the primary tables.

9.3 Runner wrapper defect (transparent)

The structured prompt requested bare canonical JSON; the execution runner incorrectly required an answer wrapper and labelled 150/150 outputs malformed. Offline parsing of frozen raw outputs recovers all 150 bare objects; no API rerun. The corrected parser accepts bare and wrapped forms (Appendix Q).

9.4 Provenance recovery does not contradict the thesis

Higher provenance Acc under B4/oracle under the generic interface is a distinct capability contrast—not evidence against representation sensitivity for transitions.

9.5 Threats to validity / scope

- Synthetic ontology/histories; no ecological validation.
- Single GPT-4.1 snapshot for structured/factorized arms; Gemini F0–F3 replication is prepared but not executed.
- Small independent human validation (n=30) is packetized but incomplete; not a human ceiling.
- F1/F2 are prospective post-result decompositions (not pre-F3 factorial preregistration); F1 is canonical state schema, not mere JSON syntax.
- Document and original-label S2 floors (0/34 every arm) remain unresolved.
- Scorer-v2 alias correction and bare-JSON offline re-parse are disclosed; raw outputs unchanged.
- No new architecture claim; no universality beyond this benchmark/packs/snapshot/interfaces.

9.6 Secondary metrics

BCR/CRR, consistency nulls, constraint coverage, and hybrid feasibility notes appear in the appendix. BCR= 0 is not used as central evidence of a new inconsistency phenomenon.

10 Conclusion

On synthetic WorldConsistMem, state-transition reconstruction is highly sensitive to how retrieved evidence and the required update are represented to the reader. With byte-matched BM25 packs and a fixed GPT-4.1 snapshot, a preregistered bundled structured transition prompt/representation intervention raises Acc from 13/150 to 109/150, and on the original S1+S2 slice from 13/93 to 52/93 (Δ+41.9 pp; CI $[31.2, 52.7]$). A prospective post-result factorization shows that the bundled intervention is not reducible to output formatting alone: canonical schema specification is the larger contributor, while explicit prev/event/next representation independently improves transition reconstruction; both increments remain positive when added after the other, and their interaction is compatible with additivity. Document transitions and original-label S2 cases remain unresolved (0/34 in every arm); CEO transitions reach 23/30 under the bundle. Forensic checks find leakage PASS, zero format-only wins, and transparent scorer/parser corrections without rewriting frozen raw outputs. WorldConsistMem therefore provides a controlled benchmark for representation-sensitive state updates—not a new memory architecture, universal LLM failure, or real-world compliance result.

Non-claims. We do not claim multi-provider generality; that retrieval is irrelevant; synergy beyond additivity; causal isolation beyond the F0–F3 arms; that F1/F2 were preregistered before F3; H0 superiority; a consistency deficit beyond nulls; ecological validity; or that 109/150 is complete reconstruction.

AI Use Statement

This statement is required by ICLR 2027 and does not count toward the page limit. We used generative AI coding assistants for manuscript editing, LaTeX packaging, experiment scripting assistance, analysis code, and repository tooling. We did not use generative AI to invent experimental results, numerical claims, citations, or benchmark labels. All headline numbers were checked against frozen artefacts. We reviewed AI-assisted text for accuracy and take responsibility for the final content.

Ethics Statement

WorldConsistMem uses fully synthetic worlds and machine-checkable constraints; the naturalistic transfer set uses fictional names only. No human subjects or personal data are involved. Real-world accountability or compliance claims are not made from this evidence.

Reproducibility Statement

We release seed-deterministic generation (seed 100; 102/1,088/7,140), frozen predictions, metric definitions (including scorer v1/v2 lineage), confirmatory scripts under results/confirmatory/, and reference implementations. Labelled confirmatory analyses are stored separately from Option A. Identifying URLs are omitted for double-blind review. Appendix D details protocols and checksums.

Acknowledgements

Omitted for double-blind review.

References

Sarim Chaudhry. LogicVault: Persistent symbolic belief states for cross-query logical consistency in LLMs. OpenReview preprint / workshop forum, 2026. URL https://openreview.net/forum?id=3pkSTLDiHo. OpenReview id 3pkSTLDiHo; associated code releases LogicBench-Cross.

Chuanrui Hu, Tong Li, Xingze Gao, Hongda Chen, Yi Bai, Dannong Xu, Tianwei Lin, Xiaohong Li, Yunyun Han, Jian Pei, and Yafeng Deng. Evaluating long-horizon memory for multi-party collaborative dialogues, 2026. URL https://arxiv.org/abs/2602.01313. Introduces the EverMemBench benchmark.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions, 2025. URL https://arxiv.org/abs/2507.05257. Introduces the MemoryAgentBench evaluation framework.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. WorldMemArena: Evaluating multimodal agent memory through action–world interaction, 2026. URL https://arxiv.org/abs/2605.29341.

Rohit Kumar Salla, Ramya Manasa Amancherla, and Manoj Saravanan. Quantifying cross-query contradictions in multi-query LLM reasoning, 2026. URL https://arxiv.org/abs/2604.14525. Accepted at the ICLR 2026 Workshop on Logical Reasoning of Large Language Models; introduces SetConsRate and related case-file consistency metrics.

Zhen Tao, Jinxiang Zhao, Peng Liu, Dinghao Xi, Yanfang Chen, Wei Xu, and Zhiyu Li. MemConflict: Evaluating long-term memory systems under memory conflicts, 2026. URL https://arxiv.org/abs/2605.20926.

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu, Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, and Jia Li. EvoMemBench: Benchmarking agent memory from a self-evolving perspective, 2026. URL https://arxiv.org/abs/2605.18421.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=pZiyCaVuti. Also arXiv:2410.10813.

Table 5: Frozen. B2 latest-only on one CEO bundle (8 queries; Acc = 0.75, Φ=0). Secondary vignette.

Query role	Gold	Prediction	Correct?
Current CEO	P_1	P_1	Y
Historical CEO	P_0	abstain	N
Succession	$P_0 \to P_1$	$P_0 \to P_1$	Y
Temporal order	true	abstain	N

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues, 2026. URL https://arxiv.org/abs/2605.12493.

Wenya Xie, Shengming Zhou, Zelin Li, Pouya Parsa, Shuang Zhou, Xinheng Ding, Chinmay Arvind, Guanchu Wang, Vladimir Braverman, Ali Payani, Yantao Zheng, and Zirui Liu. DynamicMem: A long-horizon memory benchmark in real-world settings, 2026. URL https://arxiv.org/abs/2606.22877.

A Metric Definitions (Secondary Joint Objects)

$$\mathrm{Acc}_i = \mathbf{1}[\hat{y}_i \equiv y_i^*], \tag{1}$$

$$\Phi(\hat{Y}_b; \mathcal{C}_b) = \mathbf{1}\big[\forall c \in \mathcal{C}_b : \; c(\hat{Y}_b)\big], \tag{2}$$

$$\mathrm{Acc} = \tfrac{1}{N}\sum_i \mathrm{Acc}_i, \qquad \mathrm{BCR} = \tfrac{1}{|B|}\sum_{b\in B} \Phi(\hat{Y}_b; \mathcal{C}_b), \qquad \mathrm{Gap}_{\mathrm{query}} = \mathrm{Acc} - \mathrm{BCR}. \tag{3}$$

CSR_b is the fraction of applicable constraints satisfied; BCR is the fraction of bundles with CSR_b=1. Complete reconstruction (CRR) requires every query in a bundle correct and is not identical to BCR. If $\mathcal{C}_b = \emptyset$, Φ=1.

B Frozen Transition Example

Selection rule (deterministic). Among CEO transitions that are (i) original-incorrect under scorer v2, (ii) structured-correct, (iii) original sufficiency S1, and (iv) mechanism M2 (previous-state wrong transition), take the lexicographically first query ID. This yields sl0168_000:b:ceo:0:transition.

Query text: "When and how did the CEO of org:0 change?" Gold: {from = person:4, to = person:1, t=32, org_id, role = CEO}. Matched B4 evidence (identical ordering in both conditions): three role_replacement observations at $t \in \{16, 32, 31\}$, including the gold support at t=32. Original output: {old = person:0, new = person:1, t=16}—an earlier replacement. Structured output: the gold tuple. What was repaired: previous-holder and time selection among distractors under an explicit prev/event/next+canonical schema—not missing evidence.

C Symbolic Option A Acc≠BCR (Secondary)

Frozen Option A (102/1,088/7,140) remains the primary symbolic Acc≠BCR demonstration, including identical Acc with different BCR (B2 vs. B3-k3: Acc = 0.4905 with BCR 0.219 vs. 0.000) and BM25 Acc = 0.929 with BCR = 0.591. Corruption controls and Qwen H4 BCR floors are reported in the frozen CSVs and prior tables retained below for completeness. These results support evaluation-object hygiene; they are not the LLM transition mechanism claim.

D Additional Reproducibility Details

Dataset. Scaled WorldConsistMem is generated with seed 100. Smoke validation confirms gold Acc = BCR = ConsAcc = 1 and Acc–Φ separation under controlled corruptions.

Readers. Symbolic exact-match scoring is used for the full scaled study. The LLM condition uses Qwen2.5-3B-Instruct-4bit with frozen short-form JSON decoding (temperature 0; max tokens 48) on the H4 subset (18 worlds / 192 bundles / 1,260 queries).

Systems. Headline reference systems: B1 capacity-limited compact flat; B3 recency retrieval; B4 BM25 retrieval; H0 structured hierarchical reference. Ablations and capacity sweeps are reported in the released artefacts.

Metrics. Primary consistency rate: BCR. Complementary: CSR, ConsAcc / PConsAcc grid, $\text{Gap}_{\text{query}}$, family violation rates. World-level bootstrap 95% CIs accompany BCR/CSR (and world-mean Acc) in frozen CSVs.

Labelled rescue sanity (not Option A). Re-evaluating gold answers on the H4 subset yields Acc=BCR=ConsAcc=1. Confirmatory artefacts live under results/confirmatory/ and results/new/stronger_reader_pilot/ (Appendix K). The Table 5 extract is results/new/b2_vignette_ss0100_ceo0.json. Matched-reader and transition recall=1 analyses: results/confirmatory/matched_reader_rescue/.

Reader coverage (updated). Option A completed Qwen2.5-3B-Instruct-4bit. Separately labelled confirmatory runs used two OpenAI snapshots (gpt-4.1-mini-2025-04-14, gpt-4.1-2025-04-14) on frozen H0 IDs up to 989/150 under the stated prompt/retrieval/decoding stack. Other providers were not evaluated. We do not claim multi-provider, all-LLM, or frontier coverage.

Checksum anchors. Scaled manifest sha256 37aa8604c3e6d87691332f66d2f7c0ab93578f2fe468bc29f7e8217c1675af3f; symbolic system_results.csv sha256 eb5bd5fa87496586bd830e5f685b9950e119f9fd0ac491194b3f003343f9a2a8; smoke dataset canonical sha256 3aa67ed1b6a39be3d9e1805d14f7d5ce7cc05f4655bdce0bc4ff49900c14c7ab.

E Artifact Checklist

Released artefacts accompanying this submission include: dataset generators and manifests; symbolic system results and predictions; Qwen H4 system/query-family/constraint-family results; partial-consistency recomputes; corruption-suite metrics; labelled NEW robustness JSON; and smoke/reproduction scripts. Manuscript figures are static PDFs; figure-drawing source scripts were not present in the experiment tree. Anonymized supplementary material is provided with the OpenReview submission; identifying repository URLs are omitted from this PDF.

F Extended Related Work

F.1 Long-term memory evaluation

LongMemEval evaluates chat assistants on long-term interactive memory through five ability axes—information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention—with answer accuracy as the primary metric under long-history and long-context protocols (Wu et al., 2025). LongMemEval-V2 extends the agenda toward "experienced colleague" competence, adding dynamic state tracking, workflows, and premise awareness, again with per-question accuracy (and latency-oriented reporting) as the headline outcomes (Wu et al., 2026). MemoryAgentBench evaluates agent memory via incremental multi-turn interactions across competencies such as accurate retrieval, test-time learning, long-range understanding, and selective forgetting / FactConsolidation after edits, scored primarily by task or answer accuracy (Hu et al., 2025). EverMemBench targets long-horizon

collaborative, multi-party dialogues and reports fine-grained recall, memory awareness, and profile understanding under QA accuracy protocols (Hu et al., 2026). EvoMemBench organizes agent memory from a self-evolving perspective with a scope×content taxonomy and reports task success or accuracy (and efficiency) across methods compared with long-context baselines (Wang et al., 2026). Taken together, this line provides mature protocols for long-horizon Acc and ability factorization; the measurement gap we target is joint coherence of dependent answers, not another Acc axis.

These benchmarks primarily evaluate individual query or task outcomes. WorldConsistMem does not replace ability Acc suites. It additionally evaluates whether answers within a dependent bundle jointly satisfy constraints derived from one world history. The evaluation unit is the interdependent answer set $\hat{Y}_b$ under $\Phi(\hat{Y}_b; \mathcal{C}_b)$, not a single graded QA item.

F.2 Dynamic and evolving knowledge

Evolving worlds are not novel by themselves. DynamicMem constructs long-horizon personal memory in real-world settings and evaluates temporal checkpoint reconstruction and related state-tracking accuracy for evolving user attributes, habits, and preferences (Xie et al., 2026). WorldMemArena evaluates multimodal agent memory through action–world interaction, diagnosing write, maintain, retrieve, and use stages under lifelong evolution and agentic execution protocols (Liu et al., 2026).

It is important to distinguish three notions that are easy to conflate. Tracking changing state asks whether the system recovers the correct value at a time or checkpoint. Profile or lifecycle consistency asks whether personal attributes, stage metrics, or write/use pipelines remain coherent for a user or agent trajectory. Cross-query world-history consistency asks whether interdependent answers about current state, historical state, transitions, relations, and provenance jointly satisfy constraints induced by one shared multi-entity gold history.

F.3 Conflict, temporal validity, and provenance

MemConflict evaluates long-term memory systems under dynamic, static, and conditional memory conflicts, combining black-box answer correctness with white-box retrieval and ranking diagnostics, and showing that answer correctness can diverge from retrieval fitness (Tao et al., 2026). These lines diagnose conflict handling, version validity, and temporal updates at the query, retrieval, or competency level. The WorldConsistMem contrast is evaluation grain and joint structure rather than denial of those pathologies: a conflict- or update-correct single answer can still participate in an inconsistent bundle.

F.4 Cross-query logical consistency

Salla et al. (2026) study case-file logical consistency across interdependent queries and introduce set-level metrics including Case Satisfiability and SetConsRate. Chaudhry (2026) maintain persistent symbolic belief states verified with SMT-style checking and release LogicBench-Cross. Machine-checkable constraint evaluation over multi-query answer sets is therefore not universally new. WorldConsistMem's residual is application domain and constraint origin: persistent memory systems over evolving multi-entity worlds with world-history-derived constraints.

G Fairness Caveats

Comparisons are informative but not perfectly matched:

- Internal structure differs (flat list vs multi-store H0).
- Capacity accounting charges H0 metadata (graph, provenance, indices) against the same logical-record budgets used for flat stores; absolute budgets therefore tax structured metadata.
- Retrieval budgets (k vs store-routed packs) are matched numerically where stated, but lexical BM25 is not identical to H0's routing.

Table 6: Closest related benchmarks and adjacent consistency work. Cells use Y / P / N for whether the feature is a primary evaluation object as operationalized in WorldConsistMem.

Work	LTM	Evol.	Bundles	Gold hist.	Trans.	Prov.	CQ constr.	Strict BCR	Partial CSR	Primary metric
LongMemEval	Y	P	N	P	P	N	N	N	N	Ability Acc
LongMemEval-V2	Y	Y	N	P	P	N	N	N	N	Acc (+ latency)
MemoryAgentBench	Y	P	N	P	P	N	N	N	N	Competency Acc
EverMemBench	Y	Y	N	P	P	N	N	N	N	Recall / profile Acc
EvoMemBench	Y	P	N	P	P	N	N	N	N	Task Acc
MemConflict	Y	P	N	P	P	P	P	N	P	Conflict Acc + retrieval
DynamicMem	Y	Y	P	Y	P	N	P	N	P	Checkpoint Acc
WorldMemArena	Y	Y	P	P	P	P	N	N	P	Lifecycle Acc
SetCons metrics	N	N	Y	P	P	N	Y	Y	P	SetCons / satisfiability
LogicVault	N	N	Y	P	P	N	Y	Y	P	Logical / SMT
WorldConsistMem	Y	Y	Y	Y	Y	Y	Y	Y	Y	Trans. + Acc (+BCR)

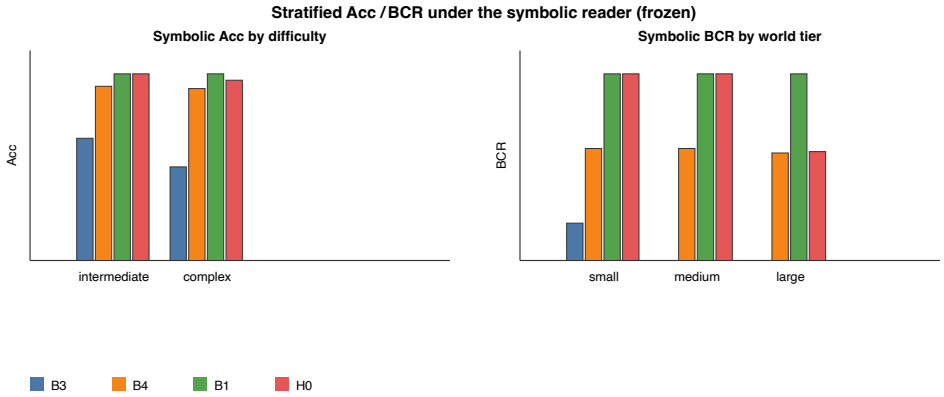


Figure 3: Symbolic Acc by difficulty and BCR by world tier.

- Deterministic vs LLM readers are separate conditions; symbolic Acc/BCR must not be pooled with Qwen Acc/BCR.
- Compact flat can retain focal lifecycle events by dropping padding, which can yield very high symbolic Acc/BCR at medium budgets; this is a fairness caveat for matched-budget interpretation, not a general ranking of store designs.

H Additional Figures and Tables

I Qwen parser sensitivity (confirmatory)

Re-scoring frozen H4 raw_output (no re-inference). Model: Qwen2.5-3B-Instruct-4bit. Across strict JSON, relaxed JSON, and answer-field regex, H0 Acc ≈ 0.187 and BCR $= 0$; B4 Acc ≈ 0.147 and BCR $= 0$; B3 Acc ≈ 0.075 and BCR $= 0$. Abstention rates $\approx$0.48–0.70; malformed rates $\leq$0.029. Cache-wide error channels: abstention $\approx$54%, substantive/unchecked $\approx$44%, malformed $\approx$1.5%. A fixed raw-output sample (seed 20260908; n=12 H0 rows) shows majority three-parser agreement; occasional field-regex recoveries of malformed strings do not change BCR. Low Acc means this rejects a single-parser artefact without supporting high-Acc Acc–BCR divergence for Qwen.

J Naturalistic transfer details (confirmatory)

Independent scenarios n=36; queries/scenario = 4; total queries = 144; scenarios sha256 b84dfa83…. Definitions: gold oracle emits gold answers; last-event uses only the last narrative sentence (current-like entity; abstains on historical/transition/temporal); entity-prior collapses to a P1/LOC1 shortcut; narrative-heuristic is full-narrative regex extraction. Scenario-level Acc CIs: last-event $[0.174, 0.236]$; gold $[1, 1]$. Role Acc (last-event): current 0.833, others 0. Correct-query counts (last-event): 0 in 6 scenarios, 1 in 30. Conjunc-

Table 7: Constraint-family violation rates (headline systems).

	Symbolic				Qwen		
Family	B3	B4	B1	H0	B1/B3	B4	H0
uniqueness	0.156	0.000	0.000	0.000	0.219	0.177	0.000
temporal	0.406	0.000	0.000	0.156	0.354	0.323	0.281
transition	0.375	0.120	0.000	0.000	1.000	1.000	1.000
relational	0.562	0.343	0.000	0.000	0.625	0.625	0.625
provenance	0.625	0.057	0.000	0.000	0.938	0.724	1.000
multi_hop	0.344	0.124	0.000	0.000	0.406	0.406	0.406
contradiction	0.000	0.000	0.000	0.000	0.052	0.073	0.188

Table 8: Query-family Acc under Qwen (H4; frozen; three-decimal display).

Family	B1	B3	B4	H0
current_state	0.010	0.010	0.083	0.542
historical_state	0.162	0.162	0.089	0.547
transition	0.000	0.000	0.000	0.000
multi_hop	0.000	0.000	0.052	0.000
provenance	0.188	0.167	0.417	0.000
temporal_order	0.128	0.128	0.205	0.308
contradiction	0.083	0.083	0.167	0.000
counterfactual	0.000	0.000	0.000	0.000
conflict_validity	0.462	0.462	0.462	0.000

tion diagnostic (reference only): last-event $p^4 \approx 0.0019$ and role-product = 0, so observed BCR = 0 is consistent with low-p conjunction rather than a high-Acc residual. Narrative heuristic Acc = 0.875 with BCR = 1 (negative $\text{Gap}_{\text{query}}$; consistent-but-not-fully-factual). No LLM reader completed on these scenarios.

K Proprietary-reader transfer check (NEW)

Sampling. Frozen stratified transfer subset of H4/H0: deterministic proportional allocation by (tier × difficulty), seed 20260909, complete bundles only (22 bundles / 149 queries). Manifest sha256 dd9790c8…. Primary report sha256 e3ac10ae….

Protocol. Requested alias gpt-4.1-mini; independent smoke resolution gpt-4.1-mini-2025-04-14; cache rows store the alias. Temperature 0; max tokens 48; prompt prompt_v3_k3_mt48; H0/LTKM evidence; same JSON parse. Statuses: 102 ok / 41 abstain / 6 malformed. All six malformed have completion_tokens=48 (truncated JSON); counted incorrect; no selective retry.

CIs. Acc: bundle-level bootstrap, 10,000 resamples, seed 20260911, percentile interval. BCR = 0/22: exact Clopper–Pearson two-sided $[0, 15.4\%]$; one-sided 95% upper 12.7%. Does not prove population BCR is exactly zero.

Matched Qwen. Exact same 149 query IDs from frozen run2 H0 predictions (sha256 5905b081…); Acc = 20.1%, BCR = 0/22. Analysis outputs under results/confirmatory/proprietary_reader_transfer/ (not pooled with Option A).

Family Acc (GPT-4.1 mini). current 17/22; historical 22/22; contradiction 6/6; temporal_order 7/11; counterfactual 6/11; conflict_validity 5/11; multi_hop 5/22; transition 0/22 (incl. 6 truncations); provenance 0/22. BCR is not reported by family (bundle-level metric).

Evidence-conditioned slice (NEW, read-only). Among GPT-4.1 mini standard-retrieval outputs, BCR = 0/8 on bundles with ≥1 retrieved evidence ID per query and no truncation; transition/provenance Acc remain 0 with evidence present. Artefacts: results/confirmatory/acceptance_push/evidence_conditioned_analysis.json.

Table 9: Frozen. Symbolic H0 component ablations (reprinted from freeze CSV). Large Acc–BCR gaps under archive/temporal/graph removals.

System	Acc	BCR	$\text{Gap}_{\text{query}}$
H0 unconstrained	1.000	1.000	0.000
H0 cap150 ret8	0.976	0.844	0.132
H–archive	0.876	0.375	0.501
H–temporal-validity	0.876	0.375	0.501
H–graph	0.938	0.594	0.344
H–provenance	1.000	1.000	0.000
H–conflict	0.800	0.812	−0.012

Table 10: Labelled NEW. Retrieval-budget sensitivity (seed 100; 6 worlds/tier; 18/192/1,260; symbolic). Identical Acc with different BCR for B2 vs. B3-k3.

System	Acc	BCR	$\text{Gap}_{\text{query}}$
B2 latest-only	0.4905	0.2188	0.2717
B3 recency-k3	0.4905	0.0000	0.4905
B3 recency-k5	0.5095	0.0208	0.4887
B3 recency-k8	0.5667	0.0625	0.5042
B4 BM25-k3	0.7968	0.3750	0.4218
B4 BM25-k8	0.9317	0.5938	0.3379

Oracle-retrieval confirmatory (NEW; completed). Same 149/22 IDs; gold supporting observations; max_tokens=256; resolved gpt-4.1-mini-2025-04-14; spend $0.0220168. Acc = 0.44295, bootstrap $[0.38356, 0.50331]$; BCR = 0/22, CP $[0, 0.15437]$; abstain 43; malformed 0; truncated 0. Family: provenance 12/22, multi_hop 0/22, transition 0/22. Artefacts: results/confirmatory/oracle_only_pilot/. Interpretation: oracle evidence did not improve BCR; failure is not solely retrieval or the prior 48-token ceiling—not a claim that retrieval is irrelevant or cross-provider.

Recommended minimal matrix (NEW; completed offline-audited). Frozen 150-bundle / 989-query expansion (sha256 7f9e50e9…); H0; k=3; max_tokens=256; temp 0; prompt prompt_v3_k3_mt48. Cumulative billed spend $1.2362336 over 3191 ledger rows. Matched standard readers: GPT-4.1 mini Acc = 0.39737 BCR = 0/150; GPT-4.1 Acc = 0.39232 (bootstrap $[0.37335, 0.41106]$) BCR = 0/150 (CP $[0, 0.02429]$; one-sided upper 0.01977); abstain/mal/trunc 368/0/0; 989 calls; resolved gpt-4.1-2025-04-14; tokens 290,207/25,748; spend $0.786398. Mini oracle Acc = 0.49949 and joint Acc = 0.49848, both BCR = 0/150 (ablations, not matched GPT-4.1 conditions). Family (GPT-4.1): historical 150/150, current 94/150, multi_hop 29/150, temporal_order 27/64, contradiction 44/58, counterfactual 44/59, transition 0/150, provenance 0/150, conflict_validity 0/58. No Gemini/Claude results. Post-hoc claim (Appendix L): zero complete reconstruction is largely consistent with compounding under family-conditioned nulls; transition/provenance are the dominant bottlenecks; GPT-4.1 does not meaningfully improve Acc vs. mini. Artefacts: MATRIX_REPORT.json.

L Post-hoc BCR / complete-reconstruction robustness

Offline analysis on frozen matrix predictions only (no API calls; no prediction edits; requested after the initial submission freeze). Primary BCR/Φ counts are unchanged.

Metric separation. Item Acc; per-bundle fraction correct; complete reconstruction (all_correct); BCR/Φ (constraint consistency). The reviewer "Acc≈0.4 conjunction" objection targets complete reconstruction; BCR remains Φ.

Null models (complete reconstruction). (A) Global Bernoulli at each condition's Acc. (B) Family-conditioned product over each bundle's query families. (C) Family-stratified

Table 11: Labelled NEW. Shortcut baselines (seed 100; 6 worlds/tier; symbolic scoring). Family-majority uses oracle gold marginals by query family.

Baseline	Acc	BCR	$\text{Gap}_{\text{query}}$
Empty answers	0.000	0.000	0.000
Family-majority (gold marginals)	0.352	0.000	0.352
Family-shuffle of gold answers	0.280	0.000	0.280
Gold oracle	1.000	1.000	0.000

Table 12: Labelled NEW. Seed sweep summary (symbolic B3/B4 k=8; 6 worlds/tier; seeds 100–105). B3 identical across seeds (limited diversity).

System	Acc range	BCR	$\text{Gap}_{\text{query}}$ range
B3 recency-k8	0.567–0.567	0.0625	0.504–0.504
B4 BM25-k8	0.929–0.932	0.5938	0.335–0.338

permutation (10^5 sims; seed 20260916) preserving per-family correct counts. Rounding: threshold met iff $n_{\text{correct}}/n + \varepsilon \geq \tau$ ($\varepsilon = 10^{-12}$; no percent rounding).

Decision. Outcome B: zero completion is largely explained by compounding; transition/provenance dominate. No "consistency deficit" wording (observed is not below the family/permutation nulls). Bundle-frac variance is not elevated vs. the permutation null (ICC exploratory ≈ -0.08). Diagnostic family exclusions are post-hoc only and do not license removing those families from the task.

Paired mini vs. GPT-4.1. Perfect ID alignment (989/150). $\Delta\text{Acc} = -0.00506$ (bundle-clustered bootstrap $[-0.0193, 0.0091]$); McNemar discordant 33 vs. 28 ($p \approx 0.61$); disagreements: both correct 360, mini-only 33, GPT-4.1-only 28, neither 568. No pre-specified equivalence margin.

Artefacts. results/confirmatory/bcr_robustness/BCR_ROBUSTNESS_ANALYSIS.json, BCR_ROBUSTNESS_REPORT.md, per-bundle JSONL, fig_bundle_completion_curves.svg.

M Post-hoc cross-query constraint consistency

Exploratory offline analysis. Constraint spec CROSS_QUERY_CONSTRAINT_SPEC.json was hashed and frozen before scoring LLM predictions. Pass/fail does not require gold match. Abstention/malformed ⇒ not_evaluable (never counted as satisfied). Prior BCR-robustness Outcome B is unchanged.

Decision. Outcome B: among answered/evaluable constraints, models are not more contradictory than matched permutation nulls (10^5 sims; seed 20260916; per-bundle aggregation). Mean coverage ≈0.32 (mini) / 0.34 (GPT-4.1). CSR is not below null (mini CSR 0.523 vs. null mean 0.476). Symbolic H0/B4 on the same IDs: CSR=1.0 with coverage > 0.9.

Title. The manuscript title emphasizes state-transition reconstruction beyond retrieval.

Artefacts. results/confirmatory/constraint_consistency/.

N Scorer correction (v1 → v2)

What v1 did. Production answers_equal required gold keys on the prediction (subset match). CEO gold tuples use from/to/t; many GPT outputs used role-replacement pay-

Table 13: Labelled NEW. Generator controls (symbolic). Density: 8 medium worlds, seed base 200. Distractors: seed 100, wpt=6.

Control	Setting	System	Acc	BCR
Relation density	0.0 / 1.0	B4	0.917 / 0.909	0.600 / 0.600
Relation density	0.0 / 1.0	B3	0.545 / 0.545	0.000 / 0.000
Distractors	on / off	B4	0.932 / 0.932	0.594 / 0.594
Distractors	on / off	B3	0.567 / 0.567	0.062 / 0.062

Table 14: Evaluation coverage and hard limits.

Dimension	Coverage
Paper shape	Benchmark + evaluation methodology
Symbolic reader	Full scaled (102 / 1,088 / 7,140)
LLM reader	Qwen2.5-3B-Instruct-4bit; H4 subset only
Stronger LLM readers	Proprietary GPT-4.1 mini transfer (22 bundles); not frontier coverage
Architecture superiority	Unsupported under Qwen (all BCR = 0)
Ecological validity	Not claimed (synthetic worlds)
Primary consistency metric	BCR (strict); CSR complementary
Labelled robustness	Seeds/density/distractors/shortcuts/retrieval-budget (separate)
B3 seed diversity	Identical Acc/BCR across seeds 100–105 (not replication)
Proprietary transfer	Frozen stratified H0 subset; Acc 45.6%; BCR 0/22

load keys old_person_id/new_person_id/time with otherwise correct values. v1 therefore scored 0/150 B4 transitions and treated 13 semantically correct CEO rows as errors.

Why that was wrong. Those aliases are field renames under the frozen event schema, not fuzzy matches or value repairs. Keeping known semantic false negatives as incorrect is not permitted for the corrected central claim.

Exact scope of v2. Only three aliases are accepted (case-sensitive, top-level only): old_person_id→from, new_person_id→to, time→t. Values are compared with the original strict rules after rename. Conflicting canonical+alias values are rejected. No fuzzy/substring keys, no per-query exceptions, no gold-dependent inference, no nested flattening. Digit-string ↔ int coercion for time is explicit and bounded.

Impact. Across all frozen transition cells, only matched GPT-4.1 B4 changes: 0/150 → 13/150 (all CEO; CI $[0.047, 0.144]$). Other conditions remain 0/150 under v2 (mini standard/oracle, GPT-4.1 standard, joint, budget k=1/8, pilots). Overall B4 Acc $0.410 \rightarrow 0.423$; bundle all-correct and BCR unchanged at 0/150. 13 rows change; raw predictions, API caches, spend ledgers, and prior freezes are untouched. No model was re-run. Artefacts: results/confirmatory/scorer_v2/ (TRANSITION_SCORER_V2_SPEC.json, SCORER_V1_RESULTS.json, SCORER_V2_RESULTS.json, changed_rows.jsonl, SCORER_CORRECTION_AUDIT.md).

Other families. A global key-remapping suspect scan found no material analogous defect meeting the freeze-stop threshold; no additional aliases were introduced.

O Evidence sufficiency strata (S1–S5)

Offline labels on B4 transition packs (evaluation-only; not used at inference): S1 explicit statement of resulting state; S2 deterministic entailment; S3–S5 underspecified/unsupported relative to the gold schema. Original counts (primary confirmatory): S1=59, S2=34, S3=28, S5=29 (S1+S2=93). Original generic-interface Acc on S1+S2 = 13/93; structured Acc on the same original slice = 52/93. Do not describe all 150 recall= 1 cases as evidence-sufficient under these labels.

Table 15: Post-hoc independence nulls for complete reconstruction (not primary BCR). Observed complete = 0 in all rows. Bern/Fam: analytical; Perm: Monte Carlo.

Condition	Acc	Bern E[comp]	Bern $P(0)$	Fam E[comp]	Perm E[BCR]	Trans/Prov Acc
mini standard	0.397	0.422	0.655	0	0	0 / 0
GPT-4.1 standard	0.392	0.390	0.677	0	0	0 / 0
mini oracle	0.499	1.754	0.171	0	0	0 / 0.733
mini joint	0.498	1.732	0.175	0	0	0 / 0.040
mini k=1 (22)	0.409	0.066	0.936	0	0	0 / 0
mini k=8 (22)	0.436	0.099	0.906	0	0	0 / 0

P Post-hoc sufficiency reclassification (forensic only)

A forensic re-audit applies written global rules (not per-item favouritism): e.g., accept person_leaves_org as employment leave (E1) and map project_ownership_transfer old/new org fields to gold from/to (O1). Proposed counts: S1=96, S2=44, S3=10, S5=0; 77/150 labels change. Original labels are retained; this taxonomy was not preregistered and is not used for the primary Outcome A gate. Artefacts: SUFFICIENCY_RECLASSIFICATION_AUDIT.md under the structured-prompt forensic audit directory.

Q Bare-JSON runner correction

The structured prompt requested bare canonical JSON. The online runner reused a parser that required an {"answer":…} wrapper and marked 150/150 outputs malformed. Offline re-parse of frozen raw outputs recovers 150/150 bare objects; no model was rerun. The prospective parser accepts bare and wrapped forms; regression tests cover bare, wrapped, abstain, malformed, schema-invalid, and conflicting representations. Artefact: BARE_JSON_RUNNER_CORRECTION.md.

R Structured-prompt forensic audit summary

Leakage gate: PASS (no gold-derived answer values in instruction/schema beyond fixed subtype constants; query/evidence byte-match B4). Independent rescoring: 109/150; second checker disagreements: 0. Wins: 97 non-format; M5 format-only: 0; M8: 0. Outcome A on original S1+S2. Artefacts under results/confirmatory/final_validation_post_scorer_v2/structured_prompt/forensic_audit/.

S Factorized ablation integrity

Prospective post-result F1/F2 execution on frozen prompts (no prompt regeneration). Aggregate prompt SHA-256: F1 4c68d42f762f5e57ab0e0c741f4666a95ce9af6b6a26af0e7cef5bc8eb68a1c5; F2 8a1a0ccb22e5fbef0ca6a0357382ce2873c6d3d63579fe26a90e2a1c8cc65e8b; F3 unchanged b3be46945eb23df93e1b084b94064f55beff7cb4868960f3153c0160dbf44882. Scorer-v2 specification SHA-256 1820c8efe0e16ac09a4b6817da18b40518dbd254425fd1fb3fbd16789f026ba8 (unchanged). Query IDs and B4 evidence packs are byte-identical across F0–F3; F1/F2 query-ID sets match (150 unique IDs). Resolved snapshot gpt-4.1-2025-04-14 for 300/300 successful calls; cumulative spend $0.440684 under the $1.50 joint ceiling; 0 transport retries on the successful run. Parser: F1/F2 emitted wrapped answer objects (150/150 ok); F3 offline re-parse remains 150/150 bare canonical JSON (execution-time wrapper-mismatch labels are not true malformedness; no API rerun). Attempt log contains one stale HTTP_401 from an earlier invalid-key probe; that attempt produced no prediction row and did not affect Acc, contrasts, or spend totals for the completed 300-call run. Analysis artefacts: FACTORIZED_ABLATION_ANALYSIS.json, FACTORIZED_ABLATION_INTEGRITY.json under results/confirmatory/final_validation_post_scorer_v2/factorized_ablation/run/.

Table 16: Factorized Acc by transition subtype (correct / n; scorer v2).

Subtype	n	F0	F1	F2	F3
CEO	30	13	10	26	23
Conflict	29	0	29	0	29
Document	34	0	0	0	0
Employment	28	0	28	0	28
Ownership	29	0	23	13	29

T Factorized qualitative extracts

Schema recovery (conflict). ID sl0168_000:b:conf:4:transition (original S1); query asks when conflict conflict:0 was resolved. Gold: {conflict_id, resolved_value=P0, t=21}. F0/F2: scalar 21 (incomplete under scorer v2). F1/F3: full canonical object matching gold. Factor: canonical state schema, not JSON syntax alone.

Representation recovery (CEO). ID sl0168_000:b:ceo:0:transition (Appendix B selection rule). Gold: CEO person:4 $\rightarrow$ person:1 at t=32. F0: earlier replacement at t=16. F1: wrong intermediate replacement (person:3 $\rightarrow$ person:4 at t=31). F2/F3: gold event under prev/event/next. Factor: explicit temporal transition representation among distractors.

Unresolved document / S2 floor. ID sl0168_000:b:doc:11:transition (original S2). Gold: docbase:0 versions 6 $\rightarrow$ 7 at t=45. All arms incorrect; F3 emits string versions v1$\nrightarrow$v2 at t=43. Schema elicitation cannot invent missing numeric version identifiers from an underspecified pack.