# TWO COUNTS OVER TEXT: WHAT A MEMORY BENCHMARK CAN SEE, MEASURED BEFORE ANYTHING IS RUN

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A long-term memory benchmark is cited for an architectural distinction: a system that maintains a composable state against one that indexes what it was told. That citation is never checked against the benchmark's own text, although it can be. Where an item's gold answer is announced verbatim inside a single logged record, returning that one record suffices and no composition is being tested – but auditing this today means running systems, a judge and an API bill, so it happens after a benchmark ships rather than before. We build an instrument that needs no model, no judge and no API call: two counts over text and a retriever. $V$ is the share of items whose gold answer appears verbatim in one record; $R^{\mathrm{str}}(k)$ is the share of that containment gap a budget-$k$ retriever fails to close, and $k^*$ is the budget where it closes, the audit's one free constant swept and priced within $0.008$ of optimal. A proved proposition makes the instrument one-sided: it can condemn a task as blind and can never acquit one. Run blind on four corpora, $k^*$ recovers class labels fixed by role before any dataset was named – single-hop dimensions close at $8$, $8$ and $16$, composition dimensions do not close at $32$, and a needle task returns the $k^*{=}1$ a working retriever must return. On a flagship benchmark built to be immune to context stuffing, $V$ runs $0.55$–$0.64$ on `knowledge_update` against $0.00$–$0.07$ elsewhere: more than half its checkable update answers are announced verbatim in one turn, and across a $74\times$ range of that benchmark's own headline axis the counts barely move. Priced on a corpus built to defeat it, $V$ travels while $k^*$ does not, since $k^*$ also measures a reader's lexical reach. The reporting rule is one-sided too: a high $V$ condemns, a low reader score proves nothing, and authors can publish $V$ and $k^*$ per dimension before shipping, at the budget the baseline was allowed.

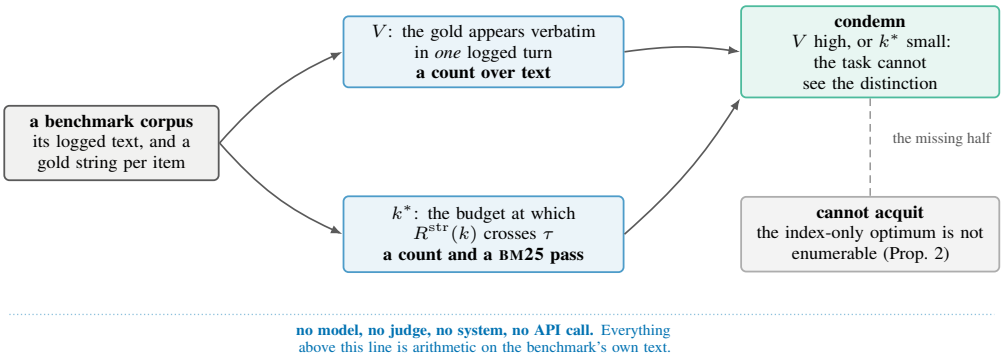
## 1 INTRODUCTION

A long-term memory system is built so that an agent can hold state across a conversation, and benchmarks are built to say whether it does: BEAM (Tavakoli et al., 2025), STATE-Bench (Microsoft Open Source, 2026), RECON (Arya, 2026), STALE (Chao et al., 2026) and the precision-aware line (Flynt, 2026), among others. The systems they rank make architectural claims that a benchmark score settles: Chhikara et al. (2025) compare ten memory approaches on one such benchmark, and Xu et al. (2025) organise memory as linked notes revised as new information arrives, reporting large gains on multi-hop questions. Between claim and score sits a question rarely asked of the benchmark itself: *can this task see the distinction it is cited for?* A task can be difficult, well curated and widely reported and still assign the same score to a system that maintains a composable state channel and one that merely indexes what it was told. If so, no care in running the benchmark recovers the distinction, which was never in the measurement.

This paper's contribution is that the question is answerable *from the benchmark's own text*, cheaply, before any system is evaluated. The theory of the underlying quantity, a task's *resolving power* on the state axis, belongs to a companion paper (Anonymous, 2026d); what this paper takes from it is restated in §3 and proved where load-bearing, so nothing here rests on work a reader cannot reach. What a rubric judge inserts between these counts and any *scored* verdict is a second companion's subject (Anonymous, 2026g).



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.




**Read the two halves of the right-hand column together.** A benchmark that fails the audit is shown not to resolve the distinction it is cited for, and that verdict is final. A benchmark that passes has not been shown to resolve it: the audit is existential, and §4 proves that no run of it can be otherwise. *What makes the asymmetry worth accepting is the cost line*, because an instrument this cheap can be run on every dimension of every corpus before a single system is evaluated.

**Figure 1. The instrument in one picture, and the asymmetry it cannot be run out of.** A benchmark corpus goes in; two counts over its own text come out. $V$ is the share of items whose gold value appears verbatim inside a single logged turn, and $k^*$ is the retrieval budget at which the share still missing crosses $\tau$. Neither needs a model, a judge, a system or an API call. What the counts support is one-sided: a benchmark they condemn does not resolve the distinction it is cited for, and a benchmark they do not condemn has not been shown to resolve it, because the index-only optimum is not enumerable (Prop. 2).

**Contributions.** **1.** An instrument that needs no model, judge, system or API call (§3–§4): $V$, the curve $R^{\mathrm{str}}(k)$, and the crossing budget $k^*$. Its one free constant is swept and shown to sit within $0.008$ of the best realisable setting; its one-sidedness is a proposition, not a caveat (Prop. 2). **2.** Blind validation on three corpora fixed by role before any dataset was named (§6): $k^*$ recovers the benchmark's own single-hop/composition labels and returns $k^* = 1$ on a needle task, a check a broken retriever cannot pass structurally. **3.** The instrument's lexical assumption, priced on a corpus built to break it (§6): every item is single-event answerable ($V \approx 1$) and the audit's own reader retrieves the needle at *no* budget up to 32, so $k^*$ does not resolve where BEAM's resolves at $8$ – the count condemns and travels; the budget read-out is a property of the reader. **4.** A four-rung audit of a deployed benchmark on its own headline axis (§5–§5): $V = 0.55$–$0.64$ on the updating dimension against $0.00$–$0.07$ elsewhere at three registered scales; across all four shipped rungs ($74\times$ in events) $V$'s fall is undecidable at shipped sizes, the ceiling barely moves, and the axis moves the budget's *reach* rather than the score. **5.** The strongest index-only adversary (§7): a budget ladder $\rho(k)$ under the benchmark's own answering prompt – two dimensions clear a frozen $0.80$ bar at $k \leq 16$, reported as *retrieval sufficiency*, not resolving power, since both sides are index channels. **6.** A channel classifier that measures instead of reading names (§8): walking the cache a runtime returns at two sequence lengths separates bounded-width carriers, indexes and local-attention windows, where a name-based check misfiles a recurrent architecture whose weights happen to be named `attention`.

The instrument's verdicts are counts, not scores: counts cannot bound the *scored* gap between two systems without a judge-mediated step priced by a companion paper (Anonymous, 2026g), and §10 states how far each claim here extends without it. The scorecard of all 60 graded claims (37 for, 23 against) is in Appendix J.

## 2 Related Work

**One paper states this paper's premise before us, in prose.** Tanguturi (2026) analyse seven memory evaluations against a property checklist and conclude, of the two this paper's corpus comes from, that "a language model with a context window large enough to fit the full benchmark passes both of these evaluations with no memory architecture at all", so "by construction" such benchmarks "cannot distinguish a memory system from a long-context LLM". That is the blindness claim, asserted rather than measured, and it is prior to us. What this paper adds is arithmetic and scope: theirs is a structural argument about context-window *sufficiency*, scoring a benchmark $0/7$ on a checklist; ours is a per-item count with a budget in it, so it applies where the context does *not* fit.

**Three independent results say the phenomenon is there, none of them using our instrument.** Sidik & Rokach (2026) score a tiered memory architecture *identically* to its unmodified baseline on LoCoMo ($\Delta F_1 = -0.005$) and read it as the blindness contrapositive; Li et al. (2026a) supply the arm our baseline is built to be – an agent-controlled *lexical* search over the unmodified archive, no structure built at all – reaching the highest mean accuracy of any system they compare ($58.2$ against $53.2$ for the strongest graph-based memory) under a matched backbone; and Wang (2026) vary one component at a time on LongMemEval-S, finding verbatim RAG statistically indistinguishable from full context ($47.2\%$ against $49.8\%$, $p = 0.34$) while swapping only the embedding model moves accuracy by $6.2$ points. Each is what $V$ and $k^*$ would have said before they ran anything. Yu et al. (2026) separate write-side from retrieval-side loss under an oracle-evidence condition – $R^{\mathrm{str}}(k)$'s decomposition on a different cut – and Huang et al. (2026a) find no substrate dominates under one harness.

**Contamination filtering removes one route to blindness and leaves others.** Liu et al. (2026) generate benchmarks whose questions cannot be answered from parametric knowledge alone, and Huang et al. (2026b) address contamination of the retrieval corpus by model-generated text. A leakage-free item is unanswerable without the state; the counts find a further class – answerable by returning the last event containing the gold, with no model at all (§5) – that passes the leakage filter while separating nothing a substring search cannot separate. Filtering on *parametric* answerability and filtering on *index-only* answerability are different filters.

**The closest prior programmes are procedural, and each states an asymmetry that stops one step short of ours.** Bean et al. (2025) review $445$ benchmarks with $42$ authors and report contested construct definitions, convenience sampling in $27\%$ and statistical tests in only $16\%$; Raji et al. (2021) make the sharper qualitative version of the point. Our contribution to that line is arithmetic rather than rhetoric. Kalaitzidis (2026) propose deriving evaluation criteria from the capability a benchmark claims, then auditing whether it discriminates that capability from proxy behaviours; our Prop. 2 says such an audit is one-sided on any benchmark harvested from natural text – condemnation is exhibitable (one index-only reader scoring high is a witness) while acquittal quantifies over every $k$-subset of the span and every function on it ($\binom{188}{8} \approx 10^{13}$ subsets on one BEAM conversation), a space no finite audit enumerates. The design half of their programme survives the limit intact, since Prop. 2's escape clause *is* construction. Li et al. (2026b) audit the audits, naming five classes of pipeline failure and shipping a six-point due-diligence gate, and state a caveat of their own: a positive verdict is necessary but not sufficient for construct sensitivity – a limitation of an instrument, discharged by a better gate. Prop. 2 is instead a statement about the quantifier, and no gate closes it. Their five classes are also why a reader should not take our own $0/10$ (§10) on trust, and the answer is that §5's counts need no model, judge or API call: the implementation detail a reader cannot see is a download away from being seen.

**What a judge-free instrument is immune to, named rather than deferred.** The judge-side literature names three defects, and this instrument is outside all three for one structural reason. *Bias*: a rubric judge carries position, verbosity and self-enhancement preferences (Zheng et al., 2023), and a count over text has none to carry. *Unreliability*: a score moves when only the evaluator changes (Yang et al., 2026), with judge rotation as the remedy (Zhu et al., 2026), where $V$ and $k^*$ have no evaluator to rotate. *Reliability that is not validity*: Norman et al. (2026) separate agreement from validity across roughly $541$K judgments and find exact-match agreement overstating discriminative ability, which is Jacobs & Wallach (2021)'s construct-validity point with a measurement behind it. The third is the one that matters here, because a perfectly reliable judge is a perfectly agreed measurement of whatever the rubric operationalises, while this paper's counts are about whether the item contains the thing at all. The immunity is bought, not free: a count cannot bound a scored difference, and the companion judge paper (Anonymous, 2026g) prices that step – on one benchmark the credit rate under perfect provenance is $0.614/0.452/0.629$, and the scorer's own share of that shortfall is $0.028$.

**The memory benchmarks are the object of study.** BEAM (Tavakoli et al., 2025), STATE-Bench (Microsoft Open Source, 2026), RECON (Arya, 2026), STALE (Chao et al., 2026), Supersede (Patel, 2026) and the precision-aware line (Flynt, 2026) are what this machinery audits, and several already move in the direction the counts reward, scoring whether an agent improves with experience or notices that a memory has been invalidated rather than whether it retrieves. Panthi & Abdelfattah

(2026) report the closest scoring-side result: changing only the credited target moves nDCG on 83.4–94.0% of shared queries. Our contribution to this group is not another benchmark; it is the pre-release report their authors can generate from their own text.

## 3 DEFINITIONS AND TWO PROPOSITIONS

Three definitions are imported from the companion theory paper (Anonymous, 2026d) and restated so this paper is self-contained; the two propositions the audit stands on are proved here.

**Definition 1** (Channels and budget, restated)**.** A *state channel* is a persistent cross-token carrier whose *width* is bounded independently of the span it serves. An *index channel* is an addressable store growing with the span it serves, read query-dependently under a per-query *access budget* of $k$ records.

What "index-only" fixes is not the span but what is *accessible* from it, a capacity in the sense of Cover & Thomas (2006) and strictly smaller: Wang & Sun (2025) show proactive interference degrading retrieval from a context the model nominally contains.

**Definition 2** (Resolving power, restated)**.** For a task $\mathcal{T}$ and budget $k$, $R_{\text{state}}(\mathcal{T}; k) = S_{\mathcal{T}}(\text{access-complete oracle}) - S_{\mathcal{T}}(\text{index-only oracle at budget } k)$, each oracle optimal within its channel class. $R_{\text{state}} = 0$ says the task is *blind* on the state axis.

**Definition 3** (The counts)**.** Over items with a short gold: $V$ is the share whose gold value appears *verbatim* inside a single logged record, a string property of the corpus with no model, judge, reader or ceiling in it. Writing $a_k$ for the share of items whose gold is contained in what a budget-$k$ retriever returns and $a_{\text{complete}}$ for the share recoverable from one record at all, $R^{\text{str}}(k) = a_{\text{complete}} - a_k$ is the share of that gap a budget-$k$ retriever fails to close, and

$$k^* = \min\{k \in K : R^{\text{str}}(k) \le \tau\}, \qquad \tau = 0.05, \quad K = \{1, 2, 4, 8, 16, 32\} \tag{1}$$

is the least budget on this fixed ladder at which the gap falls below a frozen threshold; $V$ generalises to the $k$=1 point of the same curve. The ladder makes $k^*$ a discontinuous read-out, and Schaeffer et al. (2023) show a discontinuous scoring rule can by itself produce an apparent sharp transition in a smooth quantity, so we read a one-step difference as resolution and not as effect. The budget limit itself – record length to infinity at pinned budget – is a companion paper's (Anonymous, 2026e) subject, not this one's.

**Proposition 1** (Blindness, restated)**.** *If the correct answer to every query in $\mathcal{T}$ is a deterministic function of the write record, the score distributions of any two systems differing only in the presence of a state channel coincide, and $I(\text{state channel present}; S_{\mathcal{T}}) = 0$.*

*Proof.* An index-only system can compute any function of the write record; if the answer is such a function its optimum attains the ceiling, so both terms of $R_{\text{state}}$ are equal. □

## 4 THE INSTRUMENTS AND THEIR ONE-SIDEDNESS

Two deliberately weak readers bound $R_{\text{state}}$ from below: a *last-occurrence* reader, which returns the most recent single event mentioning the query term verbatim with no composition, and *BM25-last*, which scores every event by BM25 and returns the most recent event within a fraction $c$ of the maximum. That fraction is a free parameter, priced rather than asserted: swept over $\{0.5, \ldots, 1.0\}$ on every checkable item, the returned event is identical across the sweep on only 3.4–11.6% of items and containment moves by up to 0.101, yet the frozen constant $c$=0.8 sits within 0.008 of the best realisable setting on every split. Both readers are realisable index-only strategies, so $\max_i S(\text{reader}_i) \le S(\text{index-only optimum})$ and

$$R_{\text{state}}(\mathcal{T}) \;\le\; S_{\mathcal{T}}(\text{ceiling}) - \max_i S_{\mathcal{T}}(\text{reader}_i). \tag{2}$$

**The audit condemns; it does not acquit.** A reader near the ceiling *proves* blindness; a reader far below it proves nothing, since a better index-only strategy may exist. Condemning is existential (exhibit one strategy that reaches the ceiling); acquitting is universal (show that *no* index-only strategy does), and an index-only system with budget $k$ may compute any function of the $k$ records it retrieves.

**Proposition 2** (Acquittal needs construction)**.** *A task family can be shown to have $R_{\text{state}} > 0$ only if the dependency structure of its items is known by construction, i.e. only if it was* built *so the answer provably depends on more of the span than any admissible retrieval can hold. For a task family harvested from natural text, the universal quantifier ranges over a space that cannot be enumerated, so no finite audit acquits it.*

Two consequences. **(i)** A failure to condemn with a $k{=}1$ reader is the only outcome available short of condemnation; a budget ladder answers what one-event instruments cannot – at $k{=}8$, containment coverage on the flagship's updating dimension rises from $0.194$ to $0.611$, its entire anywhere-reachable set. **(ii)** A sighted benchmark has to be *built*, not audited into existence: plant a value once, announce subsequent changes as operations rather than as resulting values, and make the answer their ordered composition. That is what the parent framework does (Anonymous, 2026a), and what §5 finds the flagship's updating dimension does *not* do.

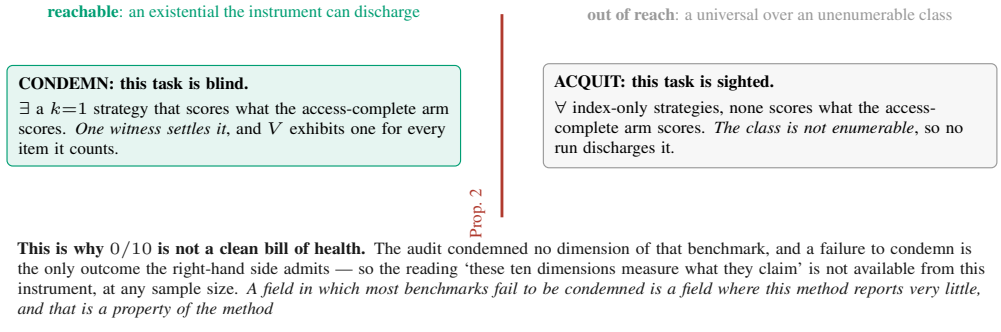



**Figure 2. The instrument is one-sided, and the asymmetry is a quantifier.** Condemning a task is existential, since one $k{=}1$ strategy that matches the access-complete arm settles it, and $V$ exhibits one for every item it counts. Acquitting is universal over a class nobody can enumerate, so no run discharges it. This is Prop. 2, and it is why a low reader score below is not a clean bill of health: a failure to condemn is the only outcome the right-hand side admits, at any sample size.

## 5 The Audit of a Deployed Benchmark

Proposition 1's condition is checked directly on text: does the gold appear verbatim inside a single conversational turn? Verbatim containment *undercounts* single-event answerability, so these are lower bounds; high rates condemn, low rates prove nothing. On the dimension that tests whether a system tracks an updated value, the majority of checkable answers are stated verbatim in one turn – the update is *announced* – at three registered scales ($V = 0.55/0.64/$ third-scale, against dashed bars of $\geq 0.5$ frozen before the splits existed) and composition dimensions stay below a frozen $\leq 0.25$ bar throughout.

**All four shipped rungs of the headline axis.** A fourth rung exists at $74\times$ in events (not tenfold: median turns per conversation run $271/1076/2088/20075$, turn length flat), and running it changes no verdict while adding one finding. $V$ falls monotonically on every rung, $0.377/0.349/0.325/0.295$ pooled and $0.611/0.636/0.547/0.438$ on the updating dimension, but the fall is not decidable at shipped sizes: the largest-rung drop is $0.90$ of its own standard error against the $2\Delta$ a band must clear under the companion rerun paper's rule (Anonymous, 2026f). Seven of ten dimensions' ceiling scores move by less than $0.04$ across the tenfold range shipped as three splits, and a fixed 32-event budget closes the gap at the smallest rung while leaving a residual at the largest two. **The context-length null reproduces with no instrument in it**: $V$ and $R^{\text{str}}(1)$ run the same pattern across all four rungs, on a quantity containing no model, no judge and no system. **The gap does not close at scale the way it closes at the smallest rung**: $R^{\text{str}}(32) = 0.000$ at the smallest split; at the two larger splits the same budget leaves a residual pooled and on the updating dimension – what the axis moves is the budget's reach, not the score.

**The benchmark's scaling axis, registered and refuted.** We predicted that an $O(1)$-state reader's precision decays as the haystack grows, froze three criteria before the 500K and 1M splits existed,

**Table 1. The count route, $(1-V)\cdot$ceiling, against the reader route of Eq. (2), on the benchmark's smallest split, same ceilings.** Lower is tighter. The count route needs the judge-mediated step a companion paper (Anonymous, 2026g) prices; the reader route needs only realisability. On nine dimensions the readers bound resolving power better; on the tenth, the one built to test updating, the benchmark's own text does, by $40\%$.

| dimension | $V$ | ceiling | $(1-V)\cdot$ceiling | ceiling − best reader |
|---|---|---|---|---|
| `knowledge_update` | **0.55** | 0.512 | **0.231** | 0.387 |
| `multi_session_reasoning` | 0.07 | 0.507 | 0.469 | **0.325** |
| `information_extraction` | 0.03 | 0.738 | 0.719 | **0.614** |
| `temporal_reasoning` | 0.00 | 0.487 | 0.487 | **0.462** |
| `summarization` | 0.00 | 0.335 | 0.335 | **0.292** |
| `event_ordering` | 0.00 | 0.437 | 0.437 | **0.157** |
| `preference_following` | 0.00 | 0.700 | 0.700 | **0.212** |
| `abstention` | 0.00 | 0.550 | 0.550 | **0.400** |
| `instruction_following` | 0.00 | 0.550 | 0.550 | **0.238** |
| `contradiction_resolution` | 0.00 | 0.219 | 0.219 | **0.113** |

and ran a ceiling arm, the readers, and 2100 fresh judged items on both unseen splits with the judge held fixed. **All three criteria failed, and the reason is that nothing moves**: not the reader, not the ceiling, not their ratio $\rho$ ($0.235/0.373/0.213$ against a frozen $\leq 0.185$ at the largest split), not the judge-free retrieval hit rate. The excess $E = (1-\rho)/(1-V)$ sits at 1.57–1.70 across all three scales rather than rising. **The thesis survives by a route the registered mechanism did not predict**: on this axis nothing crosses anything – seven of ten ceiling scores move by less than $0.04$ across the tenfold increase, and a benchmark whose headline axis is context length is reporting, on the dimension it names for updating, a quantity that axis does not move.

Nine of ten dimensions are bounded better by a weak reader than by their own text – the quantitative form of this paper's central negative, since a dimension where $V{=}0$ tells us nothing. The one dimension where text bounds it better, `knowledge_update`, is the ability the benchmark exists to provide: at most $0.63$ of the ceiling can be attributed to a state channel there, and no instrument-free route to a smaller figure exists.

## 6 The instrument run blind on three corpora

Three slots were fixed *by role* before any dataset was named: (A) conversational long-term memory, (B) document QA whose record is not a dialogue, (C) synthetic key–value retrieval. No $R^{\mathrm{str}}$ for any of them existed anywhere and we computed none before freezing, so confirmations are admissible here in the sense the companion accounting paper (Anonymous, 2026c) makes precise – which is not true of §5. **The instrument check first**: slot C gives $R^{\mathrm{str}}(1) = 0.000$ and $k^* = 1$ – a needle task is $k{=}1$ by construction, so a working retriever must close it there, a check a broken implementation cannot pass structurally. **The graded verdicts**: $k^* =$ None on both real benchmarks, so a 32-event budget does not close the gap outside the synthetic slot; one real slot clears a pooled containment bar and the other misses by $0.005$. Using LongMemEval's own dimension labels (Wu et al., 2025) and no model, single-hop dimensions close at $k^* = 8, 8$ and 16 while composition dimensions do not close at all – a budget buys the single-hop dimensions and cannot buy the composition ones, Prop. 1's content measured on a benchmark we did not write.

**The load-bearing assumption, priced.** The sharpest objection is that verbatim containment measures the wrong thing where lexical matching is exploitable. NoLiMa (Modarressi et al., 2025) is built to defeat exactly that: it plants a needle whose lexical overlap with the question is minimal by design, while the answer sits verbatim inside that needle. **The instrument's premise holds and its read-out goes to infinity, on the same items**: $V \approx 1$ on both hop depths (156 rows), while the retriever never once places the needle in a budget of thirty-two paragraphs ($a_k = 0.000$ at every $k$), so $R^{\mathrm{str}}(32) = 1.000$ and $k^*$ does not resolve anywhere, against $k^* = 8$ pooled on the flagship's smallest split. *The condemnation direction is untouched*: a high $V$ still proves single-event answerability, because containment is a witness and not an estimate. *What is removed is the comparative use of $k^*$ across corpora* – it now carries a measured confound, lexical reach, and we retire the cross-corpus reading rather than defend it. The count $V$ is the instrument that survives unrestricted.

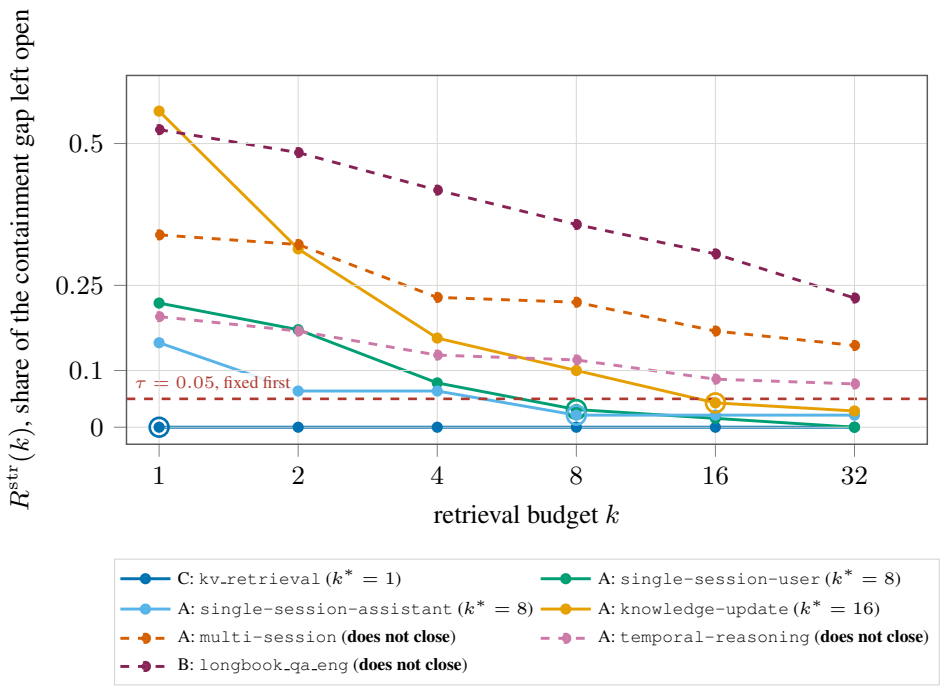



**Figure 3.** $k^*$ **recovers the class labels of the slot that carries several, without being told them.** Every slot whose answers sit in one record crosses $\tau$ and has a $k^*$; the two composition slots are still above $\tau$ at $k = 32$. The slots were fixed by role before any dataset was named, so this is a blind recovery and not a fit. The needle corpus is flat on zero: an instrument that did not return $k^* = 1$ there would be broken rather than informative.

## 7 THE STRONGEST INDEX-ONLY ADVERSARY

A retrieval-plus-model arm is itself an index channel, so it was never a ceiling – it was an adversary filed on the wrong side. A budget ladder $\rho(k) = S(\text{adversary at } k)/S(\text{full record})$, same prompt, model and judge: two dimensions clear a frozen 0.80 bar by $k \leq 16$ (`knowledge_update` at $k$=8, a second at $k \leq 16$), reported as *retrieval sufficiency* rather than resolving power, since numerator and denominator are both index channels. **Two criteria were frozen below their own resolution.** A re-run band failed at its frozen $\leq 0.02$: the per-item paired standard error at the tested sample size is 0.0287, larger than the band itself, while the pooled mean difference over all items shows no drift ($t = 0.11$). A strict-monotonicity criterion is graded NOT DECIDABLE rather than NOT SUPPORTED: none of four non-monotone dips exceeds 2 paired standard errors, and a strict inequality with no tolerance cannot be decided by any finite sample. **A one-event dump with no model beats a model-bearing adversary on half the benchmark**, which is not possible on information grounds; the companion judge paper (Anonymous, 2026g) traces the payment to rubric criteria a response satisfies by saying nothing, a floor of 17.5% of criteria on dimensions whose gold is never in one event.

## 8 THE CHANNEL CLASSIFIER

A name-based check (search the weight index for attention projections) misfiles a recurrent architecture whose time-mixing block happens to be named `attention` while holding no per-position keys. *A classifier that reads names classifies names.* Measuring instead: run a model at several sequence lengths, walk whatever cache object it returns, count tensor elements. A bounded-width carrier returns the same count at every length (1.00× growth over a 16× span); an index returns a count proportional to it (16.00×); a model with local attention over window $w$ grows to $w$ and is then flat – an index at budget $w$, exactly as the companion theory's (Anonymous, 2026d) budget-indexed definition would have it. No row in the measured table sits near the classification threshold.

## 9 THE RECORD-LENGTH AXIS

Pinning the budget and sending record length to infinity is a companion paper's (Anonymous, 2026e) limit, not this one's; three results are quoted because §5 and §10 depend on them. **The required budget is a constant share of the record**: on nested records built by deletion, $k^* \propto L^\alpha$ with $\alpha = 1.011$ over a $62\times$ ladder and $k^*/L$ pinned at $0.400$–$0.417$, so "a small retrieval budget" is a statement about record length, not about a task. **That exponent belongs to the lever, not the task**: the same resource moved without deleting anything reverses the ordering of dimensions. **The comparison is undecidable at this benchmark's size**: zero of five bootstrap intervals on the difference of two exponents exclude zero. What this paper keeps: $V$ and $R^{\mathrm{str}}$ are measured at the length the benchmark ships, and a bound read off them is a bound at that length.

## 10 DISCUSSION

**What the audit established.** $0$ of $10$ dimensions are established as blind. Prop. 2 is why that is not evidence of sight: the instrument is one-sided, so a failure to condemn is silent rather than exculpatory, and a companion theory paper (Anonymous, 2026d) shows the $0/10$ was structurally guaranteed once every arm in the apparatus is an index channel. What the audit does establish needs no instrument: the counts, the scale invariance, and the class recovery of §6. One of those three now carries a measured scope condition: the count $V$ travels to a corpus built to defeat lexical matching, while the budget read-out $k^*$ does not.

**Three deployed cases where the instrument changes the reading.** We did not find this by auditing other people's benchmarks. **(a)** Our own delay–load protocol swept a load axis with rearrangement events driving a composition query; a last-occurrence reader scored $1.00$ and was classified access-complete, because each rearrangement announced the resulting value rather than the pair of keys exchanged, so the composed answer appeared verbatim in the write record. **(b)** In an earlier study we dropped a skill for "unlearnability" after every arm landed inside the noise floor; its gold was prose with a median $154$ characters, only $8.0\%$ short enough for the exact-match scorer to grade at all – a skill pinned at its scorable fraction is indistinguishable from one that does not learn. **(c)** In a companion experiment on self-written stores (Anonymous, 2026b), a provenance marker appeared to make contamination worse; the planted contamination had been seeded as OBSERVED while the agent's own correct answers were DERIVED, so the marker instructed the model to trust the lie – graded UNTESTED (APPARATUS FAILURE), not FAILED. All three share one shape: the instrument's blindness and the phenomenon's absence produce identical readings, and no statistical power distinguishes them.

**Where zero resolving power is load-bearing, argued not graded.** A memory-poisoning defence validated by the score drop it prevents makes a claim only at the budget it was measured at, times the judge's credit rate (Anonymous, 2026g). An admission rule filtering by cross-source agreement cannot separate a fact that recurs because it is true from one that recurs because it was copied – the redundancy-for-difficulty conflation the record-length companion measured (Anonymous, 2026e). An instrument validated only on its own construction set is measuring its own construction. The measured versions of these arguments are not ours: Wang (2026)'s single-variable sweep and Sidik & Rokach (2026)'s architecture-invariant null are the checkable evidence, and a reader should weigh those above this paragraph.

**What remains owed.** Neither arm below can overturn a condemnation, since the instrument is one-sided: $V$ is a conservative floor on single-event answerability, so a paraphrase-tolerant count can only raise it. **(1)** A paraphrase-tolerant, entailment-scored $V$ – the one instrument extension in this paper that would need a model. **(2)** A budget-matched *scored* reader at the two larger splits, judge-mediated and therefore owed to the companion judge paper's (Anonymous, 2026g) frame. Four of this paper's own frozen bands could not have decided what they were asked – a minimum resolvable step of $0.301$ against a band of $0.150$, a bootstrap interval reaching $1.064$ against bands of $0.150$ and $0.519$, a paired re-run band of $0.020$ against an achievable $0.0287$, and a strict-monotonicity criterion undecidable for any nonzero gap – and are reported rather than dropped, because a label a reader cannot check is not a label.

**The reach of the audit.** The headline says ten dimensions; **the audit's actual reach is three**. The instrument needs a short gold to check containment; six of ten dimensions on the flagship have no checkable item, and only the updating dimension has enough to grade across all three context scales. The bias direction is known and favourable: containment-admissibility selects the *easy* subset for a retriever, so failures to condemn are conservative. Portability of the *instrument* is demonstrated on three corpora (plus a fourth, judged by a companion (Anonymous, 2026e), recovering the same boundary function with a different constant); portability of the *audit of this one benchmark* is three dimensions, and the ten-dimension framing borrows the wrong claim's reach.

**Limitations.** **(i)** Only the instrument-free half of the state/index dichotomy travelled to three further benchmarks; other distinctions need their own oracles. **(i′)** $V$ is a *lexical* count and over-states reachability, which makes a condemnation conservative and a failure to condemn uninformative – the safe direction; the cross-*dimension* comparison is not safe and is now measured rather than conceded (§6). **(i″)** $R^{\mathrm{str}}$ needs golds in *exactly* one event; across four benchmarks that share runs $4.3\%$ to $100\%$, and where the share is low the quantity is inflated in the auditor's favour. **(ii)** The audit did not condemn and cannot acquit – $0/10$ leaves the central question open, a property of the method. **(iii)** Our ceiling is the best of two arms we built, not the benchmark's best published system. **(iv)** One pinned judge decides every judged number; the companion (Anonymous, 2026g) measures what changing the pin does (a shift of $0.925$ on one dimension), so the four adversary-ladder criteria are conditional on the pin and do not claim to transport – what is not conditional is $V$, $R^{\mathrm{str}}$ and $k^*$, which never ask a judge. **(v)** The load-bearing-corollary arguments are reasoned, not audited: no defence evaluation, admission rule or leaderboard was checked here.

## 11 Conclusion

Whether a memory benchmark can see the architectural distinction it is cited for is a property of its text, and two counts measure it. $V$ says which items were answered at the moment they were written; $R^{\mathrm{str}}(k)$ and $k^*$ say what retrieval budget the rest demand. Run blind, the counts recover a benchmark's own labels; run on the flagship, they say that the dimension named for updating announces more than half of its checkable answers, at every scale the benchmark ships, and that the benchmark's headline axis moves the budget's reach rather than any score. The audit is one-sided by construction, so none of this acquits a benchmark, and the counts bound no scored quantity without a judge-mediated step that a companion paper (Anonymous, 2026g) prices. What a benchmark author gets is a pre-release report that costs two counts: publish $V$ and $k^*$ per dimension, name the budget the baseline was allowed, and the question "does this task measure memory" stops being a matter of opinion.

### Reproducibility statement

Every registered criterion's frozen commit is tabulated against its outcome in Appendix J (60 graded rows: 37 for the paper, 23 against). Every number in the main text is computed by an analyzer from files in `results/` at build time; every pre-registration file, grader, sweep runner and result file is included in the anonymised supplementary material. Model calls are the only cost this paper incurs beyond the adversary ladder's judge calls; every count in §5–§6 needs neither.

### Ethics statement

This paper audits publicly released benchmarks and models by computing string-containment statistics over their own text; no human data was collected, no new system was built, and no individual is identified. Where a defect is named in a benchmark's construction (e.g. a split naming inconsistency, a schema change at one rung), it is reported as a fact about the released artifact, not as a criticism of any named author. We are not aware of a Code of Ethics concern raised by this work.

### AI use statement

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. *Our use was confined to preparing the document.*

*Tasks in the required-disclosure list: none.* Generative AI was not used to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate the mathematical claims, to supply critical ingredients for the proofs or to write them, to propose or refine the hypotheses, to design the methodology or the experiments or to give feedback on them, to set or choose any pre-registered threshold, to implement methods, to clean or reformat any dataset, or to interpret the results. The counts and sweeps of §4–§6 were configured and launched by the authors, and every number reported here is computed by an analyzer the authors wrote, from the result files those runs produced; scripts, job configs, result files and verdicts are all in the supplementary material. Translation and qualitative or thematic data analysis are not applicable to this work.

*Tasks in the recommended-disclosure list.* Generative AI was used on the manuscript itself. It condensed a longer version of this work into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled and anonymised the supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims: the claims, the numbers they quote and the qualifications attached to them are the authors', carried over from the longer version.

*Verification.* Every quantitative statement in this version was checked against the graded row it cites, and every registered outcome against the freeze commit tabulated in Appendix J. The one transformation AI-written code applied to a released artefact — rewriting absolute paths and host names in the supplementary archive for anonymisation — was scoped to provenance strings and verified not to have altered any result file, score, or reported number. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artefacts.

## REFERENCES

Anonymous. The capability convergence hypothesis: Capability from access structure, not scale. *arXiv preprint*, 2026a. author list and repository withheld for double-blind review; a concurrent manuscript of the same research programme, arXiv identifier withheld.

Anonymous. Self-Cleaning and Captured Anyway: A Measured Primitive for Error in a Store an Agent Writes to Itself. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

Anonymous. Position: Preregistration Protects the Criterion, Not the Comparison: One Programme Audited Against Its Own Laws Seven Times. concurrent manuscript; author list and repository withheld for double-blind review, 2026c.

Anonymous. Resolving Power: Whether a Memory Task Can See the Distinction It Claims to Test. concurrent manuscript; author list and repository withheld for double-blind review, 2026d.

Anonymous. The Infinite-Record Limit: A Retrieval Budget That Stays Constant for the Hardest Items and Vanishes for the Median. concurrent manuscript; author list and repository withheld for double-blind review, 2026e.

Anonymous. The Infinite-Rerun Limit: Computing what a criterion can decide, before it is frozen: a resolution, a rule, and the flip rate a band buys. concurrent manuscript; author list and repository withheld for double-blind review, 2026f.

Anonymous. What a Rubric Judge Measures: Two Defects in a Shipped Scorer Change What It Measures, Not How Noisily. concurrent manuscript; author list and repository withheld for double-blind review, 2026g.

Mihir Shriniwas Arya. RECON: Benchmarking agent memory for compositional reasoning over long contexts. *arXiv preprint arXiv:2607.16716*, 2026.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Proc. 39th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv:2511.04703; 445 benchmarks reviewed by 42 authors.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. also ECAI 2025; reports a head-to-head comparison of ten memory approaches on LoCoMo.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Jeffrey Flynt. Structured belief state and the first precision-aware benchmark for LLM memory retrieval. *arXiv preprint arXiv:2605.11325*, 2026.

Wei-Chieh Huang, Weizhi Zhang, Yuchen Wu, Yankai Chen, Eric Hanchen Jiang, Wooseong Yang, Yiwei Yang, Henry Peng Zou, Hanrong Zhang, Ying Nian Wu, Haolun Wu, Kai-Wei Chang, Philip S. Yu, Xue Liu, and Aylin Caliskan. Harness the memory: A holistic evaluation of memory substrates in memory agents. *arXiv preprint arXiv:2608.15008*, 2026a.

Zhaoheng Huang, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. LLM-generated text may harm your retrieval! a robust detection strategy for retrieval-augmented generation. In *Proc. 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026b. ACL Anthology 2026.acl-long.1475.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. arXiv:1912.05511. Construct validity and construct reliability imported into computing from the measurement-modelling tradition, with the observation that an unstated operationalisation is where the assumptions hide. The frame this paper's F1 is an instance of: the operationalisation moved and nothing in the pipeline said so.

Theodore J. Kalaitzidis. The evaluation trap: Benchmark design as theoretical commitment. *arXiv preprint arXiv:2605.14167*, 2026. Introduces Epistematics: an audit procedure, a failure-mode taxonomy and design criteria for whether a proposed benchmark can discriminate a claimed capability from proxy behaviours, demonstrated on a worked audit. Procedural rather than quantitative.

Ruizhe Li, Licheng Zhang, Benfeng Xu, Mingxuan Du, Zheren Fu, and Weidong Chen. When your agent opens the chat app: Agent-controlled search over raw chat logs rivals structured memory. *arXiv preprint arXiv:2608.12888*, 2026a.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. Auditing the audit: Five failure modes in benchmark-validity audits. *arXiv preprint arXiv:2607.02586*, 2026b. Names five classes of pipeline failure in perturbation-based construct-validity audits and demonstrates each in a self-audit over five safety benchmarks and two open-weight models. F1, silent no-op perturbation, is defined as the scorer-consumed prompt being bit-identical to the unperturbed input, measured at a $100\%$ silent-no-op rate on two of the benchmarks; F4 is broken pairing in uncertainty. The six-point due-diligence gate requires a silent-no-op rate under $20\%$ (G1) and an item-level paired bootstrap (G4). Six of the seven realized failures were latent in the authors' own legacy code and the seventh was introduced while repairing F1.

Jiayi Liu, Jiaxing Zhang, Bowen Jin, and Jennifer Neville. Generating leakage-free benchmarks for robust RAG evaluation. *arXiv preprint arXiv:2605.08838*, 2026.

Microsoft Open Source. Introducing STATE-Bench: A benchmark for AI agent memory. `https://opensource.microsoft.com/blog/2026/05/19/introducing-state-bench-a-benchmark-for-ai-agent-memory/`, May 2026.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.05167. Built because models exploit LITERAL matches between needle and haystack. This is the standing objection to a containment-based instrument and we answer it in the direction it points rather than around it.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

Sugam Panthi and Rabab Abdelfattah. Same ranking, different winner: How scoring targets shape LLM memory benchmarks. *arXiv preprint arXiv:2605.24060*, 2026.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2111.15366.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. Outstanding Paper; arXiv:2304.15004.

Bronislav Sidik and Lior Rokach. MEMTIER: Tiered memory architecture and retrieval bottleneck analysis for long-running autonomous AI agents. *arXiv preprint arXiv:2605.03675*, 2026.

Samuel Sameer Tanguturi. ATANT v1.1: Positioning continuity evaluation against memory, long-context, and agentic-memory benchmarks. *arXiv preprint arXiv:2604.10981*, 2026.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. *arXiv preprint arXiv:2510.27246*, 2025. ICLR 2026.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive interference reveals working memory limits in LLMs beyond context length. *arXiv preprint arXiv:2506.08184*, 2025.

Kuan Wang. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation. *arXiv preprint arXiv:2606.29914*, 2026.

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. arXiv:2410.10813. The corpus every record-length arm in this paper runs on: 500 questions over timestamped, freely scalable chat histories, with the five ability labels — information extraction, multi-session reasoning, knowledge update, temporal reasoning, abstention — that our class split is taken from rather than invented.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.12110.

Zongyou Yang, Yinghan Hou, and Xiaokun Yang. When the judge changes, so does the measurement: Auditing LLM-as-judge reliability. *arXiv preprint arXiv:2607.08535*, 2026.

Jiangnan Yu, Kisson Songqi Lin, and Jilong Wu. WhenLoss: Diagnosing write and retrieval bottlenecks in long-context memory systems. *arXiv preprint arXiv:2605.24579*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685. The paper that named the position, verbosity and self-enhancement biases of an LLM judge and reported 80% agreement with human preference. Every bias it names is a property of how the judge READS; none is a property of what the scoring script SUBSTITUTES into the prompt.

Ziyi Zhu, Olivier Tieleman, Alexey Bukhtiyarov, and Jinghong Chen. CyclicJudge: Mitigating judge bias efficiently in LLM-based evaluation. *arXiv preprint arXiv:2603.01865*, 2026.

# SUPPLEMENTARY MATERIAL

*This appendix follows the references and is supplementary; the main text is self-contained, and reviewers are not required to read it. Appendix A carries the main text's own argument at the length the manuscript states it, one section per body section, for a reader who wants a claim's full derivation rather than its one-sentence form. The sections after it hold the full anywhere-containment curve, every registered cell of the blind three-corpus run, and the public scorecard of all* $60$ *graded rows.*

## A EXTENDED MAIN TEXT

The conference version compresses the manuscript's main text by roughly $58\%$. The sections below carry that material in full, one subsection per main-text section, in main-text order, so that a claim the body states in a sentence can be read here with the argument, the figure and the registered outcome that decide it. *Nothing here is new*: no claim, number or verdict appears that the body does not already assert, and where the body and this part differ in length they do not differ in content. Cross-references of the form §4, §5, Figure 3 and Table 1 point at the body throughout; floats the body already carries are not reprinted here.

### A.1 RELATED WORK, IN FULL

**One paper states this paper's premise before us, in prose.** Tanguturi (2026) analyse seven memory evaluations against a property checklist and conclude of the two this paper's corpus comes from: *"a language model with a context window large enough to fit the full benchmark passes both of these evaluations with no memory architecture at all"*, so "by construction, these benchmarks cannot distinguish a memory system from a long-context LLM". That is the blindness claim, asserted rather than measured, and it is prior to us. What this paper adds is not the claim but its arithmetic and its scope. Theirs is a structural argument about context-window *sufficiency*, scoring a benchmark $0/7$ on a checklist; ours is a per-item count with a budget in it, so it applies where the context does *not* fit: $V$ says which items are announced in one turn, $k^*$ says at what retrieval budget the rest become reachable, and both are computable from the benchmark's text before anything is run. Their reading of the analogous update category also matters below: they read it as testing *within-context override* rather than update handling, a complaint about what a category name invites (§A.17).

**Three independent results say the phenomenon is there, one of them in our own words.** The strongest external evidence for a resolving power of zero is a null result somebody else published. Sidik & Rokach (2026) evaluate a tiered memory architecture against its unmodified baseline on LoCoMo and score *identically* ($\Delta F_1 = -0.005$; $0.125$ against $0.120$), and their reading of it is the blindness contrapositive: "LoCoMo inserts the full conversation into context at query time, making the memory architecture irrelevant", so benchmarks that do this "cannot distinguish memory architectures". That is the quantity this paper counts, measured on a second benchmark, by other people, without our instrument, and it is what $V$ and $k^*$ would have told them before they ran anything. Li et al. (2026a) supply the arm our baseline is built to be: an agent-controlled *lexical* search over the unmodified archive, no structure built at all, reaching the highest mean accuracy of any system they compare ($58.2$ against $53.2$ for the strongest graph-based memory) under a matched backbone. An index-only system with no state carrier beating structured memory is the empirical form of the audit's asymmetry: it does not show the structures are useless, it shows the tasks do not separate them. Wang (2026) likewise vary one component at a time on LongMemEval-S and find verbatim RAG statistically indistinguishable from full context ($47.2\%$ against $49.8\%$, $p = 0.34$) while swapping only the embedding model moves accuracy by $6.2$ points, so "one variable flips the conclusion". Two further neighbours decompose the same score compatibly: Yu et al. (2026) separate write-side from retrieval-side loss with an oracle-evidence condition, which is $R^{\mathrm{str}}(k)$'s decomposition on a different cut, and Huang et al. (2026a) compare memory substrates under one harness and find no substrate dominates, with broad retrieval helping factual QA and hurting sequential decision-making, the budget argument seen from the system's side.

**Contamination and leakage work removes one route to blindness and leaves the others.** Liu et al. (2026) generate benchmarks whose questions cannot be answered from parametric knowledge alone, filtering instances solvable without retrieval; Huang et al. (2026b) address contamination

of the retrieval corpus by model-generated text. An item answerable without the state is blind by construction, and leakage-free generation removes that class. It removes one class and the count finds others: a large fraction of a shipped benchmark's items are answerable by returning the last event containing the gold value, with no model at all (§5). Those items are not leaked, the answer is genuinely in the context, and they are not solvable without retrieval, so they pass the leakage filter while separating nothing a substring search cannot separate. Filtering on *parametric* answerability and filtering on *index-only* answerability are different filters.

**Construct validity surveys establish that the problem is widespread; a count establishes it for one item.** Bean et al. (2025) review $445$ benchmarks with $42$ authors, reporting contested construct definitions, convenience sampling in $27\%$, and statistical tests in only $16\%$; Raji et al. (2021) make the sharper qualitative version of the point. Our contribution to that line is arithmetic rather than rhetoric: $V$, $R^{\mathrm{str}}(k)$ and $k^*$ turn "does this task measure what it is cited for" into two counts over the benchmark's own text, so the claim can be checked by someone who disagrees with us about everything else.

**What a judge-free instrument is immune to, named rather than deferred.** This paper's verdicts are counts, and it says so on every page; a claim of immunity is only worth something if the disease is named where a reader can check it. The judge-side literature names three, and this instrument is outside all three for the same structural reason. *Bias*: a rubric judge carries position, verbosity and self-enhancement preferences (Zheng et al., 2023), and a count over text has no preference to carry. *Unreliability*: a score moves when only the evaluator changes (Yang et al., 2026), and the remedy is to rotate judges and average the offsets away (Zhu et al., 2026), where $V$ and $k^*$ have no evaluator to rotate. *Reliability that is not validity*: Norman et al. (2026) separate agreement from validity across roughly $541$K judgments and find exact-match agreement overstating discriminative ability, which is Jacobs & Wallach (2021)'s construct-validity point with a measurement behind it. The third is the one that matters here, because it is the one a benchmark author cannot buy their way out of: a perfectly reliable judge is a perfectly agreed measurement of whatever the rubric operationalises, and this paper's counts are about whether the item contains the thing at all. The immunity is bought and not free. A count cannot bound a difference of scores, which is the reach §10 states and the price the judge companion measures (Anonymous, 2026g): on one benchmark the credit rate under perfect provenance is $0.614/0.452/0.629$, and the scorer's own share of that shortfall is $0.028$ (Anonymous, 2026g, row 205b), so an instrument that avoids the judge has avoided the smaller of the two things standing between a record and a score.

**The closest prior programme is procedural, and its procedure inherits a one-sidedness it does not state.** Kalaitzidis (2026) propose *Epistematics*: derive evaluation criteria from the capability a benchmark claims, then audit whether the benchmark can discriminate that capability from proxy behaviours, with a failure-mode taxonomy and a worked audit. That is this paper's question, asked one level up, and the two halves of it do not survive our limit equally. Prop. 2 says the audit is *one-sided* on any benchmark harvested from natural text: condemnation is exhibitable, since one index-only reader scoring high is a witness, while a pass quantifies over every $k$-subset of the span and every function on it, a space that is not enumerable ($\binom{188}{8} \approx 10^{13}$ subsets on one BEAM conversation, and the functions on each are worse). An audit procedure whose output space contains a pass reports the *absence of a witness* in a form a reader will read as the *presence of validity*, and no taxonomy of failure modes closes that gap, because the gap is in the quantifier and not in the list. We hold ourselves to the same bound and say so where it costs us: our $0/10$ (§10) is not a clean bill of health and §10 states that it cannot become one. The design half of their programme is where it survives the limit intact, and that is not a concession we make grudgingly: Prop. 2's escape clause *is* construction, so criteria applied to a benchmark being built are exactly the case where a positive verdict is available. The division of labour we would propose is theirs for design and ours for deployed artifacts, where two counts over the benchmark's own text (Def. 3) give a number before anything is run instead of a procedure to follow.

**The closest parallel work audits the audits, states an asymmetry of its own, and stops one step short of ours.** Li et al. (2026b) argue that construct-validity audits are themselves fragile (their conclusions "can be silently manufactured by implementation details that readers cannot see in the reported numbers") name five classes of pipeline failure, demonstrate each in a self-audit over five

safety benchmarks, and ship a six-point due-diligence gate. This is the same suspicion this paper acts on, one level up, and one of their gate conditions is an instrument for a defect the judge companion found by reading code (Anonymous, 2026g). They also state an asymmetry, and we should be careful not to claim it as ours: a verdict under their gate reports that an attribute delta exceeds surface-family magnitudes, which they call *a necessary but not sufficient condition for construct sensitivity*.

The difference from Prop. 2 is the kind of object, not the direction. Their caveat is about what one gate's positive verdict licenses, and it is discharged by a better gate or a direction-validation step they defer. Prop. 2 is a statement about the quantifier: on a task family harvested from natural text, acquittal ranges over every $k$-subset of the span and every function on each, so *no* finite procedure supplies it, and no gate closes that. **Their asymmetry is a limitation of an instrument; ours is a property of the question.** The consequence for this paper is one sentence: their five classes are the reason a reader should not take our $0/10$ on trust either, and the answer is that §5's counts are computable from the benchmark's own text with no model, judge or API call, so the implementation detail a reader cannot see is a download away from being seen.

**The memory benchmarks are the object of study.** BEAM (Tavakoli et al., 2025), STATE-Bench (Microsoft Open Source, 2026), RECON (Arya, 2026), STALE (Chao et al., 2026), Supersede (Patel, 2026) and the precision-aware line (Flynt, 2026) are the benchmarks this machinery audits. Several already move in the direction the counts reward: STATE-Bench scores whether an agent *improves with experience* rather than whether it retrieves, and Chao et al. (2026) probe whether an agent notices that a memory has been invalidated, both attempts to build items an index cannot satisfy. Panthi & Abdelfattah (2026) report the closest scoring-side result, changing only the credited target moves nDCG on $83.4$–$94.0\%$ of shared queries, and the judge-side mechanism behind such movements is the companion's subject (Anonymous, 2026g). Our contribution to this group is not another benchmark; it is the pre-release report their authors can generate from their own text.

### A.2 The one-sidedness argument in full

**Definition 4** (Last-occurrence reader)**.** Hold a verbatim transcript. To answer a query about $k$, return the most recent single event mentioning $k$, verbatim. No event is ever combined with another.

Both instruments are deliberately *weaker* than the retrieval systems the memory benchmarks are built to evaluate (Tavakoli et al., 2025; Arya, 2026; Flynt, 2026), which is what makes them usable as bounds: a system that cannot compose events and holds no model cannot be credited with anything a state channel supplies. It is worth being exact about what "index-only" fixes, because the natural objection is that a long-context model is index-only too; it holds the whole transcript. The quantity being held fixed is not the span but what is *accessible* from it, which is a capacity in the sense of Cover & Thomas (2006) and is strictly smaller: Wang & Sun (2025) show that proactive interference degrades retrieval from a context the model nominally contains, so accessible state is bounded below the span even when the span is free. Our readers make that bound explicit and extreme (one event, no composition) rather than leaving it to be inferred from a context length. A second instrument, *BM25-last*, scores every event with BM25 over the question's terms and returns the most recent event within $0.8$ of the maximum. Still one event, still no composition, still no model; term weighting is what the plain reader lacked.

**That $0.8$ is a free parameter, and it can be priced without spending anything**, because the reader's *selection* needs no judge: for every checkable item on the three splits, sweep the constant over $\{0.5, \ldots, 1.0\}$ and ask whether the returned event still contains the gold (`analyze_cand.py`, `results/beamcurve/CAND_SENSITIVITY.json`). *The constant is not innocuous*: the event returned is identical across the whole sweep on only $3.4$–$11.6\%$ of items, and the containment rate moves by up to $0.101$, from $0.034$ at $0.5$ to $0.143$ at $0.9$ on the 1M split. What the audit needs is weaker than innocuousness and is what we can show: each constant is a *realisable* strategy, so the bound the audit is entitled to is the best of them, and the frozen constant is within $0.008$ of the best on every split ($0.8$ is the maximum on two of three, $0.9$ beats it by $0.008$ on the third). The failure to condemn below is therefore not the retriever being handed a bad constant, which is the way this objection would have mattered. The sweep is in the gate chain and asserts that gap rather than printing it. Both are realisable index-only strategies, so $\max_i S(\text{reader}_i) \leq S(\text{index-only optimum})$ and

$$R_{\text{state}}(\mathcal{T}) \;\le\; S_{\mathcal{T}}(\text{ceiling}) - \max_i S_{\mathcal{T}}(\text{reader}_i). \tag{3}$$

**The audit condemns; it does not acquit.** A reader near the ceiling *proves* blindness. A reader far below it proves nothing: a better index-only strategy may exist. This asymmetry is why the audit's $0/10$ (§10) is not a clean bill of health (Figure 2).

**Why the asymmetry is not a defect of our instrument.** It is a difference in quantifier, and the limit makes that visible. Condemning is *existential*: exhibit one realisable index-only strategy that reaches the ceiling, and Prop. 1 does the rest. Acquitting is *universal*: show that *no* index-only strategy does. An index-only system with budget $k$ may compute any function of the $k$ records it retrieves, so acquittal quantifies over every $k$-subset of the span and every function on it. For one BEAM conversation, $n \approx 188$ turns and $k$=8 gives $\binom{188}{8} \approx 10^{13}$ subsets, and the functions on each are not enumerable at all.

> **Proposition 2 (Acquittal needs construction), restated from the body.** A task family can be shown to have $R_{\text{state}} > 0$ only if the dependency structure of its items is known by construction; that is, only if it was *built* so that the answer provably depends on more of the span than any admissible retrieval can hold. For a task family harvested from natural text, the universal quantifier ranges over a space that cannot be enumerated, so no finite audit acquits it.

Two consequences, and the second is a design rule rather than a caveat.

**(i) $0/10$ is *the only outcome this reader could have returned* here**, short of a condemnation, and the distinction between *this reader* and *this method* has to be kept explicit. Both instruments hold **one event**, i.e. $k = 1$, the weakest admissible adversary. A *higher* index-only score is exactly what condemns, by (2). A stronger reader could only have made condemnation more likely, and the budget ladder answers what the two one-event instruments cannot. At $k$=8 over the same events, the retriever's containment coverage on `knowledge_update` rises from $0.194$ to $0.611$ on the 100K split, which is that dimension's *entire* anywhere-reachable set (Appendix B); on `multi_session_reasoning` it moves $0.063$ to $0.125$ against a reachable $0.188$, and on `information_extraction` not at all. The $0/10$ is a statement about the reader's budget on one dimension and about the benchmark on the other nine, and the judged ladder of §7 agrees where it can be compared: it is at $k$=8 that `knowledge_update` crosses the blindness bar, at $\rho = 0.830$. What holds regardless is the count: $V$ is computed with no instrument at any budget (§5); what a count licenses as a bound on the *scored* quantity is judge-mediated and is the companion judge paper's subject (Anonymous, 2026g).

**(ii) A sighted benchmark has to be built, not audited into existence.** By Prop. 2 the only route to a task with demonstrable resolving power is to construct the witness: plant a value once, announce subsequent changes as operations rather than as resulting values, and make the answer their ordered composition. That is what the parent framework does (Anonymous, 2026a), it is what the soundness probes gating both readers do, each scoring $0/20$ on them by construction, and it is what §5 finds the flagship's updating dimension does not do: the update is announced, so 55–64% of checkable answers are recoverable from one turn. The prescription is therefore concrete and cheap: a benchmark that wants to measure a state channel should include a constructed witness alongside its harvested items, and report the two separately.

### A.3 The audit of a deployed benchmark, in full

Proposition 1's condition can be checked on a benchmark's text directly: for each probing question with a short gold, does the gold appear verbatim inside a single conversational turn? No reader, no judge, no model, no ceiling. Verbatim containment *undercounts* single-event answerability, since a paraphrased gold still lives in one turn, so these are lower bounds; high rates condemn, low rates prove nothing.

On the dimension that tests whether a system tracks an updated value, the majority of checkable answers are stated verbatim in one turn: the update is *announced*, so the answer is a function of the write record and Prop. 1 applies exactly. This is the same defect §A.12 reports in a protocol of our own, where a rearrangement announced the resulting value rather than the pair exchanged.

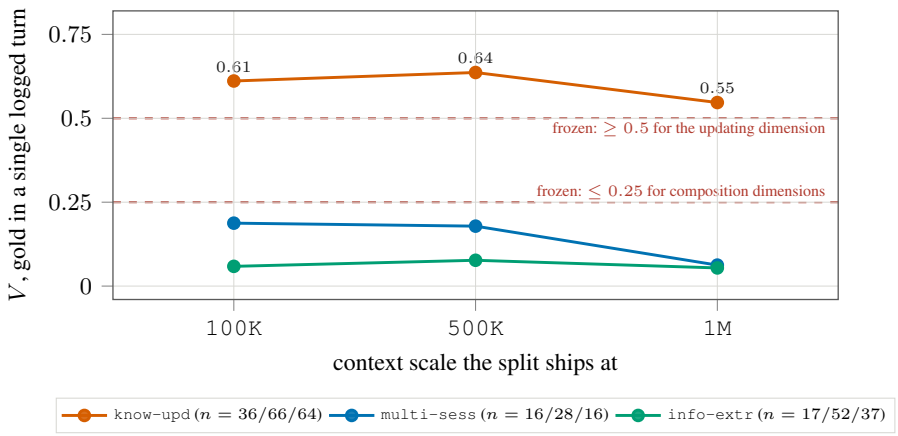



**Figure 4.** $V$ **at the three registered scales, against bars frozen before the splits were downloaded.** On the dimension named for updating, the majority of checkable answers are stated verbatim in one turn, so the answer is a function of the write record and Prop. 1 applies exactly. *The two dashed lines are the pre-registration, not a summary*: $\geq 0.5$ for the updating dimension and $\leq 0.25$ for the composition dimensions were fixed before the 500K and 1M splits existed. Every coordinate is $\text{complete}/n$ recomputed from `results/beamcurve/GRADE.json`.

### A.4 All four shipped rungs of the headline axis, in full

The audit above grades three registered scales. The benchmark's own axis has a fourth rung, and running it changes no verdict while adding the two dataset facts below and one finding: what the axis moves is the budget's reach, not the score. The anywhere-containment curve $a_{\text{complete}} - a_k$ behind these statements is Appendix B; its levels are inflated in the auditor's favour and only its decay pattern and cross-rung comparisons are used here (§6 states why).

**Three facts about the dataset that only appear once the fourth rung is run.** *(i) The split names.* BEAM's paper and project page describe its buckets as 128K, 500K, 1M and 10M; the released dataset files are named `100K`, `500K` and `1M`. Both names are in circulation for the same 20 conversations; this paper uses the file names, the names the artifact on disk carries, and states both once here. *(ii) The* 10*M rung is a separate dataset with a different record schema.* Where the three shipped together carry `chat` as a list of batches of message dicts, the 10M release carries an array of ten `plan-*` dicts whose batches hold nested `turns` arrays, some null. An instrument written against the released three does not run on the fourth: ours raised a `TypeError`. `beam_curve.py` now walks the field shape-agnostically and asserts that the walker reproduces the two-level reading element for element on the splits that carry it, so a compatibility fix cannot silently move a published number. (iii) The axis is $74\times$, not tenfold. Median turns per conversation run 271, 1076, 2088 and 20075, and turn length is flat across the four (1440, 1237, 1110, 1366 characters), so length is added as turns rather than as longer turns. "Tenfold" below always means the shipped trio 100K→1M, which is $7.7\times$ in events; the four-rung axis is $74\times$.

**What the fourth rung says is a third outcome, not either of the two that were predicted for it.** The two predictions for it were that $V$ holding near 0.5 at 10M would turn weak evidence into strong, and that $V$ falling would be more interesting than anything else in the paper. $V$ falls, monotonically, on every rung: 0.377/0.349/0.325/0.295 pooled, and `knowledge_update` runs 0.611/0.636/0.547/0.438. *The fall is not decidable at the sizes the splits ship.* The 100K→10M drop is 0.081 against a standard error of 0.090 on the difference, which is 0.90 se; by the companion's rule a band would have to clear $2\Delta = 0.180$ (Anonymous, 2026f). The reason is the rung itself: 10M ships 10 conversations $\times$ 20 questions $=$ 200 items, of which 44 carry a gold short enough for containment, spread over three of ten dimensions. The largest

> rung of BEAM's own length axis cannot support the measurement its own size invites, and the supportable report is a monotone trend with an undecidable slope, which is this programme's most common finding, arrived at here from the outside, on somebody else's benchmark, for the price of a download.

**P-T3-41 is NOT SUPPORTED and P-T3-44 is NOT SUPPORTED, and the second is the one worth having.** We predicted $R^{\text{str}}(1) \geq 0.30$ pooled and got $0.246$: a top-1 lexical retriever already reaches the gold on most checkable items, which the preregistration named in advance as strengthening §5 rather than weakening anything. We also predicted that the dimension with the largest gap would *not* be `knowledge_update`, reasoning that a high $V$ means the gold sits verbatim in some turn and is therefore the easy case for a lexical retriever. **That reasoning was backwards, and `knowledge_update` is the argmax on all three splits.** A high $V$ on an updating dimension means the gold *and its superseded predecessor* both sit verbatim in some turn, in near-identical wording, so lexical similarity to the question is exactly the signal that cannot separate them. The dimension this benchmark names for updating is the one where the access budget matters most, and the reason is a property of the task rather than of any system.

**The context-length null reproduces with no instrument in it.** §5 reports that a tenfold increase in context moves nothing, and that finding is measured with a judge, on 2100 items, at some cost. Here $V$ runs $0.377/0.349/0.325/0.295$ and $R^{\text{str}}(1)$ runs $0.246/0.226/0.188/0.205$ across *all four* of BEAM's shipped rungs, the same null, on a quantity containing no model, no judge and no system, for the price of a BM25 pass. The fourth rung was added because the audit stopped at 1M while the benchmark ships 10M. When the instrumented and instrument-free measurements of a benchmark's headline axis agree, the reading is no longer about the judge.

**The gap does not close at scale the way it closes at 100K.** P-T3-43 holds on the split it was registered for: $R^{\text{str}}(32) = 0.000$ at 100K. At 500K and 1M the same budget leaves $0.021$ and $0.017$ pooled, and $0.030$ and $0.031$ on `knowledge_update`. A fixed retrieval budget buys less of the transcript as the transcript grows, which is the one place in this paper where the headline axis does move something, it moves the budget's reach, not the score, and a benchmark reporting the second without the first is reporting the smaller half (Figure 5).

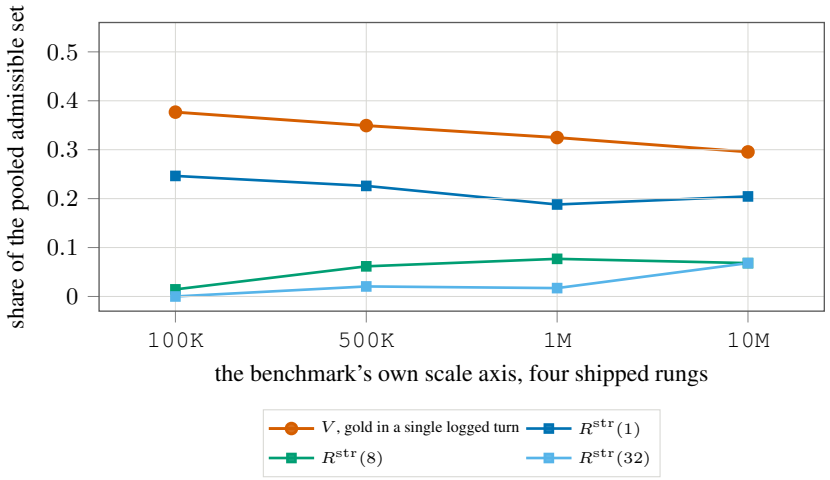



**Figure 5. Across all four shipped rungs the counts barely move, and what the axis moves is the budget's reach.** $V$ falls monotonically, from $0.377$ to $0.295$ over 4 rungs, but by less than these rungs can resolve, so its slope is undecidable at the sizes the benchmark ships. The three $R^{\text{str}}$ curves are flat rung to rung and steep in $k$: at the smallest rung the gap closes from $0.246$ to $0.000$ as $k$ grows. That is the same fact from the other side: *on these counts the benchmark's headline difficulty axis is an axis of retrieval reach.*

### A.5 THE SCALING AXIS AT ITS LARGEST RUNG, IN FULL

**The counts do not move along the axis the field is scaling.** There is one obvious reply to an audit of one benchmark: this is a fact about today's benchmark, and benchmarks grow. Take the growth to its limit. The axis being scaled is context length (100K, 500K, 1M tokens is a tenfold increase and the direction of travel is unbounded) and $V$ was measured on all three, with the last two registered before they were downloaded. Across that tenfold range no dimension crosses its own frozen bar: `knowledge_update` reads $0.61/0.64/0.55$ and stays above $0.5$ at every scale, `multi_session_reasoning` reads $0.19/0.18/0.06$ and stays below $0.25$ at every scale. The values wander; the *verdicts* do not. The same holds for the all-items containment rate, computed over every item rather than over the short-gold subset the bars were set on: $0.55/0.60/0.50$ at the three scales. The count route to a ceiling does not improve as the benchmark grows: $1 - V$ sits at $0.45 \pm 0.05$ across a tenfold range on the dimension the benchmark is cited for. *Length is not a resolving-power axis.* Whatever a longer context makes harder, it does not make the state channel necessary, because announcement is a property of how the item was written and not of how far away it now sits.

**We registered the obvious mechanism, ran it, and it is wrong.** The natural way to support the paragraph above is behavioural: if last-occurrence is an $O(1)$-state heuristic then its precision should decay as the haystack grows, so `knowledge_update` scores should fall with length while $V$ stays pinned, and the fall would be retrieval difficulty at unchanged resolving power (Figure 6). We froze that as three criteria before the data existed (`prereg/E_SCALE_PREREG.md`) and then ran the readers, a ceiling arm and 2100 fresh judged items on both unseen splits with the judge held fixed. **All three failed**, and the reason is in Table 2: nothing moves. Not the reader, not the ceiling, not their ratio, not the retrieval precision measured without a judge at all. Our prediction required the benchmark to get harder with length; it does not get harder with length.

**Table 2. A tenfold increase in context length moves nothing on this axis.** All values on `knowledge_update`, official mode, one judge throughout (cross-judge comparison is caveat C1 and this is a within-judge question). $h$ is the retrieval hit rate, the share of items where the event the reader returned actually contains the gold, given that some event does, and it needs no judge, no model and no ceiling. It is the quantity leakage filtering does not reach: an item whose gold is retrievable by a substring search is not answerable *without* retrieval, so it passes the filter of Liu et al. (2026) while separating nothing an index cannot separate. Criteria frozen before the 500K and 1M cells existed; all three failed.

| quantity | meaning | 100K | 500K | 1M | frozen criterion |
|---|---|---|---|---|---|
| $S(\text{LO})$ | index-only reader | 0.100 | 0.157 | 0.093 | — |
| $S(\text{ceiling})$ | access-complete arm | 0.425 | 0.421 | 0.436 | — |
| $\rho = S(\text{LO})/S(\text{ceiling})$ | the ratio | 0.235 | 0.373 | 0.213 | $\leq 0.185$ at 1M: **FAILED** |
| $1 - V$ | not single-record answerable | 0.45 | 0.40 | 0.50 | (measured, registered earlier) |
| $E = (1-\rho)/(1-V)$ | gap in units of the bound | 1.699 | 1.568 | 1.574 | $\geq 1.899$ at 1M: **FAILED** |
| $h$ | retrieval precision, no judge | 0.136 | 0.262 | 0.200 | $\leq 0.036$ at 1M: **FAILED** |
| events per conversation (median) | | 271 | 1076 | 2088 | the axis being scaled |

Three consequences, and the first is a loss.

**Claim 5b is FAILED, and we do not rewrite it.** The pre-registration laid out all four combinations of ($h$ falls or flat) $\times$ ($\rho$ falls or flat) with a reading assigned to each in advance, precisely so that the outcome could not be reinterpreted afterwards. The observed cell is *$h$ flat and $\rho$ flat*, whose assigned reading is "length moved nothing; claim 5b FAILED." It is recorded that way. In particular the excess $E$ is flat at $1.57$–$1.70$ rather than rising: the measured index-only deficit is a stable $1.6\times$ what single-record answerability permits to be about state, at every scale.

**The thesis survives by a route the registered mechanism did not predict.** We argued that length is not a resolving-power axis because $V$ does not cross its bar as length grows. The measurement says something stronger and simpler: on this axis *nothing* crosses anything. Seven of the ten dimen-

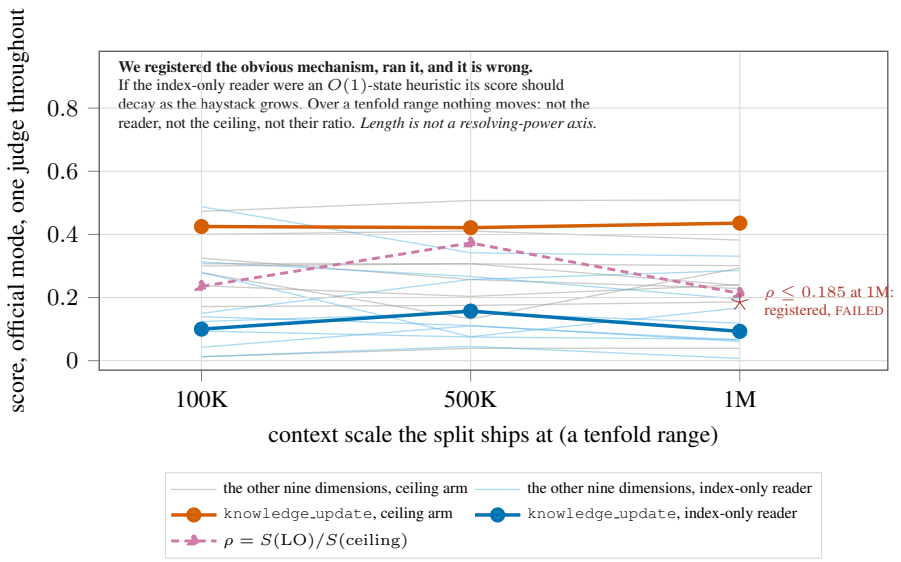



**Figure 6. We registered the mechanism, ran it on two unseen splits, and it is wrong in the direction that leaves the thesis standing.** Every dimension's ceiling arm (grey) and index-only reader (pale blue) across a tenfold context increase, with `knowledge_update` and the ratio $\rho$ drawn out. The registered prediction was that an $O(1)$-state reader's precision decays as the haystack grows, so $\rho$ should fall to $0.185$ at 1M; the star is that bar. Nothing falls, and nothing rises: not the reader, not the ceiling, not their ratio, and not the judge-free retrieval precision Table 2 reports beside them. *Length is not a resolving-power axis*, and the route to that conclusion is not the one we froze.

sions' ceiling scores move by less than $0.04$ across the tenfold increase (`knowledge_update` $0.425/0.421/0.436$, `multi_session_reasoning` $0.307/0.307/0.301$), and the reader-to-ceiling ratio ends where it started (Table 1). A benchmark whose headline axis is context length is reporting, on the dimension it names for updating, a quantity that the axis does not move. That statement is post hoc: no criterion was frozen on the ceiling's invariance, and the registered prediction that led here failed.

**The instrument ordering replicates; the inversion count does not.** P-T3-13 asked whether the weaker reader still wins at 500K and it passes at its bar, BM25-last takes $5$ of $10$; but $5/10$ is a tie, not the $6/10$ the judge companion measures under both judges (Anonymous, 2026g), and we report the weaker fact. What does replicate is the per-dimension winner: it is unchanged on $7$ of $10$ dimensions across all three splits, and the three that flip do so on the smallest margins in the table ($0.021$, $0.037$, $0.112$). That is the same pattern the companion's judge swap produced from a completely different perturbation (Anonymous, 2026g): rankings move where the scale has run out, and hold where it has not.

Three readings, and the third is the one we would keep if forced to keep one.

**(i) The two routes measure different failures, and their gap is diagnostic.** Where the structural bound is loose and the instrument bound tight, the items are not answerable from one turn and our readers are near the index-only optimum. Where the structural bound is tighter, `knowledge_update`, the answers *are* in single turns and our readers are failing to retrieve them. That gap is exactly the registered prediction of ours that failed: BM25-last recovers $0.242$ of a structural rate of $0.64$ on the unseen split. The two routes together locate the failure in retrieval rather than in the benchmark.

(ii) Nine of ten dimensions are bounded better by a weak reader than by their own text, which is the quantitative form of the paper's central negative: on this benchmark the answers mostly are *not* sitting in single turns, and that is why we could not condemn it. A dimension where $V = 0$ tells us nothing at all, and eight dimensions have $V = 0$.

**(iii) The one dimension where the text bounds it better is the one the benchmark was built to own.** `knowledge_update` is the ability a long-term memory system exists to provide, and it is the dimension where more than half the checkable answers are announced rather than composed: with no

model, judge or reader, the count says the update is announced. The step from that count to a bound on the *scored* quantity is judge-mediated, and the companion judge paper measures it: single-record answerability sits *below* the containment count at every scale, so the tempting instrument-free bound $(1-V)\cdot$ ceiling is unsupported, and the licensed replacement is instrument-dependent and looser by $1.45$ to $1.97\times$ (Anonymous, 2026g). The form this reading supports is therefore: on the dimension the benchmark is cited for, at most $0.63$ of the ceiling can be attributed to a state channel, and the reason that figure is not smaller is that no instrument-free route to a smaller one exists.

### A.6 THE BLIND THREE-CORPUS RUN, IN FULL

§5 reports the anywhere-containment curve on one benchmark, chosen knowing its outcome. Both restrictions are removable at zero API cost, and removing them is what this section does. Criteria are frozen in `prereg/E_RSTR_PREREG.md`, committed before `rstr.py` existed.

**The readout, and the one quantity this paper borrows a definition for.** The companion theory paper's limit sends the *item count* to infinity at fixed resolving power (Anonymous, 2026d). Pinning the budget and sending the *record length* to infinity instead is a different limit and a different paper (Anonymous, 2026e); what this section needs from it is only its readout, the budget at which a dimension stops being blind:

$$k^*(\tau) \;=\; \min\{\, k \in K : R^{\mathrm{str}}(k) \le \tau \,\}, \qquad \tau = 0.05, \quad K = \{1, 2, 4, 8, 16, 32\}, \tag{4}$$

restated here so that this paper is self-contained, and used here as a *level* at the one length each slot supplies. **It is a discontinuous read-out and we flag that where it is defined rather than where it is convenient**: $k^*$ moves only in ladder steps, and Schaeffer et al. (2023) show that a discontinuous scoring rule can by itself produce an apparent sharp transition in a quantity that is smooth underneath. Here the consequence is bounded and stated: a difference of one ladder step between two dimensions is a factor of two in $k^*$ and should not be read as a difference in kind, which is why the separation below is reported as `closes` against `None`, a difference between finite and infinite, rather than as a ratio of two finite budgets. §5's fixed-$k$ reading is one point on the same curve. The level separation below is not an exponent separation, said once more in the form the companion needs, *the level separation is not the exponent separation*, and reading it as one is the error the companion documents (Anonymous, 2026e).

**Blind, for the first time in this paper.** Three slots were fixed *by role* before any dataset was named: (A) conversational long-term memory, (B) document QA whose record is not a dialogue, (C) synthetic key–value retrieval (Figure 3). No $R^{\mathrm{str}}$ for any of them exists anywhere and we computed none before freezing. **Confirmations are therefore admissible here**, which is not true of §5 and not true of either outward count in the companion accounting paper (Anonymous, 2026c).

**The instrument check first, because everything else depends on it.** **P-T3-54 is SUPPORTED**: slot C gives $R^{\mathrm{str}}(1) = 0.000$ and $k^* = 1$ (Table 13). A needle task is a $k{=}1$ task by construction, so a retriever that works must close it at $k{=}1$, and unlike P-T3-42, this is a check a broken implementation cannot satisfy structurally. It is also a check a broken implementation cannot satisfy by structure alone, which Appendix B shows is not true of a monotonicity check.

**The graded verdicts.** **P-T3-52 is NOT SUPPORTED**: $k^* =$ `None` on both real benchmarks, so a 32-event budget does not close the gap outside the synthetic slot. The preregistration named P-T3-52 as the prediction we were least sure of and said which direction it would cost, and it cost that direction: "a small budget closes the gap" is not a property of the genre. P-T3-51 clears its bar on one slot only, slot B at $0.524$, while slot A pools $0.295$ against a bar of $0.30$, missing by $0.005$ and reported as missing (Figure 7). P-T3-53 was NOT TESTABLE here, each slot supplies one record length, and §9 supplies the second length and tests it.

**$k^*$ splits a benchmark's own dimensions along this paper's own distinction.** Using LONGMEMEVAL's own labels (Wu et al., 2025) and no model: `single-session-user` and `single-session-assistant` close at $k^* = 8$, `knowledge-update` at 16, and `multi-session` and `temporal-reasoning` do not close at all. A budget buys the single-hop dimensions and cannot buy the composition ones. That is Prop. 1's content on the budget axis, measured on a benchmark we did not write.

$R^{\text{str}}$ **needs a premise stronger than containment, and the direction of the gap is known.** Slot B returned $R^{\text{str}}(1) = 0.524$, high enough to look like a bug, so the retriever was checked by hand before anything was concluded. The retriever is fine; the premise is what binds. $R^{\text{str}} = a_{\text{complete}} - a_k$ means "the share of a gap a budget-$k$ retriever fails to close" only if $a_{\text{complete}}$ means *"recoverable from one event"*, and "the gold sits verbatim in *some* turn" (0.61/0.64/0.55 on BEAM's updating dimension) is a weaker statement than "the gold sits in *exactly one* turn" (4.3%, 6.1% and 5.0% there). Since $a_{\text{complete}}$ over-counts under the weaker reading and $a_k$ over-counts less, **the anywhere-form inflates** $R^{\text{str}}$, in the direction that favours the auditor.
**BEAM cannot supply the well-posed form either.** On items whose gold occurs exactly once, $a_{\text{complete}} = 1$ by construction and $R^{\text{str}}(k) = 1 - \text{hit@}k$ is well posed. No BEAM dimension has ten such items at any scale, so the exactly-once form is not computable there, and Appendix B reports the anywhere-form with this direction stated at the table rather than in a footnote. What is untouched: the curve's algebra holds for whatever $a$ is, Prop. 1 never mentioned containment, and slot C is 100% exactly-once, which is precisely why P-T3-54 passed: the instrument check and this premise are one fact seen twice.

**What the defect produced.** On the clean subset the gold is present and unique, so a budget-$k$ retriever that misses it is not missing a fact, it is failing to know which event to ask for. Ordering the cells that clear the frozen floor by $R^{\text{str}}(32)$ gives `kv_retrieval` 0.000 ($k^*$=1), `single-session-user` 0.000 (8), `single-session-assistant` 0.050 (8), `knowledge-update` 0.091, `multi-session` 0.500, `longbook_qa_eng` 0.842: a needle task, single-hop recall, an update, cross-session composition, narrative relations, **in that order**. A retrieval budget and two counts over text order tasks by how much composition they require, with no model, no judge and no system. It is post hoc and ungraded, and it is what we would register next.

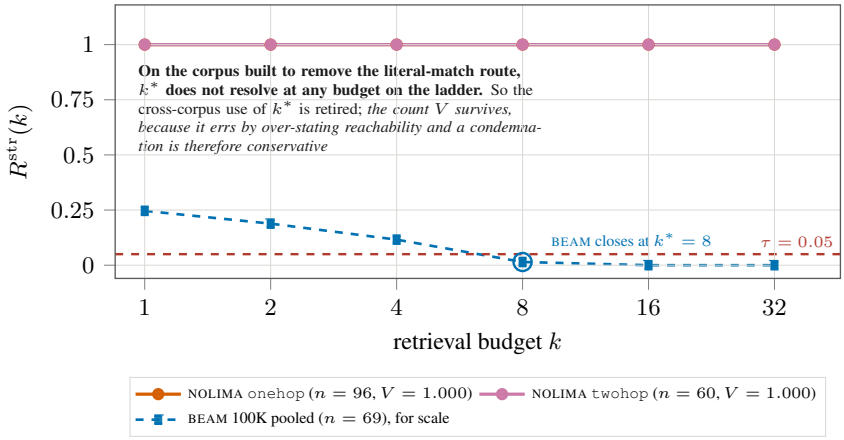



**Figure 7. The corpus built to defeat a lexical count, and what it leaves standing.** NoLiMa removes the literal-match route by construction, so it is the adversarial test of $V$: $R^{\text{str}}$ stays pinned at 1.0 at every budget on the ladder for both hop counts and $k^*$ does not resolve anywhere, against BEAM's pooled curve which closes at 8. *The cross-corpus use of $k^*$ is retired by this figure and the count survives it*: $V$ errs by over-stating reachability, so a condemnation built on it is conservative and a failure to condemn carries no information either way (§10).

### A.7 The load-bearing lexical assumption, priced in full

$V$ counts a *verbatim* occurrence, and the sharpest objection to everything above is that a long-context benchmark exists *because* lexical matching is exploitable, so an instrument made of string containment measures the wrong thing (limitation (i″)). The objection cannot be answered by argument, and it does not need to be: it names a corpus. NoLiMa (Modarressi et al., 2025) plants a needle whose lexical overlap with the question is minimal by design, while the answer, a character's name, sits *verbatim inside that needle*. Every item is therefore single-event answerable by construction while a term-matching reader has nothing to match on, which is exactly the cell this instrument's assumption is untested in. Criteria are frozen in `prereg/E_NOLIMA_PREREG.md`, committed

before `nolima_arm.py` existed; the protocol is the one used above with nothing tuned, the same BM25 ranking, the same containment rule, paragraphs as events, three placement seeds per item, haystacks truncated to 400,000 characters, no model and no judge.

**The instrument's premise holds and its read-out goes to infinity, on the same items.** $V = 1.000$ on both hop depths, 156 rows: the adapter gate P-T3-101 passes, so *every item is answerable from one event*. $a_k = 0.000$ at $k = 1, 2, 4, 8, 16$ and 32: the retriever never once places the needle in a budget of thirty-two paragraphs. Hence $R^{\mathrm{str}}(32) = 1.000$ and $k^*(0.05)$ does not resolve anywhere on the ladder, against $k^* = 8$ pooled on BEAM's 100K split. **P-T3-102 SUPPORTED**: one instrument, two corpora, both single-event answerable, and the read-out diff by the whole range of its domain. $k^*$ therefore measures the reader's lexical reach and not the task's composition depth, and limitation (i″) stops being a concession we make and becomes a measurement we report.

**Which half of this paper that costs, stated exactly.** The condemnation direction is untouched and was never at risk: a high $V$ still proves that single events answer the items, because containment is a witness and not an estimate, and §5's 0.55–0.64 on the updating dimension stands whatever a paraphrase-tolerant reader would add to it. What the result removes is the *comparative* use of $k^*$ across corpora, which §6 makes when it orders six cells by $R^{\mathrm{str}}(32)$ and reads the order as composition depth. That ordering now has a measured confound with a name, lexical reach, and a corpus where it is maximal. We therefore keep the ordering as the post hoc observation it was already labelled, and add: any cross-corpus reading of $k^*$ owes a lexical-difficulty control, and this section is that control returning *positive*. The instrument that survives unrestricted is the count $V$, which is §10's conclusion arrived at from a second direction.

*One thing this arm cannot distinguish, and we decline the stronger reading.* $a_k = 0.000$ at every $k$ means the needle is never retrieved; it does not follow that *no* index-only strategy at $k \leq 32$ could retrieve it, which is Prop. 2's one-sidedness in its own home. A reader with the needle's template could find it trivially. The claim is therefore about this reader (the audit's own, held fixed on purpose), and its consequence is about $k^*$ as a *reported quantity*, not about NoLiMa's difficulty.

**Where this paper's $k^*$ stops and the budget companion's begins, said from this side of the boundary.** The two papers share one definition ((1)) and nothing in the data, and the division is not a courtesy: what this paper reports is a *level* at the one record length each corpus supplies, and what the companion reports is a *slope* of $\log k^*$ against $\log L$ within one corpus and one reader (Anonymous, 2026e). A level is not a point on our ladder that the companion could extend, because moving between our slots moves the corpus, and this section has just measured what moving the corpus does to $k^*$: it can take the read-out across the whole range of its domain while the count $V$ stays at 1.000. No number in Table 13 is an input to any exponent, and the transport condition the companion names for its own two corpora, that the budget closes the containment gap on both, is the pre-check this section shows can fail with $V$ maximal. That is the strongest thing we can offer the companion and it is a restriction rather than a result: our contribution to its limit is a corpus on which the limit's read-out does not exist.

### A.8 The strongest index-only adversary, in full

By the channel definition of the companion theory paper, a retrieval-plus-model arm is an index channel (Anonymous, 2026d). Then it was never a ceiling; it was an adversary filed on the wrong side. Criteria P-T3-85 to P-T3-89 were frozen in `E_ADVERSARY_PREREG.md` before the runner existed, and the arm moves the adversary's budget with a model attached: BM25 over the same 4-turn chunks, same prompt, same model, same judge, same split, at `TOP_K` $\in \{1, 2, 4, \mathbf{up\ to}\ 16\}$ with the existing $k$=8 run reused. The last rung is *up to* 16 rather than 16 because the retriever drops zero-scoring chunks, so short conversations supply 12.

**P-T3-86 is SUPPORTED**: two dimensions clear the frozen 0.80 bar. *We report it as retrieval sufficiency rather than a resolving power*: numerator and denominator are both index channels, so $\rho$ is the index-only class's internal budget spread (Anonymous, 2026d) (Figure 8, Table 3). It reaches blindness through Prop. 1's contrapositive and never through (2). On `knowledge_update` that route is independently corroborated: $V = 0.55$ from text alone.

**Table 3.** $\rho(k) = S(\text{adversary at } k)/S(\texttt{vanilla})$, one judge, 400 items per rung. Both sides are index channels, so this is a retrieval-sufficiency curve.

| dimension | $k$=1 | $k$=2 | $k$=4 | $k$=8 | $k\leq$16 | clears 0.80 |
|---|---|---|---|---|---|---|
| knowledge_update | 0.56 | 0.68 | 0.78 | **0.83** | **0.90** | $k$=8 |
| information_extraction | 0.39 | 0.50 | 0.64 | 0.64 | **0.81** | $k\leq$16 |
| summarization | 0.14 | 0.22 | 0.41 | 0.51 | 0.78 | — |
| multi_session_reasoning | 0.26 | 0.37 | 0.40 | 0.61 | 0.72 | — |
| preference_following | 0.36 | 0.59 | 0.62 | 0.57 | 0.71 | — |
| event_ordering | 0.36 | 0.50 | 0.57 | 0.64 | 0.70 | — |
| temporal_reasoning | 0.21 | 0.46 | 0.54 | 0.62 | 0.67 | — |
| instruction_following | 0.25 | 0.39 | 0.48 | 0.43 | 0.66 | — |
| abstention | 0.36 | 0.50 | 0.32 | 0.59 | 0.50 | — |
| contradiction_resolution | 0.09 | 0.03 | 0.00 | 0.06 | 0.03 | — |

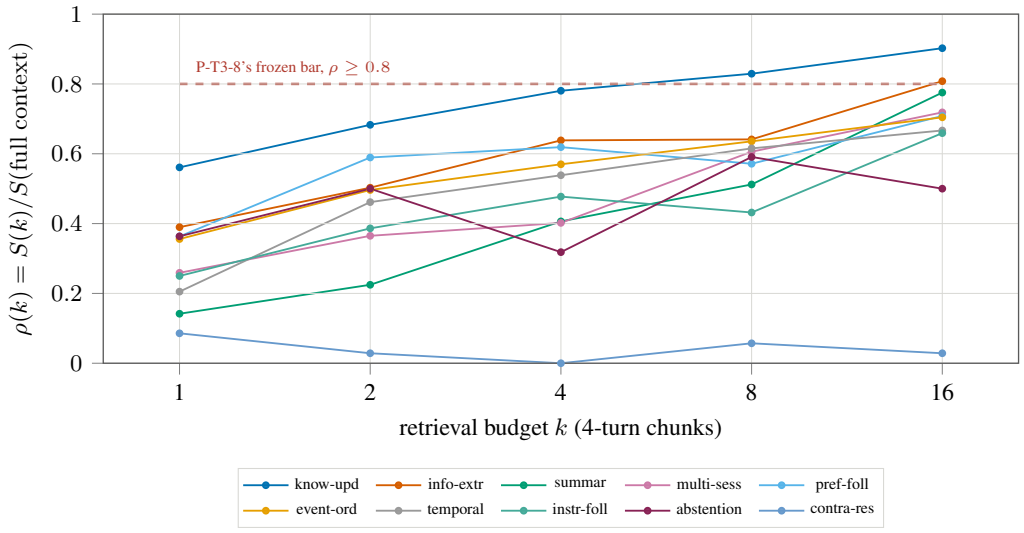



**Figure 8. Retrieval sufficiency on a benchmark we did not write.** $\rho(k)$ is the score a BM25-plus-model adversary reaches at budget $k$, as a share of what the same model reaches with the whole record. Both are index channels (Anonymous, 2026d), so this is a curve *within* one channel and not a resolving power. Two dimensions cross the frozen bar. Every coordinate is recomputed from `results/adversary/ADVERSARY.json`.

**Two of the criteria in this arm were frozen below their own resolution, and the resolution arithmetic is the criterion companion's** (Anonymous, 2026f). **P-T3-85 FAILED** at its frozen $\leq 0.02$ band: the $k$=8 re-run's largest dimension difference is 0.0500. Diagnosis before interpretation, 55 of 400 items change score on a re-run of the *identical imported code path* (Anonymous, 2026f), because `temperature`=0 hosted inference and a judge call are both resampled. The per-item paired SD is 0.182, so the paired SE at $n$=40 is 0.0287 and the frozen band was 0.70 of one standard error. The pooled mean difference over all 400 items is +0.0010 ($t = 0.11$): no drift. `knowledge_update` had 0 of 40 items change, so its result is not a resampling artefact.

**P-T3-88 is NOT DECIDABLE**, not NOT SUPPORTED. $\rho$ is non-monotone on four dimensions, and 0 *of those dips exceed* 2 SE (1.67, 1.43, 0.57, 0.56); one is a flat-at-zero line where monotonicity has no meaning. We froze a *strict inequality with no tolerance*, so no SE at all is tolerable and the criterion cannot decide. Consequently "more retrieval hurts" is **not** a finding of ours and is not reported as one.

**Those four numbers are larger than the ones this paper first printed, which is the correction and not a restatement.** A dip is $\rho(k) - \rho(k{+}1)$, a contrast, and it was compared against max of the two cells' standard errors, one arm's spread standing in for the spread of a difference. The rungs are the same 400 shards re-scored under identical filenames, so the supported treatment is *paired*, and the paired standard error of a per-item difference is *smaller* than either cell's: the deciding dip's SE falls from 0.130 to 0.109 and its $z$ rises from 1.40 to 1.67. **The correction moves every dip toward the finding we declined to make, and none of them arrives.** What prompted the check was not our own re-reading: the criterion companion's related work now engages a contemporaneous paper that prices this exact substitution in the sample-size currency, an unpaired shortcut missing $N^*$ by about a factor of two in the close-comparison regime (Anonymous, 2026f, §2), and pointing that result at our own grader found the line.

**P-T3-87 is NOT SUPPORTED, and it exposes something worse than the criterion.** Only 3 of 10 dimensions gain $\geq$ 0.15 from having a model at one chunk, and *five lose*: −0.40 on `contradiction_resolution`, −0.33 on `preference_following`, −0.32 on `instruction_following`, −0.28 on `event_ordering` and −0.10 on `multi_session_reasoning`. A one-event dump with no model beats a model-bearing adversary on half the benchmark. That is not possible on information grounds, so the dump is being paid for something other than information; the companion judge paper traces the payment to rubric criteria a response satisfies by saying nothing, a floor of 17.5% of criteria on the dimensions whose gold is never in one event, and every ratio in this section's table inherits that floor on its numerator (Anonymous, 2026g).

P-T3-89, reported not graded: on the three dimensions with short golds, containment in the retrieved chunks reaches 0.61 on `knowledge_update` at the top rung while $\rho$ reaches 0.90. The model scores well above the rate at which the gold is retrieved at all, so containment *under*-predicts here, a fourth mechanism, guessability, that neither route measures. The forecast table scored 2 hits and 2 misses.

### A.9 THE CHANNEL CLASSIFIER, IN FULL

A definition that has to be applied by reading an architecture description is not yet an instrument. A name-based check, search the weight index for attention projections, disqualifies RWKV, whose time-mixing block HUGGINGFACE happens to name `attention` while it holds `time_decay` and `time_mix_*` and retains no per-position keys at all. *A classifier that reads names classifies names.*

`channel_probe.py` measures instead, and the measurement is three lines of principle: run the model at several sequence lengths, walk whatever cache object it returns, count the tensor elements. A carrier of bounded width returns the same count at every length; an index returns a count proportional to it (Table 4). Nothing in it needs to know what the architecture is called, and it gives the interesting intermediate case its right name: a model with local attention over a window $w$ grows to $w$ and is then flat, which is *an index at budget $w$*, exactly as the budget-indexed definition of the companion theory paper would have it (Anonymous, 2026d) (Figure 9).

**Table 4. Channel class, measured over a $16\times$ span of sequence length rather than read off a model card.** Carrier elements are counted by walking whatever cache object the runtime returns, in millions, at 64 and 1024 tokens. A bounded-width carrier returns the same count at every length, i.e. growth $1.00\times$; an index returns the span. A name-based check disagrees with this measurement on the `rwkv` rows, which is why the check has to be a measurement. Generated from `results/ssm/CHANNEL.json` by `channel_probe.py`.

| model | what it is called | carrier, $10^6$ elements at 64 | at 1024 | growth | class by Def. 1 |
|---|---|---|---|---|---|
| `gpt2` | transformer, KV cache | 1.18 | 18.87 | $16.00\times$ | index (at whatever $k$ the stack exposes) |
| `falcon-mamba-7b-instruct` | selective SSM | 10.49 | 10.49 | $1.00\times$ | state channel |
| `rwkv-raven-14b` | linear attention | 1.02 | 1.02 | $1.00\times$ | state channel |
| `rwkv7-2.9B-world` | linear attention | 5.41 | 5.41 | $1.00\times$ | state channel |
| *local attention, window $w$* | *(not run)* | — | — | *$w$ then flat* | *index at budget $w$* |

Two consequences worth stating separately from the arm below. First, this is the check a *reader* of this paper can run on a system we never touched, so the classification our audit depends on is not a claim of ours. Second, the width axis is the one the channel definition names and the only one this probe sees: a model can have a bounded carrier and still cost $O(n)$ per token to update it, which is precisely the gap between a prompt-level scratchpad carrier and a genuine state-space model, the gap the companion theory paper's SSM arm measures (Anonymous, 2026d).

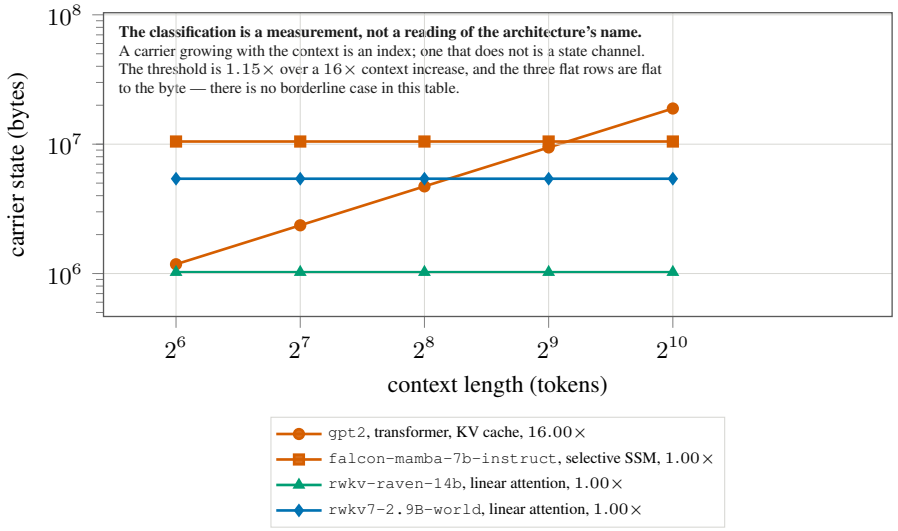



**Figure 9. Def. 1's dichotomy is a measurement, not a reading of the architecture's name.** What each model must carry forward per token, as the context grows $16\times$. A transformer's KV cache grows with the context and is therefore an index; the three recurrent carriers are flat *to the byte*, which is what puts them on the other side of the definition. The classification threshold is drawn in Table 4 and no row in this table is near it, the measured growths being $16.00\times$ and $1.00\times$, with nothing between.

### A.10 The record-length axis, in full

(1) pins the budget and lets the record length go to infinity, and this paper cannot follow it: §6's slots each supply one length. Six arms were built to follow it, two record lengths of the same 500 questions, a nested deletion ladder of seven rungs, a registered and powered re-run of that ladder, an undeleted lever, a skill-free retrieval baseline and a recurrence-controlled repair, and they are a paper of their own (Anonymous, 2026e) for a reason this paper should state rather than imply: *their object*

*is the budget a task demands, not whether a benchmark can see a distinction.* Three of their results are quoted here because §5 and §10 depend on them.

- **The required budget is a constant share of the record.** On nested records built by deletion the exponent in $k^* \propto L^{\alpha}$ is $\alpha = 1.011$ over a $62\times$ ladder and $k^*/L$ is pinned at $0.400$–$0.417$, so "a small retrieval budget" is a statement about a record length and not about a task, and every fixed-$k$ number in this paper is a number at this benchmark's shipped length.
- **That exponent belongs to the lever, not to the task.** The same resource moved without deleting anything gives an exponent that reverses the ordering of the dimensions, and no lever on a fixed corpus moves length alone; deletion takes copies of the gold with it, a second draw takes the distractors with it. This is the same shape as the companion theory paper's non-identifiability corollary, one level down (Anonymous, 2026d).
- **The comparison is undecidable at this benchmark's size.** Zero of five bootstrap intervals on the difference of the two exponents exclude zero, and closing the gap at the band we kept reaching for would take 12 to $54\times$ the items. Three of those criteria were frozen below the resolution of the statistic they graded, which is audited in (Anonymous, 2026f) and not reported here as a null.

What this paper keeps from that programme is one sentence, and it is a restriction on §5: $V$ and $R^{\mathrm{str}}$ are measured at the length the benchmark ships, and a bound read off them is a bound at that length.

### A.11 What the audit established

The audit returns $0$ of $10$ dimensions established as blind. Prop. 2 is why that is not evidence of sight: the instrument is one-sided, so a failure to condemn is silent rather than exculpatory, and the companion theory paper shows the $0/10$ was structurally guaranteed once every arm in the apparatus is an index channel, the comparison being the index-only class's internal budget spread rather than a resolving power (Anonymous, 2026d). What the audit does establish needs no instrument: the counts of §5 and §5, the scale invariance of §5, and the class recovery of §6. **One of those three now carries a measured scope condition**: the count $V$ travels to a corpus built to defeat lexical matching, while the budget read-out $k^*$ does not (§6), so the instrument that survives unrestricted is the count and not the ladder computed from it. The scorer-side findings on the same benchmark, two defects in released code and a judge that is strict and lenient at once, are the judge companion's (Anonymous, 2026g).

### A.12 Three deployed cases where the instrument changes the reading

We did not find this by auditing other people's benchmarks.

**A protocol that could not see the channel it was built to test.** Our own delay–load protocol swept a load axis with rearrangement events driving a composition query. A last-occurrence reader scored $1.00$ on all three query axes and was classified access-complete. The defect was one design decision: each rearrangement announced the *resulting value* rather than the *pair of keys exchanged*, so the composed answer appeared verbatim in the write record and Prop. 1 applied exactly. This is the same shape Table 12 measures in someone else's benchmark.

**An exclusion decided by an instrument already condemned.** In an earlier study of this programme we dropped a skill for "unlearnability" after every arm landed inside the noise floor. Its evaluation set was prose, median gold $154$ characters, only $8.0\%$ short enough for the exact-match scorer; every number lay at the fraction of items scorable at all. A skill pinned at its scorable fraction is indistinguishable from one that does not learn, and extra seeds never separate them.

**An intervention whose label was anti-correlated with truth.** In a companion experiment on self-written stores (Anonymous, 2026b), a provenance marker appeared to make contamination worse. The planted contamination had been seeded as OBSERVED while the agent's own correct answers were DERIVED, so the marker instructed the model to trust the lie. Graded UNTESTED (APPARATUS FAILURE), not FAILED.

All three share the shape this paper generalises: *the instrument's blindness and the phenomenon's absence produce identical readings*, and no amount of statistical power distinguishes them. Power gates measure whether an effect of a given size would be detected if the instrument could see it. They are silent on whether it can.

### A.13 WHERE ZERO RESOLVING POWER IS LOAD-BEARING

Three places where a blind item would change a published conclusion, *none of them graded here*, stated in one sentence each because an objection one has not paid for is worth what it cost. **(a) Defence evaluations**: a memory-poisoning defence is validated by the score drop it prevents, and on an item answerable from one turn the drop is bounded by the retrieval budget it was measured at, times the judge's credit rate (Anonymous, 2026g), so a defence measured at one budget makes a claim only at that budget. **(b) Admission rules that decide by agreement**: a rule filtering a store by cross-source agreement cannot separate a fact that recurs because it is true from one that recurs because it was copied, which is the redundancy-for-difficulty conflation the record-length companion measured in our own exponent (Anonymous, 2026e). (c) Instruments validated on their own construction set: an instrument scoring perfectly on the set it was built from and zero off it is measuring its own construction, which is the situation of the companion theory paper's witness (Anonymous, 2026d) and why that paper reports it as a positive control. A defence author may reply that their attack is evaluated on the subset where composition *is* required, which would be correct and would make the verbatim count their tool rather than our objection; we have not checked whether any cited work does this, and saying so is cheaper than the check. The measured versions of this argument are not ours: for (a) and (b) the closest checkable evidence is Wang (2026)'s single-variable sweep and Sidik & Rokach (2026)'s architecture-invariant null, and a reader should weigh those above the three arguments in this paragraph.

### A.14 WHAT REMAINS OWED, IN FULL

*Neither arm below can overturn a condemnation, because the instrument is one-sided by Prop. 2.* $V$ is a conservative floor on single-event answerability, so a paraphrase-tolerant count can only raise it and can only make a task *more* blind than reported; the arms bear on what the instrument may additionally condemn, never on what it has already condemned.

**(1) The lexical-difficulty control is run, and what it leaves owed is a paraphrase-tolerant count.** §6 prices the instrument's central assumption on a corpus built to defeat it and finds $k^*$ unresolved where BEAM's is 8, so the cross-corpus use of $k^*$ is retired and the count $V$ survives. What that arm does not supply is a $V$ that tolerates paraphrase. Verbatim containment lower-bounds single-event answerability, so a low $V$ proves nothing and §5's null dimensions stay null by assumption rather than by measurement. The arm is an entailment-scored $V$ over the same items, with the verbatim count kept beside it as the conservative floor, and it is the one instrument extension in this paper that would need a model, which is why it sits here rather than in the body.

**(2) A budget-matched *scored* reader at the larger splits.** The counting side of this question is closed: Appendix B's ladder holds a BM25 reader's containment coverage at every $k$ up to 32 on all three splits, and §4(i) now reads the $0/10$ off it. What is owed is the scored counterpart at a matched budget: §7's judged ladder runs on 100K only, so a dimension whose gold comes into reach at $k{=}8$ on 1M is still graded there by a $k{=}1$ reader. The arm is one pass of the existing ladder at the two larger splits. It is owed rather than done because it is judge-mediated, which puts its verdicts in the companion's frame rather than this paper's zero-judge one (Anonymous, 2026g).

**Four of this paper's own frozen bands could not have decided what they were asked**, which is a fact about the bands rather than about the world. We do not report those four as nulls anywhere above. The arithmetic is one line each and it is here rather than in a companion, because a label a reader cannot check is not a label:

- **P-T3-71**, a two-point exponent whose $k^*$ moves only in doublings over a lever of ratio $10.05$: the smallest non-zero value it can take is $\log 2 / \log 10.05 = 0.301$, against a frozen band of $0.150$.
- **P-T3-67** and **P-T3-68**, exponents fitted over the ladder: the widest measured $95\%$ bootstrap interval over items reaches $[0.000,\ 1.064]$, so $\Delta = 1.064$ against bands of $0.150$ and $0.519$.

- **P-T3-85**, a per-dimension mean difference on a re-run at $n = 40$, paired: the per-item paired SD is $0.182$, so $\Delta = 0.182/\sqrt{40} = 0.0287$ against a band of $0.020$.
- **P-T3-88**, a strict monotonicity: a strict inequality has $B = 0$, so it is undecidable for any $\Delta > 0$, and here the paired $\Delta$ of the deciding dip is $0.109$.

Each of those is a division a reader can do; the general treatment, the composition rule and the audit of the whole programme's bands are a companion's (Anonymous, 2026f), and this paper imports the vocabulary and not the verdicts. Two of the four were frozen *after* the check that catches this existed, which is the criterion companion's argument that the check must be a script rather than a habit.

### A.15 What this paper needs from its companions

Companion manuscripts are cited throughout, and a reader is entitled to ask whether this one stands on its own. The split is clean enough to state, and Table 5 carries it.

**Table 5. What is self-contained and what is imported.** The left column is everything a reader can check without leaving this paper. In every imported row the import is a *restriction* on what may be claimed rather than a premise a result rests on.

| self-contained | imported, and what it costs |
|---|---|
| The instrument: $V$, $R^{\mathrm{str}}(k)$, $k^*$, their definitions, the CAND sweep, and every count in §5–§6 | **Nothing.** Two counts over text and a retriever |
| The one-sidedness of the audit (Prop. 2) and the blindness condition (Prop. 1, restated and proved in §3) | **Nothing**; proved here. The general theory around them is the companion's (Anonymous, 2026d) |
| The channel classifier and its table (§8) | **Nothing** to run; the width definition it operationalises is the companion's |
| — | The *record-length* restriction: every level here is measured at the lengths the corpora ship, and the budget's growth with length is (Anonymous, 2026e). Without it we would over-read $k^*$ |
| — | The step from any count to a bound on a *scored* difference: judge-mediated, priced by (Anonymous, 2026g), and used for no verdict here |
| — | The state/index distinction itself (Anonymous, 2026a) |

### A.16 The reach of the audit, in full

The headline says ten dimensions. **The audit's actual reach is three**, and stating that plainly is worth more than the headline. Our instrument needs the gold string to be short enough to check for containment; on the flagship benchmark six of ten dimensions have no checkable item at all, and only `knowledge_update` has enough of them to grade across all three context scales. Every containment-based number in the audit is therefore about a subset, the subset is reported per cell in the tables of §5 and Appendix B, and a reader should discount the ten-dimension framing accordingly.

Two things keep this from being fatal, and they are different in kind. The direction of the bias is known: containment-admissibility selects items whose answer is verbatim present, which is the *easy* subset for a retriever, so a dimension that looks reachable on it may be less reachable overall, our failures to condemn are therefore conservative and our condemnations are not weakened. The instrument-free route does not have this problem at the same severity: §6 runs blind on three slots fixed by role before any dataset was named, and the companion runs the same ladder on a second corpus of 198 items whose records are paragraphs rather than dialogue turns, recovering the same boundary function with a different constant (Anonymous, 2026e). **Portability of the *instrument* is therefore demonstrated on three corpora; portability of the *audit of* BEAM is three dimensions.** Those are different claims, and the ten-dimension framing borrows the wrong one's reach.

### A.17 ONE OPEN QUESTION

**Is the criticism of the flagship that the dimension fails, or that the field misreads it?** We measure that on `knowledge_update` the gold string is verbatim present in one turn on $V = 0.55$ of checkable items. That is a fact about the corpus. It becomes a criticism only under the reading that the dimension tests *composition across records*, and that reading is one the benchmark's own documentation does not assert in those words; it is supplied by citation practice, including by the parent framework and by us. The finding is therefore: *a dimension whose name invites a compositional reading is, on more than half its checkable items, answerable by retrieving one turn*. Whether its authors intended a compositional test is not something a count can measure, and we do not claim it; we searched for external work reading the dimension as cross-record composition and found none, and the nearest external reading goes the other way (Tanguturi, 2026). The count stands in place of any stronger sentence.

**Does a benchmark owe transfer tests to itself?** The flagship labels ten dimensions and is cited for statements of the form "a system that does well on `knowledge_update` handles `multi_session_reasoning`". That is a transfer claim between two regimes inside one artifact, and the accounting companion shows that such claims multiply as $P(P-1)$ while the tests of them do not (Anonymous, 2026c). For ten dimensions that is **ninety ordered pairs**, and we are aware of no published test of any of them; the dimensions are reported side by side and the reader supplies the generalisation. This paper measures something adjacent and does not resolve it: $k^*$ separates the dimensions into the classes the benchmark's own labels assert (§6), which is evidence that the labels track *something*, and says nothing about whether performance transfers between them. Applying the companion's accounting inside a single benchmark rather than across a programme is a straightforward extension and neither paper has done it.

### A.18 LIMITATIONS, IN FULL

**(i) One distinction, and now four benchmarks, but only the instrument-free half travelled.** Everything *instrumented* here is the state/index dichotomy on one benchmark at one split family. §6 takes the structural half to three more, blind, and that half is exactly the half needing no oracle. Other distinctions need their own oracles and nothing guarantees each has a baseline as cheap as this one.

**(i′) $V$ is a *lexical* count, and the direction it errs in is the safe one.** NoLiMa was built to remove the literal-match route, and the strongest model there falls from $99.3\%$ to $69.7\%$ once it is gone (Modarressi et al., 2025). $V$ counts an item as reachable whenever its gold string appears, so it *over*-states reachability and is an upper bound on what a one-record reader can be handed rather than a measure of what it can answer. **That direction makes a condemnation conservative and a failure to condemn uninformative**, which is the one-sidedness Prop. 2 states. Where it is not safe is the comparison across dimensions, and that half is now measured rather than conceded: on NOLIMA itself this instrument's $k^*$ does not resolve at any budget on the ladder while BEAM's resolves at $8$ (§6), so the cross-corpus use of $k^*$ is retired and the count is kept. What remains is that $V$ has no paraphrase-tolerant form here (§A.14(1)).

**(i″) The containment proxy is weak where a benchmark's golds are common strings, and the price is measured.** $R^{\mathrm{str}}$ requires golds sitting in *exactly* one event; across the four benchmarks that share runs from $4.3\%$ to $100\%$, and the two conversational memory benchmarks sit at the bottom. Where the share is low the quantity is inflated in our favour (§6). A model-free proxy is cheap because it does not ask what an event *means*, and that is also the assumption it is silently making.

**(ii) The audit did not condemn, and cannot acquit.** $0/10$ leaves the central question open. A field in which most benchmarks fail to be condemned is a field where this method reports very little, and that is a property of the method.

**(iii) Our ceiling is ours.** It is the best of two arms we built, a BM25 retrieval arm using the benchmark's own answering prompt and a full-context arm, and not the benchmark's best published system, whose answers are not released. A stronger ceiling would only lower every ratio further.

**(iv) One judge decides the judged numbers, and the counting chain never asks one.** The judged arms (§5, §7) use one pinned judge, and the companion measures what changing the pin does: on `abstention` a judge change carries the score from $0.075$ to $1.000$, a shift of $0.925$ (Anonymous, 2026g, row 204). Because $\rho$ is a ratio of two judged scores, a judge that carries both sides toward $1.000$ flattens the curve rather than renumbering it, and `abstention` is a row of Table 3 sitting at $0.36/0.50/0.32/0.59/0.50$, mid-range at every rung, exactly the headroom a saturating judge removes. The four criteria graded in §7 are therefore conditional on the pin and we do not claim they transport. What is not conditional is the part that never asked a judge: $V$, $R^{\mathrm{str}}$ and $k^*$ are counts over text, and the $0/10$ of §5 and every verdict in §6 and §6 read out of them.

**(v) §A.13 is an argument.** Its three consequences are reasoned from Prop. 1 and one structural count on one dimension. No defence evaluation, no admission rule and no leaderboard was audited here, and none of the three is graded. They are stated because a bound whose consequences are not named is easy to agree with and hard to use.

**Artifact availability.** Every preregistration, grader, sweep runner and result file named in this paper exists in a repository released publicly *after* acceptance; until then a reader has this document's timestamp and nothing else. A criterion a reader cannot reach is indistinguishable, to that reader, from an untested claim; the term for it is a *locatable* test, and a companion develops the accounting around it (Anonymous, 2026c). The locatable share of every criterion in this paper is $0$ for everyone but us while the repository is private, whatever our own commit history says. What the preprint buys is the other half, a timestamp issued by someone who is not the claimant. The ordering is therefore: criteria frozen in private commits, anchored by this preprint's date, made locatable on release.

One notational point, because this paper uses it. A citation of the form `[row NN]` names a row of the cited paper's claim ledger, `<paper>/claims_all.tex` in that repository, and not a row of any table printed in the cited PDF, no paper of this family prints its ledger, which is a deliberate choice that costs exactly this: until release, a row pointer is locatable to us and not to a reader. We use it anyway because the alternative is an imported number with no identifier at all.

## B THE ANYWHERE-CONTAINMENT CURVE ON BEAM

$R^{\mathrm{str}}(k)$ is well posed when $a_{\mathrm{complete}}$ means "recoverable from exactly one event"; on BEAM only $4.3$–$6.1\%$ of containment items satisfy that premise, below the grading floor on every dimension, so the exactly-once form is not computable there (§6). This appendix reports the *anywhere* form, $a_{\mathrm{complete}} = V$ (Table 6): an upper-bound variant whose levels are inflated in the auditor's favour, and whose decay pattern and cross-rung comparisons are what the body uses (Figure 10). Criteria are frozen in `prereg/E_BEAMCURVE_PREREG.md`, committed before `beam_curve.py` existed.

**A monotonicity check that a broken implementation satisfies structurally is not a test of that implementation.** `bm25_topk` returns its $k$ best re-sorted into *transcript* order, because §4's reader needs recency; so scoring once at $k$=32 and slicing takes the *earliest* $k$ of the top 32, not the top $k$. P-T3-42, registered as a monotonicity check on the implementation, cannot catch that class of fault: sliced nested sets are monotone whatever they contain. What catches it is content, $R^{\mathrm{str}}(32)$ coming out $0.000$ on all three dimensions at once is too clean to be a measurement, and `nested_topk` therefore ranks by score and asserts on a sample that its sets equal `bm25_topk`'s element for element.

## C EVERY CRITERION THIS PAPER GRADES, WITH ITS FREEZE DATE

Read from `git log --reverse` at build time. The claim a preregistration makes is a claim about *time*, so it is the one claim in this paper that must not be transcribed. Amendments are counted and never overwrite: P-T3-26 was added by amendment to `E_ORACLE` specifically so that the extract arm's criterion was frozen before its data, after P-T3-21 turned out to measure paste-answerability rather than the quantity the paper calls $A$ (§5). The criteria graded in *this* paper are the `E_RSTR`, `E_BEAMCURVE`, `E_SCALE`, `E_NOLIMA` and `E_ADVERSARY` rows.

**Table 6.** $R^{\mathrm{str}}(k)$ **on BEAM, computed with no instrument.** For every probing question with a gold of at most 80 characters, the same containment subset §5 uses, $a_k$ is the share whose gold appears verbatim in the top-$k$ turns under the same BM25 scorer §4 condemns, and $a_{\text{complete}} = V$ is the share whose gold appears anywhere. **That last phrase is weaker than $R^{\mathrm{str}}$ requires**: only 4.3–6.1% of these items have a gold occurring in *exactly* one turn, so every level in this table is inflated in the auditor's favour, and the decay pattern and cross-rung comparisons are what it supports (§6). Six of the ten dimensions have *no* checkable item on any split and are excluded by the preregistration's floor clause; one more is excluded on two splits. The instrument therefore reaches **three dimensions of ten**, which is the size of the portable part. **The whole tabular is generated from `results/beamcurve/GRADE.json`**, and the generator found four rows the hand-typed version had dropped with no clause saying so: `information_extraction` and `multi_session_reasoning` clear the floor on *all four* splits and appeared only at 100K and 10M. The pooled row always included them, so nothing was overstated, what was missing is that *the decay is one dimension's*: the added rows are flat, and every falling curve here is `knowledge_update`'s. Bold is each split's largest $R^{\mathrm{str}}(1)$, a rule replacing three hand-placed cells.

| split | dimension | $n$ | $V$ | $k$=1 | $k$=2 | $k$=4 | $k$=8 | $k$=16 | $k$=32 |
|---|---|---|---|---|---|---|---|---|---|
| 100K | information_extraction | 17 | 0.059 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | knowledge_update | 36 | 0.611 | **0.417** | 0.306 | 0.167 | 0.000 | 0.000 | 0.000 |
| | multi_session_reasoning | 16 | 0.188 | 0.125 | 0.125 | 0.125 | 0.062 | 0.000 | 0.000 |
| | *pooled* | 69 | 0.377 | 0.246 | 0.188 | 0.116 | 0.014 | 0.000 | 0.000 |
| 500K | information_extraction | 52 | 0.077 | 0.019 | 0.019 | 0.019 | 0.019 | 0.019 | 0.019 |
| | knowledge_update | 66 | 0.636 | **0.439** | 0.288 | 0.182 | 0.121 | 0.061 | 0.030 |
| | multi_session_reasoning | 28 | 0.179 | 0.107 | 0.071 | 0.071 | 0.000 | 0.000 | 0.000 |
| | *pooled* | 146 | 0.349 | 0.226 | 0.151 | 0.103 | 0.062 | 0.034 | 0.021 |
| 1M | information_extraction | 37 | 0.054 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | knowledge_update | 64 | 0.547 | **0.328** | 0.266 | 0.203 | 0.141 | 0.094 | 0.031 |
| | multi_session_reasoning | 16 | 0.062 | 0.062 | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 |
| | *pooled* | 117 | 0.325 | 0.188 | 0.154 | 0.111 | 0.077 | 0.051 | 0.017 |
| 10M | information_extraction | 17 | 0.118 | 0.059 | 0.059 | 0.059 | 0.000 | 0.000 | 0.000 |
| | knowledge_update | 16 | 0.438 | 0.250 | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 |
| | multi_session_reasoning | 11 | 0.364 | **0.364** | 0.364 | 0.273 | 0.273 | 0.273 | 0.273 |
| | *pooled* | 44 | 0.295 | 0.205 | 0.136 | 0.091 | 0.068 | 0.068 | 0.068 |

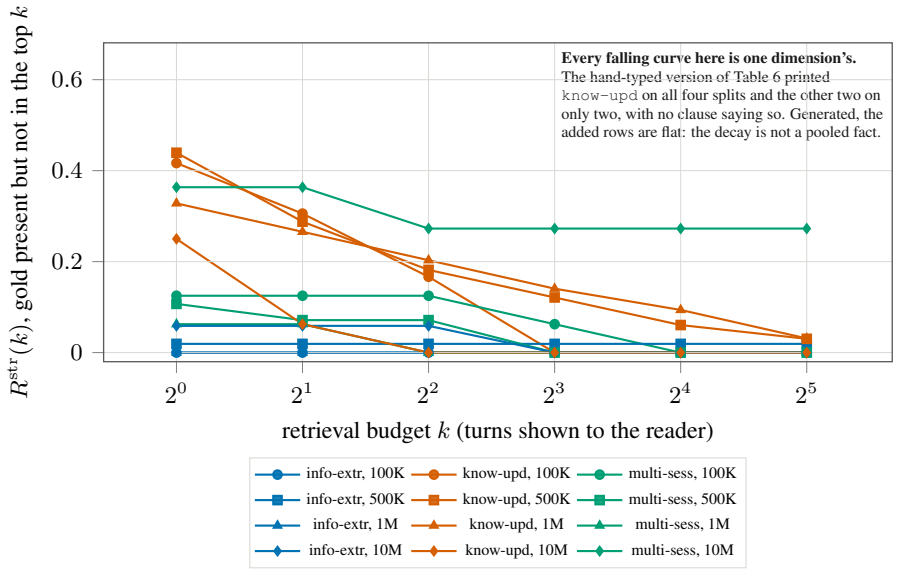



**Figure 10. Every falling curve on this axis belongs to one dimension.** $R^{\mathrm{str}}(k)$ is the share of items whose gold sits somewhere in the transcript but not in the top $k$ turns the reader is shown: text and a BM25 pass, no model, no judge, no ceiling. The generated Table 6 added four rows the hand-typed version had dropped without a clause, and the figure is why that mattered; the added rows are flat, so the decay the appendix reports is `knowledge_update`'s and not a pooled artefact. Rows below the preregistration's floor of ten checkable items are absent rather than plotted small.

**Table 7.** The preregistration record, generated from git.

| experiment | frozen at | criteria | amend. | words |
|---|---|---|---|---|
| E_ADVERSARY | ff4fa5cc, 2026-08-12 09:02 | P-T3-85, P-T3-86, P-T3-87, P-T3-88, P-T3-89 | 0 | 1646 |
| E_ALPHA | c8949a38, 2026-08-09 01:33 | P-T3-65, P-T3-66, P-T3-67, P-T3-68 | 0 | 1132 |
| E_BASELINE | 5c096499, 2026-08-09 04:12 | P-T3-75, P-T3-76, P-T3-77, P-T3-78, P-T3-79 | 0 | 1539 |
| E_BEAMCURVE | 972d23a0, 2026-08-08 07:14 | P-T3-41, P-T3-42, P-T3-43, P-T3-44 | 0 | 628 |
| E_BUDGETFIX | b1bc1d9b, 2026-08-26 07:23 | P-T3-116, P-T3-117, P-T3-118, P-T3-119, P-T3-120 | 0 | 605 |
| E_BUDGET | aa605151, 2026-08-08 03:36 | P-T3-37, P-T3-38, P-T3-39, P-T3-40 | 1 | 408 |
| E_IRT | 5ddb9f4e, 2026-08-07 06:35 | P-T3-27, P-T3-28, P-T3-29, P-T3-30, P-T3-31 | 0 | 365 |
| E_K8BEAM | cbf71422, 2026-08-19 04:28 | P-T3-104, P-T3-105 | 0 | 280 |
| E_LENGTH | 8d3ceb62, 2026-08-08 14:24 | P-T3-55, P-T3-56, P-T3-57, P-T3-58 | 0 | 1512 |
| E_NOLIMA | cbf71422, 2026-08-19 04:28 | P-T3-101, P-T3-102, P-T3-103 | 0 | 267 |
| E_ONCE | efd01888, 2026-08-11 09:13 | P-T3-80, P-T3-81, P-T3-82, P-T3-83, P-T3-84 | 0 | 1263 |
| E_ORACLE | 4c65475e, 2026-08-06 16:08 | P-T3-21, P-T3-22, P-T3-23, P-T3-24, P-T3-25 | 1 | 463 |
| E_OX | cbf71422, 2026-08-19 04:28 | P-T3-106, P-T3-107 | 1 | 296 |
| E_REASONS | cbf71422, 2026-08-19 04:28 | P-T3-108, P-T3-109 | 0 | 207 |
| E_RSTR | d9ed0cd8, 2026-08-08 10:30 | P-T3-51, P-T3-52, P-T3-53, P-T3-54 | 0 | 1177 |
| E_SCALE | 24969f6f, 2026-08-06 05:26 | P-T3-10, P-T3-11, P-T3-12, P-T3-13, P-T3-14, P-T3-15, P-T3-17, P-T3-18, P-T3-19, P-T3-20 | 2 | 684 |
| E_SSM2 | cbf71422, 2026-08-19 04:28 | P-T3-114, P-T3-115, P-T3-116, P-T3-117 | 0 | 152 |
| E_SSM | 77b26454, 2026-08-16 07:33 | — | 0 | 298 |
| E_STATE | 86a68c2b, 2026-08-15 01:26 | P-T3-90, P-T3-91, P-T3-92, P-T3-93, P-T3-94, P-T3-95, P-T3-96 | 0 | 943 |
| E_SWEEP | 1f4225f8, 2026-08-08 15:21 | P-T3-59, P-T3-60, P-T3-61, P-T3-62, P-T3-63 | 0 | 1324 |
| E_SYMN | cbf71422, 2026-08-19 04:28 | P-T3-110 | 1 | 303 |
| E_UNBOUND | cbf71422, 2026-08-19 04:28 | P-T3-111, P-T3-112, P-T3-113 | 2 | 429 |
| E_WITNESS | a4095483, 2026-08-07 16:19 | P-T3-32, P-T3-33, P-T3-34, P-T3-35, P-T3-36 | 0 | 376 |
| E_XCORPUS | 99533e47, 2026-08-15 09:24 | P-T3-97, P-T3-98, P-T3-99, P-T3-100 | 0 | 323 |
| E_XLEN | d822c872, 2026-08-09 02:26 | P-T3-69, P-T3-70, P-T3-71, P-T3-72, P-T3-73, P-T3-74 | 0 | 1883 |

## D EVERY DIMENSION AND SPLIT, GRADED AND UNGRADED

P-T3-24's scope clause restricts the graded cells, and the restriction is material: of the 9 dimension-by-split cells in the containment subset, only 3 carry enough items to be graded, and all 3 are the same dimension. The excluded cells are printed here in full *with their values*, because a scope clause that hides the numbers it excludes is indistinguishable from a scope clause chosen after seeing them. Ours was frozen before the data; the table lets a reader confirm that excluding them was not doing us a favour.

**Table 8.** The containment subset by dimension and split, from `analyze_oracle.py`. $S_{\text{oracle}}$ is the index-only optimum's score, $S_{\text{ceil}}$ the access-complete score, and $R_{\text{upper}}$ their difference, the measured upper bound on $R_{\text{state}}$ for those items.

| dimension | split | $n$ | $S_{\text{oracle}}$ | $S_{\text{ceil}}$ | $R_{\text{upper}}$ | graded |
|---|---|---|---|---|---|---|
| `information_extraction` | 100K | 1 | 0.000 | 0.000 | 0.000 | ✗ |
| `information_extraction` | 500K | 4 | 0.000 | 1.000 | 1.000 | ✗ |
| `information_extraction` | 1M | 2 | 0.000 | 1.000 | 1.000 | ✗ |
| `knowledge_update` | 100K | 22 | 0.091 | 0.364 | 0.273 | ✓ |
| `knowledge_update` | 500K | 42 | 0.310 | 0.460 | 0.151 | ✓ |
| `knowledge_update` | 1M | 35 | 0.257 | 0.429 | 0.171 | ✓ |
| `multi_session_reasoning` | 100K | 3 | 0.000 | 0.333 | 0.333 | ✗ |
| `multi_session_reasoning` | 500K | 5 | 0.000 | 0.400 | 0.400 | ✗ |
| `multi_session_reasoning` | 1M | 1 | 0.000 | 1.000 | 1.000 | ✗ |

The $n$ column is why: 16 items are spread over the 6 ungraded cells, against 99 in the 3 graded ones. *Note also that every ungraded cell has $S_{\text{oracle}} = 0$.* The index-only optimum scores nothing there, so including them would have made P-T3-21 fail harder, not more gently.

## E THE GRADERS' VERDICTS AGAINST THE GRADE THE LEDGER CARRIES

Each grader was committed before the data it grades and prints its own verdict, and the second column is lifted from that output rather than restated: a verdict a reader cannot reproduce by running the script is not a verdict, and the scripts are in the repository under the names given. *The third column is the grade the same criterion carries in this programme's claim ledger*, looked up at build time rather than typed here, and the fourth says whether the two agree *as classes*, meaning a bar missed, a bar met, or a criterion that cannot be evaluated at all. The two vocabularies differ by design (a preregistration freezes its own words), so only the class comparison means anything, and it is the one that separates the case this table exists for: a criterion printed as a missed bar where the ledger says the criterion could not be evaluated, or the reverse.

**Table 9. Every verdict a grader prints, against the grade its criterion carries in the claim ledger.** The second column is the grader's own stdout; the third is looked up in `claims_all.tex` at build time. The vocabularies differ because each preregistration froze its own words, so only the class comparison is meaningful, and a **differ** in the last column fails the build.

| criterion | the grader prints | the ledger grades | ledger | class |
|---|---|---|---|---|
| `P-T3-21` | `FAILED` | **not supported** | `T6_judge` | agree |
| `P-T3-26` | `FAILED` | **not supported** | `T6_judge` | agree |

All 2 agree by class. The qualification the graders add is not a disagreement but a scope statement, and it is carried as its own ledger row: missing `P-T3-21`'s bar refutes this programme's *proof* of $V \leq A$ and not the inequality, since the assumption it needs may hold under a different judge, which is a different experiment, not a different wording.

## F WHAT IS ON DISK

Every arm this paper reports, counted from `results/` at build time. The *oracle* and *extract* arms are the two this paper added; *ceiling*, *judge* and *vanilla* are the arms the earlier sections grade.

**Table 10.** The experimental record, counted at build time.

| arm | files | records | arm | files | records |
|---|---|---|---|---|---|
| adv_k1 | 0 | 0 | oracle | 0 | 0 |
| adv_k16 | 0 | 0 | quantfloor | 2 | 2 |
| adv_k2 | 0 | 0 | reasons | 1 | 1 |
| adv_k4 | 0 | 0 | rstr | 2 | 2 |
| adv_k8 | 0 | 0 | ssm | 8 | 8 |
| adversary | 1 | 1 | state | 2 | 2 |
| alpha | 1 | 1 | sweep | 2 | 2 |
| baseline | 2 | 2 | unbound | 2 | 2 |
| beamcurve | 2 | 2 | vanilla | 0 | 0 |
| budget | 240 | 240 | witness | 123 | 123 |
| ceiling | 0 | 0 | xcorpus | 1 | 1 |
| cot | 0 | 0 | xlen | 2 | 2 |
| extract | 0 | 0 | BUDGET | 1 | 1 |
| extract2 | 3 | 3 | FIGURES | 1 | 1 |
| extract2b | 0 | 0 | beam_100K_bm25last | 1 | 1 |
| extractb | 0 | 0 | beam_100K_lo | 1 | 1 |
| factor | 1 | 1 | beam_1M_bm25last | 1 | 1 |
| judge | 0 | 0 | beam_1M_lo | 1 | 1 |
| length | 1 | 1 | beam_500K_bm25last | 1 | 1 |
| nolima | 1 | 1 | beam_500K_lo | 1 | 1 |
| once | 4 | 4 | | | |
| **total** | **409** | **409** | | | |

## G NOTATION, INCLUDING THE LETTERS A COMPANION REUSES

Every symbol this paper uses, with the place it is defined. The second half of the section exists because this family reuses letters across papers and no paper said so: a reader holding two of them at once has had no way to discover it except by being confused first. *The collisions below are computed from this table rather than asserted*, so a symbol added to any paper of the family cannot silently collide with another.

**Table 11.** The symbols of this paper.

| symbol | what it denotes | defined |
|---|---|---|
| $V$ | *verbatim containment*: the share of items with a short gold whose gold appears verbatim inside a single logged record. A string property of the corpus, with no model, judge, reader or ceiling | Def. 3 |
| $k$ | per-query access budget | Def. 1 |
| $a_k$, $a_{\text{complete}}$ | reachable at budget $k$; recoverable from one record at all | Def. 3 |
| $R^{\text{str}}(k)$ | $a_{\text{complete}} - a_k$, the structural miss curve | Def. 3 |
| $k^*$ | the smallest budget on the ladder at which $R^{\text{str}}$ reaches the threshold. `None` when no rung crosses | (1) |
| $R_{\text{state}}$ | resolving power, restated from the theory companion so this paper is self-contained | Def. 2 |
| $h$ | *retrieval hit rate*: the share of items where the event the reader returned actually contains the gold | §7 |
| $\rho$ | the index-only class's *internal budget spread*, whose numerator and denominator are both index channels, which is why it is reported as retrieval sufficiency and *not* as a resolving power | §7 |
| $w$ | the width of a local-attention window, which the classifier reads as an index at budget $w$ rather than as a state channel | §8 |

**The same letter, a different object.** Each line below names a symbol of this paper that a companion uses for something else. None of these is a notational slip to be repaired by renaming: the objects are each named after their own literature's convention, and renaming one to suit another paper would cost more than it saves. What was missing is the list.

- $\rho$, an index-only budget spread here; in Anonymous (2026e) it is the retriever itself.

**What is deliberately shared.** $k$ is a per-query access budget in every paper of the family that has one; $V$ is verbatim containment in all four papers that count it; $\iota$ is the credit rate in all three that use it; $n$ is a count of items throughout. Those four are shared *because* they are the same object, and a paper that imports one imports the definition with it.

## H GOLD-IN-A-SINGLE-TURN, WITH THE REGISTERED OUTCOMES

The registered outcome per dimension (Table 12), and the same count over every dimension of the benchmark rather than the three the audit can grade (Figure 11). The gaps in the second are the specification: a dimension with no short gold on any split has no containment rate to measure, and the preregistration's floor clause records that as *untestable* rather than as a low number.

**Table 12.** Gold-in-a-single-turn rate. The 500K and 1M columns were registered before those splits were downloaded (frozen bars: $\geq 0.5$ for knowledge update, $\leq 0.25$ for the composition-type dimensions).

| dimension | 100K | 500K | 1M | short golds | registered outcome |
|---|---|---|---|---|---|
| `knowledge_update` | 0.61 | **0.64** | **0.55** | 36/66/64 | supported |
| `multi_session_reasoning` | 0.19 | **0.18** | **0.06** | 16/28/16 | supported |
| `information_extraction` | 0.06 | 0.08 | 0.05 | 17/52/37 | — |
| `summarization` | no short golds | | | 0 | untestable, reported |

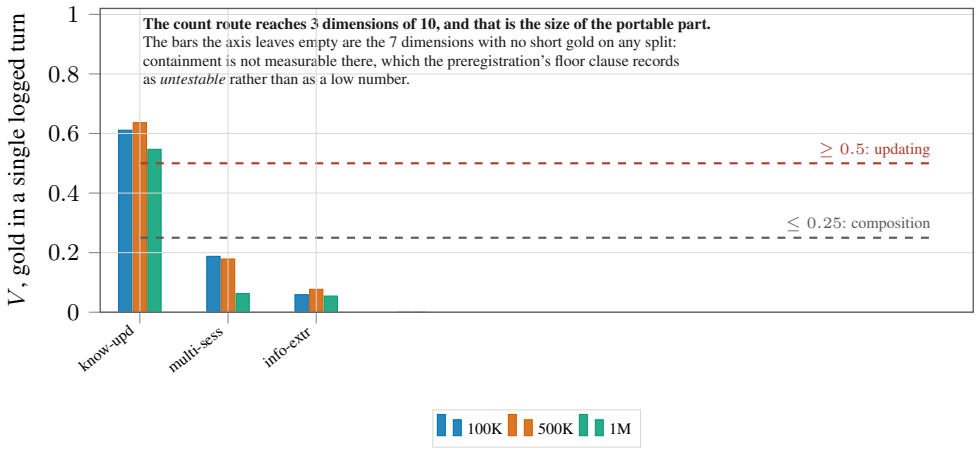



**Figure 11. Where the count route reaches at all.** $V$ on every dimension of the benchmark at the three registered scales, with the two bars frozen before the 500K and 1M splits were downloaded. The main text (§5) draws the three dimensions the audit can grade; this draws all ten, and the gaps are the point: the dimensions with no bar have no short gold on any split, so containment is not measurable there and the preregistration's floor clause records them as *untestable* rather than as a low number. *An instrument's reach is part of its specification*, and this one reaches three dimensions of ten.

## I E-RSTR, EVERY SLOT AND BUDGET

Every cell of E-rstr, graded against its frozen criteria (Table 13), and the three slots' pooled curves on one axis (Figure 12). The curves are what makes `None` legible as a result rather than as a gap: a slot whose $R^{\text{str}}$ never reaches $\tau$ has no $k^*$ to report, at any budget.

**Table 13.** E-rstr, graded against frozen criteria. Slot C is not coverage; it is the instrument check. $k^*$ is (1) at $\tau = 0.05$, and `None` is a result. The whole tabular is generated from `results/rstr/GRADE.json`; $k^*$ is recomputed per cell from its definition, which the artifact stores only for the pooled rows, and the row order is $(k^*, R^{\mathrm{str}}(32))$ ascending rather than a hand ordering.

| slot | cell | $n$ | $V$ | $R^{\mathrm{str}}(1)$ | $R^{\mathrm{str}}(32)$ | $k^*$ |
|---|---|---|---|---|---|---|
| C | `kv_retrieval` | 500 | 1.000 | 0.000 | 0.000 | 1 |
| A | `single-session-user` | 64 | 0.891 | 0.219 | 0.000 | 8 |
| A | `single-session-assistant` | 47 | 0.660 | 0.149 | 0.021 | 8 |
| A | `knowledge-update` | 70 | 0.857 | 0.557 | 0.029 | 16 |
| A | `temporal-reasoning` | 118 | 0.347 | 0.195 | 0.076 | `None` |
| A | `multi-session` | 118 | 0.610 | 0.339 | 0.144 | `None` |
| A | *pooled* | 417 | 0.626 | 0.295 | 0.070 | `None` |
| B | `longbook_qa_eng` | 347 | 0.571 | 0.524 | 0.228 | `None` |

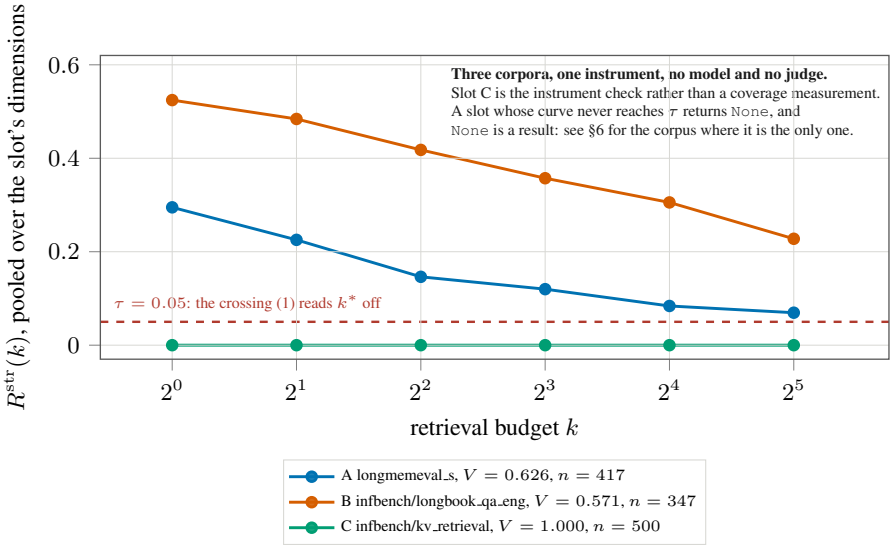



**Figure 12. E-rstr on all three slots, at every budget.** The pooled curve per slot, with $\tau$ (the level (1) reads $k^*$ off), drawn once. Slot C is the instrument check rather than a coverage measurement, and a slot whose curve never reaches $\tau$ returns `None`: Table 13 prints that outcome per cell, and §6 is the corpus where it is the only outcome at any budget. *The three curves are the same instrument on three corpora*, which is the claim the slot design exists to support.

## J THE SCORECARD, EVERY GRADED ROW

This paper's claims are graded here rather than in its sections, so that the sections can be about what was found and this appendix can be complete. Criteria were frozen in `prereg/` before the corresponding data existed, and one generator writes all three tables from one source, so the index below cannot disagree with the rows it counts.

**Table 14. The scorecard in eight lines, counted from the scorecard.** Every graded row of Table 16 and Table 17 falls in exactly one class below. Where a family's status names two classes at once, the row is counted in the *more serious* one, which is the only direction of error a summary of one's own scorecard may have. It is an index, not a substitute, and the rows are all still there: 21 claim families in Table 16, 40 sub-rows in Table 17. *What the unflattering rows are for.* 30 of the 60 rows carry a positive grade and 23 an unflattering one, and the second number is not a defect count: each of those rows narrows a claim made above it, names a number of ours that moved, or marks a question this apparatus cannot decide. **None of them is an experiment left undone**: work nobody has run is not graded here at all, and is bounded where it is stated.

| grade | rows | what the grade means in this paper |
|---|---|---|
| withdrawn / superseded | 0 | asserted, then taken back or re-assigned; the sequence is in the post-mortems |
| not supported / failed | 7 | a frozen criterion was missed, on our own data |
| self-correction | 6 | our own arithmetic or our own criterion was wrong and is repaired here |
| not decidable / not evaluable | 1 | the apparatus or the band cannot separate the outcomes: not a null |
| acknowledged limit | 9 | a boundary of the result, stated where the result is |
| formal | 3 | proved from the definitions; no measurement can move it |
| measured | 20 | an artifact on disk, and a reader may dispute the artifact |
| argued / reported | 7 | no grade is claimed: an argument, or a number printed without a criterion |
| *other* | 7 | *status words outside the eight classes above* |
| total graded rows | 60 | every numbered row of both tables |

**Table 16.** Claims by evidential status. Grades: **formal** (proved, no data), **measured**, **supported / not supported / failed** (frozen criterion met or missed), **not decidable** (the criterion's band is below the measurement's resolution), **not testable** (the design cannot pose the question), **self-correction**, **acknowledged limit**, **argued**, **reported** (not graded). Rows recording grades we issued and then corrected are kept: the sequence is the contribution.

| Claim | Status | Evidence |
|---|---|---|
| 300. An index-only reader's score upper-bounds what a benchmark's items require of a state channel | **formal** | Prop. 1; (2) needs realisability and no judge |
| 301. The audit condemns and can never acquit | **formal** | Prop. 2: acquittal quantifies over $\binom{188}{8} \approx 10^{13}$ subsets and every function on each |
| 302. $V$, $R^{\mathrm{str}}(k)$ and $k^*$ are computable from a benchmark's text with no model, judge, system or API call | **measured** | the whole audit's marginal cost is a download; Def. 3 |
| 303. The audit's verdict on the deployed flagship is $0/10$ dimensions blind | **measured** | $V = 0.55$–$0.64$ on the updating dimension against 0.00–0.07 elsewhere |
| 304. P-T3-41: $R^{\mathrm{str}}(1) \geq 0.30$ pooled | **not supported** | 0.246; a top-1 lexical retriever already reaches the gold on most checkable items |
| 305. P-T3-43: a 32-event budget closes the gap at the registered split | **supported** | $R^{\mathrm{str}}(32) = 0.000$ at 100K |
| 306. The counts do not move along the axis the field is scaling | **measured** | over a tenfold range no dimension crosses its own frozen bar; $1 - V = 0.45 \pm 0.05$ on the cited dimension |
| 307. P-T3-13: the weaker reader still wins at 500K | **supported** | BM25-last takes 5 of 10, which clears the bar |

*Table 16, continued.*

| Claim | Status | Evidence |
|---|---|---|
| 308. P-T3-54: the instrument check a broken retriever cannot pass by structure | **supported** | slot C gives $R^{\mathrm{str}}(1) = 0.000$ and $k^* = 1$ on a needle task |
| 309. P-T3-52: a 32-event budget closes the gap outside the synthetic slot | **not supported** | $k^* =$ `None` on both real benchmarks; the prereg named this as the least certain prediction |
| 310. $k^*$ recovers a benchmark's own single-hop/composition labels with no model | **measured** | 8, 8, 16 against `None`, `None` on LONGMEMEVAL's own labels |
| 311. $R^{\mathrm{str}}$ needs a premise stronger than containment, and the gap's direction is known | **acknowledged limit** | anywhere-containment ($0.61/0.64/0.55$) against exactly-once ($4.3\%$, $6.1\%$, $5.0\%$): the anywhere form inflates $R^{\mathrm{str}}$ |
| 312. P-T3-101: the NoLiMa adapter gate | **supported** | $V = 1.000$ on both hop depths, 156 rows: every item is answerable from one event |
| 313. P-T3-86: two dimensions clear the frozen 0.80 bar under the strongest index-only adversary | **supported** | `knowledge_update` at $k$=8, `information_extraction` at $k \leq 16$ |
| 314. P-T3-85: the $k$=8 re-run reproduces within 0.02 per dimension | **failed**; *below: 1 not supported* | largest difference 0.0500; and the frozen band was 0.70 of one standard error |
| 315. P-T3-88: $\rho$ is monotone in the budget | **not decidable**; *below: 1 not evaluable* | non-monotone on four dimensions and 0 of the dips exceed 2 SE; a strict inequality with no tolerance has $B = 0$ |
| 316. P-T3-87: a model at one chunk beats a no-model dump on most dimensions | **not supported** | 3 of 10 gain $\geq 0.15$ and five lose, to $-0.40$ |
| 317. Channel membership is decided by measurement, not by reading an architecture description | **measured** | carrier elements counted over a $16\times$ span; a name-based check misfiles `rwkv` |
| 318. The judged arms are conditional on one pinned judge | **acknowledged limit** | a judge change carries `abstention`, a row of this paper's own ladder, to saturation (Anonymous, 2026g, row 204) |
| 319. P-T3-88's dips were compared against one cell's standard error, not a difference's | **self-correction** | the rungs are the same 400 shards, so the supported treatment is paired: SE $0.130 \to 0.109$, worst dip $1.40 \to 1.67$ |

*One row per claim family. Each* STATUS *cell names the grades of that family's sub-rows, so a failure is counted here even though its row is elsewhere: all 40 sub-rows are in Table 17 in full, of which 7 acknowledged limit, 5 self-correction, 3 not supported. Nothing is dropped, and nothing is summarised into a kinder word than it had.*

**Table 17.** Supporting refinements, moved from Table 16 by the stated rule. The direction is deliberate: a selection rule for a scorecard should be biased *against* its authors, so unflattering rows stay in the main table and supporting detail moves here. `claims_all.tex` is the source both tables are generated from.

| Claim | Status | Evidence |
|---|---|---|
| *The 40 rows below were moved here from Table 16 by the rule in `papers/split_claims.py`: a lettered sub-row whose status is* supportive *moves; every top-level claim and every row carrying a withdrawal, failure or self-correction stays. Nothing is dropped.* | | |
| 300a. …and the scaled form is anti-conservative in the ceiling | **acknowledged limit** | a stronger ceiling lowers every ratio; the direction is stated at the equation |
| 301a. …so $0/10$ is the only outcome this reader could have returned, short of condemnation | **argued** | both instruments hold one event, the weakest admissible adversary |
| 301b. …and a procedural audit whose output space contains a pass inherits the same bound | **argued** | absence of a witness read as presence of validity; no failure-mode taxonomy closes a quantifier |

*Table 17, continued.*

| Claim | Status | Evidence |
|---|---|---|
| 301c. …and the closest parallel work states an asymmetry of its own, which is not this one | **argued** | theirs is a necessary-not-sufficient caveat on one gate's verdict; Prop. 2 is about the quantifier (Li et al., 2026b) |
| 302a. BM25-last's one free constant is priced without spending anything | **measured** | within 0.008 of the best realisable setting on every split; 0.8 is the maximum on two of three |
| 302b. …and the constant is not innocuous, which is why the bound is the best of the sweep | **measured** | the returned event is identical across the sweep on 3.4–11.6% of items; containment moves up to 0.101 |
| 303a. …and it is not a clean bill of health | **acknowledged limit** | Prop. 2: the count route cannot acquit, and §10 says so |
| 303b. On nine dimensions the reader route bounds resolving power better than the count route | **measured** | the tenth, built to test updating, is bounded better by the benchmark's own text, by 40% |
| 304a. P-T3-44: the largest-gap dimension would not be `knowledge_update` | **not supported** | the reasoning was backwards, and the preregistration named the direction it would cost |
| 305a. …and a fixed budget buys less of the transcript as the transcript grows | **measured** | 0.021 and 0.017 pooled at 500K and 1M; the axis moves the budget's reach, not the score |
| 306a. …and the registered behavioural mechanism for that is wrong | **not supported** | the prediction that scores fall with length while $V$ stays pinned; three criteria, frozen first |
| 306b. The ceiling's invariance across the axis | **reported**, post hoc | no criterion was frozen on it, and the registered prediction that led there failed |
| 306c. The largest shipped rung cannot support the measurement its own size invites | **measured** | a monotone trend with an undecidable slope, on somebody else's benchmark |
| 307a. …and $5/10$ is a tie rather than the $6/10$ a companion measures | **self-correction** | we report the weaker fact (Anonymous, 2026g, row 203a) |
| 307b. What replicates is the per-dimension winner, not the inversion count | **measured** | unchanged on 7 of 10 across three splits; the three flips are the smallest margins (0.021, 0.037, 0.112) |
| 308a. …and P-T3-42, registered as a monotonicity check on the implementation, cannot catch that class | **acknowledged limit** | a monotone curve is producible by a broken reader; App. B |
| 309a. P-T3-51: $R^{\mathrm{str}}(1)$ clears its bar on the blind slots | **not supported** | clears on slot B only (0.524); slot A pools 0.295 against 0.30, missing by 0.005 and reported as missing |
| 309b. P-T3-53: the record-length axis, on the blind corpora | **not testable** | each slot supplies one record length; §9 supplies the second |
| 310a. …and the ordering of six cells by $R^{\mathrm{str}}(32)$ | **reported**, post hoc | ungraded, and it is what we would register next |
| 310b. $V$ on the blind slots, which is the count the companions import | **measured** | 0.571 on `longbook_qa_eng` over 347 items, 0.626 pooled on slot A, 1.000 on the needle slot |
| 310c. …and $V \cdot n$ is therefore the item count a budget-$k$ reader could in principle close | **formal** | one multiplication, which is how the budget companion derives its own figure rather than quoting ours |
| 310d. Table 6 had dropped four dimension rows that clear the preregistration's own floor | **self-correction** | `information_extraction` and `multi_session_reasoning` at 500K and 1M, 52, 28, 37 and 16 items against a floor of 10, printed at the other two splits and absent here with no clause; the pooled row always included them |
| 310e. …and the added rows sharpen the appendix's own reading rather than soften it | **measured** | both added dimensions are flat at every $k$, so every falling curve in the table is `knowledge_update`'s and the pooled decay is one dimension's |
| 311a. …and the flagship cannot supply the well-posed form | **measured** | no dimension has ten exactly-once items at any scale; the direction is stated at the table |
| 312a. P-T3-102: one instrument, two single-event-answerable corpora, and the read-out differs by its whole domain | **supported** | $a_k = 0.000$ at every $k \leq 32$, so $k^*$ does not resolve where BEAM's resolves at 8 |

*Table 17, continued.*

| Claim | Status | Evidence |
|---|---|---|
| 312b. ... so $k^*$ measures the reader's lexical reach and not the task's composition depth | **measured** | the comparative cross-corpus use of $k^*$ is retired rather than defended |
| 312c. ... and the count $V$ is what travels | **measured** | containment is a witness, not an estimate; the condemnation direction is untouched |
| 312d. ... and no level in this paper is an input to the budget companion's exponent | **acknowledged limit** | a level at one length is not a slope across lengths (Anonymous, 2026e, row 121) |
| 312e. What this arm cannot distinguish | **acknowledged limit** | that no index-only strategy at $k \leq 32$ could retrieve the needle; Prop. 2 in its own home |
| 313a. ... and $\rho$ is retrieval sufficiency, not a resolving power | **self-correction** | numerator and denominator are both index channels (Anonymous, 2026d, row 15) |
| 313b. Retrieval precision at the three scales, with no judge in it | **measured** | $h = 0.136/0.262/0.200$ against a frozen $\leq 0.036$ at 1M: FAILED, and the failure is the count moving the wrong way |
| 314a. ... with the diagnosis before the interpretation | **measured** | 55 of 400 items change on a re-run of identical code; pooled mean difference $+0.0010$ ($t = 0.11$), no drift |
| 315a. ... so "more retrieval hurts" is not a finding of ours | **self-correction** | first graded NOT SUPPORTED; the resolution arithmetic reclassified it (Anonymous, 2026f, row 29) |
| 316a. ... and what it exposes is worse than the criterion | **argued** | a no-model dump beating a model-bearing adversary is impossible on information grounds, so the dump is paid for something else |
| 316b. P-T3-89: containment underpredicts where golds are short | **reported**, not graded | $\rho$ reaches 0.90 where retrieval containment reaches 0.61; guessability, a fourth mechanism |
| 317a. ... and local attention over a window $w$ classifies itself as an index at budget $w$ | **measured** | the budget-indexed definition of the theory companion (Anonymous, 2026d, row 100) |
| 317b. ... and the probe sees width only | **acknowledged limit** | a bounded carrier can still cost $O(n)$ per token to update |
| 318a. ... and that magnitude was typed into this paper by hand | **self-correction** | it is now the companion's generated macro, so it moves when its artifact moves |
| 318b. This paper owns two counts and a budget read-out, and no resolving power | **acknowledged limit** | $R_{\text{state}}$ is the theory companion's, $\iota$ the judge companion's, $\alpha$ the budget companion's |
| 319a. ... and the correction runs against the reading it supports | **measured** | a paired SE is smaller, so every dip grows; still 0 of 4 reach 2 SE and NOT DECIDABLE stands (Anonymous, 2026f, row 97c) |