MACHINE CAUSALITY STRATIFICATION: DERIVED AUTHORITY, REVERSIBLE CAUSAL LEDGERS, AND COVERAGE-GATED CAPABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory systems for agents increasingly store causal claims, yet the causal lines surveyed in Section 2 report capability as a single level rather than a graded, evidence-derived one. That framing cannot answer what matters when a system's own memory encodes causal claims: how deep is the claim, why should it be believed, can its authority be revoked, and can one high-level edge make the whole system look causally mature. We present **CKB**, a causal-memory substrate with a machine-governed causal-authority lifecycle: a graded **CDL** (Causal Depth Level, associative to interventional), a predicate-derived and reversible **CAL** (Causal Authority Level, a strict accumulation over evidence predicates), a per-edge capability coordinate **CCC**, and a coverage-gated envelope **CCE** that stops one high-depth edge from inflating the reported depth. Authority is captured in an append-only, reversible witness DAG whose transitions are receipted and replayable. We separate *what kind of causal claim an edge expresses* (CDL) from *how much authority the system may assign to it* (CAL). An implemented Rust substrate and a cross-language Python reference oracle execute the calculus at the substrate and conformance layer; we report structural certificates, exact indexed recall, a registered counterfactual edge audit, and downstream measurements including a null result for incremental explicit-structure presentation.

Downstream, a preregistered paired test (N = 80, a LongMemEval-S public-development subset) finds that external memory helps substantially (+32.5 pp over no memory, p = 3.0e-8), but that presenting the same evidence as typed causal/temporal structure yields no detectable gain over a flat presentation (+1.25 pp, p = 1.0); the null persists under a matched-token placebo arm. On the full 500-question public development set, the frozen official gpt-4o-2024-08-06 evaluator scores a frozen candidate at 493/500; with no baseline and no recorded answer generator, we report this as an observability result, not a comparative or generalization claim (Section 8).

1 INTRODUCTION

1.1 PROBLEM

An agent is a system that persists, acts, and must answer for its actions over time, and persistence is memory: without it there is nothing to learn from, no commitment to keep, and perception collapses into reflex. Memory alone, however, is not a foundation. A store that cannot say why it holds a claim, that cannot separate what was observed from what was inferred, and that cannot withdraw a conclusion when its evidence collapses manufactures confidence its contents never earned, and any agent built on such a store inherits that manufactured authority. Stable causal memory, in which every claim carries an explicit and revocable warrant, is therefore the minimal substrate on which agency can be built at all. Today those questions are answered, when at all, by a boolean, which we argue is the wrong representation: the missing object is an explicit, evidence-governed **authority lifecycle** over a persistent epistemic state.

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.

1.2 HYPOTHESIS (M1)

Persistent memory as revisable epistemic state. We hypothesise a stronger representation: an explicit, provenance-grounded epistemic state over temporal and causal dependencies, in which derived consequences stay linked to the evidence they depend on, so retrieval is one operation over that state rather than the definition of memory. The hypothesis does not assume that a causal representation is ground truth: the maintained state is the system's current best-supported resolution under the evidence available at that time.

1.3 SCOPE NOTE (M0)

CKB is not a retrieval system, a graph database, or a graph-based RAG Lewis et al. (2020) variant; it is a hypothesis about the representation of an agent's epistemic state. Our claims are at the level of (i) what may enter memory and under what source constraints (epistemic admission), (ii) which conclusions remain warranted under the currently supported state (authority and revision), and (iii) evidence preservation through answer generation. We do not position the contribution as an increment over graph-structured retrieval: baseline comparisons test this representation hypothesis, **not** an improvement of retrieval. **CKB is one foundational substrate within a broader agent architecture**; this paper scopes itself to persistent epistemic memory and does not claim to solve full agent autonomy.

- **CKB CLAIMS:** epistemic admission; persistent epistemic state; authority/provenance semantics; temporal-causal dependency; revision of derived consequences; evidence preservation.

1.4 CONTRIBUTIONS

1. **CDL - Causal Depth Level (1).** A per-edge graded ladder from associative to interventional: *what was claimed about the world*.
2. **CAL - Causal Authority Level (2).** Predicate-derived, strictly accumulated, reversible authority assigned *by the system*, never by the claim's author.
3. **CCC - per-edge capability coordinate (3).** `CCC = (CDL, CAL)`, the coordinate at which each stored edge is reported.
4. **CCE - coverage-gated system envelope (4).** `system_verified_cdl(`τ`) = max{ d: Pr(CDL` $\geq$ `d` $\wedge$ `CAL` $\geq$ `a_min)` $\geq \tau$ `}`, so one isolated high-depth edge cannot raise the system's verified level.
5. **Authority calculus + reversible witness DAG (5).** Authority derived from auditable evidence in one pipeline, with revocable/demotable/fork-mergeable transitions over a ledger-minted, content-bound snapshot.
6. **Implemented and audited substrate (6).** Typed causal/epistemic edges, belief/world separation, and exact indexed recall on a Rust engine.

Evaluation frame: *Causal capability here is treated as an architectural property of the memory substrate and its authority calculus. The reader is held fixed only in §7.2; on LongMemEval-S Wu et al. (2025) the answer generator is UNKNOWN and no such control is claimed (§7.3). The contribution concerns learned-system capability, not parametric learning*.

Standing qualifiers: the LongMemEval-S outcome is a *post-hoc, evidence-adjudicated* result on a *public development set*, not a hidden-set measurement; the CCE arms are *synthetic injections, not a deployed capability*; the temporal receipts are reviewer receipts with zero confidence intervals; third-party evidence is still pending.

2 RELATED WORK

2.1 ANTECEDENTS

Each line we credit is a **conceptual antecedent**, not something CKB re-derives; we claim none of them "missed" causality: the causal hierarchy of Pearl Pearl (2019), the OCEBM Centre for

Evidence-Based Medicine, University of Oxford / CEBM Levels of Evidence, W3C PROV-O Lebo et al. (2013), AGM Alchourrón et al. (1985) belief revision, CMMI (SEI). Pearl grades the kind of *query* while CKB grades the *depth of a stored claim* (CDL); CMMI measures organisational maturity while CCE gates a memory system's causal capability. Antecedent map and verification status come from a prior-art audit whose primary pages were all opened on 2026-09-18.

2.2 BASELINES

- **LongMemEval** is our **baseline**: 500 public-development questions over five long-term-memory abilities. It measures "retrieved and answered", not "why a stored causal claim is currently warranted"; CKB's question does not live in its metric space.
- **Retrieval-side adjacent lines**: RAG and GraphRAG Edge et al. (2024) add structure or capacity **to retrieval**; recent agent-memory systems (MemGPT Packer et al. (2023), Generative Agents Park et al. (2023), HippoRAG Jiménez Gutiérrez et al. (2024), mem0 Chhikara et al. (2025)) extend retrieval with structured or persistent context. CKB differs in kind: memory is an **evidence-governed epistemic state** with derived authority and revision, not a richer retrieval index.
- **Closest temporal-KG memory line**: **Zep Rasmussen et al. (2025) / Graphiti**, the highest novelty-risk item: only its abstract was read, so its internal authority/revision semantics are **UNREAD**; we claim no priority and do **not** compare accuracy.
- `TimeRAG` and `arXiv:2010.07487` (not xERTE) are excluded as out of scope (Appendix D). Any LoCoMo Maharana et al. (2024) *score* is unusable: its judge-tolerance mechanism fixed but **never re-run**, so LoCoMo may appear only as a dataset citation. The paper therefore cites no LoCoMo and no MuSiQue Trivedi et al. (2022) number; CKB's own MuSiQue supporting-evidence-recovery attempt was an exact tie with its BM25-only arm, the number withheld pending an audit of the external evidence pack.

2.3 NOT-CLAIMS

1. Not an increment of RAG.
2. Not GraphRAG or "GraphRAG+": **capability inclusion, not sibling classification**.
3. Not a graph database replacement.
4. Not a temporal-KG completion/prediction method.
5. Not an accuracy-improvement claim.

That audit did **not** run a systematic literature search: the boundary is "**no counterexample found**", not "**no precedent exists**". We adopt that weaker wording, make no global-uniqueness or priority claim, and qualify it as *to our knowledge, subject to a complete primary-source audit*.

3 ARCHITECTURE: EPISTEMIC CONSTITUTION AND CAUSAL DYNAMICS (M3)

The knowledge base separates epistemic admission from causal dynamics. Two responsibilities are distinct. **Epistemic Constitution** determines what may enter persistent memory and under what evidential status: it records provenance, source bindings, authority, contradiction, temporal validity, and the further conditions that distinguish an observation from an admissible memory claim. **Causal Dynamics** governs how admitted memory state influences other state: causal dependencies propagate support, competing evidence may alter resolved attitudes, and downstream consequences are recomputed when their upstream basis changes. This separation prevents the causal graph from becoming its own source of authority: **causal structure governs dependency and revision, while evidential admission governs why a relation is currently warranted**.

4 FRAMEWORK: MACHINE CAUSALITY STRATIFICATION (MCS) — CDL, CAL, CCC, CCE

4.1 TWO ORTHOGONAL AXES: CDL AND CAL

CDL (Causal Depth Level) measures *what kind of causal claim an edge expresses*: associative (CDL-0, no causal claim made), direct (CDL-1), mechanistic (CDL-2, intermediate mechanism known), compositional (CDL-3, multi-hop with conjoined conditions), regime-contextual (CDL-4, holds only under a stated world regime), counterfactual (CDL-5, answers "if not A, how would B change"), and interventional (CDL-6, do(A) manipulation).

CAL (Causal Authority Level) measures *how much authority the system is permitted to assign to that claim*: it is not a field written by the claim's author. CAL is derived from an evidence predicate chain with strict accumulation,

```
DerivedCAL = max({0} U { k: P1 AND P2 AND ... AND Pk })
```

so CAL = k holds only if every predicate up to k holds, and a missing intermediate predicate caps the level. The five predicates **P1–P5** are defined in Appendix B.4. A CDL-5 claim with CAL-1 is a deep claim under weak evidence; a CDL-1 claim with CAL-3 is a shallow claim under strong evidence.

4.2 CCC AND THE COVERAGE-GATED ENVELOPE (CCE)

Each stored edge carries its **capability coordinate** `CCC = (CDL, CAL)`. A system cannot report its own causal depth as the best single edge, so three distinct quantities are defined: `max_supported_cdl` (representability anywhere), `max_demonstrated_cdl` (per-edge best at CAL $\geq$ a_min), and the coverage-gated envelope

```
system_verified_cdl(tau) = max{ d: Pr(CDL >= d AND CAL >= a_min) >= tau }
```

The envelope is the only one that answers "at what causal depth is the system *verified*". Throughout this paper `a_min = CAL-1`, the lowest non-zero authority level, so the second gate of the envelope carries its value wherever the formula is instantiated. The anti-inflation property is conditional on $\tau > 1/N$ (Appendix B.4).

The CCE experiment injects **synthetic** high-CDL edges into a frozen causal memory (A/B/C three arms at $\tau = 0.30$) and shows that a single isolated CDL-5 / high-CAL edge raises `max_demonstrated_cdl` to 5 while `system_verified_cdl(`τ`)` stays at **1**, and that the envelope rises to **5** only after the coverage threshold is crossed. The injected edges are synthetic and are used only to demonstrate the anti-boasting semantics of the capability envelope; **they are not evidence that the deployed system attains CDL-5**, and the experiment validates the **envelope rule**, not a deployed CDL-5 capability claim. *Arm table and one stated discrepancy* in Appendix B.4. Read this as a **definition check / worked example** under the envelope formula (one edge gives $1/|\text{edges}| < \tau$), **not** as a refutable empirical finding.

4.3 WHY STRATIFICATION IS NOT DECORATION

The stratification does two jobs that a boolean cannot. Both jobs are exercised over a real engine (contribution 6, Section 6) rather than only in prose.

5 AUTHORITY CALCULUS AND THE REVERSIBLE WITNESS DAG

Authority is computed along a single pipeline with **no re-derivation at the core**:

```
AuthenticatedReceipts -> G1b-P1 (predicate state) -> G1b-P2 (transition legality) -> CAL ledger
```

`G1b-P1` evaluates g(Evidence, Provenance, Receipts) $\rightarrow$ (P1..P5) in pure Rust; the transition core (`G1b-P2`) consumes the already-derived predicate state and never recomputes evidence roots, audit,

or do(X); every transition is appended to a reversible witness DAG with exact per-boundary receipt coverage. Two distinct layers must not be conflated: the **framework layer is a specification, not an implementation** — the audited companion framework tree contains no implementation (its own README is 854 bytes and self-declares `SOURCE_VIEW_ONLY` / `NOT_CANONICAL_AUTHORITY`) — while executable authority semantics exist only in the companion runtime (a Python reference model) and the Rust transition core. The organisation-level load-bearing contract applies to this section and is carried verbatim: *"Layer B — review-only and non-runtime in this paper"*, with `Production Rust governance runtime` on the forbidden list. That ledger is absent from the paper-cited companion tree and was opened first-hand only in the audit tree.

Five invariants are named and instantiated in the Rust transition core: an adversarial test file and a transition-adversarial test file carrying 24 and **17** `#[test]` attributes over 24 `fn` definitions (a prior internal count of 22 was not reproducible), including `derived_cal_strict_accumulation`, `exact_receipt_multiset`, `forged_snapshot_rejected`, `partial_and_full_resolution`, `multilevel_demotion_and_cal5_cert`, bound to the invariant corpus rows 8-13, with a cross-language vector comparison in the conformance script. **Executed in the completion revision**: `cargo test --offline -- --test-threads=1` in the Rust transition core ran the corpus to completion — the adversarial file **24 passed / 0 failed** and the transition-adversarial file **17 passed / 0 failed** (**41** total, exit 0, 2.7 s wall, 2026-09-18; log on file) — and the conformance script reported 16 `[PASS]` checks (15 cross-language vectors plus the G1a-T promote witness), `ALL PASS`, exit 0. The invariants are therefore **specified, adversarially instantiated, and executed**.

- `CAL_DERIVED_NOT_CALLER` (CKB-INV-008; counterexample `WITNESS_SELF_DECLARED_CAL`) — a witness cannot self-declare its own CAL; a caller-supplied `source_grounded` flag cannot substitute for predicate P1.
- `TRANSITION_RECEIPTS_EXACT_COVER` (CKB-INV-009; counterexample `UNRECEIPTED_CROSS_LEVEL`) — a cross-level transition requires an exact multi-set of receipts covering every crossed boundary.
- `SNAPSHOT_CONCURRENCY_CONTENT_BOUND` (CKB-INV-011; counterexample `FORGED_SNAPSHOT`) — a forged or self-reported snapshot cannot carry authority.
- `RESOLUTION_EATS_ALL_HEADS` (CKB-INV-012; counterexample `PARTIAL_RESOLUTION`) — a partial resolution cannot consume only some heads.
- `CAL5_LOCKED_BEHIND_CAL4` (CKB-INV-013; counterexample `CERT_AUDIT_BYPASS_P4`) — strict accumulation locks CAL-5 behind CAL-4.

Reversibility. Because every transition is appended rather than overwritten, authority can be **revoked**, **demoted**, and **fork-merged** over a *ledger-minted, content-bound* snapshot. Precisely: reversal appends a new `Direction::Demotion` witness and nothing is deleted or mutated; the ledger is *supersedable and auditable*, not state-restoring (Appendix B.4).

Boundary (binding). The witness DAG is not a production governance runtime. The specification and bounded reference artifacts do **not** establish automatic promotion, demotion, or fork resolution for arbitrary deployed stores.

6 SUBSTRATE

6.1 WHAT IS IMPLEMENTED

The substrate is implemented and audited. The Rust workspace (`ckb_engine` v0.1.0, with a 1001-line engine module) carries the engine, the Python reference models carry conformance, and a fresh-process CLI returned `version 0.1.0` with exit 0.

6.2 EXACT INDEXED RECALL (E3)

On the frozen snapshot-bound workload (N=2000 candidates, K=100, `ckb-engine` redb — a **synthetic** fixture of 2000 generated incoming evidence edges on one subject, not a document corpus), exact indexed recall reproduces the full-scan result at a 26.3x speedup with `exact_`

`match = true` and a 28.9% scan ratio (579 of 2000 eligibility items read, 558 exact-scored, 21 rejected on eligibility without scoring), on a frozen receipt: `T_build = 117.362 ms`, `T_reference = 143.536 ms/query`, `T_indexed = 5.453 ms/query`, `break_even_queries = 0.85`. The 28.1x fresh re-run and decomposition are withdrawn: the fresh half was never located as a receipt, and is neither frozen nor verified. Non-claims: this machine and workload only, not a universal speedup, not an O(K) bound, not a pruning victory; worst case can still scan $\sim$100%, and the guarantee is exact plus potentially early-terminating.

6.3 Engine conformance

The experiment ledger's per-surface results — Rust semantic suite (E7) 7/7 + 1/1; authority reference model (RUNTIME) 35/35 by a standalone checker (not pytest: `pytest --collect-only` exits 5); epistemic reference model (G3) all PASS; temporal minimal repair (E8) 50/1, 13/13, 13/13, 260/260 as reviewer receipts without confidence intervals; the typed-runtime candidate at 31/0 and engine 48/1/0 with a fresh-process CLI PASS, staying report-only with status `payload_drifted` (the receipt verifies; its payload no longer matches the tree — superseding the earlier "receipt not reproduced" reading); and the R4E suite at 287 passed / 0 failed, exit 0 (2026-09-18) on a byte-copy plus a disclosed taxonomy reconstruction, establishing executability under that fixture, not byte-identity — are tabulated in Appendix B.0.

The VERIFIED label above is the ledger's source-level status, not an execution claim: only the 287-test run and the section 5 corpus (41 passed / 0 failed plus 16 PASS checks, exit 0) were independently logged; the 7/7 + 1/1, 35/35 and G3 lines are reported from source. The temporal numbers are frozen reviewer receipts, with no merge, canonical, promotion or C-evaluation authority.

7 Experiments

7.1 Structural certificate gap: sufficiency is not structural certification (E5, E1, E2)

Set relation (limitation 14; full disclosure in B.1): DEV100 (100 queries) is a subset of the LongMemEval-S 500; fresh20 copies 20 of them; the E9 cohort is 80 = DEV100 minus the 20 r4d queries, homogeneous on frozen metadata, with an empty non-causal control. The memory benefit holds only on this pre-selected causal/temporal subset (100/100 corpus_id intersection).

On rebuilt DEV100 the gold-blind exact-match coverage is 100/100 and the 4096-token bundle reaches 100/100 retrieval sufficiency, yet the structural certificate accepts only 28/100, overlapping the frozen historical certificate 20/20 (10/20 structural). fresh20 repeats it (gold 20/20, certification 10/20); the certificate is run as-is, never pushed to 100/100. That gap is the finding, not a broken instrument: where a query genuinely lacks a causal-antecedent card or a resolvable epistemic state, rejection is the correct outcome (B.1). The recovery shows input-package custody, not graph recovery or certificate-graph replication, with zero answer or judge calls.

Counterfactuals are registered as instances, not interventions (11,831 registered edge-query instances over 20 queries). The no-effect count is reported as-is, not as evidence those edges are causally unnecessary; answer_effect / evidence_effect / judge_effect remain pending a model rerun; the ledger decomposes arithmetically as 14 necessary + 21 redundant-alternative + **243** causal `NO_EFFECT` + **10,050** epistemic `NO_EFFECT` + **1,503** not-evaluable = **11,831** (Appendix B.1.1).

7.2 Downstream effect: external memory supported; incremental explicit structure NOT ESTABLISHED (E9)

A narrow downstream experiment (fixed `gpt-5.6-sol` through the same Codex harness for every arm, one deterministic generation per query, graded by one fixed deterministic mechanical scorer applied blind and uniformly, zero judge calls) asks whether explicit causal/temporal structure raises accuracy over the same evidence presented flat. Four arms: **A** (typed structure), **B** (same evidence, flat), **C** (no memory), and **D** (the matched-token placebo: B's evidence verbatim plus content-neutral filler padded to A's length, added later under its own pre-registration). primary cohort 80 queries, corrective set 20, both subsets of the same 500 public development set as above (limitation 14).

Preregistered: exact paired binary test, alpha = 0.05, delta_min = 0.1 (10 pp), inference unit the 80 paired queries; only A-vs-B and A-vs-C primary.

- A vs C: 80 paired units (54 concordant), 26 / 0 discordant, +32.5 pp, p (exact, two-sided) = 3.0e-8, primary (preregistered), SUPPORTED.
- B vs C: 80 (55), 25 / 0, +31.25 pp, p = 6.0e-8, secondary (post-hoc, not preregistered), SUPPORTED (direction only).
- A vs B (incremental explicit structure): 80 (77), 2 / 1, +1.25 pp, p = 1.0, primary (preregistered), NOT ESTABLISHED.
- Corrective replication A vs C and B vs C: 20 (11), 9 / 0, +45.00 pp, p = 0.0039, secondary (post-hoc corrective; not_preregistered_independent_confirmation), repeats the primary (direction only).

Corrective arm accuracies: A 0.50 / B 0.50 / C 0.05 (A-vs-B 0 / 0 discordant); five p-values uncorrected.

A/B evidence parity was the binding construction gate; the total-token gate was withdrawn before any model output, so arm D re-tests token length post hoc. A matched-token placebo arm (D) was therefore added later, under its own pre-registration written before any of its model outputs: D reproduces B's evidence verbatim and pads it with content-neutral filler to A's length (80 queries, same model and harness, deterministic scorer, zero judge calls). At matched tokens the structure effect does **not** reappear: A 0.3250 vs D 0.3250 (Δ = 0.0; 2/2 discordant; exact McNemar p = 1.0), and the filler itself is inert (D vs B Δ = +0.0125; 2/1 discordant; p = 1.0). The token mismatch is therefore excluded as the explanation of this null; the null remains a low-powered no-detection and is not an equivalence claim (no equivalence margin is powered here).

Budget: 360 execution calls + 60 corrective = 420 (the 300 unique-analysis / 60 duplicate / 60 bug-invalidated counts are three views of the same 360, not additive); reader fixed gpt-5.6-sol, temperature not guaranteed 0, so arm accuracies are **single-draw** estimates (no repeats, no CI).

Two conclusions: either form of external memory substantially improved performance over none (the defensible claim); the incremental benefit of typed causal/temporal presentation over the same evidence flattened was NOT ESTABLISHED (the honest negative). Primary cohort (N=80) accuracies: A 0.325, B 0.3125, C 0.0.

Boundary (binding): one presentation treatment; the corrective run is not independent replication, and the primary A-vs-B (2 / 1) is a low-powered no-detection; the corrective A-vs-B (0 / 0) carries no power either way, not an equivalence proof. We report this null as a first-class result: it was produced by a preregistered gate under hard A/B evidence parity, it survives a matched-token placebo control (arm D: A vs D p = 1.0, Δ = 0.0), and publishing it is part of what keeps the positive claims honest.

7.3 LongMemEval-S (public development set): evidence-adjudicated correctness

Seven qualifiers hold at every use: (1) post-hoc hybrid evidence-adjudicated; (2) public development set, explicitly not hidden-set generalization; (3) the 493 and the 7 are measured on the frozen pre-adjudication predecessor candidate, the accepted successor having no official score; (4) the two judge outputs are not combined into a consensus label; (5) not an official benchmark score; (6) third-party replay still pending; (7) **no same-question control arm**, so the figure is **not subtractable** from any system's accuracy and supports no treatment-effect claim. Never interchangeable: 500/500 (hybrid evidence-adjudicated), 493/500 (frozen gpt-4o, predecessor), 500/500 (Luna snapshot, same candidate), 445/500 (M5 RUN2, different candidate, 55 wrong rows), 449/500 (M6-SOL, receipt-only, 51 wrong rows); the 55 and the 51 are different candidates' wrong-row counts intersecting in only 38 rows.

7.3.1 Protocol. LongMemEval-S, the 500-question public development set (500 unique question_id, duplicate check in B.1), three-layer adjudication (evidence, contract, gold as comparator only).

7.3.2 Frozen evaluator. The frozen gpt-4o-2024-08-06 evaluator (temperature 0, max_tokens 10) scored 493/500 (98.6%) on the frozen pre-adjudication predecessor candidate: 7 wrong rows out of 500, 98.6% belonging only to that score and candidate.

7.3.3 Dual-judge snapshot (not consensus). On the same frozen candidate Luna (gpt-5.6-luna; 500 rows / 500 singleton batches, true 500 / false 0) scores 500/500 against the frozen gpt-4o evaluator's 493/500; no consensus quantity exists, and the 7 disagreements are all one-directional (luna_yes_gpt4o_no = 7, luna_no_gpt4o_yes = 0).

7.3.4 Successor unscored. The accepted successor's eval-results file is 0 bytes and successor_official_score_available: false; no statement in this paper attaches 493/500 to the accepted successor.

7.3.5 Adjudicated outcome. Post-hoc hybrid evidence-adjudicated **500/500** = **493** frozen-gpt-4o-2024-08-06 accepts + **7** source-level adjudications (493 + 7 = 500) on that predecessor; the seven: five judge disagreements, one source/gold conflict, one synthesis bug repaired. The seven are author-assigned under a rule written after the judge's outputs were seen (rule_defined_posthoc: true, fully_manual_blind_review: false), being exactly the difference between 493 and 500; the evaluator emits only a boolean, so the withdrawn "∼1% (5/500)" becomes five of seven adjudicated rejections (5/7), never proof a judge was wrong.

7.3.6 Lineage (internal, not results). M6's taxonomy gives evidence_loss 24 / execution_failure 19 / disputed_gold_or_judge 8 over 51 of the 500 public questions; the 55-row and 51-row sets union to 68 rows (13 matrix-only, 17 M5-only, intersection 38); M6's raw output is absent (receipt-only, UNREAD); the scored predecessor and successor have no generator field (answer model UNKNOWN).

7.3.7 Error classes. F1: an "evidence-retrieved-but-dropped-in-synthesis" class was repaired by a fail-closed evidence-anchor gate; F2: five of seven adjudicated rejections were false negatives.

7.3.8 Tracks. Track 2 (production) needs a frozen gold-blind run plus a third-party replay receipt that does not exist; Track 1 (development) is the hybrid 500/500, and the paper keeps the two tracks separate because third-party evidence is still pending.

7.3.9 semantic_sufficiency_automatically_proven: false, and no structural-certificate claim follows.

8 Discussion and Limitations

8.1 The three-layer distinction

These results motivate a distinction between retrieving evidence, maintaining an epistemic state, and preserving evidence through answer generation: failure at one layer cannot be repaired by raising performance at another, the E2AP error class being the instance. This interpretation, not a measured three-way ablation, joins the two cited findings with our own inference.

8.2 Limitations (binding, each with its evidence status)

1. CDL/CAL are first-cut graph heuristics, not a validated per-edge depth estimator; CCE arms are synthetic.
2. The taxonomy is a governance framework, not a law: CDL does not claim strictly nested causal knowledge.
3. The governance layer (Layer B; Layer A is the substrate, Section 6) is not a production runtime: no automatic promotion, demotion or fork resolution.
4. Internal evidence is bounded: fresh20 is 20 frozen queries, not an accuracy result; the 11,831 counterfactual instances test graph computation, not interventions.
5. The CCE demonstration is synthetic (a_min = CAL-1, tau = 0.30) and validates no general edge classifier.
6. Model-mediated steps are not byte-deterministic; reproducibility requires raw outputs and deterministic replay.
7. The external-treatment result is an observability result, not an official score: no hidden-set generalization, no judge consensus, no fully blind manual review, and no same-question control arm; third-party replay remains pending.

8. MuSiQue (conventional-retrieval judge) did not discriminate: the governed arm tied its BM25-only arm exactly on supporting-evidence recovery — a sanity result, not a missed advantage: governance did not degrade conventional retrieval, and none was claimed.
9. The temporal surface (E8) has zero CI check runs, so those frozen receipts are not confidence intervals.
10. Documentation trails the implementation (a receipt-hash mismatch and a stale test-count self-report); details in Appendix B.2.
11. Two internal discrepancies previously logged here are reconciled by arithmetic, printed rather than left to the reader: **243 + 10,050 + 1,503 = 11,796**, and **0.50 - 0.05 = +45.00 pp** (the §7.2 corrective triple at N=20, the r4d subset, against the primary N=80 C-arm **0.0**).
12. Novelty coverage is bounded: no systematic literature search, abstracts rather than full texts for adjacent lines (notably Zep/Graphiti, UNREAD).
13. The workspace inventory is an artifact-hygiene audit, not a research limitation; details in Appendix B.2.
14. No held-out surface: DEV100 is a subset of the 500, fresh20 a copy of 20 inside DEV100, the E9 cohort DEV100 minus r4d(20).
15. LME500 correctness has an unknown answer generator; E9's 420 calls exceed its post-hoc-corrected analysis surface, with seven p-values uncorrected (five campaign + two arm-D).
16. Executed but not third-party: 41 #[test] attributes (24 + 17) and the conformance script - 41 passed / 0 failed (2.7 s), 16 [PASS] checks, 2026-09-18, on the author's machine only.

8.3 What evidence would change these conclusions

- Answer correctness: measuring answer_effect / evidence_effect / judge_effect and showing structure contributes causally to correct answers would make the claim an efficacy claim.
- A judge that scores our surface: governed memory against a retrieval baseline on a benchmark judging temporal state reconstruction or causal semantics.
- Runtime closure: wiring the authority oracle into the production engine rather than a Python reference model would make Layer B runtime-integrated.
- Per-edge depth validation: an empirically validated CDL/CAL depth estimator would remove the largest internal caveat on CCE.
- Third-party replay: an independent replay receipt would close the third-party-evidence gap.
- A hidden-set result: a held-out, non-public split rather than the 500 public-development questions would license the generalization language avoided here.

8.4 Future work (explicitly not claimed)

Retraction propagation is an open problem: the layout is specified and receipted, but no live propagation experiment is claimed; Milestone B governed retrieval and AdmissionPolicy key rotation are not started.

9 Conclusion

We presented a causal-memory substrate with a machine-governed causal-authority lifecycle: causal capability should not be a boolean, and stable causal memory is the minimal substrate of agency. CDL/CAL stratification, CCC, CCE and the reversible witness DAG make over-claiming detectable and authority derived and auditable, the calculus executing at the substrate and conformance layer (governance is specified and review-only, not a deployed runtime). The empirical picture is mixed: external memory helps (+32.5 pp and +31.25 pp, $p \sim$ 3e-8 / 6e-8; post-hoc corrective replication +45.0 pp, p = 0.0039), while explicit causal/temporal presentation yields a bounded negative (NOT ESTABLISHED, p = 1.0). On LongMemEval-S the frozen official evaluator scores a frozen candidate at 493/500, an observability result with no baseline and no recorded answer generator (§7.3); the transferable output is a mechanism, not a number.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used generative AI tools in this work for the four purposes selected on the submission form, and no others. (i) To aid or polish writing: AI-assisted language editing was applied to the manuscript. (ii) For retrieval and discovery: we used AI to search for and identify relevant literature; every bibliographic identifier obtained this way was verified against the primary source, and identifiers that could not be verified are marked NOT_FOUND and are not cited. (iii) For research ideation or execution: this is the category the ICLR policy treats as requiring disclosure; our use fell within it. (iv) To draft sections of the paper: AI produced drafts of parts of the manuscript. No other AI-use category was selected on the submission form, and none is disclosed here. We have reviewed all AI-assisted output before inclusion: every quantitative claim is bound to an audited source anchor, values that could not be traced to one were left as explicit placeholders rather than filled, and unresolved discrepancies are reported as open findings rather than smoothed. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work analyses an existing public benchmark (LongMemEval-S) and does not involve human subjects, personal data collection, or model deployment. The failure taxonomy and evaluation artefacts are reported with their provenance, including the cases in which the original raw evaluator output is unavailable. We see no application of this work that is intended to cause harm; the limitations that bound its empirical claims are stated in §8.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Every quantitative claim in the main text is tied to a named artefact and a content hash; the claim-to-evidence mapping is given in the appendix. Where an artefact is unavailable on this machine – no third-party replay of the reported evaluation, and no raw evaluator output for the M6 lineage – the paper prints the gap rather than substituting a value, and those numbers are marked as receipt-only in §8.2. Provenance of the audit itself: the evaluation and the evidence audit were all produced by the sole author within a single repository; no independent researcher ran, replicated or externally audited any reported result, and the *reviewer receipts* cited in §7.3 and §8.2 are the author's own frozen self-audit logs, not external peer review. This is the sense in which third-party verification remains pending. The frozen stores, evaluation logs, and analysis scripts that produced the reported counts were not included with this submission and will be released upon publication.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, June 1985. doi: 10.2307/2274239. URL https://doi.org/10.2307/2274239.

Centre for Evidence-Based Medicine, University of Oxford. OCEBM levels of evidence. Web resource. URL https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence. Live page returned HTTP 403 (CloudFront) on 2026-09-18; framing content observed via Internet Archive snapshot id_20240224195759. Version year and working-group member list NOT_VERIFIED in the fetched text.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A

graph RAG approach to query-focused summarization, 2024. URL https://arxiv.org/abs/2404.16130.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Main Conference Track*, 2024. doi: 10.52202/079017-1902. URL https://papers.nips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV ontology. W3C Recommendation, April 2013. URL https://www.w3.org/TR/2013/REC-prov-o-20130430/. W3C Recommendation 30 April 2013.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020. doi: 10.48550/arXiv.2005.11401. URL https://arxiv.org/abs/2005.11401.

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

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019. doi: 10.1145/3241036. URL https://doi.org/10.1145/3241036.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

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 (ICLR 2025)*, 2025. doi: 10.48550/arXiv.2410.10813. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

NOTE ON NUMBERING

The appendix reproduces, verbatim, the source-anchor layer, the table inventory, and the claim-to-evidence mapping of the working manuscript. In the working manuscript those blocks are labelled §10–§13; here they are Appendices A–D. Table labels (`T1`, `T2`, `T3`) and claim ids (`C-xx`) are unchanged. Nothing in the appendix is evidence that the main text does not already bound.

A. Appendix A: Evidence inventory (figures and table inventory)

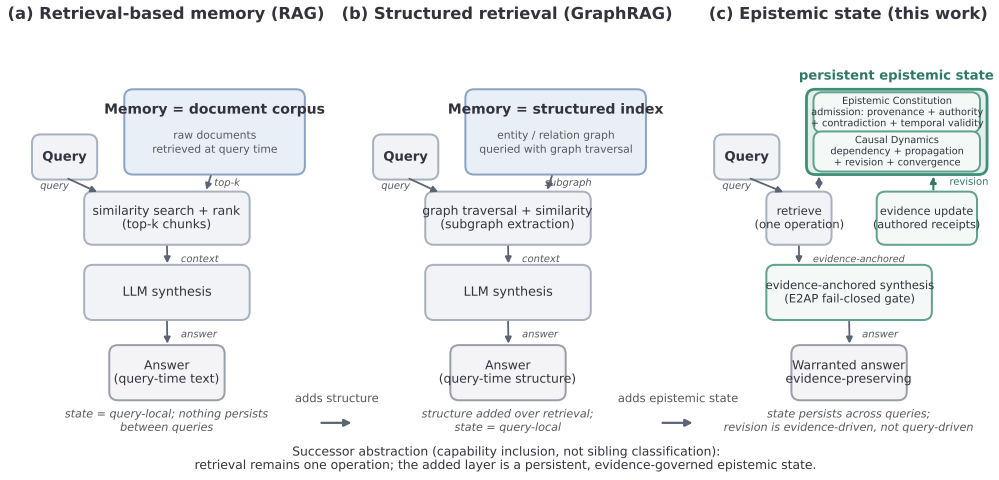


FIG1 — From retrieval to epistemic state . **(a)** Retrieval-based memory (RAG): query-time ranking over a document corpus; the state is query-local. **(b)** Structured retrieval (GraphRAG): retrieval over an entity/relation index; structure is added *over retrieval* and the state remains query-local. **(c)** The proposed epistemic state: provenance-grounded admission (Epistemic Constitution), dependency-driven revision (Causal Dynamics), and evidence-preserving synthesis (E2AP); retrieval persists as **one operation** over a persistent state. The progression is a **capability-inclusion**

relation (the added layer is a persistent, evidence-governed epistemic state), **not a sibling taxonomy**.

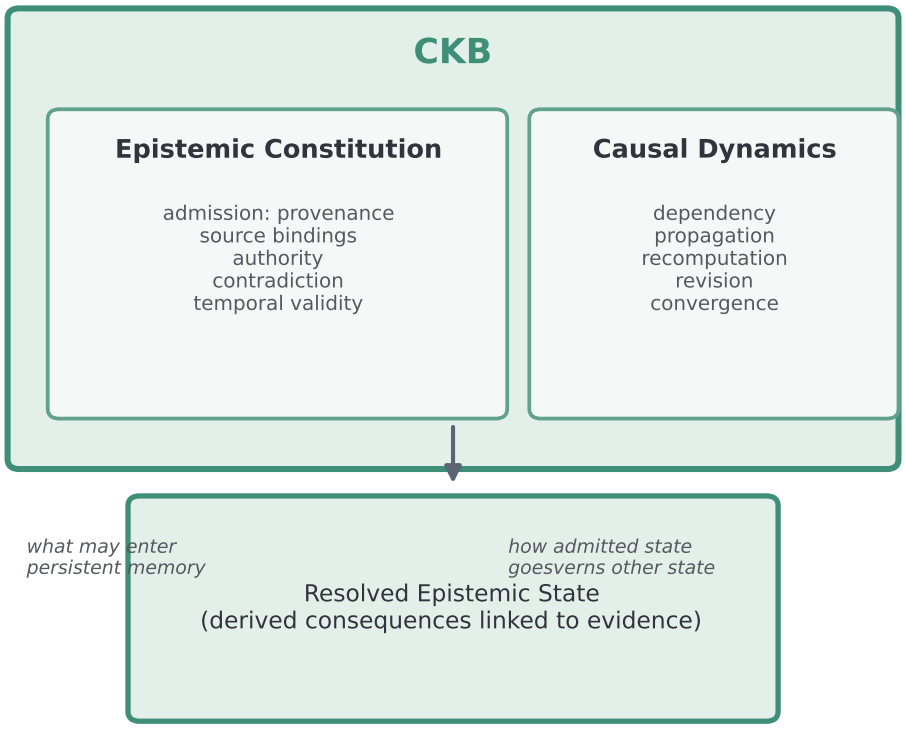


FIG2 — CKB two-layer architecture . CKB separates **Epistemic Constitution** (provenance, authority, contradiction, temporal validity, admission) from **Causal Dynamics** (dependency propagation, recomputation, revision, convergence), converging on a *Resolved Epistemic State*. The figure carries **no retrieval box**: retrieval is an operation inside CKB, not a layer (`PROSE_ARCHITECTURE_LAYERS.tex:6-24`). The separation is the mechanism that prevents the causal graph from becoming its own source of authority.

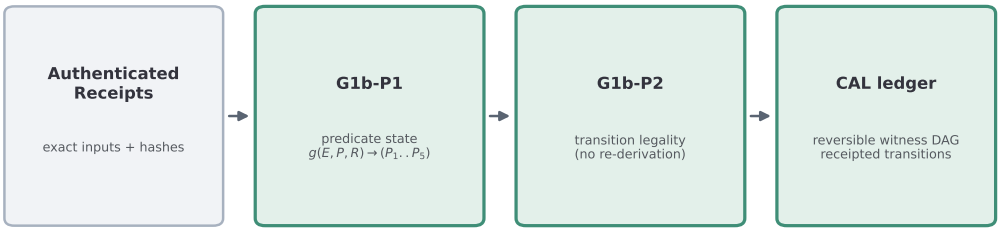


FIG3 — Authority semantics pipeline . `AuthenticatedReceipts` → `G1b-P1 (predicate state)` → `G1b-P2 (transition legality)` → `CAL ledger` → `reversible witness DAG`. The pipeline is drawn as **single-pass and re-derivation-free**: the transition core consumes an already-derived predicate state and never recomputes evidence roots, audit, or do(X) (`PROSE_p2_p7_v0_1.md:66-74`). Adversarial invariants (`CAL_DERIVED_NOT_CALLER` (CKB-INV-008), `TRANSITION_RECEIPTS_EXACT_COVER` (CKB-INV-009, counterexample `UNRECEIPTED_CROSS_LEVEL`), `SNAPSHOT_CONCURRENCY_CONTENT_BOUND` (CKB-INV-011), `RESOLUTION_EATS_ALL_HEADS` (CKB-INV-012), `CAL5_LOCKED_BEHIND_CAL4` (CKB-INV-013)) are labelled on the transition boundary **as specified**

#	Table	Content	Status of its numbers
T1	Exact indexed recall (E3)	**26.3× frozen** + 28.9% scan + `exact_match = true` (receipt: `T_build 117.362 ms`, `T_reference 143.536 ms/q`, `T_indexed 5.453 ms/q`, `break_even 0.85`); `28.1×` fresh and ARMS ≈81× / ≈0.3× are withdrawn to REPORT_ONLY	**FROZEN** for the frozen half (`PERF_RECEIPT.txt:17-33`, `TABLES.tex:4-20`); the fresh half is **NOT_FOUND** as a receipt (printed only in prose `EXPERIMENTS.md:53-71` and in the hash-less rows `TABLES_P6_P8_v0_1.md:11-14`, whose Source row `:15` still reads `[VERIFY]`)
T2	Structural certificate vs answer coverage (E5/E1)	DEV100 100/100 coverage, 100/100 bundle-sufficient, **28/100** certificate (49/86/58/100); fresh20 20/20 and **10/20**; historical overlap 20/20; `DEV100` ⊆ the same 500-question public set (§7.1)	FROZEN (`TABLES.tex:22-40`, `EXPERIMENTS.md:95-106`)
T3	Registered counterfactual audit (E2)	11,831 instances over 20 queries; 14 necessary / 21 redundant-alternative / 11,796 irrelevant causal; 10,050 no-effect; 1,503 not-evaluable epistemic	FROZEN (`TABLES.tex:42-59`, `EXPERIMENTS.md:27-44`); the §7.1 row-vs-aggregate reading is reconciled by arithmetic (`14+21+243+10,050+1,503 = 11,831`; `243+10,050+1,503 = 11,796`); the earlier `EVIDENCE_REGISTRY.json` anchor is **NOT_FOUND** (v0.2 row C-12)
T4	Conformance summary (E7/E8/Milestone A)	7/7 + 1/1; **35/35** (standalone `python test_properties.py`); 50/1 + 13/13 + 260/260 (**reviewer receipts, zero CI**); **Milestone A 31/0 + engine 48 (1 ignored)** on-disk receipt; R4E **287 collected → 287 passed / 0 failed, exit 0**	FROZEN/VERIFIED; **Milestone A row = REPORT_ONLY / payload_drifted**; E8 row = **FROZEN reviewer receipts, zero CI check runs** (`EXPERIMENTS.md:141-153`; `TABLES.tex:61-75`); its status is *receipt-backed but not CI-verified*, which is a different thing from the Milestone A candidate row above

semantics: the transition core rejects these counterexamples in the specification and in the Rust test corpus, which are now **executed and passing** (41/41 Rust tests, 16/16 conformance checks; §5). The figure therefore illustrates **rejection semantics, not test counts**.

The table inventory of the working manuscript, verbatim:

#	Table	Content	Status of its numbers
T5	Causal-memory effectiveness + null result (E9)	A vs C **26/0 of 80** (54 concordant) +32.5 pp, p=3.0e-8 — **primary**; B vs C 25/0 of 80 (55) +31.25 pp, p=6.0e-8 — secondary; A vs B 2/1 of 80 (77), p=1.0 **NOT ESTABLISHED** — primary; corrective 9/0 of 20 (11) +45.0 pp, p=0.0039 — secondary; **matched-token placebo A vs D 2/2 of 80 (Δ = 0.0), p=1.0 — primary *within arm D's own later preregistration*, secondary for this table's original family**; D vs B 2/1 of 80 (Δ = +0.0125), p=1.0 — secondary; **no multiple-comparison correction**	FROZEN (`TABLES.tex:77-93`, `EFFECT_RECEIPT.md:21-43`); the four E9-arm labels are per `PREREGISTRATION_FREEZE.json:29-37`, while the two arm-D labels are per `PREREGISTRATION_PLACEBO_ARM_D.md:33-34`, a **separate later** preregistration written before arm D's outputs — the original freeze did **not** contain arm D
T6	LongMemEval-S public-development outcome	hybrid **500/500** = 493 (frozen gpt-4o, predecessor) + 7 adjudications (5 judge / 1 benchmark source / 1 real bug); predecessor snapshot 493/500 (98.6%, the 493 rows **not** manually re-audited); Luna snapshot 500/500 (comparability only, **no baseline**, not subtractable); successor official score **not available**; answer generator of the scored candidate **UNKNOWN**	receipt-level per `A4...:27-32`, `:129-140`; all rows nested in the same public set (**no held-out surface**, §7.1); third-party replay **PACK_PENDING**
T7	Candidate lineage (internal only, not results)	M5 RUN2 445/500 (`deepseek-v4-flash`) with 55 wrong rows; M6-SOL 449/500 (`gpt-5.6-sol`) receipt-only with 51 wrong rows; taxonomy 24/19/8 over 51; set algebra 38/13/17/68	receipt-only / **UNREAD** raw output (`A4...:25-26`, `:33-37`, `:64`) — **must never be tabled with T6**

B. Appendix B: Extended experimental detail

B.0 Engine and reference-model conformance

`VERIFIED` in this table is the ledger's source-level status, not an execution claim. **The two surfaces whose execution was independently logged in the completion revision are the R4E suite** (287 passed / 0 failed, exit 0) and the §5 `G1B_P` corpus (**41 passed / 0 failed** plus 16 `[PASS]` **conformance checks, exit 0). The other quantitative rows —** 7/7 + 1/1**,** 35/35 and the G3 all-cases-PASS line — remain `OBSERVED_SOURCE`: they are reproduced as the repository states them (with its standalone-runner caveat), and no run log, runner or exit code was located for them in this work. They are therefore **not** upgraded to `OBSERVED_RUNTIME`, and the difference between "the repository reports this" and "this was run here" is preserved row by row.

Evidence surface	Result	Status	Anchor
Rust semantic suite (E7): `causal_probe`, `causal_benchmark_300`	**7/7 + 1/1**	VERIFIED	`EXPERIMENTS.md:125-130`
Authority reference model (`RUNTIME`) property tests	**35/35** (executed with `python test_properties.py`, a standalone checker — **not** a pytest suite; `pytest --collect-only` on that tree returns exit 5 / no tests collected)	VERIFIED	`EXPERIMENTS.md:131-134`; `A2...:139-141`
Epistemic reference model (G3)	all cases PASS (same standalone-runner caveat; the repository's stale self-reported count is **not** used here)	VERIFIED	`EXPERIMENTS.md:135-137`; `A2...:162`
Temporal minimal repair (E8)	**50 pass / 1 ignored**; temporal **13/13**; fresh A/B **13/13**; stress **260/260**	reviewer receipts, **zero CI**	`EXPERIMENTS.md:141-153`
Milestone A typed runtime	**31 passed / 0 failed**; engine **48 passed / 1 ignored / 0 failed**; fresh-process CLI rebuild/query PASS	**REPORT_ONLY / payload_drifted** (`CANDIDATE_NOT_CANONICAL`)	`[repo]/CKB_TYPED_RUNTIME\CANDIDATE_RECEIPT.json:5-6, :37-38`
R4E 287-test suite	`287 tests collected in 1.93s`, then **executed: 287 passed / 0 failed, exit 0** (2026-09-18). The run used a byte-copy of the frozen R4E snapshot **plus the disclosed M6-taxonomy reconstruction**, so it establishes the suite's executability under that fixture — not byte-identity with an original that is absent everywhere on this machine (the frozen snapshot's own state was 285 passed / 2 `FileNotFoundError`)	**OBSERVED_RUNTIME**	`A2...:137, :272`; log `pytest_r4e_full_after_fixture2.log` (sha256 `446AF615...6927`)

The temporal row must be read with its provenance wall. Those numbers were read from the frozen `TEMPORAL_EVIDENCE_FREEZE.md` receipt, **not** independently re-run in this work; GitHub reports **zero CI check runs** for them, so they are frozen **reviewer receipts, not native approvals**, and carry **no merge, canonical, promotion, or C-evaluation authority** [C-14] (`EXPERIMENTS.md:147-153`; restated as a binding non-claim in `DISCUSSION.md:57-59`). This paper never presents a reviewer receipt as a passing CI run.

The Milestone A row is a candidate whose printed identity does not reproduce on this machine. The only candidate receipt on disk states `payload_file_count: 21`, `payload_tree_sha256: 4f4d50437da5583370313859fabe3462399d30b7ce2d03d1a9f8236d518233e9`, `typed_runtime_cargo_test_locked = ''31 passed; 0 failed''`, `ckb_engine_full_regression = ''48 passed; 1 ignored; 0 failed''` and `status: CANDIDATE_NOT_CANONICAL`; the receipt file itself hashes to `47EE955A267A58CBA817AC77DE218437D1C175A804B5E4A7FC5AF1A8A6DC2650`. The four surfaces are reconciled into **two artifacts from two different times**, and every discrepancy

is now accounted for. (i) MILESTONE_A_STATUS_LOG.md:5–7 is a 2026-08-30 transcription of an author report, and the verification it claims was never performed on a file: the log's own open item reads "verify receipt + counters", the receipt sha it cites (F44A8031...) matches no file in the tree, the string occurs nowhere in the code audit, and the paper-cited companion tree contains **no** CANDIDATE_RECEIPT.json; its 13 passed / 0 failed and 45 passed / 1 ignored therefore describe that unverified report, not the receipt. (ii) The on-disk CANDIDATE_RECEIPT.json (47EE955A...2650, schema v4, payload_file_count: 21) is **internally consistent**: recomputing its payload_tree_sha256 from its own 21 entries under its declared algorithm reproduces 4f4d5043... exactly (independent recomputation performed for this paper). (iii) What no longer reproduces is the **payload, not the receipt**: 9 of the 21 recorded file hashes still match the working tree and **12 differ**, so the receipt describes a revision that has since drifted — independently confirming the code audit's 12 of 21 (A2...:258–262). This paper therefore prints the **on-disk** values with status **REPORT_ONLY / payload_drifted** (the receipt verifies; its payload no longer matches the tree), and the previously printed set (13/13, 45 (1 ignored), f44a8031..., 16/16, f31a305e...) is **withdrawn as unreproduced here** (the underlying cause — different command set or different revision — is not established). Milestone A **does not upgrade any claim label** in this paper, and its CANDIDATE_NOT_CANONICAL positioning is unchanged (MILESTONE_A_STATUS_LOG.md:46 still lists "verify receipt + counters" as an open box) [C-15]. The independent review for that milestone ended REVIEW_INCOMPLETE after three platform misjudgements and **was not dressed up as a pass**; that record is **REPORT_ONLY** (manual report only, not a receipt field) (MILESTONE_A_STATUS_LOG.md:8, :19; CLAIM_EVIDENCE_LEDGER.md:25, C-16) [C-16].

Cross-language equality is not semantic coverage. A CLOSED Python oracle (G1a-T), a Rust predicate evaluator (G1b-P1), a Rust transition core (G1b-P2) and a cross-language conformance runner (G1b-P3) execute the same canonical op-script and are required to agree exactly on accept/reason/state and on the SHA identity of the resulting ledger. But two implementations may agree on inputs that both accept while the intended trace semantics — a legal fork that must produce a *false* pass, or a snapshot that must be content-bound rather than authority-bound — remain uncovered. The gate is therefore on **semantic coverage**, not merely on equality, and it spans the whole Rust + Python corpus rather than the Python runner alone: the conformance runner contributes the 16 cross-language [PASS] checks, while the two semantics named just above are exercised by the Rust adversarial suite (legal_fork_via_ledger_snapshot, forged_snapshot_rejected) [C-13]: CrossLanguageEquality $\neq$ SemanticCoverage.

B.1 Experiments (extended)

B.1.1 Structural certificate gap: answer/evidence sufficiency $\neq$ structural certification (E5, E1, E2)

Set-relation disclosure (all empirical surfaces in this paper draw on one public question set). DEV100 (100 queries) $\subseteq$ the LongMemEval-S public development set (500 questions, gold sha d6f21ea9d60a0d56f34a05b609c79c88a451d2ae03597821ea3d5a9678c3a442); fresh20 is a **second copy of 20 questions already inside DEV100**; the E9 primary cohort is **80 = DEV100 minus the 20 r4d queries**; and that cohort was selected on frozen task metadata (temporal-reasoning|temporal_version $\cup$ knowledge-update|contradiction_supersede), so it is homogeneous and its NON_CAUSAL_CONTROL subgroup is EMPTY**. The surfaces presented across §7.1–§7.3 are therefore** not independent evidence faces**: they are nested subsets of the same 500 public questions, there is** no held-out surface anywhere in this paper, and the E9 memory benefit holds only for this pre-selected causal/temporal subset. Evidence: ARM_CONSTRUCTION_SPEC.json:31 (homogeneity and empty control), EFFECT_MEASURE_001_FREEZE\QUERY_PRIMARY_DEV100.jsonl (item_id: lme-dev:... / corpus_id), and the intersection of the 100 analysis corpus_id values with the 500 question_id values = **100/100** (OBSERVED_TOOL_RESULT, empirical reviewer). This disclosure is restated at the head of §7.2 and §7.3.

Answer coverage does not entail structural certification. On the rebuilt DEV100 surface, the gold-blind exact-match coverage is **100/100** and the 4096-token bundle achieves **100/100** retrieval sufficiency, yet the structural certificate accepts only **28/100** items (causal-closure containment 49,

temporal eligibility 86, universe scope 58, evidence identity 100), with a per-item overlap match of **20/20** against the frozen historical certificate (10/20 structural) [C-11].

The same pattern holds on the independent fresh20 surface: gold coverage **20/20**, structural certification **10/20** [C-11]. The certificate is run as-is and **is not pushed toward 100/100**: if a query genuinely lacks a causal-antecedent card or a resolvable epistemic state, the correct outcome is to keep rejecting it. Non-claims: the DEV100 recovery establishes **input-package custody** (100 rows, 100 unique `item_id`, 0 duplicates from 40+20+20+20), **not** historical graph recovery, **not** certificate-graph replication, and **zero** answer or judge calls were made on this surface [C-11].

Counterfactuals are registered as instances, not interventions. The frozen audit evaluates **11,831** registered edge-query instances over **20** queries: 14 necessary, 21 redundant-alternative, and 11,796 irrelevant causal instances, plus 10,050 no-effect and 1,503 not-evaluable epistemic instances (2 queries) [C-12]. The no-effect count is reported **as-is** and is **not** evidence that those edges are causally unnecessary: a graph-computation ablation tests structural dependence and cannot alone satisfy intervention authority, and `answer_effect` / `evidence_effect` / `judge_effect` remain `PENDING_MODEL_RE_RUN`. Wording discipline: never write "10,050 edges are causally unnecessary" and never write "35 causal interventions" — only *registered edge-query counterfactual instances*.

A previously logged cross-document discrepancy is **now reconciled**: the two documents report **different rows of one table, not competing readings**. `EXPERIMENTS.md:36` gives `CAUSAL · NO_EFFECT` = **243**, and `EXPERIMENTS.md:37` separately gives `EPISTEMIC · NO_EFFECT` = **10,050**. The flat breakdown in `TABLES_P6_P8_v0_1.md:36-41` is a coarser view of the same ledger: `14 + 21 + 243 + 10,050 + 1,503 = 11,831`, and its `IRRELEVANT` row (**11,796**) decomposes as `11,796 = 243 + 10,050 + 1,503`, so `14 + 21 + 11,796 = 11,831` holds as well. **This paper prints the three named classes (14 / 21 / 11,796) and also the 243 / 10,050 breakdown.** The `11,831` total's printing basis is `TABLES_P6_P8_v0_1.md:36-40`; the E2 table in `PAPER\EXPERIMENTS.md:32-38` carries **no total row**, and the ledger row for this claim pointed at an `EVIDENCE_REGISTRY.json` that is **NOT_FOUND** in both trees (v0.2 row C-12). Evidence: OBSERVED_DIFF; NOT_FOUND (registry).

B.1.2 Downstream effect: external memory supported; incremental explicit structure NOT ESTABLISHED (E9)

A narrow downstream experiment (fixed `gpt-5.6-sol` through the Codex agentic harness, deterministic mechanical scorer, blind scoring before unblinding) asks one question: does showing the causal/temporal structure explicitly to a downstream model increase answer accuracy over the same evidence presented flat? Three arms — A (same evidence as typed causal/temporal structure), B (same evidence, flat), C (no memory) — on a primary 80-query cohort and a 20-query corrective set [C-17][C-18]. **Set relation (restated here as promised in §7.1):** the cohort is an 80-query subset of the same 500-question public development set used in §7.3 (DEV100 minus the 20 r4d queries), selected on frozen causal/temporal metadata, `NON_CAUSAL_CONTROL` EMPTY. **Test identity and inference unit:** the preregistered test is an exact paired binary (McNemar-style) test at `alpha = 0.05` with `delta_min = 0.1` (10 pp), and the **inference unit is the 80 paired queries**, not the discordant pairs; each p-value is therefore printed with both its discordant counts and its concordant count, because the `n` values previously printed here were discordant counts. Only **A vs B** and **A vs C** were preregistered as primary; every other contrast is **secondary (post-hoc, not preregistered)** and carries directional weight only.

Corrective arm accuracies, printed once with their receipt: **A 0.50 / B 0.50 / C 0.05**, with the corrective A-vs-B contrast **0 / 0 discordant**.

Multiple-comparison declaration (mandatory). Seven p-values are reported — the **five** of the original campaign (A vs B, A vs C, B vs C $\times 2$, corrective pair) plus the **two** arm-D contrasts added later — with **no multiple-comparison correction** (no family-wise definition, no Holm, no Bonferroni). The two preregistered primary tests are A vs B and A vs C; all others, including both arm-D contrasts, are secondary and directional only. A-vs-D is primary **only inside arm D's own later preregistration** (`PREREGISTRATION_PLACEBO_ARM_D.md:33`). The observed A-vs-B difference of **+1.25 pp** is an order of magnitude below the preregistered `delta_min = 0.1`.

Comparison	Paired units (concordant of them)	Discordant pairs (A / other)	Effect	p (exact, two-sided)	Label	Verdict
A vs C (external memory vs none)	80 (54)	26 / 0	**+32.5 pp**	**3.0e-8**	**primary (preregistered)**	**SUPPORTED**
B vs C (flat memory vs none)	80 (55)	25 / 0	**+31.25 pp**	**6.0e-8**	secondary (post-hoc, not preregistered)	SUPPORTED (direction only)
A vs B (incremental explicit structure)	80 (77)	2 / 1	+1.25 pp	**1.0**	**primary (preregistered)**	**NOT ESTABLISHED**
Corrective replication, A vs C	20 (11)	9 / 0	+45.00 pp	0.0039	secondary (post-hoc corrective; `not_preregistered_independent_confirmation`)	repeats the primary (direction only)
Corrective replication, B vs C	20 (11)	9 / 0	+45.00 pp	0.0039	secondary (post-hoc corrective)	repeats the primary (direction only)

A/B parity: what was gated, and what was removed before execution. Evidence-*content* parity between A and B was and remains a **hard gate**: both arms carry the same evidence. The preregistered **total-token** parity gate ($\pm$`5%` of A's ground-truth token count, with the rule that "a token-budget failure invalidates the A-vs-B comparison", `ARM_CONSTRUCTION_SPEC.json:23-25`) was **removed by amendment before any model output was produced**, because the pilot showed parity and the structure treatment to be mechanically in conflict (A $\approx$ 33k characters vs B $\approx$ 14k: $\approx 2.3\times$ when fully dumped, $1.39\times$ in compact form). The amendment re-classified structure-token overhead as **treatment metadata — measured and reported per query, not a gate** — explicitly forbade the claim that "causal semantics itself, excluding all extra-token effects, causes the improvement", and added **no** matched-token placebo arm *at that time*. A whole-file size ratio on disk was measured earlier (2,639,603 vs 2,022,017 bytes, $\approx$ **1.31**$\times$) but is **superseded** and is not used here, because it pools invalidated rows; the executed per-query measurement appears in the next paragraph. That gap was closed afterwards. A matched-token placebo arm (D) was added under its own pre-registration written before any of its model outputs: D reproduces B's evidence verbatim and pads it with content-neutral filler to A's length, over the same 80-query cohort, model, harness and deterministic scorer. At matched tokens the structure effect does **not** reappear — A 0.3250 vs D 0.3250 (Δ = 0.0; 2/2 discordant; exact McNemar p = 1.0) — and the filler itself is inert (D vs B Δ = +0.0125; 2/1 discordant; p = 1.0). Consequences, stated rather than implied: A-vs-B `p = 1.0` is **not** a causal statement that causal structure has no effect independent of extra tokens *in general*, but the token mismatch is now excluded as the explanation of this null, and the null still cannot be used to argue where the value of structure *does* lie.

Per-query structure-token overhead, as the amendment requires. The amendment's requirement is per-query measurement, not a file-size proxy. On the valid 100-query analysis surface (`kind=DEV` in `A_CONTEXT_FREEZE.jsonl` / `B_CONTEXT_FREEZE.jsonl`, that is the 80-query primary cohort plus the 20 valid r4d queries) arm A carries **1,925,473** characters of frozen `context` against arm B's **1,377,409**, an overhead of **548,064** characters, mean **5,481** per query (median 5,169, range 1,291–13,631), **positive for 100/100** queries and **28.7%** of arm A's context on average (range 13.3–41.1%). Pooled, **A/B = 1.398**$\times$, which places the executed treatment in the amendment's *compact annotation* regime ($\approx$1.39$\times$) rather than its full-dump regime ($\approx$2.3$\times$). The evidence-content hard gate was re-checked per query on the same pass: identical card sets, byte-identical card text and identical question text for **100/100** queries. The 20-query corrective replication carries **51,889** characters of overhead, mean **2,594** per query, **A/B = 1.140**$\times$, with the corrected question line present in **20/20** arm-A contexts. The whole-file ratio previously printed

Quantity	Value	Anchor
Unique analysis queries / unique analysis calls	100 / 300	ANALYSIS_SET_CORRECTION_001.json
Duplicate technical calls (same 20 queries, same prompt)	60	same
Invalidated calls (fresh20 confirmatory partition; prompt bug question == query_id, 20/20)	60	EFFECT_RECEIPT.md:16-17
Corrective-replication calls	60 (20 queries × 3 arms)	CORRECTIVE_REPLICATION_001_RECEIPT.md:5-9
Stray calls quarantined	4	ANALYSIS_SET_CORRECTION_001.json
E9 campaign total	**420** = 360 execution (240 primary + 60 invalidated + 60 duplicate) + 60 corrective; the **4** stray calls are quarantined **outside** the 360, and the 300/60/60 rows are three views of the same 360, **not additive**	ANALYSIS_SET_CORRECTION_001.json (execution.total_calls: 360, stray_records_quarantined: 4)
Reader identity	fixed gpt-5.6-sol via the Codex CLI agentic harness, read-only tools, temperature **not** guaranteed 0 on the raw API	PREREGISTRATION_AMENDMENT_001.md

here (2,639,603 vs 2,022,017 bytes, 1.31×) is **superseded**: it pools the 20 **invalidated** fresh20 rows, whose context carries question == query_id (the PROTOCOL_DEVIATION_002 defect), so it does not measure the treatment. Per-query rows and both input hashes: 06_MEASUREMENT\A_B_STRUCTURE_OVERHEAD.json (06_MEASUREMENT\measure_structure_overhead.py; OBSERVED_TOOL_RESULT, this revision).

Model-call budget (E9).

Read this budget before the p-values: the primary 80-query cohort rests on 240 calls, but the campaign also contains **60 technical duplicates** of 20 questions (their third execution in the same campaign) and **60 calls invalidated by a prompt bug**, so effective sample size and execution volume must not be conflated, and the analysis set was itself corrected **after** execution (ANALYSIS_SET_CORRECTION_001.json, discovered_after_execution: true; unique_analysis_queries: 100).

Two conclusions, kept in this order: **providing either form of external memory substantially improved performance over no external memory** (the defensible claim), and **the incremental benefit of typed causal/temporal presentation over a flat presentation of the same evidence was NOT ESTABLISHED** (the honest negative) [C-17][C-18]. *Anchors:* companion/\EFFECT_MEASURE_001_FREEZE\EFFECT_RECEIPT.md:21-43 (primary N=80: A 0.325, B 0.3125, C 0.0; A vs B 2/1 p = 1.0; A vs C 26/0 p = 3.0e-8; B vs C 25/0 p = 6.0e-8), :69-92 (§5 binding claim boundary), CORRECTIVE_REPLICATION_001_RECEIPT.md:18-38 (A 0.50, B 0.50, C 0.05; A vs C and B vs C 9/0, +45.00 pp, p = 0.0039; A vs B 0/0 discordant), EFFECT_FINDING_DISCUSSION.md:16-24. Evidence level: OBSERVED_SOURCE (the two receipts were opened and read in this work unit).

Corrective-replication discipline (mandatory qualifiers). The corrective run was *motivated by an execution bug* (the fresh20 context was built with question == query_id, 20/20), so it is a **corrective replication, not independent external replication**, it is **not preregistered independent confirmation**, and small-N (20) means the exact A-vs-B tie is a **low-powered no-detection**, not an equivalence proof. The invalidated fresh20 confirmatory partition and the invalidated duplicate-stability diagnostic **must not be quoted as results**.

Cross-surface numerical consistency (resolved). The corrective arm accuracies appear as 0.50/0.50/0.05 in CORRECTIVE_REPLICATION_001_RECEIPT.md:20-24, while TABLES_P6_P8_v0_1.md:63-64 summarises the *same* run with only the two +45.00 pp contrasts. These are **not conflicting**: 0.50/0.50/0.05 is the corrective **r4d subset (N=20)** and 0.0 is the primary **N=80 C-arm**, and 0.50 - 0.05 = +45.00 pp is exactly the printed contrast. Evidence: OB-

SERVED_DIFF (CORRECTIVE_REPLICATION_001_RECEIPT.md:20–24; TABLES_P6_P8_v0_1.md:56–64).

Boundary (binding, carried verbatim in substance). This tests **one downstream presentation treatment** (identical evidence plus explicit graph/temporal annotations versus the same evidence flattened). It is **not** evidence that causal structure is useless, and it does **not** probe CKB's system-function surface (retrieval, provenance, temporal consistency, historical reconstruction, multi-hop traversal, governed updates). The result is expressed as *no detectable benefit* — **not** an equivalence proof. Absolute accuracy in both arms is low under the mechanical scorer (∼0.31–0.33 primary), and the *relative* memory-versus-no-memory effect is what is large.

B.1.3 LONGMEMEVAL-S (PUBLIC DEVELOPMENT SET): EVIDENCE-ADJUDICATED CORRECTNESS

Read this subsection with all seven qualifiers attached, every time: (1) *post-hoc hybrid evidence-adjudicated*; (2) *public development set*, PUBLIC_DEVELOPMENT_SET_NOT_GENERALIZATION / NOT_HIDDEN_NOT_GENERALIZATION; (3) the 493 and the 7 are measured on the **frozen pre-adjudication predecessor candidate**, and the **accepted successor has no official score**; (4) the two judge outputs are **not** combined into a consensus label; (5) this is **not** an official benchmark score; (6) third-party replay is **PACK_PENDING**, so the manuscript does not claim the ready label; (7) there is **no same-question control arm**, so the figure is **not subtractable** from any system's accuracy and supports no treatment-effect claim.

Note on the numbering contract of this subsection. Five numbers circulate in this campaign family and are **never interchangeable**: 500/500 (hybrid evidence-adjudicated), 493/500 (frozen official gpt-4o, predecessor), 500/500 (Luna snapshot, same frozen candidate), 445/500 (internal lineage only, different candidate), 55 / 51 (wrong-row counts, **different candidates**, intersection only 38). Every use below states the **candidate** × **judge** × **denominator**.

Set relation (restated). These 500 questions are the same public development set whose nested subset supplies the E9 cohort of §7.2: DEV100 ⊆ this set (500 unique question_id), fresh20 is a 20-question second copy inside DEV100, and the E9 primary cohort is DEV100 minus its 20 r4d queries. There is **no held-out surface** anywhere in this paper (see the disclosure at the head of §7.1).

Answer-generator identity (what is recorded, and what is not). The M5 4096 run records answer_model: ''deepseek-v4-flash'', prompt version v6, answer_max_tokens: 16384, answer_model_calls: 500 and **0 judge calls** at generation time, and the M6-SOL run states that **only the answer model changed** (DeepSeek v4-flash → gpt-5.6-sol high) over the same M5 answer packets and the same v6 prompt. For the **predecessor candidate actually scored here** (e13f866d...9a64) and for the accepted successor (c23e2f0b...c63557), the pack contains **no generator field at all**: LUNA_SUCCESSOR_RECEIPT.json records schema/base/output sha and changed_rows only. The answer model behind the reported 493/500 and 500/500 is therefore **UNKNOWN — not recorded in the pack**, those two numbers are **not reproducible as to who answered**, and the independence of candidate generation from judging is **unverified** (judge gpt-5.6-luna, candidate named LUNA_SUCCESSOR; whether they are the same model family cannot be confirmed from the receipts, and this is registered as an unexcluded risk, **not** as an established self-preference bias).

Model-call budget (LME500). Answer generation costs **≥500 calls per candidate pass** (M5 4096: 500 answer-model calls, 2,538,385 prompt tokens and 366,751 completion tokens; M6-SOL: 2,507,357 generation-usage tokens over the same 500 packets, **receipt-only**: the M6 raw freeze is absent from the pack, so this figure rests on a single usage receipt rather than on frozen raw output. State this precisely, because it is checkable: the receipt names eight artifacts by hash, and the two that would directly support the 449/500 result and its 51 error rows — official_hypotheses (B152BDBFA2FA4CD8B405839CE6075F5FF6B10BAB4417027092663D61A038C674) and official_eval_results (8197753D603C053E9E0D31A0FC7ACE4DC1D863ADB59F0CC00677CFCD405994EE) — are **not present anywhere in the evidence tree**; a scan of its 462 files found no matching digest, and four files of the same shape are 0 bytes, i.e. no output was produced, not that output exists) and **0 judge calls inside any generation run** (rendering

is deterministic). Scoring is a separate pass: 500 rows by the frozen `gpt-4o-2024-08-06` evaluator plus 500 singleton-batch rows in the `gpt-5.6-luna` ledger, i.e. **2 × 500 scoring calls** for the dual-judge comparison, with `gold_file_accessed: false` at answer time.

7.3.1 Protocol. Task: LongMemEval-S, the **500-question public development set** (500 unique `question_id`; gold sha `d6f21ea9d60a0d56f34a05b609c79c88a451d2ae03597821ea3d5a9678c3a442`) [C-LME-01]. Pipeline: source-grounded typed extraction → evidence-anchored answer synthesis → deterministic rendering, **answer-time gold-free** (a gold-aware path exists in development and is disclosed; `gold_used_at_answer_time = false`). Repair discipline is **mechanism-keyed** (no question-id or answer hardcoding), and `formal_runtime_contains_question_specific_rules: false` is receipt-supported. Adjudication is three-layered — evidence layer (does the source support this answer), contract layer (the answer must be anchored to retrieved evidence and consistent with the question), and gold reference (gold is only a comparator, never an answer source).

7.3.2 Frozen official evaluator on the predecessor candidate. The frozen official evaluator **gpt-4o-2024-08-06** (evaluator source sha `5085eb9ae08b91c6c4f97c6bea9b8c2d6e55f9ec72c27734cafa0a04f00a430a`, temperature 0, max_tokens 10) scored **493/500 (98.6%)** on the **frozen pre-adjudication predecessor candidate** `LUNA_SUCCESSOR_HYPOTHESES.jsonl` (sha256 `e13f866d45aac9dbac52df0a9480c7723e7b9ed95d4e9f31fc275af832649a64`, 500 rows, `fixed_before_both_evaluations: true`), i.e. **7 wrong rows out of 500** [C-LME-02]. Per-category: ku 78/78, ssu 70/70, tr 133/133, ssa 56/56, ms 130/133, ssp 26/30 [C-LME-02]. The `98.6%` figure belongs **only** to this score and to this candidate, and is a public-development number, never a generalization claim (`A4...:27-28`, `:133`, `:144`).

7.3.3 Dual-judge comparability snapshot (not consensus). As a **comparability snapshot** on the **same frozen candidate**, a dual-judge run reports **Luna (gpt-5.6-luna**, route Codex OAuth, context policy `one_question_per_luna_context`, ledger sha256 `a2b095e7b3568f8fbf3b7bf616ce11bb1390a7c861e8424051ab22f9f8c34a0d`, 500 rows / 500 singleton batches / `score_bearing_true` 500 / judgments true 500, false 0) at **500/500**, and the frozen official gpt-4o evaluator at **493/500** [C-LME-05]. The two judge outputs are **not combined into a consensus label**; there is no consensus quantity in this campaign (`A4...:106-110`). The disagreeing set is **7 rows**, all one-directional (`luna_yes_gpt4o_no` = 7, `luna_no_gpt4o_yes` = 0), enumerated with ids in `DUAL_JUDGE_DISAGREEMENT_AUDIT.md` (sha `d3d0cc90...e4e2`) [C-LME-05].

Two integrity records belong beside that snapshot rather than in a footnote: (a) the candidate rewrite that would have satisfied the Luna judge on all seven rows was **explicitly rejected as judge gaming** — `DUAL_JUDGE_DEV7_RECEIPT.json` records status `REJECTED_DIAGNOSTIC_JUDGE_DIRECTED_REWRITES` with `official_evaluator result: null` and blocking error `429 credit_balance_exhausted`, its eval-results file being 0 bytes; and (b) the accepted successor (`DUAL_JUDGE_SUCCESSOR_HYPOTHESES.jsonl`, sha256 `c23e2f0bb225e54145613d8d4e30550caabb6f49082e7e951fef372081c63557`, 500 rows) **changes exactly one row**, `a89d7624` [C-LME-05][C-LME-04] (`A4...:30`, `:37`, `:39`, `:73-74`).

7.3.4 The accepted successor has no official score (A4-B1, P0). `DUAL_JUDGE_SUCCESSOR_HYPOTHESES.jsonl.eval-results-gpt-4o` is **0 bytes** (sha256 `e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855`), and `EVIDENCE_ADJUDICATION_RECEIPT.json:29-35` records `successor_official_score_available: false` with `applies_to_predecessor_candidate: true` [C-LME-04]. **No statement in this paper attaches 493/500 to the accepted successor.** The absence is by design, not an oversight (`A4...:32`, `:147`, `:166`).

7.3.5 The adjudicated outcome and its decomposition. We report the **post-hoc hybrid evidence-adjudicated** outcome as **500/500**, computed as **493 rows accepted unchanged by the frozen official gpt-4o-2024-08-06 evaluator plus 7 source-level adjudications** (493 + 7 = 500, arithmetic re-verified), all evaluated on the frozen pre-adjudication predecessor candidate named in §7.3.2 [C-LME-03]. The seven adjudications decompose as: **five judge disagreements** (answers that satisfied their gold requirements **under our post-hoc source-level rule** but were rejected by the judge), **one benchmark source/gold conflict** (the gold answer requires information absent from the

retrievable source), and **one genuine synthesis bug in our own pipeline**, which we repaired [C-LME-03]. The adjudication rule is `rule_defined_posthoc: true` and `fully_manual_blind_review: false`; the frozen evaluator emits only a boolean {`model, label`} with **no rationale**, so the paper states what our source-level adjudication concluded, and **never** that a judge was proven wrong (`A4...:31`, `:116`, `:137`, `:150`, `:167`). This developmental figure has **no same-question control arm**: no memory-ablated, retrieval-only, or public-system baseline was run on these 500 questions under this evaluator, so the number is **not subtractable** from any system's accuracy and supports **no treatment-effect claim**; it is a single-arm descriptive observation on a public development set. The only memory-ablated evidence on these same question identifiers comes from a *different* endpoint on a nested subset (§7.2: arm C, the no-memory arm of the same 80-query nested cohort, deterministic scorer, answer model separately recorded as `gpt-5.6-sol`), and the two must not be subtracted from one another.

An earlier draft's "∼1% (5/500) judge false-negative rate" is **withdrawn**: only the seven judge rejections were adjudicated, and the 493 accepts were **not** manually re-audited, so no benchmark-wide judge false-negative rate is estimable. The defensible form is: *five of the seven adjudicated judge rejections were false negatives (5/7 of the adjudicated rejections)* (`A4...:121`, `:151`).

7.3.6 Candidate lineage (internal, not results). The campaign contains several candidate generations — M5 RUN2 scored **445/500 (89.0%)** with **55 wrong rows** out of 500 (28 complete + 27 incomplete) under the same gpt-4o-2024-08-06 evaluator; M6-SOL scored a receipt-only **449/500 (89.8%)** with **51 wrong rows** out of 500 (22 complete + 29 incomplete); an M6 error taxonomy gives evidence_loss 24 / execution_failure 19 / disputed_gold_or_judge 8 over **51 of the 500 public questions** (the three class counts are read from the per-case `failure_class` column of a gold-aware matrix; because the M6 raw evaluator output is absent from the pack, that matrix is a disclosed reconstruction — the taxonomy file is a faithful rebuild whose contract tests pass, but its byte-identity with any original is **UNKNOWN**); and M5's 55-row and M6's 51-row error sets intersect in only **38** rows (13 matrix-only, 17 M5-only, union 68) [C-LME-07]. These belong to **pipeline lineage and error analysis only**: each is bound to a different candidate, M6's raw evaluator output is **absent from the pack** (receipt-only, `UNREAD`), and none of them may be placed in the same table as the successor or the hybrid figure (`A4...:25-26`, `:33-36`, `:64`, `:140`, `:193`). **The 55 and the 51 are never interchangeable, and neither is a subset of the other** (`A4...:36`, `:146`). M8 (442/500, `REJECT_M8_ANSWER_LANE`) and M9-100 (`new_score: null`, self-described "not an official benchmark arm") are withheld entirely (`A4...:40`, `:152`). An older public attachment reporting 500/500 with artifact sha `bc692b10877d...` has **no sha relationship** to this pack and is excluded (`A4...:42`, `:153`). Generator identity for these lineage rows: M5 = `deepseek-v4-flash` (recorded); M6-SOL = gpt-5.6-sol, high (recorded); the scored predecessor and the accepted successor = **UNKNOWN** (no generator field in the pack, see the identity note above).

7.3.7 Error-class findings (the paper-valuable part). Two findings outrank the score itself. **F1 (mechanism):** we identified and repaired an "evidence-retrieved-but-dropped-in-synthesis" error class; the fix is that an answer must be anchored to retrieved evidence (evidence-reference validation plus mandatory fields), and the production v7 path implements it as a **fail-closed evidence-anchor gate** where missing or invalid evidence indices cause rejection rather than an apparently well-formed but insufficiently grounded answer [C-23]. **F2 (benchmark methodology):** five of the seven adjudicated judge rejections were false negatives — answers semantically satisfying gold but rejected on literal/format strictness — and one row required gold information absent from the source material; this argues that an API judge's label is a **weak proxy** for correctness in this setting, while evidence adjudication is more re-checkable [C-23]. We present F2 as a **methodological observation on 7 adjudicated rows**, not as a benchmark-wide rate and not as proof that any judge erred.

7.3.8 The two reporting tracks must not be merged. Track 1 (development) = the **post-hoc hybrid evidence-adjudicated 500/500** described above. Track 2 (production) = a **frozen gold-blind production run plus a third-party replay receipt**, and **that receipt does not exist**: the audit searched `[workspace]`, `[workspace]/companion/`, `[workspace]/[repo]` and `[abs]/[repo-ai]` for replay/third-party artifacts and found no LME500 replay package (**NOT_FOUND**) [C-20] (`A4...:78`, `:164`, `:183`). Therefore: the paper carries **PACK_PENDING** for third-party evidence, **does not** claim "independently replayed" or "third-party verified", **does not** claim the ready label (the pack's own state machine requires third-party verification for it), and never writes a single merged number across the two tracks.

7.3.9 The 500/500 hybrid figure is not a structural-certificate claim. `semantic_sufficiency_automatically_proven: false`, and §7.1's anti-self-certifying result (28/100 structural certificate against 100/100 answer coverage) stands beside it; the two must not be presented as mutual corroboration (`A4...:156`).

7.3.10 Known internal contradiction, carried forward then repaired at source. `LME500_EVALUATION_SECTION_v0_1.md` originally stated **"0 disagreements"** for the Luna snapshot at `:21` while `:22` and `:31–35` of the same file enumerate **7 disagreements**, alongside a duplicated evidence-adjudicated row (`:19–20`) and a withdrawn FN rate (`:39`). The audit ruled the "0 disagreements" wording an error: the ledger's 500 singleton rows each carry a judgment payload whose only keys are `label` and `question_id` (the row itself carries 13 keys — `correct` / `incorrect` integer counts summing to 500 / 0, `score_bearing`, a single `model_id openai-codex:gpt-5.6-luna`, and hashes — all of them single-judge state), so the ledger cannot support that claim; disagreement is a two-judge relation and is recorded only in `DUAL_JUDGE_RECEIPT.json` (`judge_disagreements`: `luna_yes_gpt4o_no` = 7, `luna_no_gpt4o_yes` = 0), so the corrected form is "Luna's own ledger records 0 false judgments; there are 7 disagreements **against** the frozen gpt-4o label" (`A4...:120`, `:119–122`, `:194`). **The source file has since been revised to exactly that corrected form** (all three defects repaired; `LME500_EVALUATION_SECTION_v0_1.md`, sha256 `199E3F3DD20AE643 4026229BC00F6FD4D13860A768898B468B102F42C1856F16`, 4,568 B, 55 lines: the duplicate row removed, "0 disagreements" replaced by "0 false judgments, not 0 label disagreements", and the withdrawn ∼1% benchmark-wide rate replaced by the defensible `5 of 7` adjudicated form). **This manuscript uses the corrected form throughout, and the source file now agrees with it; the contradiction is closed at source, not merely restated here.**

B.2 Limitations (binding, each with its evidence status)

1. **CDL/CAL assignment is first-cut.** The per-edge CDL/CAL values are first-cut graph heuristics, not a validated per-edge depth estimator; the CCE arms are **synthetic demonstrations, not real-edge maturity evidence**. Replacing the heuristics with an empirically validated depth estimator is the largest remaining internal caveat on the CCE envelope.
2. **The taxonomy is a governance framework, not a law of nature.** CDL does not assert that causal knowledge is strictly nested, and confidence, strength, regime and invariance remain distinct fields.
3. **Layer B is not a production runtime.** The specification and bounded reference artifacts do not establish automatic promotion, demotion, or fork resolution for arbitrary deployed stores.
4. **Internal evidence is bounded and workload-specific.** fresh20 is 20 frozen queries and is not an accuracy result; the 11,831 counterfactual instances evaluate graph-computation effects, not real-world interventions; the 26.3× observation is workload-specific, neither universal nor asymptotic.
5. **The CCE demonstration is synthetic** with a_min = CAL-1 and $\tau = 0.30$, and does not validate a general edge classifier.
6. **Model-mediated steps are not byte-deterministic**; reproducibility requires raw outputs, retry histories and deterministic replay.
7. **The external-treatment result is an observability result.** The public-development-set outcome is evidence-adjudicated, **not** an official score; there is **no** hidden-set generalization claim, **no** judge consensus, **no** fully blind manual review, **no** same-question control arm (not subtractable, no treatment-effect claim), and **no** SOTA/human-level comparison anywhere in this paper; the frozen gold-blind production run and its third-party replay receipt remain **PACK_PENDING**.
8. **MuSiQue (conventional-retrieval judge) did not discriminate.** Under a shared BM25 lexical candidate generator, the CKB-governed arm tied its BM25-only arm exactly on supporting-evidence recovery, both above a no-retrieval baseline — a bounded negative showing that governed causal memory does not automatically improve *conventional paragraph-retrieval* metrics, because those paragraphs are independent, untimed and carry no causal edges. **Value withheld:** the MuSiQue support-F1 triple is not reported, because

its evidence chain and pack anchor were not completed in the audited artifact surface used by this paper.

9. **Reviewer receipts are not CI.** The temporal surface (E8) is frozen reviewer receipts with **zero CI check runs**, and Milestone A remains a candidate; neither is a native approval, and neither grants merge or canonical authority.
10. **Documentation trails the implementation.** The audited repository carried a 12-of-21 candidate-receipt hash mismatch and a non-collectable "769 tests pass" claim; the `769` figure is a **stale repository self-report** (`pytest` exits 2 with 717 collected + 16 errors) and is *not* the same surface as the R4E collection of **287 tests**, which was subsequently **executed: 287 passed / 0 failed, exit 0** (2026-09-18; log `pytest_r4e_full_after_fixture2.log`, sha256 `446AF615...6927`; run against the frozen snapshot plus the disclosed M6 reconstruction) — the two must not be merged. This paper cites audited numbers only (`A2...:137`, `:161`, `:258-272`).
11. **Two previously logged internal number discrepancies are now reconciled, not merely carried.** In §7.1 the 243 and 11,796 figures were two rows of one ledger, not competing readings; in §7.2 the corrective 0.50/0.50/0.05 triple is the N=20 r4d subset while 0.0 is the primary N=80 C-arm. Both reconcile by arithmetic on the cited surfaces and remain printed, not smoothed.
12. **Coverage limit of our novelty claim.** The prior-art audit performed no systematic literature search and read abstracts rather than full texts for the adjacent lines (notably Zep/Graphiti, **UNREAD**), so our boundary is "no counterexample found" (`A5...:86-88`, §F-11).
13. **The workspace inventory was produced by deterministic scripts after an earlier delivery attempt failed.** `A3_INVENTORY_AUDIT.md` is now present, produced with deterministic scripts rather than by an autonomous agent. It reports a CKB-related footprint of **33,892 files / 13,295,481,889 B** under `[workspace]` (57 top-level items matching `CKB`), dominated by `companion/` (**32,816 files / 12,608,626,886 B**), and finds **four** byte-identical **2,156,408,832 B** redb snapshots (`D1D1DBA5A07FF45D...`) — plus, as a negative result, that none of the five large snapshots matches the recorded `base.frozen.redb` digest `65762C6D1A8BDBF7...`, so the "frozen artifact unreachable" verdict is **not** overturned. No claim in this paper rests on it, and no claim should: the finding is that a deterministic scripted audit is the correct instrument here, and an autonomous agent was the wrong one (see Governance).
14. **Every empirical surface is nested inside one public question set, and there is no held-out surface.** `DEV100` $\subseteq$ `LongMemEval-S 500`; `fresh20` is a second copy of 20 questions already inside DEV100; the E9 primary cohort is `DEV100` $\setminus$ `r4d(20)` and was selected to be homogeneous, so its `NON_CAUSAL_CONTROL` subgroup is **EMPTY**. The §7.1–§7.3 surfaces are therefore **not independent evidence faces**, and the external-memory benefit is established only on the pre-selected causal/temporal subset of that public development set (OBSERVED_TOOL_RESULT: 100/100 intersection of the 100 analysis `corpus_id` values with the 500 `question_id` values; `ARM_CONSTRUCTION_SPEC.json:31`).
15. **The reported LME500 correctness numbers have an unknown answer generator, and the E9 execution surface is larger than its analysis surface.** The pack records `answer_model: deepseek-v4-flash` for M5 and gpt-5.6-sol (high) for M6-SOL, but records **no** generator field for the predecessor candidate actually scored (493/500) or for the accepted successor (500/500): that identity is **UNKNOWN / not recorded**, so those figures are not reproducible as to who answered, and candidate–judge independence is **unverified**. On the E9 side the campaign ran **360** execution calls plus **4** stray calls quarantined outside that count, and then **60** corrective calls: **420** campaign calls in total (**300** unique analysis calls, **60** technical duplicates, **60** calls invalidated by a prompt bug, **4** quarantined stray, **60** corrective; the unique-analysis, duplicate and invalidated figures are three views of the same 360, not additive) — with the analysis set corrected **after** execution. **Seven** p-values are reported (**five** from the original campaign plus the **two** arm-D contrasts) **without any multiple-comparison correction**.
16. **The §5 adversarial invariants and the cross-language conformance check are executed, but only on the audited tree and only on the author's machine.** The **41** `#[test]`

Predicate	Definition
P1	source_grounded_support >= 1
P2	independent_evidence_chains(ei) >= 2
P3	structural_validated AND (counterfactual_independently_validated OR replication_independently_validated)
P4	interventional_valid (the CAL-34 gate)
P5	governance_certificate_valid AND independent_audit_valid

Arm	N	max_demonstrated_cdl	printed Pr	system_verified_cdl
A (original)	620	3	0.0000	1
B (single injection)	621	5	0.0016	1
C (coverage injection)	886	5	0.3002	5

attributes across G1B_P\tests\adversarial.rs (24) and tests\transition_adversarial.rs (**17** — recounted in this work unit; the systems review's 22 is not adopted), and conformance_test.py, are now **executed and passing**: **41 passed / 0 failed** across the two Rust files (exit 0, 2.7 s) and 16 [PASS] conformance checks (ALL PASS, exit 0), both run 2026-09-18 and both with dated log shas (§5 execution-status paragraph). The standing limitation is scope, not status: these are the audited tree's own tests on the author's machine, not third-party CI, and those paths lie outside the audited core-code scope (A2...:45).

B.4 Predicate definitions (P1–P5) and two boundary conditions

Why this is here. §4 computes DerivedCAL = max({0} ∪ { k: P1 ∧ P2 ∧ ... ∧ Pk }) and §5/§B.0 refer to P1..P5, but the five predicates are stated only in the companion contract. They are reproduced here so the strict-accumulation claim can be checked from the paper alone. Each Pi is a predicate over (edge, evidence_set, provenance); no predicate may rely on an ungrounded model judgement as primary evidence.

CAL >= k iff P1 AND P2 AND ... AND Pk. The reference implementation evaluates this as a first-false break, so CAL = k holds iff every predicate up to k holds and a false predicate at position j caps CAL at j−1. Invariant CKB-INV-013 tests exactly the non-monotone case (P1..P3 true, P4 false, P5 true ⇒ CAL = 3). Contract anchor: CKB_CAL_ESCALATION_CONTRACT.md:347-359; reference implementation G1B_P/src/eval.rs.

Two qualifiers that the §4 prose does not carry, stated here.

1. **Monotonicity is in the predicate chain, not in the evidence set.** Adding evidence can flip P4 or P5 to false — for example an audit that shares a validator with the issuer fails P5 — so CAL may *decrease* as evidence grows. This is intended: it is the reason demotion exists as a first-class transition rather than an error case.

CCE arm table (the §4.2 experiment), with one discrepancy stated rather than smoothed.

The verified column is computed by the envelope function under the headline a_min = CAL−1. The printed Pr column was emitted by the experiment script's *diagnostic* using CAL >= 3, so it is the coverage of the stricter predicate, not of a_min = CAL−1. The receipt is otherwise item-by-item consistent with the three arms stated in §4.2; this is the one respect in which the printed table mixes two predicates, and it is disclosed here rather than smoothed. Anchor: CCE_EXPERIMENT_RECEIPT.md:52-56.

1. The coverage-gated envelope's anti-inflation property is conditional on tau > 1/N. The guarantee that one isolated high-depth edge cannot raise system_verified_cdl holds because that edge contributes at most 1/N of the coverage mass. If tau <= 1/N, a single

edge *does* satisfy the threshold and raises the reported level. The property is therefore relative to the tau/N ratio and is not unconditional.

B.3 What evidence would change these conclusions

- **Answer correctness.** If answer_effect / evidence_effect / judge_effect are measured and the causal structure is shown to contribute causally to correct answers, the substrate claim upgrades to an efficacy claim.
- **A judge that scores our surface.** Running CKB and a retrieval baseline on a benchmark whose judge directly scores temporal state reconstruction, causal semantics, or provenance governance (rather than conventional paragraph overlap), reporting both accuracy and the auditable-authority properties, would turn "different design centre" into a measurable difference; the MuSiQue attempt shows the conventional-retrieval judge does not provide that surface.
- **Runtime closure.** Wiring the authority oracle into the production Rust engine.rs (rather than a Python reference model) would upgrade Layer B from "formally closed + reference executable" to "runtime-integrated governance".
- **Per-edge depth validation.** An empirically validated CDL/CAL depth estimator would remove the largest internal caveat on CCE.
- **Third-party replay.** An independent replay receipt would close the PACK_PENDING row and would be the first evidence that the campaign is reproducible outside our own account (A4...:164, :192).
- **A hidden-set result.** Any evaluation on a held-out, non-public split — rather than the 500 public-development questions — is what would license the generalization language this paper deliberately avoids (A4...:41, :135).

C. Appendix C: Claim anchors

Each claim id used in this manuscript, with the file actually read and the anchor inside it. Evidence grades follow the audit vocabulary (OBSERVED_SOURCE / OBSERVED_TOOL_RESULT / OBSERVED_DIFF / INFERENCE / UNKNOWN / UNREAD / STALE / NOT_FOUND / NOT_RUN). Paths are absolute; A1..., A2..., A4..., A5... abbreviate the four audit files under [paper package]/01_AUDIT\.

Claim-id entries in this manuscript that are NOT backed by a whitelisted number (placeholders only, deliberately not filled): the MuSiQue support-F1 triple (§2.2, §8.2 item 8). Everything else quantitative in §1–§9 resolves to a row above.

Evidence status summary

Counts below are derived from the two status surfaces read in this work unit — ...\SUBMISSION_VIEW\CLAIM_EVIDENCE_LEDGER.md:8-34 (22 ledger rows with an explicit Status column) and the dry-run spine's status markers [v] / [VERIFY] / [PACK_PENDING] / [REPORT-ONLY] at ...\PAPER_FAMILY\INTERNAL_MANUSCRIPT_SPINE.md:6 — plus the NOT_RUN-class rows identified by the A1/A2 code audits. Counts are stated with their counting basis so they cannot be read as a claim about the whole repository.

Ready label not claimed. Per the pack's own state machine, that label requires third-party verification, which is **not** achieved (A4_LME500_EMPIRICAL_AUDIT.md:154-155, :192); the third-party row is PACK_PENDING and the public-development track remains a post-hoc hybrid evidence-adjudicated observability result, PUBLIC_DEVELOPMENT_SET_NOT_GENERALIZATION.

Number-surface statement (whole manuscript). The only LongMemEval-S values printed in this paper are the ones the empirical audit classified as writable, each with its mandatory qualifier: 493/500 (98.6%) bound to frozen official gpt-4o-2024-08-06 × frozen predecessor candidate sha e13f866d...9a64 × denominator 500 (C-LME-02); 500/500

Claim id	Claim (short)	Anchor	Evidence level
C-01	Causal capability is treated as boolean; CKB provides a governed causal-authority lifecycle	...\SUBMISSION_VIEW\CLAIM_EVIDENCE_LEDGER.md:10 (C-01)	OBSERVED_SOURCE (paper framing; FROZEN-draft)
C-02	Identity frame (reader model held fixed; no attribution to parametric learning); contributions 1–6	...\PAPER_FAMILY\INTERNAL_MANUSCRIPT_SPINE.md:42-43; ...\CLAIM_EVIDENCE_LEDGER.md:11	OBSERVED_SOURCE
C-03	Epistemic Constitution / Causal Dynamics separation; causal graph is not its own authority	...\SUBMISSION_VIEW\PROSE_ARCHITECTURE_LAYERS.tex:3-4; ...\CLAIM_EVIDENCE_LEDGER.md:12	OBSERVED_SOURCE (framework VERIFIED)
C-04	CDL-0..6 ladder per edge; DerivedCAL strict accumulation; CAL is not an author-written field	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:11-32; ...\CLAIM_EVIDENCE_LEDGER.md:13	OBSERVED_SOURCE (framework VERIFIED)
C-05	CCC = (CDL, CAL); three distinct depth quantities; coverage-gated envelope definition	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:36-45; ...\CLAIM_EVIDENCE_LEDGER.md:14	OBSERVED_SOURCE (VERIFIED)
C-06	CCE A/B/C: one isolated synthetic CDL-5 edge $\Rightarrow$ max_demonstrated 5 but system_verified_cdl(τ) = 1; rises to 5 past threshold; **synthetic, not deployed capability**; a_min = CAL-1 and τ = 0.30 stated with the definition	[repo]/CKB_CAUSAL_STRATIFICATION_FRAMEWORK\CCE_EXPERIMENT_RECEIPT.md:21, :52-56; [workspace]/companion/\PAPER\EXPERIMENTS.md:79-91 (E4); ...\CLAIM_EVIDENCE_LEDGER.md:15-16	OBSERVED_SOURCE (definition check / worked example; receipt opened this round)
C-07	Implemented substrate: typed causal/epistemic edges, belief/world separation, counter-evidence retained	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:57-60; ...\CLAIM_EVIDENCE_LEDGER.md:16	OBSERVED_SOURCE (conformance VERIFIED)

bound to gpt-5.6-luna × one-question-per-context × same frozen candidate × denominator 500, comparability snapshot only (C-LME-05); 500/500 bound to post-hoc hybrid evidence-adjudicated × public development set × 493 + 7 (C-LME-03); the 7 disagreements with ids (C-LME-05); the 5 / 1 / 1 disposition with rule_defined_posthoc: true / fully_manual_blind_review: false (C-LME-03); the 24 / 19 / 8 taxonomy explicitly over 51 of the 500 public questions of a different candidate (C-LME-07); and 445/500 / 449/500 as internal lineage only, with candidate sha and receipt-only marking (C-LME-07). Every audited blacklist item — the withdrawn ∼1% (5/500) judge false-negative rate, Luna's shared-context diagnostics 494/500 and 497/500 (two separate non-scoring entries), the M8 442/500 and M9-100 arms, the DEV7 judge-directed rewrite, and the AGI_Core attachment's 500/500 with artifact bc692b10877d... — appears nowhere in this paper as a result, and in the two places where the campaign's internal record must be acknowledged (the withdrawn FN rate in §7.3.5 and the lineage in §7.3.6) it is named **as excluded**, never as a quantity (A4_LME500_EMPIRICAL_AUDIT.md:142-156, :25-26, :39-40, :42).

Open findings carried by this manuscript (all negative or unresolved, none suppressed): (1) LME500_EVALUATION_SECTION_v0_1.md originally contained, at :19-21 and :39, a duplicated evidence-adjudicated row, a "0 disagreements" statement contradicted by :22, and a ∼1% judge false-negative rate that this manuscript **withdraws** — the 493 accepts were never manually re-audited, so no benchmark-wide rate is estimable and the defensible form is *five of the seven ad-*

Claim id	Claim (short)	Anchor	Evidence level
C-08	E3 recall **26.3× frozen**, exact_match=true, 28.9% scan, T_build 117.362 ms, break_even 0.85 (**FROZEN**). The 28.1× fresh value and the ARMS decomposition (A→B ≈81×, B→C ≈0.3×) are **REPORT_ONLY / withdrawn from the result surface**: no located receipt	FROZEN half: [repo]/CKB_COMPLETENESS_CLOSURE_v0_1\ckb-engine\PERF_RECEIPT.txt:17–33; ...\SUBMISSION_VIEW\TABLES_P6_P8_v0_1.md:6–15. Fresh half: EXPERIMENTS.md:53–71 (prose) and ...\SUBMISSION_VIEW\TABLES_P6_P8_v0_1.md:11–14 (hash-less rows; Source :15 = [VERIFY]), **NOT_FOUND** as receipt	OBSERVED_SOURCE (receipt read this round) + **NOT_FOUND** (fresh re-run)
C-09	Authority derived, not written as a field; reference model property tests	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:66–83; ...\CLAIM_EVIDENCE_LEDGER.md:18	OBSERVED_SOURCE (governance proof VERIFIED)
C-10	Witness DAG revocable/demotable; fork/merge via ledger-minted content-bound snapshots	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:76–83; ...\CLAIM_EVIDENCE_LEDGER.md:19	OBSERVED_SOURCE (VERIFIED)
C-11	Structural certificate gap: DEV100 coverage 100/100, bundle 100/100, certificate **28/100** (49/86/58/100), overlap 20/20; fresh20 20/20 vs **10/20**	...\SUBMISSION_VIEW\PROSE_P6_EMPIRICS.tex:3–12; EXPERIMENTS.md:95–106; ...\CLAIM_EVIDENCE_LEDGER.md:20	OBSERVED_SOURCE (structural certificate FROZEN)
C-12	11,831 registered counterfactual edge-query instances over 20 queries; 14 / 21 / 11,796; 10,050 no-effect; 1,503 not-evaluable; effects PENDING_MODEL_RE_RUN	...\SUBMISSION_VIEW\TABLES_P6_P8_v0_1.md:36–40; EXPERIMENTS.md:27–44; ...\CLAIM_EVIDENCE_LEDGER.md:21. The ledger's earlier EVIDENCE_REGISTRY.json anchor is **NOT_FOUND** in both trees (v0.2 row C-12)	OBSERVED_SOURCE (**VERIFIED**, downgraded from FROZEN for lack of the registry) + **NOT_FOUND**
C-13	Executable governance: Python oracle = Rust transition core on accept/reason/state/SHA; cross-language equality ≠ semantic coverage	...\SUBMISSION_VIEW\PROSE_p2_p7_v0_1.md:85–114; ...\CLAIM_EVIDENCE_LEDGER.md:22–23	OBSERVED_SOURCE (conformance VERIFIED)

judicated judge rejections (A4...:151 blacklist item 8, :194 A4-B4, :121). **Status: repaired at source, not merely restated.** The file has since been revised (sha256 199E3F3DD20AE643 4026229BC00F6FD4D13860A768898B468B102F42C1856F16, 4,568 B, 55 lines): the duplicate row was removed and the "0 disagreements" wording replaced by "0 false judgments, not 0 label disagreements", so A4...:119–122, :194 is now **closed** rather than pending. This manuscript uses the corrected forms throughout, and the source file now matches them; the finding is retained here as a record of what was found and repaired, not as an open defect; (2) the causal NO_EFFECT reading 243 vs 11,796 was an open discrepancy at drafting time and is now **reconciled** as two rows of one ledger (14 + 21 + 243 + 10,050 + 1,503 = 11,831; 243 + 10,050 + 1,503 = 11,796) — kept here rather than deleted (§7.1, OBSERVED_DIFF); (3) the corrective arm-accuracy figures 0.50/0.50/0.05 (N=20 r4d) vs the primary 0.0 (N=80) were likewise carried as a transcription difference and are now **reconciled**, their difference being exactly the printed +45.00 pp (§7.2, OBSERVED_DIFF); (4) the accepted successor has no official score and never will under the current receipts (A4–B1, P0); (5) no third-party replay exists (A4–B2, P0); (6) the M6 raw evaluator output is absent — the two artifacts the M6-SOL receipt names

Claim id	Claim (short)	Anchor	Evidence level
C-14	E8 temporal surface: 50 pass / 1 ignored, 13/13, 13/13, 260/260 — **reviewer receipts, zero CI**, frozen file only	[workspace]/companion/\PAPER\EXPERIMENTS.md:141-153	OBSERVED_SOURCE (cited-as-frozen; provenance note at ...:147-149)
C-15	Milestone A typed runtime: on-disk receipt CANDIDATE_RECEIPT.json = 31 passed; 0 failed, engine 48 passed; 1 ignored; 0 failed, payload 21 / tree 4f4d5043..., receipt sha 47EE955a...2650, CANDIDATE_NOT_CANONICAL; the previously printed 13/13 + 45 (1 ignored) + f44a8031... + 16/16 + f31a305e... is **withdrawn as unreproduced**	[repo]/CKB_TYPED_RUNTIME\CANDIDATE_RECEIPT.json:5-6, :37-38; contradiction surfaces ...\PAPER_FAMILY\MILESTONE_A_STATUS_LOG.md:5-8, :20-26; A2...:258-262	OBSERVED_SOURCE (**REPORT_ONLY / payload_drifted**; four mutually inconsistent surfaces)
C-16	Milestone A review ended REVIEW_INCOMPLETE (3 platform misjudgements), not dressed up as a pass	MILESTONE_A_STATUS_LOG.md:8, :19; ...\CLAIM_EVIDENCE_LEDGER.md:25	REPORT_ONLY (manual report; not a receipt field)
C-17	E9 effectiveness: A vs C 26/0 (+32.5 pp, p=3.0e-8); B vs C 25/0 (+31.25 pp, p=6.0e-8) → SUPPORTED	[workspace]/companion/\EFFECT_MEASURE_001_FREEZE\EFFECT_RECEIPT.md:21-43; EFFECT_FINDING_DISCUSSION.md:20-22	OBSERVED_SOURCE (FROZEN)

by hash, official_hypotheses and official_eval_results, are not present in the evidence tree — so 449/500, 51 and 24/19/8 are receipt-only (A4-B3, P1; the exact hashes and the scan are in the LME500 model-call-budget paragraph); (7) A3_INVENTORY_AUDIT.md was **NOT_FOUND** when this manuscript was drafted and is now present — but it was produced with deterministic scripts after an earlier delivery attempt failed, so no inventory-based claim is grounded on it (see §8.2 finding 13); (8) the repository's own documented test count is not collectable and a candidate receipt mismatches 12 of 21 hashes, so documentation is not cited as implementation truth (A2...:258-262, :270-272); (9) the MuSiQue number surface is outside the audited whitelist and is therefore withheld pending a completed evidence-chain audit; (10) **all four empirical surfaces are nested subsets of one public question set** (DEV100 $\subseteq$ LME500, fresh20 a second copy inside DEV100, E9 primary = DEV100 \ r4d(20), NON_CAUSAL_CONTROL EMPTY), so they are not independent evidence faces and no held-out surface exists (§7.1, limitation 14); (11) the **answer generator of the scored candidate is UNKNOWN** — not recorded in the pack — so the 493 and the two 500/500 figures are not reproducible as to who answered, and candidate–judge independence is unverified (§7.3 identity note, limitation 15); (12) the **§5 adversarial invariants are executed on this machine but not third-party-verified**: **41** #[test] attributes (24 + 17; recounted this round, the review's 22 is not adopted) plus conformance_test.py were run on 2026-09-18 with **41 passed / 0 failed** and 16 [PASS], exit 0 (§5, limitation 16) — OBSERVED_RUNTIME on the author's machine, **not** CI; (12b) the R3 calibration bands (C-19) stay VERIFY with no band figure located, and the withdrawn C-08f fresh numbers remain REPORT_ONLY (Appendix C C-08f/C-19); (13) the A/B **total-token parity gate was removed pre-execution**, and the gap was later closed by a **matched-token placebo arm (D)** added under its own pre-registration (A vs D $\Delta = 0.0$, 2/2 discordant, exact McNemar p = 1.0; D vs B $\Delta = +0.0125$, 2/1 discordant, p = 1.0), so the A-vs-B p = 1.0 is not attributable to extra-token effects (§7.2, C-18; PLACEBO_ARM_D_001_RECEIPT.md section 2); (14) ...\SUBMISSION_VIEW\TABLES_P6_P8_v0_1.md:62-64 still carries an unfilled [VERIFY] place-holder in the corrective-replication row — this manuscript does not print that cell and flags the file for repair; (15) the ledger's own status summary line does not sum

Claim id	Claim (short)	Anchor	Evidence level	
C-18	E9 negative: A vs B 2/1 discordant of **n=80** (77 concordant), +1.25 pp, p=1.0 → NOT ESTABLISHED (**primary, preregistered**); corrective replication 9/0 of n=20 (11 concordant), +45.0 pp, p=0.0039, motivated by the `question == query_id` bug, **not** independent replication; A/B **total-token parity gate removed pre-execution**; a matched-token placebo arm (D) was added later under its own preregistration (`EFFECT_MEASURE_001_FREEZE\PLACEBO_ARM_D_001\PREREGISTRATION_PLACEBO_ARM_D.md`) and returns A 0.3250 vs D 0.3250 (A-vs-D 2/2 discordant, Δ = 0.0, exact McNemar p = 1.0) with D-vs-B Δ = +0.0125 (2/1 discordant, p = 1.0)	`PLACEBO_ARM_D_001_RECEIPT.md` section 2; `D_BLIND_SCORES.json`	`EFFECT_RECEIPT.md:29-43, :69-92`; `CORRECTIVE_REPLICATION_001_RECEIPT.md:5-9, :18-38, :44-50`; `EFFECT_MEASURE_001_FREEZE\PREREGISTRATION_AMENDMENT_001.md:24-29, :38-43`; `PREREGISTRATION_FREEZE.json:29-37`	OBSERVED_SOURCE (FROZEN negative result) + OBSERVED_DIFF (frozen spec vs amendment)	
C-19	R3 named-model calibration bands: **no band figure located**; the row stays `VERIFY` and **no** R3 band value is printed in this manuscript	`...\CLAIM_EVIDENCE_LEDGER.md:28` (v0.1 row; its digits are **withdrawn rather than restated** — author decision, 2026-09-18); `CLAIM_EVIDENCE_LEDGER_v0_2.md:40`; `MILESTONE_A_STATUS_LOG.md:36-40`	OBSERVED_SOURCE / **NOT_FOUND** (bands); `VERIFY` (unresolved by design)	
C-22	Genome/invariant corpus: 15 architecture-genome invariants, contract `CAL-rev0.5`, `G1b` adversarial corpus (**24** + **17** `#[test]`)	`companion/\G1B_P\CKB_INV_CORPUS.jsonl` sha256 `FD81EDF1...A3F1` / 7,393 B; `tests\adversarial.rs` sha256 `071DE35B...E677` / 19,188 B; `tests\transition_adversarial.rs` sha256 `328A4D5E...E006` / 19,896 B; `...\CLAIM_EVIDENCE_LEDGER.md:31, :43`	OBSERVED_SOURCE (**VERIFIED**, source-level) + **OBSERVED_RUNTIME** (41 passed / 0 failed, exit 0, 2026-09-18)	

to its rows (`CLAIM_EVIDENCE_LEDGER.md:33` vs `:8-32`) — recorded in the evidence-status table above and corrected in `CLAIM_EVIDENCE_LEDGER_v0_2.md`.

D. Appendix D: Source anchors for the main text

Under the 9-page submission budget the inline `file:line` anchors that accompanied individual sentences in §1–§9 were consolidated here. **Nothing was deleted or altered:** every entry below is the exact text that stood in the main text of the pre-consolidation revision (705 lines, sha256 `808221974a542012df41a395450e98aacd3f3f5b3ecbbf054a7eba06b62c77f4`), and the bracketed line numbers refer to that revision. Claim ids (`[C-xx]`, `[C-LME-xx]`) stay in the main text and continue to resolve through Appendix C.

Claim id	Claim (short)	Anchor	Evidence level
C-08f	E3 **fresh half**, withdrawn from the result surface: 28.1×, A→B ≈ 81×, B→C ≈ 0.3×, 180.71 / 2.22 / 6.43 ms/q — prose and hash-less table rows only	...\PAPER\EXPERIMENTS.md:53-71 (prose); ...\SUBMISSION_VIEW\TABLES_P6_P8_v0_1.md:11-14 (its source row :15 reads SHA [VERIFY]), sha256 8B6C499E34DC51D3... / 3,423 B; CLAIM_EVIDENCE_LEDGER_v0_2.md:29	OBSERVED_SOURCE (prose/rows) + **NOT_FOUND** (no receipt; the on-disk PERF_RECEIPT.txt files carry the frozen values only) ⇒ REPORT_ONLY
C-23	E2AP: evidence-retrieved-but-dropped-in-synthesis; fail-closed evidence-anchor gate; 5 of 7 adjudicated judge rejections were false negatives (5 judge disagreements / 1 benchmark source conflict / 1 real bug); one gold/source conflict. **New id** — this content previously shared the colliding id C-19 with the R3-band row	...\SUBMISSION_VIEW\PROSE_E2AP_FINDING.tex:4-7; ...\PAPER_FAMILY\LME500_EVALUATION_SECTION_v0_1.md:38-39; A4...:31, :121	OBSERVED_SOURCE; the "5 of 7" form is the audit-corrected wording; the parent doc **has been repaired at source** (A4...:119-122; sha256 199E3F3D...856F16, 4,568 B), so it now carries the corrected form; status **VERIFIED** (source-level)
C-20	Third-party replay does not exist on this machine ⇒ replay pending on this machine, ready label not claimed	A4_LME500_EMPIRICAL_AUDIT.md:164, :183, :192; ...\CLAIM_EVIDENCE_LEDGER.md:29	OBSERVED_TOOL_RESULT (searches) / NOT_FOUND
C-21	Novelty boundary: antecedents credited; closure claim only; no "missed causality"; capability inclusion not sibling. **Id collision resolved:** the E2AP material previously sharing C-19 now has its own id C-23, and the layer-separation claim has C-24	...\SUBMISSION_VIEW\PROSE_RELATED_WORK.tex:23-27; ...\PAPER\RELATED_WORK.md:60-68; A5...:59-88; collision surfaces A2...:100-115	OBSERVED_SOURCE + OBSERVED_TOOL_RESULT; novelty judgement INFERENCE on a limited sample

§1

- **[1:29]** (anchor line) *Anchor for the M1 wording above:* the paragraph is the frozen dry-run text [workspace]/companion/\SUBMISSION_VIEW\PROSE_INTRO_HYPOTHESIS.tex:4, carried over verbatim in substance from [workspace]/companion/\PAPER_FAMILY\INTERNAL_MANUSCRIPT_SPINE.md:28 (M1 = persistent memory as revisable epistemic state). Evidence level: OBSERVED_SOURCE.
- **[1:40]** (anchor line) *Anchor:* [workspace]/companion/\PAPER_FAMILY\INTERNAL_MANUSCRIPT_SPINE.md:59-70 (§M0 Scope note; the two-line ontology is at :67-68) and [workspace]/companion/\SUBMISSION_VIEW\PROSE_INTRO_HYPOTHESIS.tex:6. Evidence level: OBSERVED_SOURCE. Note the contrast with retrieval is offered as **evidence, not classification** — being compared to a class is not being a member of it (INTERNAL_MANUSCRIPT_SPINE.md:70).
- **[1:55]** (inline anchor) (companion/\PAPER\EXPERIMENTS.md, EFFECT_MEASURE_001_FREEZE*.md, TABLES_P6_P8_v0_1.md)

§2

- **[2:77]** (anchor line) *Anchors:* A5_PRIOR_ART_AND_POSITIONING.md:40-55 (the 15-row map), :70-74 (§2.2 novelty judgement per item). Evidence level: OBSERVED_TOOL_RESULT. Two bibliographic honesty notes carried from the same audit: the CMMI technical-report identifier is **NOT_FOUND** (A5...:300-308), and Pearl's 2009 *Causality* (2nd ed.) book entry is **NOT_FOUND** (A5...:310-313), so the causal

Claim id	Claim (short)	Anchor	Evidence level
C-24	Three-layer separation: retrieving evidence / maintaining an epistemic state / preserving evidence through synthesis — failure at one layer is not repaired by performance at another; E2AP is the instance, the E9 null is the other side	`...\SUBMISSION_VIEW\PROSE_DISCUSSION_LAYERS.tex:2; PROSE_E2AP_FINDING.tex:4; EFFECT_FINDING_DISCUSSION.md:38-44`	OBSERVED_SOURCE (the two findings) + **INFERENCE** (their joint reading; **no** three-way ablation exists); status **REPORT_ONLY** (framing claim; new id)
C-LME-01	LongMemEval-S = 500 public-development questions, five abilities, gold sha `d6f21ea9...3a442`; not named a hop-count causal benchmark by its authors	`A4_LME500_EMPIRICAL_AUDIT.md:21; A5_PRIOR_ART_AND_POSITIONING.md:40, :346`	OBSERVED_TOOL_RESULT
C-LME-02	Frozen official gpt-4o-2024-08-06 scored **493/500 (98.6%)** on the frozen predecessor candidate sha `e13f866d...9a64` (7 wrong of 500); per-category ku 78/78, ssu 70/70, tr 133/133, ssa 56/56, ms 130/133, ssp 26/30	`A4...:27-28`	OBSERVED_TOOL_RESULT + OBSERVED_SOURCE (independently recounted from the raw JSONL by the A4 empirical audit)
C-LME-03	**post-hoc hybrid evidence-adjudicated 500/500** = 493 frozen-judge accepts + 7 source-level adjudications; decomposition 5 judge disagreements / 1 benchmark source conflict / 1 real bug; `rule_defined_posthoc: true`, `fully_manual_blind_review: false`	`A4...:31, :70, :89, :137; ...\LME500_EVALUATION_SECTION_v0_1.md:31-35`	OBSERVED_SOURCE + OBSERVED_TOOL_RESULT (493+7=500 re-verified)
C-LME-04	Accepted successor (sha `c23e2f0b...c63557`) changed exactly 1 row (`a89d7624`); its official evaluator output is **0 bytes** ⇒ `successor_official_score_available: false`	`A4...:32, :68, :74, :147`	OBSERVED_TOOL_RESULT / NOT_FOUND (by design)

hierarchy is carried by the fully verified 2019 CACM article. No DOI, page range, or venue is invented anywhere in this paper.

- **[2:85]** (inline anchor) (`[repo]/README.md:1189`)
- **[2:89]** (inline anchor) (`INTERNAL_MANUSCRIPT_SPINE.md:64`)
- **[2:90]** (inline anchor) (`INTERNAL_MANUSCRIPT_SPINE.md:89-90`)
- **[2:95]** (inline anchor) (`PROSE_RELATED_WORK.tex:23-27`)

§3

- **[3:103]** (anchor line) *Anchor:* `[workspace]/companion/\SUBMISSION_VIEW\PROSE_ARCHITECTURE_LAYERS.tex:3-4` (frozen dry-run text), mapping to `INTERNAL_MANUSCRIPT_SPINE.md:28` (M3 = the two-layer separation is a formal term pair). Evidence level: OBSERVED_SOURCE. The two terms — *Epistemic Constitution* and *Causal Dynamics* — are formal terms of this framework, not borrowed labels.

§5

- **[5:156]** (inline anchor) (`A2_CKB_CODE_IMPLEMENTATION_AUDIT.md:94-96`)

Claim id	Claim (short)	Anchor	Evidence level
C-LME-05	Luna comparability snapshot **500/500** on the same frozen candidate (ledger sha `a2b095e7...4a0d`, one question per context, 500 singleton batches); 7 disagreements, `luna_yes_gpt4o_no`=7 / `luna_no_gpt4o_yes`=0; **no consensus label**; judge-gaming rewrite rejected (`REJECTED_DIAGNOSTIC_JUDGE_DIRECTED_REWRITES`, 429)	`A4...:29-30, :39, :73, :106-110`	OBSERVED_TOOL_RESULT + OBSERVED_SOURCE
C-LME-06	Boundary: public development set, `NOT_HIDDEN_NOT_GENERALIZATION` / `PUBLIC_DEVELOPMENT_SET_NOT_GENERALIZATION`; not an official score, not a blind manual review, not consensus, not generalization, ready label not claimed; diagnostic partition 250/250 (gold-aware post-hoc dev/regression split of the same 500 public questions, not a scoring surface)	`A4...:41, :135, :139, :154-155`	OBSERVED_TOOL_RESULT
C-LME-07	Internal lineage only: M5 RUN2 **445/500** (`answer_model: deepseek-v4-flash`) with **55** wrong rows of 500 (28+27); M6-SOL **449/500** (`gpt-5.6-sol`, high) receipt-only with **51** wrong rows of 500 (22+29); taxonomy 24/19/8 over **51 of the 500** questions; 55∩51 = 38 (13 matrix-only, 17 M5-only, union 68); raw M6 output absent; generator of the scored predecessor/successor **UNKNOWN**	`A4...:25-26, :33-37, :64, :140, :165; M5_4096_RAW_FREEZE.json:8; M6_SOL_RESULT_RECEIPT.json:45`	OBSERVED_SOURCE / OBSERVED_TOOL_RESULT; `UNREAD` raw output; **UNKNOWN** generator for the scored candidate; **never merged with C-LME-03**

- **[5:168]** (anchor line) *Anchor:* `PROSE_p2_p7_v0_1.md:76-83`; genome/invariant corpus VERIFIED (`CLAIM_EVIDENCE_LEDGER.md:31`, C-22: 15 architecture-genome invariants, CAL-rev0.5 contract, G1b adversarial corpus); implementation source OBSERVED_SOURCE with execution `OBSERVED_RUNTIME` (**41/41** Rust tests and **16/16** conformance checks, exit 0, 2026-09-18 — see the execution-status paragraph above); `A2...:45` still excludes these paths from its core-code scope.
- **[5:172]** (inline anchor) (`DISCUSSION.md:95-97`)

§6

- **[6:182]** (inline anchor) (`PROSE_p2_p7_v0_1.md:57-60`)
- **[6:194]** (inline anchor) (`EXPERIMENTS.md:72-75`)
- **[6:211]** (inline anchor) (`[repo]/CKB_TYPED_RUNTIME\CANDIDATE_RECEIPT.json`)
- **[6:213]** (inline anchor) (`PROSE_p2_p7_v0_1.md:85-114`)

Status	Count	What is counted (basis)	Anchor
FROZEN	**8 → 7**	Counting basis: literal first token of the Status column in ...\SUBMISSION_VIEW\CLAIM_EVIDENCE_LEDGER.md:8-34. **v0.1 = 8**: C-01, C-02, C-08, C-11, C-12, C-17, C-18, C-21 (C-01/C-02 are FROZEN (draft)). **v0.2 = 7** = v0.1 minus C-12, which moves to VERIFIED (its EVIDENCE_REGISTRY.json anchor is **NOT_FOUND**); the C-08 split leaves the frozen half FROZEN and creates the separate REPORT_ONLY row C-08f, and the two new rows C-23/C-24 enter as VERIFIED/REPORT_ONLY	CLAIM_EVIDENCE_LEDGER.md:10-11, :17, :20-21, :27, :32-33; CLAIM_EVIDENCE_LEDGER_v0_2.md §3
VERIFIED	**11 → 12**	v0.1 = 11: C-03, C-04, C-05, C-06, C-07, C-09, C-10, C-13, C-14, C-15, C-22 (C-06's qualifier is restated where the label is used). v0.2 = 12 after adding C-12 and the new C-23 (E2AP) and removing C-15	CLAIM_EVIDENCE_LEDGER.md:12-16, :18-19, :22-24, :31, :33; CLAIM_EVIDENCE_LEDGER_v0_2.md §3

§7

- **[7:225]** (inline anchor) (EXPERIMENTS.md:97-106)
- **[7:225]** (inline anchor) (PROSE_P6_EMPIRICS.tex:3-12; EXPERIMENTS.md:95-106)
- **[7:227]** (inline anchor) (TABLES_P6_P8_v0_1.md:43-45)
- **[7:227]** (inline anchor) (EXPERIMENTS.md:27-44, :190-203)
- **[7:229]** (inline anchor) (EXPERIMENTS.md:36 vs TABLES_P6_P8_v0_1.md:38)
- **[7:233]** (inline anchor) (PREREGISTRATION_FREEZE.json:29-37)
- **[7:243]** (inline anchor) (CORRECTIVE_REPLICATION_001_RECEIPT.md:20-24)
- **[7:245]** (inline anchor) (PREREGISTRATION_FREEZE.json:30-33)
- **[7:247]** (inline anchor) (PREREGISTRATION_AMENDMENT_001.md:9-17, :24-29, :38-43)
- **[7:267]** (inline anchor) (EFFECT_RECEIPT.md:16-17, :54-66)
- **[7:267]** (inline anchor) (EFFECT_RECEIPT.md:54-66; CORRECTIVE_REPLICATION_001_RECEIPT.md:5-9, :44-50)
- **[7:271]** (inline anchor) (EFFECT_RECEIPT.md:40-43)
- **[7:271]** (inline anchor) (EXPERIMENTS.md:180-186; EFFECT_FINDING_DISCUSSION.md:38-44)
- **[7:281]** (inline anchor) (M6_SOL_RESULT_RECEIPT.json:45; A4...:47)
- **[7:281]** (inline anchor) (...\posthoc\M5_RUN\M5_4096_RAW_FREEZE.json:8)
- **[7:283]** (inline anchor) (M5_4096_RAW_FREEZE.json; M6_SOL_RESULT_RECEIPT.json; DUAL_JUDGE_RECEIPT.json)

Status	Count	What is counted (basis)	Anchor
REPORT_ONLY	**1 → 4**	v0.1 = 1: C-16 (Milestone A review REVIEW_INCOMPLETE, manual report only, no receipt field). v0.2 adds the Milestone A artifact row (C-15; the ledger's own wording at that time was "receipt unreproduced", superseded in this manuscript by payload_drifted after the receipt was independently re-verified as internally consistent and only its payload found to have drifted), the fresh-recall half of C-08 (C-08f) and the new C-24 (three-layer separation — framing only, INFERENCE, no three-way ablation). The E8 temporal surface is **also** reviewer-receipt-only — its ledger status is VERIFIED while its *provenance* is "reviewer receipts, zero CI" (EXPERIMENTS.md:147-153); it is not double-counted here	CLAIM_EVIDENCE_LEDGER.md:25, :33; CLAIM_EVIDENCE_LEDGER_v0_2.md §3; MILESTONE_A_STATUS_LOG.md:8; A2...:258-262
VERIFY	**1 → 1**	C-19 (R3 named-model calibration bands): the band figures were **not located** in the report, so the row stays VERIFY **unresolved** and this paper uses **no** R3 band value. The digits formerly carried by the v0.1 ledger row :28 are **withdrawn rather than restated** (author decision, 2026-09-18), and no R3 band figure is printed anywhere in this manuscript (new Appendix C row C-19)	CLAIM_EVIDENCE_LEDGER.md:28, :33; MILESTONE_A_STATUS_LOG.md:36-40

- **[7:285]** (inline anchor) (LME500_EVALUATION_SECTION_v0_1.md:11-14)
- **[7:285]** (inline anchor) (LME500_EVALUATION_SECTION_v0_1.md:9)
- **[7:285]** (inline anchor) (LME500_EVALUATION_SECTION_v0_1.md:6-9; A4...:41)
- **[7:301]** (inline anchor) (LME500_EVALUATION_SECTION_v0_1.md:39, A4...:121)
- **[7:301]** (inline anchor) (PROSE_E2AP_FINDING.tex:4; LME500_EVALUATION_SECTION_v0_1.md:38)
- **[7:303]** (inline anchor) (A4...:154-155, :164; INTERNAL_ANONYMITY_CHECK.md:27)

- **[7:305]** (inline anchor) (EVIDENCE_ADJUDICATION_RECEIPT.json:48)

§8

- **[8:315]** (inline anchor) (EFFECT_FINDING_DISCUSSION.md:38–44)
- **[8:315]** (inline anchor) (PROSE_E2AP_FINDING.tex:4)
- **[8:319]** (inline anchor) (DISCUSSION.md:98–100)
- **[8:319]** (inline anchor) (DISCUSSION.md:60–63)
- **[8:320]** (inline anchor) (PROSE_LIMITATIONS.tex:3–5)
- **[8:321]** (inline anchor) (PROSE_LIMITATIONS.tex:5–7)
- **[8:322]** (inline anchor) (PROSE_LIMITATIONS.tex:7–10)
- **[8:323]** (inline anchor) (PROSE_LIMITATIONS.tex:10–12)
- **[8:324]** (inline anchor) (PROSE_LIMITATIONS.tex:12–13)
- **[8:325]** (inline anchor) (PROSE_LIMITATIONS.tex:13–17; A4...:135)
- **[8:326]** (inline anchor) (DISCUSSION.md:64–78)
- **[8:327]** (inline anchor) (EXPERIMENTS.md:147–153, MILESTONE_A_STATUS_LOG.md:41–42)
- **[8:333]** (inline anchor) (ANALYSIS_SET_CORRECTION_001.json)
- **[8:338]** (inline anchor) (DISCUSSION.md:85–87)
- **[8:339]** (inline anchor) (DISCUSSION.md:88–94)
- **[8:340]** (inline anchor) (DISCUSSION.md:95–97)
- **[8:341]** (inline anchor) (DISCUSSION.md:98–100)
- **[8:347]** (inline anchor) (MILESTONE_A_STATUS_LOG.md:48–49)
- **[8:347]** (inline anchor) (PROSE_DISCUSSION_LAYERS.tex:4–5)

Counts. 5 whole anchor lines and 53 reference-only parentheticals were consolidated (58 entries, 4,223 characters). Parentheticals that carry substantive nouns — for example those listing an invariant count or a contract name — were **not** moved and remain in the main text, where their claims are still stated.

Status	Count	What is counted (basis)	Anchor
PACK_PENDING	**1 → 1**	C-20 third-party independent benchmark evidence (no replay package found on this machine)	CLAIM_EVIDENCE_LEDGER.md:29, :33; A4...:164
draft (in addition to FROZEN)	**2 → 3**	Rows that also carry draft: v0.1 C-01/C-02; v0.2 adds C-21 (novelty boundary, FROZEN (draft framing)); counted once each, under FROZEN	CLAIM_EVIDENCE_LEDGER.md:33; CLAIM_EVIDENCE_LEDGER_v0_2.md §3
counting basis check	—	v0.1 = 22 C- rows and 8 + 11 + 1 + 1 + 1 = 22 on the literal basis. The ledger's own summary line prints FROZEN 6 · VERIFIED 12 · REPORT_ONLY 1 · VERIFY 1 · PACK_PENDING 1 · draft 2, which does **not** sum to its own 22 rows: the summary line is recorded as a **defect of the ledger**, not as a second basis. v0.2 carries **25 framework rows** with 7 + 12 + 4 + 1 + 1 = 25, and its §3 also prints a **whole-file** count of 37 rows (25 framework + 12 LongMemEval-S rows, whose extra PACK_PENDING is C-LME-10)	CLAIM_EVIDENCE_LEDGER.md:33–34 vs the 22 rows at :8–32; CLAIM_EVIDENCE_LEDGER_v0_2.md §3
NOT_RUN (execution-status basis, not a ledger status)	**2 → 0**	Both previously NOT_RUN surfaces are now executed. (i) **R4E suite**: 287 tests collected in 1.93s, then **287 passed / 0 failed, exit 0** (frozen snapshot plus the disclosed M6 reconstruction). (ii) §5 G1B_P corpus**:** 41 #[test] attributes → **41 passed / 0 failed, exit 0**, with conformance_test.py at 16 [PASS] / ALL PASS, exit 0. Still not executed anywhere: the repository's own self-reported 769-test claim, which is **not collectable** (pytest **exit 2**, 717 collected + 16 errors) and remains a different surface from both of the above	A2_CKB_CODE_IMPLEMENTATION_AUDIT.md:137, :161, :272, :45; A1_CKB_REPO_DOC_AUDIT.md:41, :317; run logs g1bp_test.log, g1bp_conformance.log, pytest_r4e_full_after_fixture2.log