Resolving Power: Whether a Memory Task Can See the Distinction It Claims to Test

Anonymous authors
Paper under double-blind review

Abstract

Memory benchmarks are cited to settle an architectural claim: that a system maintaining a compressive state does something a scalable verbatim index cannot. Whether a given task can see that distinction is never checked, and a task that cannot scores the two systems identically, however large, leakage-free or reliably judged it is. The natural estimator is unavailable too: a difference of two realised scores bounds the quantity from neither side, since both realisations bound their own optima in the same direction. We define a task's resolving power at access budget k as the score gap between an access-complete and an index-only oracle, give a checkable sufficient condition for blindness and a condition on cardinality at a budget and a segmentation for sightedness, and prove that a sound positive result requires a task whose subtracted optimum is bounded by construction. We then build one: a 240-character scratchpad, rewritten once per record and never shown the question, whose probe answers appear in no record at all. That bounds the index-only optimum at k=1 by 1/900 by design rather than by any claim about retrievers; the carrier scores 0.695, putting resolving power at 0.694 or above against a band of 0.151 fixed in advance, while a reader with one record and unlimited chain-of-thought writes the answer on 0 of 200 probes. Sending the cardinality to 32, the bounded-width carrier holds where the access-complete arm collapses, so from $m = 8$ upward a system that sees every record scores below one that sees only a fixed-width summary. Eight recurrent checkpoints screened against a floor fixed in advance all fail it, so the architecture-level question stays not decidable at current scale. The residue is a two-clause design rule – make the answer depend on more records than the budget holds, and let it appear verbatim in none of them – shipped as a generator with 500 probes that re-derives both premises on demand.

1 Introduction

A long-term memory system holds state across a conversation, and benchmarks say whether it does. The systems they rank make architectural claims: a compressive, bounded-width state channel is credited with what a scalable verbatim index could not have done (Chhikara et al., 2025; Xu et al., 2025). Between claim and score sits a question rarely asked of the task itself: can it see the distinction it is cited for? A companion builds the instrument and audits deployed benchmarks (Anonymous, 2026e); a second measures what a rubric judge inserts between a count and a scored verdict (Anonymous, 2026h). This paper is the theory those audits stand on, and the construction answering its hardest consequence.

The consequence is easy to state and uncomfortable to hold: a difference of two realised scores bounds resolving power from neither side (Prop. 5); realised ceilings and readers under-estimate their optima in the same direction, so no precision, band or rerun count turns a measured gap into a bound. The repair is a different object: a one-sided bound from a realisable member, or a task whose subtracted optimum is bounded by construction. §5 supplies both; the second is the first memory task, to our knowledge, whose resolving power is positive by proof rather than by plausibility.

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.

What this paper contributes.

1. The theory of the estimand (§3, §4). Definitions of channel, budget and resolving power; the blindness condition (Prop. 1); sightedness as cardinality at a budget and a segmentation (Prop. 3) and its unbounded, chunking-proof form (Prop. 4); non-identifiability from realised comparisons (Prop. 5), with the clause that returns a valid bound.
2. A task with $R_{\text{state}} > 0$, built and measured (§5). A bounded-width carrier scores 0.695 on a no-record family against an index-only optimum of 1/900 fixed by the construction, $R_{\text{state}}(\mathcal{T}; 1) \geq 0.694$ against a preregistered band 0.151; controls built in: an oracle-at-k=2 row beats the carrier, a chain-of-thought arm moves nothing (0 of 200 traces contain the gold).
3. A design rule with both clauses load-bearing, and its generator (§5). R1: more records than the budget holds, at every segmentation a designer must anticipate. R2: the answer verbatim in none of them. Table 3 is R2's evidence: three k=1 strategies reach 0.359/0.485/1.000 against a candidate bound of 1/7200 on a family violating it.
4. The rule's robust form measured, a premise of our own broken by it (§6). With cardinality m running $2 \to 32$ the carrier holds $0.917 \to 0.783$ while complete access falls $1.000 \to 0.033$, ending 0.750 above it: a realised access-complete arm is not an upper bound on what a task admits (Prop. 5 on the leading side).
5. The construction attempted on genuine state channels (§7). Two architecturally distinct state-space models fail the protocol the same way, at a tenth of their own ceilings, so the verdict is not decidable; eight pure recurrent checkpoints were screened against a floor fixed before any ran, none clears it, the best reaching 0.535.

The audit built on this quantity is one-sided (Prop. 2): agreement is consistent with both readings, so an instrument can only condemn. The R_{state} diagram is in Appendix A (Figure 3); the scorecard of every registered claim and grade is in Appendix O, generated from a single source.

2 Related work

Psychometrics has a name for this quantity, and the sign is the other way round. An item exhibiting differential item functioning (DIF) responds to a nuisance group and is repaired or dropped (Holland & Wainer, 1993; Millsap, 2011). Resolving power is DIF with the sign flipped and the two roles exchanged: here the group is the target (a compressive state channel against a scalable index), and $R_{\text{state}}(\mathcal{T}; k) = 0$ says the benchmark exhibits no differential functioning across the channel axis, in psychometric terms a clean bill of health and in ours the defect. We cannot borrow their detection methods: Mantel–Haenszel and IRT-based DIF condition on the latent trait estimated from a sample of each group, and §4 shows our second group's optimum is not enumerable, so there is no sample to condition on. That is why our positive result is a construction (§5) rather than a detection.

Construct validity, and where a per-item quantity goes past a survey. Bean et al. (2025) review 445 benchmarks and report contested constructs as systemic; Raji et al. (2021) make the sharper qualitative point. R_{state} takes a value on one item, and where it is zero the item cannot separate the systems it was written to separate, whatever the construct clarification; the arithmetic form of this argument is the companion instrument paper's (Anonymous, 2026e), and Tanguturi (2026) anticipate the premise in prose: "by construction, these benchmarks cannot distinguish a memory system from a long-context LLM".

The long-context literature already builds families whose answer is in no record. Hsieh et al. (2024) introduce multi-hop tracing and aggregation families in the shape of our own two; the contribution here is the classification Prop. 3 adds, since no bound in that literature says what an index-only strategy scores on either. Xia (2026) report ONCU, a difference of two realised scores over a realised denominator that §4 says bounds neither side, exhibited on the

increment family, where a carrier that is none of ONCU's four conditions outscores the oracle condition. Nadali et al. (2026) carry a key–value cache across chunk boundaries with 100% exact-match retrieval, calling it internal state; their own abstract settles the classification instead (the cache "passes the enlarged cache forward," an index under Def. 1), a caution against the vocabulary a saturating cache invites.

Partial identification is the correct frame, borrowed knowingly. The econometric tradition from Manski (1995; 2003) rejects point identification bought with unfalsifiable assumptions, in favour of the set the evidence admits; §4 places the audit's one-sided region in that frame, and the ordinary remedy, add a realisable member on the unbounded side, is what §5 does. Kalaitzidis (2026) ask the same question procedurally, on a benchmark proposal before any item exists, which R_{state} cannot; R_{state} condemns a shipped item with a number, which a taxonomy cannot. Anonymous (2026a) supplies the distinction itself, state against index; Appendix G carries the full import accounting.

3 Resolving power and the blind limit

Definition 1 (Channels, budget, and resolving power). 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. All three are fixed ex ante, and the admissible budget for every verdict in this paper is $k = 1$, because that is what the companion instrument paper's readers hold (Anonymous, 2026e): a larger k makes condemnation easier, so a failure to condemn at k=1 is a failure at the most favourable budget for the benchmark.

> A KV cache is not a state channel under this definition. It grows linearly in the tokens seen ($O(n)$ storage), so it satisfies the definition of an index; and it is destroyed between episodes, so it cannot carry anything into the next conversation, which every benchmark this paper audits requires. Architectures whose recurrent state is genuinely fixed-width, state-space models included, are state channels by this definition, and §7 runs one as the leading arm.

Definition 2 (Resolving power at a budget). $R_{\text{state}}(\mathcal{T}; k) = S_{\mathcal{T}}(\text{access-complete oracle}) - S_{\mathcal{T}}(\text{index-only oracle at budget } k)$, each oracle optimal within its channel set at that budget. The budget is in the definition because leaving it out makes the definition vacuous: §4.1 measures one task at four budgets and the answer changes sign across them.

Proposition 1 (Blindness). 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_{state} are equal. □

Proposition 2 (Acquittal needs construction; restated from Anonymous, 2026e). A task family can be shown to have $R_{\text{state}} > 0$ only if the dependency structure of its items is known by construction. Condemning is existential (exhibit one index-only strategy reaching the ceiling); acquitting quantifies over every k-subset of the span and every function on it, a space that cannot be enumerated; so no finite audit acquits.

Proposition 3 (Sightedness, at a budget and a segmentation). Let σ be the segmentation: the partition of the write stream into the records a retriever may address. $R_{\text{state}}(\mathcal{T}; k, \sigma) > 0$ requires at least one query whose answer is determined by no set of k records of σ, and hence requires more records than the admissible budget holds, at that segmentation.

The segmentation argument is measured, not decorative. Merge two bearing records into one chunk and a k=1 index holds both: the same task is sighted under one segmentation and blind under a coarser one. In Table 2, index@2-oracle reaches 0.990 against index@1's 0.000

on the same probes, and every bound below holds at the shipped segmentation. Sightedness is a cardinality requirement and composition is not the operative property: on the witness family, score is retrieval coverage times a constant, and the model composes whenever it has the records (§4.1), which is measured rather than assumed against the tempting stronger reading that an index "retrieves k records and does not compose them".

Proposition 4 (Unbounded cardinality). Let a stream of L records determine the answer through $m(L)$ of them and no fewer, with $m(L) \to \infty$. Fix any budget k and any segmentation whose chunks hold at most w records. Then k chunks hold at most kw records, so for every L with $m(L) > kw$ the answer is determined by no k chunks, and Prop. 3's necessary condition holds at every fixed (k, w) for all sufficiently long streams.

This is the form R1 should have; the parent framework's own construction already has it (Anonymous, 2026a), a value stated once and a stream of exchanges then announced, so $m(L)$ grows with L. Two qualifications: it is Prop. 3's necessary condition that holds at every (k, w), not positivity, and the price is paid on the leading side, since a carrier of fixed width must eventually fail as $m(L)$ grows. §5's family has $m = 2$, the cheapest witness; §6 measures the robust form.

The limit. Take a memory benchmark with infinitely many items, every one of whose answers is a function of the write record. Send the number of items to infinity. Resolving power on the state channel is identically zero, independently of size. A benchmark can be large, leakage-free, and scored by a perfectly reliable judge, and still return the same number for a system with a composable state channel and one without: seeds, intervals and power analyses price whether an effect would be detected if the instrument could see it, and are silent on whether it can, and Norman et al. (2026)'s finding that agreement does not establish validity extends the same point to a judge's reliability.

4 Why R_{state} is bounded from neither side

Prop. 2 sends a quantifier to infinity on the index-only side; sent to infinity on both sides it decides what any realised comparison can establish (drawn in Appendix D, Figure 6).

Proposition 5 (Neither side, and the one condition that returns one). Write $\mathcal{A}$ for the access-complete class at budget k and $\mathcal{B}$ for the index-only class, both unenumerable; $A \in \mathcal{A}$, $B \in \mathcal{B}$ for members we can build and run; and $R = S(\mathcal{A}^\star) - S(\mathcal{B}^\star)$ for Def. 2. Then (i) $S(A) - S(B)$ bounds R from neither above nor below, whatever the precision of S; (ii) if a construction supplies $U \geq S(\mathcal{B}^\star)$, then $S(A) - U$ is a valid lower bound on R; symmetrically for the upper side.

Proof. Running a system gives $S(A) \leq S(\mathcal{A}^\star)$ and $S(B) \leq S(\mathcal{B}^\star)$: two inequalities in the same direction. A lower bound on R needs a lower bound on $S(\mathcal{A}^\star)$, which $S(A)$ supplies, and an upper bound on $S(\mathcal{B}^\star)$, which no realisation supplies. Under (ii), $S(A) - U \leq S(\mathcal{A}^\star) - S(\mathcal{B}^\star) = R$. The obstruction is the missing bound on the subtracted optimum, not the precision of S, so no band, sample size or rerun count removes it. □

Which clause applies decides both halves of this programme. A ceiling-minus-reader difference fails clause (i) in exactly one term: $\max_i S(\mathcal{R}_i)$ is a legitimate lower bound on the subtracted optimum, while S(ceiling) is a realised arm standing where an upper bound on the access-complete optimum is required, so no better retriever fixes it. §5 satisfies clause (ii): (2) bounds the index-only optimum by the construction of the probe family, which is why (3) is the only two-term statement in this paper that is legal.

Every arm in this apparatus except our own carrier is an index channel. The companion audit's readers, retrieval arm and full-context arm all have $O(n)$ storage (Anonymous, 2026e), so the full-context arm sits on the subtracted side of Def. 2, not the leading side. A ceiling-versus-reader audit therefore computes $S(\text{index@all}) - S(\text{index@1})$, the index-only class's internal budget spread, which bounds nothing about the true R_{state}: the audit's 0/10 was structurally guaranteed and would return 0/10 on a maximally sighted benchmark too.

What the non-identifiability does not touch. V, $R^{\text{str}}(k)$ and k^*, the counting instruments of the companion paper (Anonymous, 2026e), never need an access-complete arm: V is a share of items whose gold is verbatim present in one event, a property of the text and a retriever with no model or ceiling, so Prop. 1's contrapositive gives a condemnation route the non-identifiability result does not touch.

We then ran the missing arm on ourselves. Our composition probes are already built to the recipe of §5: a value stated only for entity A, a swap announced as a pair of keys, the query about B, so the gold appears in no event mentioning B. Both readers score 0 on them by construction; the missing half is an access-complete arm on the same probes, scored by exact containment with no judge (Table 1).

Table 1. A task with $R_{\text{state}} > 0$, exhibited rather than argued for. 60 composition probes, three arms on the same probes under one scoring rule. The extract arm receives only the two load-bearing events, measuring ι for this (family, scorer) pair.

arm	what it sees	score	
index-only (lo)	one event	$0/60 = 0.000$	by construction
index-only (bm25-last)	one event	$0/60 = 0.000$	by construction
access-complete	all events, in order	$\mathbf{52/60 = 0.867}$	P-T3-32 bar 0.50: supported
extract	the two bearing events only	$55/60 = 0.917$	this is ι
$\widehat{R} = 0.867$, Wilson 95% [0.758, 0.931]		lower bound > 0: P-T3-33 supported, and near-certain by construction, see below	
$\widetilde{R} = \widehat{R}/\iota = 0.945$		$\iota \geq S(\text{complete})$, which the factorisation $S = a\iota$ requires, see below	
se inflation 1.376 (Anonymous, 2026h, row 213b), effective n 31.7		$1/\iota$ alone would say 1.091 and 50.4	

On a task built so that composition is required, an access-complete system attains 94.5% of what a perfect-credit scorer would grant it while every index-only strategy attains none. Here $\iota = 0.917 \geq 0.867 = S(\text{complete})$, so the consistency requirement of the factorisation $S = a\iota$ holds, which is the one check this table can give it. This does not overturn the companion's 0/10: BEAM's dimensions are unaffected by a result about our own probes; it shows only that $R_{\text{state}} > 0$ is attainable and that the recipe produces it.

4.1 The budget axis

The audit's instruments impose two limits at once: a budget (the reader sees k=1 event) and an addressability limit (the answer must be readable off a single event). Separating them is the companion's programme (Anonymous, 2026f); two of its results are quoted here. First, on the witness probes, coverage of the bearing pair runs 0.000/0.383/0.867/1.000 at $k = 1/2/4/8$ and score divided by coverage is flat at 0.87–0.93: what binds is retrieval coverage, not composition, ruling out the tempting "ordered composition" reading of the probes' difficulty and confirming Prop. 3's cardinality statement. Second, the score in this

apparatus factors: writing a_k for the coverage a budget-k retriever attains,

$$R_{\text{state}}(k) \;=\; \big(a_{\text{complete}} - a_k\big)\,\iota, \qquad R^{\text{str}}(k) \;=\; a_{\text{complete}} - a_k, \tag{1}$$

tested on the retrieval-budget axis the construction was not built from, and every arm falls within half a binomial standard error of ι (Appendix D, Figure 7). R^{str} needs no model or judge and is the quantity the record-length companion sends to its own limit (Anonymous, 2026f).

5 A task with positive resolving power

Everything to this point is one-sided: every realised arm in this apparatus is an index channel, so the leading term of Def. 2 is never occupied by a measurement. This section supplies both a realisable leading-side member and a constructively bounded subtracted side; it exhibits one task with $R_{\text{state}} > 0$ and says nothing about whether any deployed benchmark contains such a task, which the companion answers separately and negatively (Anonymous, 2026e).

The leading side, occupied. We build an arm that is a state channel by Def. 1 rather than by courtesy: it consumes the record one event at a time and after each event rewrites a scratchpad capped at 240 characters, the event then dropped and never available again. Three properties are frozen and checked in the artifact: the pass is single and forward; the cap binds (209 truncations counted); and the update prompt contains no question field at all, so the scratchpad cannot be a query-dependent retrieval.

The subtracted side, bounded by the construction rather than by an arm. A realisable index-only arm bounds an optimum only from below, so substituting one's own retriever overstates resolving power. On a family whose gold sits verbatim in one record, the tempting bound is a uniform-guess prior of 1/7200 (Table 3, Appendix A), and three k=1 strategies built against it reach 0.359 (retrieve any attribute-matching value record and echo it), 0.485 (the same, preferring the rarest entity by document frequency), and 1.000 (read the link record first, then fetch the value record it names): a sequence with no evident last member, Prop. 5(i) seen from the inside.

The repair removes the object of the quantification. If the answer occurs in no record, there is no record for a retriever to be good at finding. The family states a base for one entity and an offset in the link record, and asks for their sum; the generator asserts per probe that the answer occurs in no record and regenerates if it does (0 of 200 leaked). Any k=1 strategy must supply one operand from the prior, uniform on 900 values:

$$S(\text{index-only optimum at } k{=}1) \;\le\; \tfrac{1}{900} \;=\; 0.00111, \tag{2}$$

by the construction rather than by an enumeration of retrievers.

The bar a reader should not skip in Figure 1 is index@2 with an oracle retriever at 0.990: handed both bearing records, an index needs no state at all, Prop. 3's cardinality condition seen at its own boundary.

$$R_{\text{state}}(\mathcal{T}; k{=}1) \;\ge\; S(\text{state}) - \tfrac{1}{900} \;=\; 0.694 \;>\; 0, \tag{3}$$

against a frozen band of 0.151, P-T3-96 supported. This paper therefore exhibits a task with positive resolving power, with a realisable member on the leading side and a subtracted side bounded by a construction. Why §4's 0.867 needed the qualification it carried: at k=8, 57% of the transcript, the same family gives a gap in the other direction, -0.033 — a resolving power is a joint property of a task and a budget.

Table 2. E-state, $n = 200$, exact containment, no judge. The oracle row is unregistered and labelled so.

arm	score	what it is, and which side of Def. 2 it sits on
index@1	0.000	index-only at the admissible budget: the subtracted side
index@2	0.180	index-only at k=2, our own retriever
index@2-oracle	0.990	index-only at k=2 handed both bearing records [unregistered]
state	**0.695**	bounded-width carrier, 240-char scratchpad, single forward pass: the leading side
complete	0.990	index@all, a ceiling reference, and not the leading side
optimum at k=1	≤ 0.00111	(2), from the construction



Figure 1. Table 2, drawn. Five arms on one family of 200 probes, one scoring rule, no judge.

The design rule, released.

R1. The answer is determined by more records than the budget holds, at every segmentation a designer must anticipate (Prop. 3).

R2. The answer occurs verbatim in none of them. Not implied by R1, and the half easy to omit: Table 3 shows what an answer sitting in some record costs.

The cheapest construction satisfying both is the one measured above. R2 is the test a deployed benchmark fails: the companion measures V, the verbatim-presence share, at 0.55 on the flagship dimension against 0.00–0.07 elsewhere (Anonymous, 2026e), exactly the rate at which R2 is violated, and it costs a benchmark author two counts over their own text. We release the generator and 500 probes (make_probe_set.py, release/sighted_probes.jsonl), with a –verify mode re-deriving the bound's three premises per probe.

Three scope caveats (existence not isolation; the advantage is a property of the budget, not a general one; and the carrier's cost κ relative to the ceiling) are in Appendix A.

Chain-of-thought does not relax the budget, measured rather than argued. k counts records exposed to the reader, not tokens emitted, so a k=1 reader may reason about one operand and must still supply the other from the prior. The same 200 scratchpads re-read by a reader licensed to reason step by step (Table 4, Appendix A) do not move index@1, and the gold appears nowhere in its reasoning on any probe. On the carrier, CoT costs 0.165, unpredicted: the scratchpad has already composed at write time, leaving the reader nothing to reason about.

The order-sensitivity prediction is not settled, after three attempts. §3 argues the non-abelian structure of the parent's construction does no work; three successive runs (Appendix B), each removing a confound the last one's code review found, never produced the embarrassing direction, but none cleared its own frozen premise gate, so we still decline the verdict.

6 The robust form of the rule, measured

Prop. 4 is the form of R1 no budget or chunking defeats; §5's family sits at its cheapest point, $m = 2$. This section puts m on an axis. One entity is given a value once; $m - 1$ later records each state an increment to that same entity, as an operation and never a resulting value; the answer occurs in no record, asserted per probe, so the index-only optimum at $k = 1$ is bounded by the uniform prior, weaker at small m (1/218 at $m = 32$). Everything else matches §5.

Figure 2 draws the two crossing curves against m, with the first (confounded) run shown dashed for contrast, where both arms reach the floor together, why that run was graded not evaluable rather than negative.

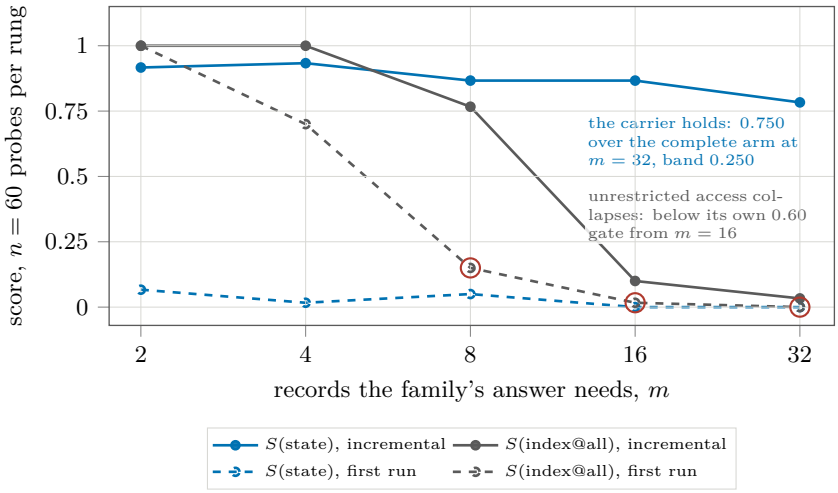

Figure 2. A bounded carrier holds where an arm with unrestricted access collapses. Solid: the incremental family, carrier 0.917 → 0.783 over a 16× range in m while the complete arm falls to the floor. Dashed: the first run (Appendix F), where both arms reach the floor together.

The carrier does not fail, and the arm with complete access does. Over a 16× range in m the carrier moves 0.917 → 0.783: the price clause of Prop. 4 is not reached at $m = 32$, the direction opposite to what we registered. Meanwhile complete falls 1.000 → 0.033, and from $m = 8$ upward the bounded-width carrier scores above the arm that sees every record, by 0.750 at the deepest rung. The mechanism is the parent framework's own claim: the carrier composes on write, one addition per record, while the access-complete reader must compose $m - 1$ increments at read time in a single forward pass with no scratchpad of its own.

Two consequences. (i) An access-complete arm is not an upper bound on what a task admits: our own frozen gate required $S(\text{complete}) \geq 0.60$ before the carrier could be graded, which at $m \geq 16$ would refuse to grade rungs the carrier answers four-fifths of the time — Prop. 5 applied to the leading side, in the safe direction, since a realised leading-side arm gives a lower bound. (ii) The witness family generalises and the bound tightens as it does: $0.783 - 1/218$ at $m = 32$ is a positive resolving power on a family whose cardinality requirement defeats every fixed (k, w) by Prop. 4.

7 The same claim on state-space models

§5's carrier is a proxy: bounded width is enforced by a character cap and an update prompt with no question field, both violable. A state-space model needs no such enforcement, so we ran two. falcon-mamba-7b has 643 weight tensors of which zero are attention projections, carrier 10,485,760 values whatever the sequence length, a state channel by Def. 1: the hidden-state carrier scores 0.035 against a bound of 1/900, read alone P-T3-S1 not supported. The control refuses that reading: driving the same weights with §5's scratchpad protocol scores 0.005, a fortieth of the model's own ceiling (0.340), so a carrier cannot be blamed for a failure that follows the model, and under the corrected premise gate (half the model's own ceiling, not whether it can add given both operands) the verdict is not decidable, not answered against us. rwkv7-2.9B (linear attention) fails the same way: hidden state 0.065 (12% of ceiling), scratchpad 0.005, both carriers answer with the offset record, keeping the last bearing record and losing the one before it. Eight pure recurrent checkpoints have now been screened against a floor fixed before any ran (Table 7, Figure 4); none clears it, the best reaching 0.535, so the architecture-level form of the claim is not decidable at the largest checkpoints we can run, blocked by whether such a checkpoint can do two-addend composition at all, which is not the channel question. The mis-specified premise gate, the two-model summary, and a four-defect code review are in Appendix C; none of the four can strengthen the verdict.

8 Discussion, limitations and conclusion

Every proof and the construction are self-contained; imports are restrictions, never premises (Appendix G).

What remains owed. Neither arm below changes this paper's claim of existence, not isolation. An order-sensitivity comparison whose gate can hold ($0–2); a tighter bound at small m; and a pure-recurrent checkpoint that can do the task at all, best of eight so far 0.535. Full pricing in Appendix I.

Limitations and conclusion. One carrier model per witness family, a proxy for an architecture (§7); exact containment as the scorer, removing the judge axis by construction; and segmentation sensitivity measured at one point, the leading side a lower bound only. Against those limits: blindness has a checkable sufficient condition, sightedness a growing-cardinality counting requirement, and no difference of realised scores bounds resolving power either way, so the positive route is a construction or nothing. That construction exists, a bounded-width carrier at 0.695 against an index-only optimum at 1/900 by construction, surviving to a cardinality of 32 while complete access falls to 0.033, and it costs a benchmark author two clauses: more records than the budget holds, verbatim in none of them.

Reproducibility statement

Every plotted coordinate and every cross-paper number is written by a script from the result JSONs listed in the supplementary material, and no figure or table contains a typed number. The preregistration for every arm graded in this paper, the probe generator (with its –verify mode re-deriving the construction's premises per probe), and the scoring code are included in the anonymised supplementary material, with sha256 hashes so a reader can verify a downloaded archive matches the one this paper was computed from. Every criterion graded in Appendix O was frozen in a preregistration file before its data existed; the appendix states the counting rule and the freeze commit for each.

Ethics statement

All probe families are synthetic, generated from fixed-seed scripts over fictitious entities and numeric values; no human data was collected and no individual is identified or evaluated.

The two state-space checkpoints run locally are openly released weights used under their standard terms. 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. Every result reported here is computed by a script the authors wrote, and every number is checked against the result file it cites. 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.

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. The Underdetermined Stream: Two of the Three Walls Are Resource Bounds; the Third Is the Task. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

Anonymous. The Structural Content of a Transfer Relation: What a Benchmark's Own Labels Can Predict About Its Other Dimensions, Measured on Eight Public Corpora, and the Refinement Limit That Bounds It. concurrent manuscript; author list and repository withheld for double-blind review, 2026c.

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, 2026d.

Anonymous. Two Counts over Text: What a Memory Benchmark Can See, Measured Before Anything Is Run. concurrent manuscript; author list and repository withheld for double-blind review, 2026e.

Anonymous. The Infinite-Record Limit: A retrieval budget that is a constant share of the record for the hardest items and a vanishing one for the median, and an exponent that belongs to the lever. concurrent manuscript; author list and repository withheld for double-blind review, 2026f.

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, 2026g.

Anonymous. What a Rubric Judge Measures: Two Defects in a Shipped Scorer, the Ceiling on Perfect Provenance, and the Credit Rate That Prices the Step from a Count to a Bound. concurrent manuscript; author list and repository withheld for double-blind review, 2026h.

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.

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.

Paul W. Holland and Howard Wainer (eds.). Differential Item Functioning. Lawrence Erlbaum, Hillsdale, NJ, 1993.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In Conference on Language Modeling (COLM), 2024. arXiv:2404.06654. Introduces multi-hop tracing (variable tracking: a chain of X2=X1 bindings) and aggregation (common- and frequent-words extraction over a list) as categories beyond retrieval. Both families have the property this paper's construction needs – the answer is in no single record – and neither is accompanied by a bound on what an index-only strategy could score.

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.

Charles F. Manski. Identification Problems in the Social Sciences. Harvard University Press, 1995.

Charles F. Manski. Partial Identification of Probability Distributions. Springer, 2003.

Roger E. Millsap. Statistical Approaches to Measurement Invariance. Routledge, 2011.

Alireza Nadali, Patrick Cooper, Ashutosh Trivedi, and Alvaro Velasquez. Kv-fold: One-step kv-cache recurrence for long-context inference. arXiv preprint arXiv:2605.12471, 2026. Treats the KV cache as the accumulator of a left fold over sequence chunks, appending the new keys and values at each step and passing the enlarged cache forward, with no retraining. Reports 100% exact-match retrieval over 152 needle-in-a-haystack trials from 16K to 128K tokens at chain depths to 511, and contrasts itself with streaming methods that trade fidelity for bounded memory.

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.

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.

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

Haizhou Xia. Diagnosing evidence utilization in long-context and retrieval-augmented language models under matched evidence conditions. arXiv preprint arXiv:2606.06758, 2026. A four-condition protocol (no-evidence, full-context, retrieved, oracle-evidence) over 18,000 predictions, with ONCU as a denominator-valid estimate of the oracle-reference advantage a model recovers. A difference of two realised scores, normalised.

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.

Supplementary Material

This appendix follows the references and is supplementary. The main text is self-contained: every claim it makes is stated with the measurement that decides it, and the sections below hold the extended related work, the full order-sensitivity comparison across all three runs, the state-space arm in full, the budget-axis factorisation in full, the cardinality axis (both the incremental family and the first, confounded run), the full import accounting, two open questions, what remains owed in full, the notation table, the preregistration record, and the public scorecard of every graded row. Reviewers are not required to read it.

A Extended Related Work

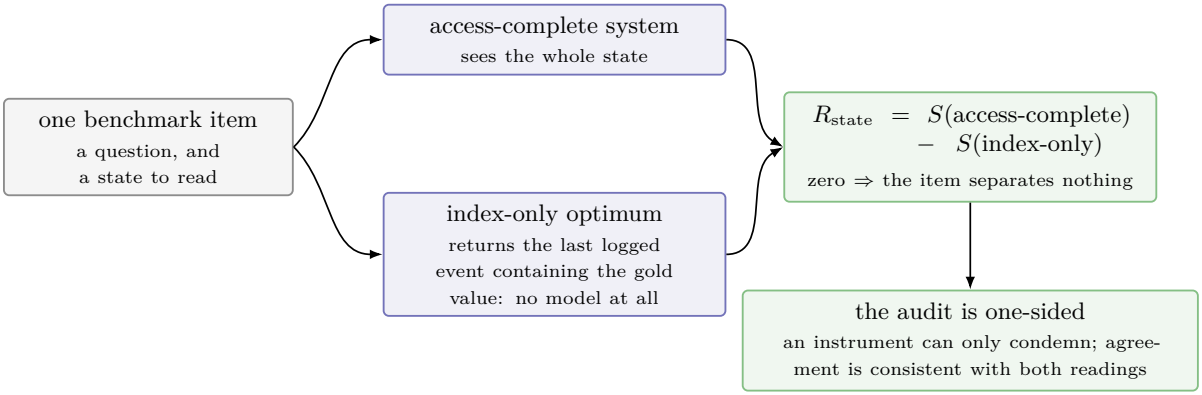

Figure 3. What a resolving power is. R_{state} is defined per item as the score a system with full access to the state achieves minus the score achieved by an optimum that uses no model at all. Where that difference is zero the item cannot separate the systems it was written to separate, and no judge, however reliable, changes that. The audit built on this quantity is one-sided (Prop. 2): agreement between the instrument and a system is consistent with both readings, so an instrument can only condemn. What a rubric judge inserts between any scored comparison and this estimand is measured by a companion paper (Anonymous, 2026h).

Figure 1 (the bar-chart form of Table 2) and Figure 2 (the cardinality axis) are given in the main text (§5, §6) and are not repeated here.

Three things §5's construction does not say. (i) It is existence, not isolation: storage cannot be matched between an $O(1)$ carrier and an index, and what is matched is the read budget, 240 characters against one retrieved record. (ii) The advantage is a property of the budget: handed both bearing records, an index scores 0.990 and beats the state arm's 0.695. (iii) $\kappa = S(\text{state})/S(\text{complete}) = 0.702$, and the pre-flight band was too wide to freeze, so κ is reported and not graded.

Psychometrics has a name for this quantity, and the sign is the other way round. An instrument is measurement invariant across two groups when an item's response distribution depends on the latent trait and not on group membership; where invariance fails at the item level it is differential item functioning, and DIF is the standard object of a large detection literature (Holland & Wainer, 1993; Millsap, 2011). Resolving power is DIF with the sign flipped and the two roles exchanged. There the group is a nuisance (language, gender, test form) and an item responding to it is biased and should be repaired or dropped. Here the group is the target: the architectural class, a compressive state channel against a scalable index. An item that responds to it is the only kind worth having, and $R_{\text{state}}(\mathcal{T}; k) = 0$ says precisely that the benchmark exhibits no differential functioning across the channel axis, in psychometric terms a clean bill of health and in ours the defect.

Naming the inversion also says why we cannot borrow their methods. Mantel–Haenszel, logistic-regression and IRT-based DIF detection all condition on the latent trait estimated

from the responses of both groups, so each requires a sample from each group. Our second group is the class of systems with a fixed-width carrier, and §4 shows its optimum is not enumerable: there is no sample to condition on, and no statistical machinery manufactures one. That is why this paper's positive result is a construction (§5) rather than a detection; we build a member of the group the detection literature assumes it already has.

Construct validity, and where a per-item quantity goes past a survey. Bean et al. (2025) review 445 benchmarks and report contested constructs as systemic; Raji et al. (2021) make the sharper qualitative version of the point. A survey can say a field's constructs are often unclear; R_{state} takes a value on one item, and where that value is zero the item cannot separate the systems it was written to separate, whatever the construct clarification. The arithmetic form of this argument, two counts a benchmark author can run on their own text, is the companion instrument paper's (Anonymous, 2026e), and the premise appears in prose before either of us: Tanguturi (2026) conclude of two deployed memory evaluations that "by construction, these benchmarks cannot distinguish a memory system from a long-context LLM".

The long-context literature already builds families whose answer is in no record, and this paper's propositions classify them. Hsieh et al. (2024) introduce two categories beyond retrieval: multi-hop tracing, whose variable-tracking family is a chain of X2=X1 bindings, and aggregation, whose common-words family asks for the most frequent items of a list. Our own two families are those two, in form: the chain family of §6 is variable tracking and the increment family is an aggregation. That is what a construction wants to hear, the shape being published two years before this paper, and it is why the contribution here cannot be the task. What Prop. 3 adds is a classification those families do not carry. An aggregation whose answer needs c records is sighted at $k < c$ and blind at $k \geq c$, so a budget nobody states decides whether such a family measures composition or retrieval; and we know of no bound in that literature on what an index-only strategy scores on either family, the quantity §5 bounds by construction rather than by claim. The increment family has a second reason to be taken seriously that is not ours either: it is a proactive-interference construction, and Wang & Sun (2025) measure that repeated updates to one entity defeat working memory in current models. We did not build it from that result, which makes it better evidence that the family is not a private artifact.

A normalised oracle gap is now being proposed as a diagnostic, and Prop. 5 is about exactly that quantity. Xia (2026) compare four matched conditions (no evidence, full context, retrieved, and oracle evidence) over 18,000 predictions, and report ONCU, the share of the oracle-reference advantage a model recovers. That is a difference of two realised scores over a realised denominator, and §4 says what such a difference can bound: neither side, whatever the precision, since both realisations bound their optima in the same direction. The protocol is not thereby wrong, matched conditions removing confounds a single score cannot, but its denominator is not a ceiling on what the task admits. That is now exhibited rather than argued: on the increment family the access-complete arm falls to 0.033 while a bounded-width carrier holds 0.783, so from $m = 8$ upward a carrier scores above the arm that sees every record. Such a carrier is none of the four conditions, so the oracle condition is not the maximum available and a ratio against it is not a share of anything.

A 2026 long-context method supplies a live instance of the distinction Prop. 3 says a needle task cannot resolve. Nadali et al. (2026) carry a key–value cache across chunk boundaries as the accumulator of a left fold, with no retraining, and report 100% exact-match retrieval over 152 needle-in-a-haystack trials from 16K to 128K tokens at chain depths to 511. The paper's own language for what this shows is internal state reused across segments and a stable form of recurrence. Two facts from that same abstract settle the classification without any measurement of ours: the update "appends the newly produced keys and values, and passes the enlarged cache forward", so the carrier grows with the sequence and is an index under Def. 1; and the evidence is a needle task, which is a k=1 task by construction, so a perfect index-only score is exactly what Prop. 3 says cannot separate a state channel from an index.

This is not a correction of their result, and we say which claim we are addressing. They do not claim a bounded carrier, they contrast themselves with streaming methods that "trade fidelity for bounded memory", and on their own terms the result is what it says: long-range retrieval survives a chunked forward pass. What the vocabulary invites, and what a reader takes from internal state plus a saturating drift plateau, is a claim about accumulated state the reported evidence cannot carry. The instrument for saying so costs nothing: walk the cache the runtime returns at two sequence lengths and count the elements, and a count proportional to length is an index whatever the method is called (Anonymous, 2026e). That the classification is available by measurement rather than by reading a method description is the point of the companion's probe, and this is the first published method we can point it at that uses recurrence language for a cache the paper itself describes as enlarging.

The same question was asked three months earlier, procedurally. Kalaitzidis (2026) argue that a benchmark operationalises theoretical commitments, and propose an audit for whether a proposed benchmark can discriminate a claimed capability from proxy behaviours, this paper's question, in almost this paper's words, arriving first. That work supplies a procedure, a failure-mode taxonomy and a worked audit. What a procedure cannot supply is what makes a verdict checkable by someone who disagrees: a quantity taking a value on a single item, and a proof about which side of it can be bounded. The boundary is worth stating in both directions: an audit procedure runs on a benchmark proposal, before any item exists, which R_{state} cannot; and R_{state} condemns a shipped item with a number, which a taxonomy cannot.

Partial identification is the correct frame, and it is borrowed knowingly. The econometric tradition beginning with Manski (1995) rejects point identification bought with unfalsifiable assumptions in favour of reporting the set the evidence admits (Manski, 2003); §4 places the audit's one-sided region in that frame, and the ordinary remedy there, add a realisable member on the unbounded side, is exactly what the construction of §5 does.

Where the parent framework enters. The access-structure account of Anonymous (2026a) supplies the distinction itself, state against index, and the construction pattern the witness generalises: a value stated once, subsequent changes announced as operations rather than as resulting values. The theory here neither needs nor imports its empirical claims; the dependency is the estimand's definition, and Appendix G carries the full import accounting.

Table 3. Three k=1 strategies on a single-record family, and the reason such a family cannot carry a proof. Enumeration over 200 probes; the candidate bound is the first row.

index-only strategy at $k = 1$	score	
guess an [A-H]NNN code uniformly, the candidate bound	0.00014	
retrieve any attribute-matching value record, echo its value	0.359	
the same, preferring the rarest entity by document frequency	0.485	
read the link record first, then fetch the value record it names	1.000	$R = 0$

Table 4. Chain-of-thought at query time, against the same arms without it. Scored on the ANSWER: line; the unscored "gold anywhere in the trace" column is the credit a lenient scorer would have handed out.

arm	without CoT	with CoT	gold anywhere in the trace
index@1	0.000	**0.000**	**0.000**
state	0.695	0.530	0.635

B The order-sensitivity comparison, all three runs

The prediction §3 left owed is not settled, after two attempts that are worth more than a settlement would have been. §3 argues that the non-abelian structure of the parent's

construction does no work and leaves it untested. Our first attempt (P-T3-93) compared the order-insensitive family against sanity.py's order-sensitive one and returned 0.265 against a band of 0.151. Reading the generator rather than the number showed the two families differ in three ways of which only one is order, the question's suffix, singular against plural in the link record, and whether the entity carrying the gold may also appear in the background chatter, so the verdict was not about its own axis. Amendment 1 cut two families from one random stream differing in exactly one sentence, and on the preregistered analysis $|S(\text{state}, \text{sym}) - S(\text{state}, \text{exchanged})| = |0.530 - 0.400| = 0.130 \leq 0.151$: supported.

Then the code review of §5 found that 35 of those 200 probes carry a decoy value record for the same (attribute, entity) pair, inherited from the background rule of the generator we were matching. On the 165 uncontaminated probes the same comparison gives $|0.576 - 0.406| = 0.170 > 0.151$. Two defensible analyses of one dataset disagree, so the verdict is not decidable and we print both numbers rather than choosing. That is the report the companion argues for and the one we would demand of anybody else (Anonymous, 2026g); choosing the preregistered number because it is the preregistered one, when we now know the preregistered set is contaminated, would be the forking path in its most defensible-sounding form. The confound P-T3-93 tripped on is nevertheless visible rather than argued in the matched run: index@2 is 0.435 on both families, the oracle 0.980 against 0.970, complete 0.920 on both, and only index@1 moves (0.150 against 0.000), which is the singular/plural retrieval artifact we disclosed in the amendment before running it.

A third run, on a generator built to remove that artifact, settles what the comparison can be about and not the comparison. The decoy came from one clause of the background rule, so the rule was changed (when the background draws the entity that carries the gold, its attribute is drawn from the complement of the queried one), and the pair was regenerated at $n = 200$ under a fresh seed, with decoys $= 0$ asserted per probe and the sym/exchanged difference still exactly one sentence. The result is that the frozen premise gate fails, and it fails on the axis the amendment had already identified: index@1 is 0.170 on the symmetric family against 0.000 on the exchanged one, so the criterion is graded not evaluable as frozen. The mechanism is measured rather than argued. On 34 of the 200 probes the two families' bm25 top-1 differs, always in one direction: the link sentence's own tokens are part of the query's evidence, so a one-sentence difference reaches every index arm through retrieval, and only the carrier (which retrieves nothing, consuming the record once in order) is immune to it. That is the same lesson P-T3-93 taught one level up, and it is a reason to doubt any matched-pair design whose arms retrieve.

What the decoy-free pair shows once the design's own controls replace the frozen gate, labelled post hoc because that is what it is. The controls the comparison needs are all satisfied and are not the gate that failed: decoys 0/200 by construction, the gold in exactly one record on both families, index@2 equal at 0.470 on both, the oracle-at-k=2 within 0.010, and complete saturated on both. Under those, and at the same frozen band of 0.151, the carrier scores 0.630 on the symmetric family against 0.400 on the exchanged one: a gap of 0.230, with the symmetric family higher. The direction that would have embarrassed the cardinality reading did not occur (§??(b) registered it as an exchanged-family advantage); what occurred is its opposite, and the ordered composition costs the carrier. We do not promote that to a verdict, because the criterion it would answer was graded not evaluable and a post hoc analysis cannot inherit a frozen band. It is reported as the one direction the three runs agree on, and §8 carries the version that could be graded.

C The state-space arm in full

§7 reports what the arm establishes and what it cannot decide, on two models; Table 5 is the first model's four arms in full, which the body reports as one sentence. This is the rest of it; every defect below moves the arm toward less evidence, which is what makes "not decidable" the supportable verdict rather than a convenient one (Figure 5).

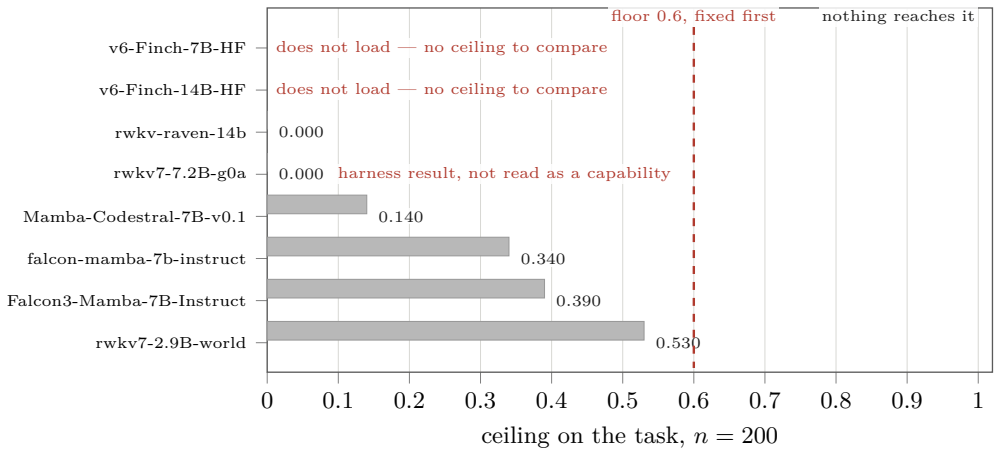

Figure 4. Every candidate against the floor fixed before any of them was run, and none clears it. A ceiling below the floor makes the channel question unanswerable rather than answered negatively. Recomputed from results/ssm/SCREEN.json and SSM2.json.

Table 5. E-ssm on the arith family, same probes and seed as §5. Criteria frozen in prereg/E_SSM_PREREG.md before the runner existed. Greedy decoding on local weights is deterministic and we checked it: two runs are byte-identical, so $\Delta_{\text{app}} = 0$ here rather than assumed. ssm_arm.py, ssm_scratch.py, analyze_ssm.py.

arm	score	of ceiling	what it is
ssm-extract, both bearing records given	0.340	—	the model's own ceiling
ssm-all, records through the hidden state	**0.035**	10%	the state channel
scratchpad protocol, same weights	0.005	1%	the control
ssm-one, one record	0.005	—	index at k=1, bound 1/900

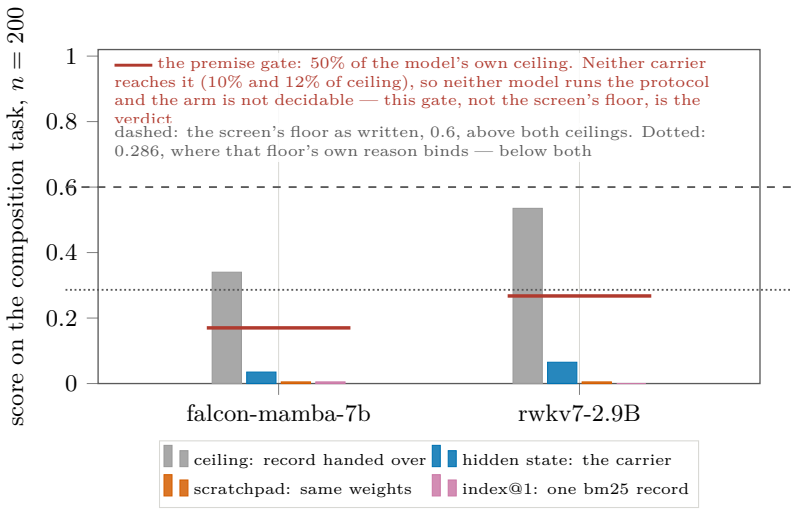

Figure 5. The four arms of Table 5, and why the verdict is not decidable rather than negative. Both local carriers score about a tenth of their own ceiling when the records must pass through the hidden state, and the scratchpad control (the same weights writing to text) scores lower still. The reason the arm cannot be graded is the red rule: the premise gate asks for half the model's own ceiling, neither carrier reaches it, and analyze_ssm.py returns not decidable from exactly that test, so neither model runs the protocol at all and the channel comparison is confounded with scale. The screen's floor is context and not cause, and it is drawn twice because the number frozen as 0.6 does not follow from its own stated reason, which binds at 0.286: above both ceilings on the first reading, below both on the second. What these bars measure is whether a checkpoint can do the task at all, which is not the question about channels the paper set out to ask.

P-T3-S1 is not supported: $R = 0.035 - 1/900 = 0.034$ against a band of 0.142 set from the statistic's resolution before the run. Read alone that is a hit against §5, and the preregistration says to report it in that direction.

The preregistered premise gate was mis-specified, and the corrected gate decides the verdict. P-T3-S2 asked whether the model can do the arithmetic when handed both operands. It can, at 0.340, so the gate passed and declared P-T3-S1 evaluable. The question it should have asked is whether the model can run a carrier protocol at all, and the scratchpad control answers that and it fails. A premise gate that tests a capability the arm does not exercise is not a premise gate. Under the corrected gate, the best carrier must reach half the model's ceiling, 0.170, and the best reaches 0.035. The architecture question is therefore not decidable at this scale, not answered against us. What this arm does establish is that the extension is cheap and runnable, that Def. 1 applies to a real recurrent model without any prompt-level enforcement, and that a 7B pure state-space model does not reproduce §5's result in either carrier. What would settle it is a stronger pure-recurrent model, not more probes and not a better prompt.

Table 6. Two pure state channels of different architectures, same probes, same scorer.

model	growth	ceiling	hidden state	of ceiling	scratchpad	k=1
falcon-mamba-7b (selective SSM)	1.00×	0.340	0.035	10%	0.005	0.005
rwkv7-2.9B (linear attention)	1.00×	0.535	0.065	12%	0.005	0.000

We then looked for that stronger model, and the search is the result. A carrier arm can only be read against a model whose ceiling leaves room for it, so we fixed a screening floor before running anything: the ceiling on this family must reach 0.60, below which a carrier at the corrected gate's half-of-ceiling sits under the band and no outcome is decisive (Table 6, Table 7). Every open pure-recurrent checkpoint we could obtain was screened on that one cheap arm.

Two provenance facts about this table, because it is the one place in the paper where a result file was overwritten. (i) ssm_screen.py wrote SCREEN.json whole on every invocation, so screening row 4 deleted rows 1–3 from the artifact this caption cites, and the paper printed rows the file no longer held. The script now merges by model id, the three rows were recovered from the commit that first wrote them, and the file records that recovery. A caption that cites an artifact has to be re-checkable against it, which is this programme's own rule turned on its own screen. (ii) Row 4's ceiling is 0.530 here and 0.535 in Table 6: the screen and the full arm build the same prompt by two code paths and disagree on one probe of 200. Both are printed with their source rather than reconciled, because reconciling them by choosing would hide a 1/200 apparatus difference that is real.

One candidate then cleared the floor's reason while failing its number, and that is worth stating before the result. rwkv7-2.9B scores 0.535 on the full arm. The floor was written as 0.60 and justified by the claim that below it a carrier at the corrected gate's half-of-ceiling sits under the band; but half of 0.60 is 0.300 and the band is 0.142, so the constraint actually binds at $2 \times 0.1431 = \mathbf{0.286}$. The frozen number does not follow from its stated reason; the constraint binds at 0.286. Both are reported, and the direction matters: the corrected floor admits more models and demands more arms, including one that had already failed, so it is not a door opened for the hypothesis. Under it, rwkv7-2.9B is in and its full arm was run.

What a code review of this arm found. Four defects, none discovered by the numbers looking wrong.

Table 7. Screening pure-recurrent models on the ceiling, floor fixed at 0.60 before any was run. The channel column is measured, not read off an architecture description (channel_probe.py). Rows 1–4 are the first screen; rows 5–8 are a second round whose candidate list was frozen before any of them ran (§7). ssm_screen.py, results/ssm/SCREEN.json, results/ssm/SSM2.json.

	model	channel	ceiling	outcome
1	falcon-mamba-7b	state chan.	0.340	below floor; full arm run above
2	rwkv-raven-14b	state chan.	0.000	below floor. It runs and is coherent; it subtracts where it should add, writing "766 − 149 = 567"
3	v6-Finch-7B	—	—	does not load: its custom code names a parameter in an fp32 keep-list this transformers rejects
4	rwkv7-2.9B	state chan.	0.530	below floor, and the best of the eight. It did not load until the fla package was installed, and this table's first version recorded that failure rather than this score
5	Falcon3-Mamba-7B-Instruct	state chan.	0.390	below floor
6	Mamba-Codestral-7B-v0.1	state chan.	0.140	below floor
7	rwkv7-7.2B-g0a	state chan.	0.000	below floor, and read as a harness result rather than a capability: the library warns that its own rwkv-7 path may be wrong, and row 4 is the same family at less than half the size
8	v6-Finch-14B-HF	—	—	does not load, for row 3's reason at a later library version. Clearing the offending keep-list lets the classes construct, and the checkpoint's own loading path then reads 4.6 TB from a 28 GB file without finishing in an hour, so this row is a fact about the release rather than a screen result

(1) The purity check read names. It searched parameters for attention projections and disqualified RWKV, whose time-mixing block is merely named attention. Replaced by measurement: the channel probe of the companion instrument paper (Anonymous, 2026e).
(2) The premise gate tested the wrong capability, asking whether the model can add when handed both operands rather than whether it can run a carrier protocol at all (§7 above).
(3) The screening floor did not follow from its own justification, 0.60 written where the reason implies 0.286.
(4) The index arm was handed the wrong record. ssm-one used the record fixing the base. A k=1 retriever picks the record naming the queried entity, which is the offset record in 200 of 200 probes and the base record in 0. Re-run correctly, the index scores 0.000 on both models against 0.005 and 0.000 as run: the mistake was conservative, since it gave the subtracted term its better shot, and R is unchanged.
Two checks came back clean. bearing() is correct on all 200 probes, and greedy decoding on local weights is byte-identical across runs on both models, so $\Delta_{\text{app}} = 0$ is measured rather than assumed. One came back dirty and is not repaired here: left padding is not neutral for a recurrent model, because pads enter the recurrence before the content. Comparing each prompt alone against the same prompt in a padded batch, the first token's argmax agrees 11/12 on falcon-mamba with a largest logit shift of 14.25, and 12/12 on rwkv7 with 1.75. No verdict here moves, because the margin being judged, 0.035 against a band of 0.142, is four times the perturbation; but the correct treatment for a recurrent model is right padding with the true end indexed, and it is owed.

Nothing in the body rests on any of the above: the verdict §7 reports is not decidable, and every defect listed here would have to be repaired in our favour before the arm could say anything stronger.

D The factorisation on the budget axis

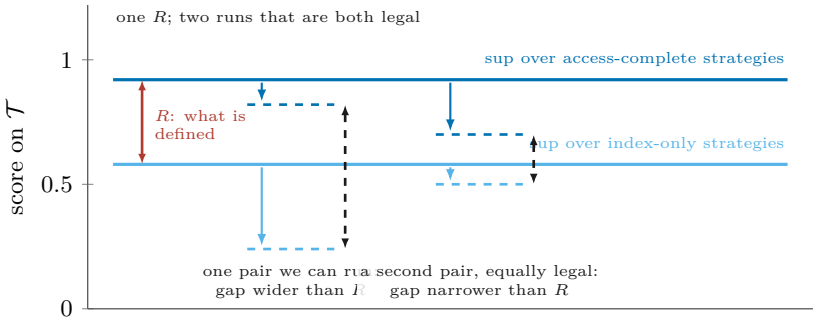

Figure 6. Why a difference of two realised scores bounds resolving power from neither side. R is the difference of two optima (solid rules). Every arm we can actually run sits at or below its own optimum (dashed), and both errors point the same way, so their difference does not. The escape is not a better measurement: it is a subtracted optimum bounded by a construction, which is what §5 builds.

§4 states the factorisation and says which axis can test it. The arms are here in full (Table 8), and the test itself is one axis (Figure 8): if ι is a property of the (family, scorer) pair and carries no system index, coverage may move as far as it likes and S/a may not follow it.

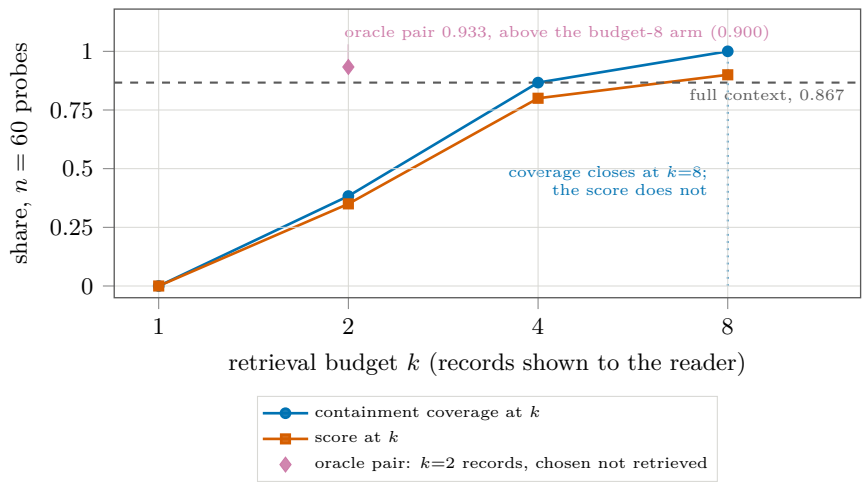

Figure 7. Coverage saturates a rung before the score does, and an oracle pair beats a reader that has already seen everything. At this budget the binding constraint is which two records, not how many. Recomputed from results/BUDGET.json.

Table 8. The factorisation tested on the budget axis, from local artifacts only. Coverage moves by a factor of 2.6 and S/a does not follow it: every arm sits within half a binomial standard error of ι. Computed from local artifacts; nothing a reader cannot reach is needed, and the whole tabular is generated from results/factor/FACTOR.json, which already held every cell of it.

arm	coverage a	score S	S/a	$\lvert S/a - \iota \rvert$ in SE
bm25-2	0.383	0.350	0.913	0.10
bm25-4	0.867	0.800	0.923	0.18
bm25-8	1.000	0.900	0.900	0.47
oracle-2	1.000	0.933	0.933	0.47
$\iota = 55/60 = 0.917$, binomial SE 0.0357 at $n = 60$				worst 0.47

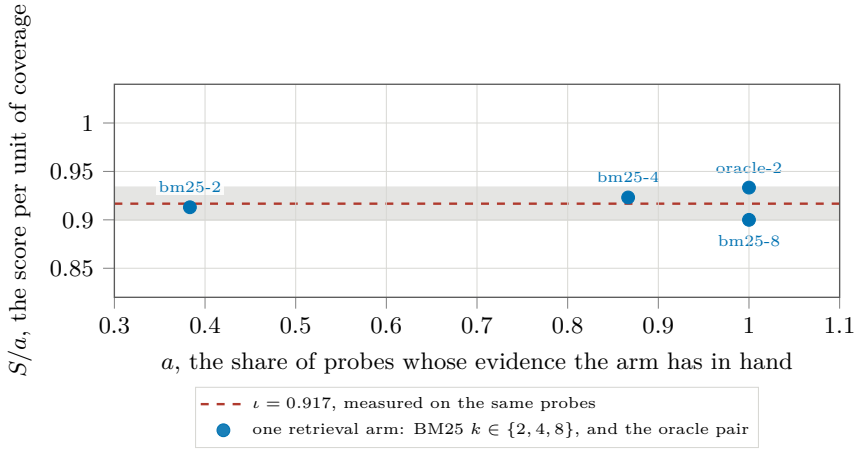

Figure 8. The factorisation tested on the one axis this paper can move. Each point is one retrieval arm (BM25 at $k \in \{2, 4, 8\}$, and the oracle pair, which is not a budget at all but the two bearing events handed over directly), plotted as its coverage a against the score it converts that coverage into. If $S = a\iota$ with ι a property of the (family, scorer) pair and not of the arm, the points lie on a horizontal line at ι, and they do. The shaded band is half a binomial standard error and every arm is inside it, the worst at $0.47\,\sigma$, while coverage moves by 2.6×. This tests the factorisation on coverage and on one family; the other axes are not tested here, and §4 says which claims that restricts.

Table 9. The cardinality axis at $n = 60$ probes per rung. state is the bounded-width carrier, complete the same model handed the whole record in order, and index@1 one bm25 record. Every rung asserts that the answer occurs in no record. Read the two middle columns against each other: the carrier is nearly flat while the arm with unrestricted access collapses. The whole tabular is generated from results/unbound/UNBOUND_INCR.json, the incremental family. That directory holds two graded artifacts whose state columns differ by an order of magnitude at $m = 2$, the deepest row already read the incremental macros, and nothing in the table said which file the four rows above it came from.

m	state	complete	index@1	k=1 bound	
2	0.917	1.000	0.000	1/8	
4	0.933	1.000	0.000	1/22	
8	0.867	0.767	0.000	1/50	the carrier overtakes
16	0.867	0.100	0.000	1/106	
32	**0.783**	**0.033**	0.000	**1/218**	

E The cardinality axis, incremental family in full

F The cardinality axis: the first run

§6's family is the second one built for that section. The first put the m records on a chain of distinct entities (a base value for e_0, then e_i's value stated as an offset from e_{i-1}'s), and it returned a result its own frozen gate declares not evaluable: at $m \geq 8$ the access-complete arm itself falls below the gate's 0.60, and at $m = 2$ the carrier scored 0.067 against 0.695 on §5's two-record family. Both numbers stand and neither is about cardinality.

The confound was ours and it is visible in the artifact rather than inferred. A 32-entity chain needs more names than the generator's pool holds, so the pool was extended with compound names, and the background rule draws from that same extended pool, so almost every noise record introduced a new entity. Reading the stored scratchpads shows what the carrier then spends its width on: at $m = 2$ the surviving notebooks hold thirteen to twenty distinct entity–value entries in 240 characters, the chain's own records among them only sometimes. The first run therefore measured the width of the live value set, not the depth of the composition, because it moved two things at once. What it does measure cleanly is a price the proposition's own qualification predicts in a different coordinate: a fixed-width carrier fails at roughly thirteen concurrent entries regardless of how shallow the composition is (Table 10). §6's family returns the pool to §5's twelve names and puts every increment on one entity, so depth is the only axis that moves, and the amendment naming it was frozen before that run.

The amendment also wrote down three predictions before the run: that the access-complete arm would collapse in m, that the carrier would stay flat, and therefore that the carrier would end above the arm with complete access, a conjunction which, if it held, refutes the premise the first gate was built on (Figure 9). All three held (Table 9), which is why §6 states the third as its finding and not as a surprise.

G What this paper needs from its companions

Table 11 carries the accounting; the two rows to check first are the two marked as restrictions.

H Two open questions

(a) What happens to resolving power on an underdetermined task? Suppose a memory task admits several equally correct conventions (a state recorded as json or as xml, an event ordered by mention or by timestamp) so that the answer is not unique. Then two things

Table 10. The first cardinality run, and the width pressure its scratchpads record. Recomputed from results/unbound/m*/state_*.json at build time. The notebook is capped at 240 characters and sits within a few characters of that cap at every rung, while the number of rewrites the cap forced rises with m: the carrier is width-saturated throughout, which is the mechanism the paragraph above describes. The count of distinct entity–value entries is deliberately absent from this table. It is the quantity the paragraph reads off the scratchpads, and it is a reading rather than a statistic: the notebooks are written in formats the model invented and they differ between probes, so any regular expression counting "entries" over them would be a definition we had introduced. The two columns here are what the runner stored.

m	S(state)	S(complete)	rewrites truncated	notebook chars	n
2	0.067	1.000	8.7	238	60
4	0.017	0.700	9.4	228	60
8	0.050	0.150	13.5	234	60
16	0.000	0.017	21.4	237	60
32	0.000	0.000	34.3	232	60

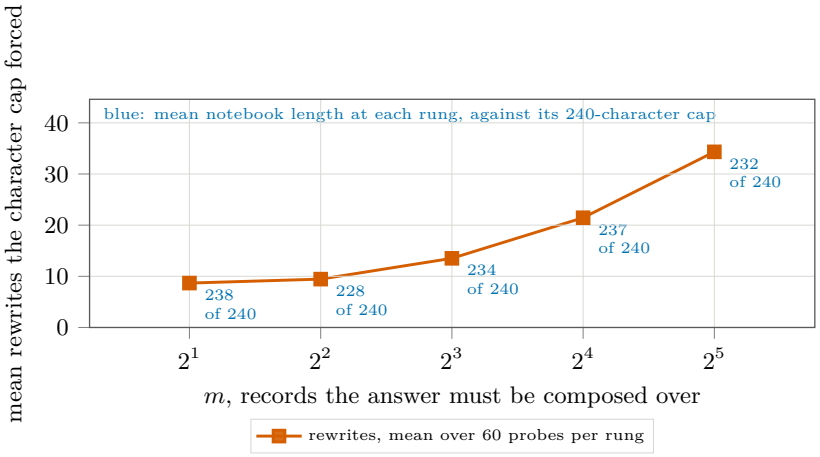

Figure 9. Width, not depth: what the first cardinality run actually measured. The carrier's notebook is capped at 240 characters and sits within a few characters of that cap at every rung, while the rewrites the cap forces rise fourfold in m. The blue figure under each point is that rung's mean notebook length against the cap; the rising curve is the rewrites. A carrier that is full at $m = 2$ and still full at $m = 32$ is not being tested on composition depth; it is being tested on how many concurrent entries it can hold, which is what the extended name pool changed. The mechanism is in the artifact rather than inferred from the scores, which is why the amendment naming it was frozen before the second family ran.

Table 11. What is self-contained and what is imported. In every imported row the import is either a restriction on what may be claimed or a corroboration, never a premise a proof rests on.

self-contained	imported, and what it costs
Every definition and proposition (§3, §4), with proofs	Nothing. Prop. 2 is restated from the companion and its proof sketch is carried in the restatement
The constructive existence proof, $R_{\text{state}} \geq 0.694$ (§5), its controls, and the design rule with its generator	Nothing for the measurement: exact containment, no judge, local weights or one pinned model
The factorisation check on the budget axis (Table 8)	Nothing: computed from files this paper holds. The factorisation's construction and its price are the judge companion's (Anonymous, 2026h)
—	The audit of deployed benchmarks: V, R^{str}, k^* and the four-corpus runs are (Anonymous, 2026e), quoted here only as motivation
—	The record-length limit: every number here is at one record length, and the budget's growth with length is (Anonymous, 2026f). Without it we would over-read the witness's k^*
—	The credit rate ι as a general axis, and the $\Theta(1/\iota)$ price of removing it (Anonymous, 2026h); this paper's own extract arm supplies the one value it uses
—	The state/index distinction itself (Anonymous, 2026a)

happen at once and they are not independent. The scorer's credit rate falls, because a correct response in the unrewarded convention is refused credit (Anonymous, 2026h); and the estimand itself becomes ambiguous, because "the access-complete optimum" is now an optimum over a set of conventions the task does not choose between. We think the second is the worse of the two and that our own machinery cannot see it: R_{state} is a difference of scores, and a difference of scores under a convention-sensitive scorer measures the convention as much as the channel. The stream companion makes underdetermination its subject and states the residual as a conditional entropy (Anonymous, 2026b); the composition of that quantity with a refusal rate is not worked out anywhere in this family, and we flag it as the most likely place where two of these walls turn out to be one.

(b) What result would most embarrass the cardinality reading? §5's order-sensitivity comparison is reported as not decidable, and the reading we currently favour, following §3, is that the non-abelian structure does no work and sightedness is a condition on cardinality alone. The result that would most embarrass it is a matched pair, cut from one random stream and free of decoy records, in which the order-sensitive family scores materially above the symmetric one on the carrier at k=1. That would mean the carrier is exploiting order rather than cardinality, and Prop. 3, which states the requirement as counting, would have to be re-opened. Three runs have now looked for it and none found it: the decoy-free subset of the second run moves 0.170 toward the symmetric side, and the third run, on a generator built to remove the confound both earlier ones carried, gives 0.630 symmetric against 0.400 exchanged, the same direction again and larger, 0.230. None of the three can be graded on its own frozen criterion (§5), so we still decline the verdict; but the result that would embarrass the cardinality reading is now the one direction the evidence has never moved in, and what would settle it is the graded arm of §8(1) rather than a fourth analysis of the same runs.

I What remains owed, in full

Neither arm below can change this paper's claim, which is existence and not isolation. $R_{\text{state}}(T;1) > 0$ is bounded from the construction rather than from any measurement these

arms would supply; what they bear on is how far the result separates composition from cardinality.

(1) An order-sensitivity comparison whose premise gate can hold. $0–2. Three runs have measured this comparison and none could be graded on its own frozen criterion: the first confounded three things with order, the second removed two of them, and the third (§5) removed the last and then failed a premise gate requiring index@1 to be blind on both families, which retrieval makes it impossible to guarantee, because the one sentence that distinguishes the families is part of the query's evidence. The arm that is owed is not another sample; it is a design whose gate survives its own manipulation. The one we would freeze next removes retrieval from the premise: replace index@1 in the gate with the construction's own bound, which is what §5's witness does and what makes its criterion gradeable. The outcome that would count against this paper is stated in §??(b), and it is the direction three runs have not moved in.

(2) The bound on the unbounded-cardinality family, tightened at small m. §6 measures the family Prop. 4 calls for and finds the carrier holding to $m = 32$. What it does not have is a tight subtracted side at the shallow rungs: the construction bounds the k=1 optimum by supplying one missing addend from a uniform prior, which is $1/8$ at $m = 2$ against $1/900$ for §5's family, so the deep rungs carry the witness and the shallow ones carry only the trend. Widening the offset alphabet is one line of the generator and would put every rung on a bound comparable with the witness's; nothing in the reported result depends on it, which is why it is owed rather than blocking.

(3) A pure-recurrent checkpoint that can do the task at all. Cost unknown. §7's verdict is not decidable because every model measured sits near a tenth of its own ceiling, and eight checkpoints have now been screened against a floor fixed before any of them ran (Table 7). What is owed is not more probes and not a better prompt: it is a pure recurrent model that clears 0.60 on two-addend composition when handed both operands. The best of the eight reaches 0.535. This is the one owed item in this paper that we cannot buy, and if the ceiling is a property of the class rather than of these checkpoints, it is itself a finding about state channels, one this apparatus is positioned to make and has not earned yet.

I.1 Limitations

J 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 (§4). The criteria graded in this paper are the E_STATE, E_WITNESS, E_UNBOUND, E_SYMN, E_SSM, E_SSM2 and CoT rows.

K 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 12. 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

Table 13. The containment subset by dimension and split, from analyze_oracle.py. S_{oracle} is the index-only optimum's score, S_{ceil} the access-complete score, and R_{upper} their difference, the measured upper bound on R_{state} for those items.

dimension	split	n	S_{oracle}	S_{ceil}	R_{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.

L 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 14. 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.

M 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 15. 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			

N 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 16. The symbols of this paper.

symbol	what it denotes	defined
k	per-query access budget: how many records an index channel may read. Every verdict in this paper is at $k = 1$	Def. 1
$S(\cdot)$	the score a system attains on a task, under the stated scorer	§3
$R_{\text{state}}(\mathcal{T}; k)$	resolving power: access-complete oracle minus index-only oracle at budget k. Zero means the task separates nothing	Def. 2
$\mathcal{A}^\star$, $\mathcal{B}^\star$	the two optima the definition names. A and B are realised arms standing in for them, which is the substitution that makes R_{state} bounded from neither side	§4
c	cardinality: the number of records an answer depends on. A family needing c records is sighted at $k < c$ and blind at $k \geq c$	§3
m	that same cardinality, read as the axis it is grown along in the robustness arm	§6
a_k, a_{complete}	the share of items whose gold is reachable at budget k, and the share reachable anywhere in the record	(1)
$R^{\text{str}}(k)$	$a_{\text{complete}} - a_k$: the instrument-free form of the curve. Two counts over text and a retriever, with no model and no judge	(1)

Table 16, continued

symbol	what it denotes	defined
$\widetilde{R}$	the judge companion's $\widehat{R}/\hat{\iota}$, which reaches the same estimand through two scored arms. Kept as a separate symbol on purpose: where the two diverge, the factorisation is falsified	§4.1
ι	credit rate, imported: the rate at which the scorer credits an answer produced from a record that contains the gold verbatim	(Anonymous, 2026h)
V	verbatim containment, imported: the share of items whose gold appears verbatim inside a single logged record	(Anonymous, 2026e)

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.

- B, a realised comparator arm here; in Anonymous (2026g) it is a criterion's band; in Anonymous (2026h) it is the share on which the state channel separates nothing.
- m, a cardinality here; in Anonymous (2026c) it is the number of parts a subject is split into; in Anonymous (2026g) it is the number of executions averaged.

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; ι 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.

O 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 17. The scorecard in eight lines, counted from the scorecard. Every graded row of Table 19 and Table 20 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: 40 claim families in Table 19, 76 sub-rows in Table 20. What the unflattering rows are for. 61 of the 115 rows carry a positive grade and 54 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	6	asserted, then taken back or re-assigned; the sequence is in the post-mortems
not supported / failed	10	a frozen criterion was missed, on our own data
self-correction	28	our own arithmetic or our own criterion was wrong and is repaired here
not decidable / not evaluable	2	the apparatus or the band cannot separate the outcomes: not a null
acknowledged limit	8	a boundary of the result, stated where the result is
formal	16	proved from the definitions; no measurement can move it
measured	35	an artifact on disk, and a reader may dispute the artifact

Table 17, continued.

grade	rows	what the grade means in this paper
argued / reported	10	no grade is claimed: an argument, or a number printed without a criterion
total graded rows	115	every numbered row of both tables

Table 19. Claims by evidential status. Grades: formal (proved, no data), measured, not supported / failed (frozen criterion missed), withdrawn, self-correction, not evaluable / not decidable (apparatus or resolution limit, not a null), acknowledged limit, reported (not graded). Rows recording grades we issued and then withdrew are kept: the sequence is the contribution.

Claim	Status	Evidence
70. A two-line recipe for a task with positive resolving power	formal	R1 cardinality (necessary), R2 the answer in no record (what makes the bound closable)
71. The audit's honest reach is three dimensions, not ten	acknowledged limit	six of ten have no containment-checkable item at all
72. The instrument and the audit stand without any companion	code-verifiable	Table 11; the imports are restrictions, not premises
73. Our criticism of beam is about the reading, not the authors' intent	argued	$V = 0.55$ is a fact about the corpus; the compositional reading is supplied by citation
74. What would most embarrass the cardinality reading, named in advance	argued	a decoy-free matched pair where the ordered family scores materially higher
60. Chain-of-thought does not relax the access budget	measured	index@1 0.000 with CoT; gold in 0.000 of traces
61. A KV cache is an index under Def. 1, not a state channel	formal	$O(n)$ growth, and destroyed between episodes
62. Whether $\iota < 1$ is generic or beam's is untested	acknowledged limit	exact match removes only the scorer's share of $1 - \iota$, measured at 0.028; the reading share survives it (Anonymous, 2026h)
63. On an underdetermined task ι and the estimand fail together	argued	flagged as where two of the family's walls may be one
64. A ten-dimension benchmark carries 90 internal transfer pairs, none tested	reported	§??(c); the accounting is (Anonymous, 2026d)
42. A design rule for benchmarks with provably positive resolving power, released	formal + code	answer in no record; 500 probes, --verify re-derives the bound per probe
40. A realisable $O(1)$ carrier exhibits $R_{\text{state}}(\mathcal{T}; 1) > 0$	measured, P-T3-96; below: 1 withdrawn	$\widehat{R} = 0.694$ against a band of 0.151 set before the criterion
41. The order-sensitivity prediction is not settled	not decidable, P-T3-95; below: 1 not evaluable	0.130 preregistered against 0.170 decoy-free, band 0.151
1. Answers that are functions of the write record give zero resolving power	formal	Prop. 1
2. One-sidedness: the audit condemns, never acquits	formal	Prop. 2
3. Soundness of the two index-only readers	measured	40/40, 0/20
4. Verbatim one-turn containment on knowledge_update	measured, registered; below: 1 not supported, 1 withdrawn	0.61/0.64/0.55 over 100K/500K/1M

Table 19, continued.

Claim	Status	Evidence
5. Composition dimensions lack that structure	measured, registered; below: 1 not supported	0.18, 0.06
13. The three active fixes share one voided assumption	formal	Prop. 1; §2
6. The scorer never substitutes the question (F1)	code-verifiable	Anonymous (2026h, §2)
7. The scorer truncates half credit on 9 of 10 (F2)	code-verifiable	Anonymous (2026h, §2); nine evaluators cast through int(), the tenth uses float()
8. F1's magnitude	measured, 100K only; below: 1 not supported	0.222 to 0.003 (BM25-last); 0.280 to 0.000 (last-occurrence)
9. Cross-judge comparison against published ceilings	measured, withdrew our grades	offset +0.194, +0.297
10. A retrieval arm loses to full context on every dimension	measured	10/10; every dimension
11. BEAM is blind on any dimension	withdrawn; below: 1 withdrawn	0/10, four tightenings
12. Blindness is establishable under this protocol	failed; below: 1 not supported	the four failures are the result
15. Every arm here is an index channel, so the headline was within-class	re-derived, against us	all four $O(n)$; vanilla on the subtracted side
16. The needed adversary was on disk, filed as the ceiling	measured, P-T3-86 supported; retrieval sufficiency	$\rho = 0.83$ at k=8, 0.90 at $k \leq 16$; 0.81 elsewhere
75. V and R^{str} on BEAM's fourth and largest rung	measured, no model, no judge, no API; below: 1 not evaluable	V pooled 0.377/0.349/0.325/0.295 over a 74× axis; monotone
100. Sightedness is a property of (task, budget, segmentation) and not of the task	formal	Prop. 3, restated; index@2-oracle 0.990 against index@1 0.000 measures it
109. A 2026 long-context method supplies a live instance of what Prop. 3 says a needle task cannot resolve	argued	recurrence and internal-state language over a cache its own abstract describes as enlarged, evidenced on needle retrieval (Nadali et al., 2026)
101. A cardinality that grows with the stream is sighted at every fixed budget and chunk width	formal	Prop. 4: k chunks hold $\leq kw$ records
102. R_{state} is bounded from neither side by realised scores, proved in this paper	formal	Prop. 5, four lines; nothing load-bearing is imported
103. The second instrument's candidate constant is a real free parameter	measured	selection identical across the sweep on 3.4–11.6% of items; containment moves up to 0.101
104. P-T3-33 is near-certain by construction and is not counted as evidence	self-correction	both index arms are 0 by construction, so the Wilson bound clears whenever the ceiling is not near zero
105. $\iota \geq S(\text{complete})$ is a consistency requirement of $S = a\iota$, not a finding	formal	$a \leq 1 \Rightarrow S(\text{complete}) \leq \iota$ identically; measured $0.917 \geq 0.867$
106. P-T3-30 should not have been registered	self-correction	registered and declared weak in one document; no verdict rests on it

Table 19, continued.

Claim	Status	Evidence
107. "The field misreads knowledge_update" is not a sentence this paper is entitled to	withdrawn; below: 1 withdrawn	we searched for external compositional readings and found none; the count stands in its place
108. Prop. 4 of Anonymous (2026h)'s $\Theta(1/\iota)$ rate holds at fixed a_c	acknowledged limit	a rubric can move a_c and ι together; the joint case is not worked out

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 76 sub-rows are in Table 20 in full, of which 25 self-correction, 6 not supported, 5 acknowledged limit, 3 failed, 2 withdrawn. Nothing is dropped, and nothing is summarised into a kinder word than it had.

Table 20. Supporting refinements, moved from Table 19 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 76 rows below were moved here from Table 19 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.		
70a. V is exactly the rate at which a benchmark violates R2	measured	0.55 on the flagship dimension against 0.00–0.07 elsewhere
71a. The admissibility bias is conservative for our failures to condemn	argued	it selects the subset a retriever finds easy
71b. Instrument portability is three corpora; audit portability is three dimensions	measured	and we had been letting one borrow the other's reach
60a. CoT costs the carrier 0.165	measured	$0.695 \to 0.530$; 0.635 even scoring the whole trace
60b. The carrier composes at write time, which is why reasoning adds nothing	argued	the scratchpad already holds $915 = 766 + 149$
61a. Fixed-width recurrent architectures would qualify and we did not run one	acknowledged limit	the natural extension of §5
42a. The rule's second clause is the one our own first family omitted	self-correction	a verbatim-present answer makes the bound a claim about retrievers
40a. The subtracted term is bounded by the construction, not by a retriever	formal	the answer occurs in no record; 0 leaks of 200
40b. Our first bound on the same quantity was wrong by three orders of magnitude	withdrawn	asserted 1/7200; three k=1 strategies reach 0.359, 0.485, 1.000
40c. A best-known strategy was twice read as an optimum, which is what makes a bound look proved	self-correction	vanilla, then 1/7200; neither was a minimum over strategies
40d. The claim is existential and does not isolate the channel axis	acknowledged limit	an oracle index at k=2 scores 0.990 and beats the carrier
40e. The figure of this arm printed four of its five bars under the wrong arm's name	self-correction	pgfplots splits yticklabels on commas and four labels contain one, so the list of five became a list of eight and the first five were used: the 0.990 bar was labelled our retriever, and S(state) = 0.695, this paper's one positive result, was labelled S(index@2)

Table 20, continued.

Claim	Status	Evidence
40e†. The carrier pays a third of the ceiling for being bounded	measured	$\kappa = 0.702$ at a 240-character cap
40f. …and the carrier's cap was drawn at 900 characters, which is not a cap	self-correction	CAP is 240, frozen in the prereg; 900 is the gold vocabulary and the denominator of (2). The figure then concluded "the channel ran well inside its own cap": against the real cap the maximum is 240 exactly and 209 of 3643 updates were truncated. The cap binds, which is what the prose always said
40g. …and this arm's artifact was one convention behind the family	self-correction	analyze_state.py has been unrunnable since the migration (criterion_resolution moved to audits/) and was in no gate, so the Z correction of 2026-08-18 never reached it. Recomputed: Δ halves and the margin goes $2.15 \to 4.30\,\Delta$. No verdict moves, and the direction is away from the resolution limit
41a. 35 of 200 matched probes carry a decoy value record	self-correction	inherited from the background rule we were matching
41b. The soundness gate passed vacuously because it looked for the wrong records	self-correction	bearing records located by gold in e; now family-independent and hard-failing
2a. A task with $R_{\text{state}}(\mathcal{T}; k{=}1) > 0$	measured	0.867 vs 0; Wilson $[0.758, 0.931]$; P-T3-32/33
2a†′. That witness is blind at k=8	measured, against us	$\widehat{R} = -0.033$; k=8 is 57% of the span
2a†††″. It shows addressability, not composition	re-attributed	$S \approx \text{coverage} \cdot \kappa$, $\kappa \approx 0.9$
2a†††‴. …and the figure drawn for that sweep plotted a coverage curve that falls	self-correction	the oracle arm's k is the string "2", so an isdigit() filter drew it inside the retrieval curve: coverage printed as 0.000/0.383/1.000/0.867/1.000, and the saturation label read k=8 in the caption and k=2 in the figure. Coverage cannot fall when the reader is shown more records
2b. ι's estimation inflates the standard error	measured	1.376, not 1.091; effective n 31.7 of 60; P-T3-34
2f. Structural and scored deconfounding agree	measured, within 1.4 se	$\widetilde{R} = 0.945$ vs $R^{\text{str}}(1) = 1.000$; residual 0.055
2g. R^{str} runs on a benchmark we did not write	not supported, P-T3-41	pooled $R^{\text{str}}(1) = 0.246$ against 0.30
2h. The largest gap is on knowledge_update	not supported; our rationale was backwards	all three splits; P-T3-44
2i. The instrument reaches 3 of 10 dimensions	acknowledged limit, floor as frozen	six have no checkable item on any split
2j. The context-length null reproduces with no instrument	measured, not registered	V = 0.377/0.349/0.325; $R^{\text{str}}(1)$ = 0.246/0.226/0.188
2k. A fixed budget reaches less as the transcript grows	measured, P-T3-43 supported	$R^{\text{str}}(32) = 0.000/0.021/0.017$

Table 20, continued.

Claim	Status	Evidence
2m. First outside use of the definition convicts a sibling	reported, not ours	20/20 and 0/30; Anonymous (2026d)
2n. That conviction produced a repair, not a concession	reported, not ours	spread to 0.133; Anonymous (2026d)
4a. the containment chain $V \le A \le B$'s first inequality $V \le A$	failed, two frozen bars	0.091/0.310/0.257 and 0.614/0.452/0.629 vs 0.80
4b. …so the containment chain $V \le A \le B$ is unsupported	measured	$V - A = 0.24/0.35/0.20$
4b$^{\dagger\prime}$. …and was never true: $V \le A \iff \iota \ge 1$	formal	the gap is the identity $V(1-\iota)$; Prop. 1 of Anonymous (2026h)
4c. The replacement bound is looser and not instrument-free	measured, post hoc	1.45–1.97×; the licensed repair
4d. We spent an instrument-free count on an instrumented inference	self-correction	Anonymous (2026h, §5)
4e$^{\prime}$. ι is removable, not merely reportable	formal	Prop. 3 of Anonymous (2026h), the deconfounding identity
4e$^{\dagger\prime\prime}$. Its price: identified but unaffordable	formal; narrows our own claim	$\Theta(1/\iota)$; 99 items behave like 54.5; Prop. 4 of Anonymous (2026h)
4e$^{\dagger\dagger\prime\prime\prime}$. We published the $n\iota^2$ column alone	self-correction	30.7 items, a 3.2× loss, while proving $n\iota$
4e†††. A raw score difference is confounded by $1-\iota$	formal	$\iota = 0.614/0.452/0.629$
4f. One perfect record beats our ceiling on 2 of 3 splits	measured, not registered	-0.250, -0.200
4g. $1-\iota$ contaminates item-response difficulty	not supported, withdrawn	partial $\rho = +0.126$, $p = 0.18$, bar 0.30
4h. Adaptive selection prefers scorer-induced items	not supported, reversed	high-discrimination ι higher by 0.230
5a. Tenfold context does not move either across its bar	measured, three scales	0.61/0.64/0.55 vs 0.5; 0.19/0.18/0.06 vs 0.25
5b. Scores fall with length at unchanged resolving power	failed, three criteria	ρ 0.235/0.373/0.213; E 1.70/1.57/1.57; h 0.136/0.262/0.200
5c. Nothing moves on that axis at all	measured, post hoc	7/10 shift <0.04; 2100 judged items
13a. The framing is carried in prose and in no figure of ours	self-correction	a schematic drawn for it carried no data and is not printed; §2
8a. F2's magnitude is stable across length	failed	4/10 against a bar of 8/10; 2800 judged items
8b. The defect manufactures event_ordering's length effect	measured, post hoc	corrected ≤0.013 at every scale; official ranges 0.204
8c. …and both figures row 8 carried were wrong	self-correction	0.277 matched no arm and 0.925 is a judge change on another dimension
8d. …and the count in row 7 was published as eight	self-correction	nine functions contain int(response['score']); "eight, and one uses float" is nine
15c. What survives is the half that never needed a ceiling	relocated	V, $R^{\mathrm{str}}(k)$, k^*; V is ceiling-free
16a. A no-model dump beats a model-bearing adversary	measured, P-T3-87 not supported	five of ten dimensions, -0.40 to -0.10

Table 20, continued.

Claim	Status	Evidence
16b. The adversary side carries a floor unrelated to information	measured; a third scorer defect	17.5% passed where $V \leq 0.03$; 18.9% on a negation
16c. The control for that floor	acknowledged limit, free to fix	vanilla/ceiling shards discard judge reasons
75a. …and the fall across that axis is not decidable at the sizes shipped	not decidable	drop 0.081 against 0.090 se, 0.90 se; the rung ships 200 items, 44 checkable
75b. …and this paper named the smallest rung by its file rather than its paper	self-correction	files say 100K, the paper says 128K, same 20 conversations
75c. …and the 10M release carries a different record schema, so the instrument did not run	measured, code-verifiable	nested plan-*/turns arrays; walker now asserted equivalent on the three shipped splits
75d. …and "tenfold" understated the axis by an order of magnitude	self-correction	median turns 271/1076/2088/20075; turn length flat
100a. …and the earlier statement of it left the segmentation out	self-correction	every constructed bound, (2) included, holds at the shipped segmentation
109a. …and the classification needs no argument of ours, only the companion's probe	code-verifiable	walk the cache at two lengths and count elements (Anonymous, 2026e, row 317)
101a. …but positivity still needs a realisable carrier, and a fixed-width one must eventually fail	acknowledged limit	the necessary condition is met at every (k, w); the leading side is not shown
102a. …and clause (ii) is what licenses this paper's one two-term bound	formal	(2) is constructive, so (3) is valid; (??) fails on the ceiling term
102b. …and the figure drawn for it showed only one of the two cases	self-correction	one realised pair cannot exhibit a two-sided failure, and the pair first drawn put the measured gap above R, the one reading the proposition exists to forbid relying on; Fig. 6 now draws two pairs against the same optima
103a. …and the constant we froze is within 0.008 of the best available	measured	analyze_cand.py; the failure to condemn is not our retriever's
107a. …and the authors of the code whose defects we report were contacted before submission, repaired one of the two and dispute the other	reported	disclosed to all six authors 2026-08-24, naming the commit, the files and the lines; answered 2026-08-25 with the offer to withhold declined, and repaired 2026-08-31 at b2da22e, which fixes F2 and leaves F1 in place
14a. Structural instrument check on a synthetic needle	measured, P-T3-54 supported	$R^{\mathrm{str}}(1) = 0.000$, $k^* = 1$
14b. A 32-event budget closes the gap	not supported, P-T3-52	$k^* =$ None on both benchmarks
14c. P-T3-51's bar on the second slot	not supported, missed and reported as missed	pooled 0.295 against 0.30
14e. R^{str} needs exactly-once; claim 4 measured at-least-once	self-correction	exactly-once 4.3/6.1/5.0%; no dimension clears the floor

Table 20, continued.

Claim	Status	Evidence
111b. The figure of the two cardinality families ringed the rung its own verdict is read from	self-correction	gate_P-T3-112 is a per-rung field computed for both families by one loop, but it is the chain family's criterion. Amendment 1 gates the incremental family at m=2 and grades it on the margin at m=32, so ringing m=16 and m=32 there marked as unread the rung the 0.750 margin is measured at
114a. The screen's figure said every candidate and drew six of eight	self-correction	the row builder skipped anything whose ceiling is None, which is exactly the two checkpoints that do not load, v6-Finch-7B-HF and v6-Finch-14B-HF. SSM2.json's own verdict reads "3 of 4 could be measured". A checkpoint that could not be run is part of the result
114b. …and it plotted the one row the paper spends a paragraph declining to read	self-correction	rwkv7-7.2B-g0a returns 0.000 where the 2.9B of its family returns 0.530 under a library that warns about its own rwkv-7 path. The prose calls this a harness result and refuses it as a capability ordering; the figure put both on one axis untagged
115a. The figure of the four arms named the wrong cause for its own verdict	self-correction	analyze_ssm.py returns not decidable from best_carrier $\geq$ 0.5× ceiling, and the screen's floor appears nowhere in that computation. Figure and caption both attributed it to the floor; the gate that decides was two 5pt dashes with a throwaway label
115b. …and rested on the floor number this paper says does not follow from its reason	self-correction	0.600 as written against 0.286 where its justification binds; the figure drew 0.600 alone and concluded from it. Both are drawn now, and the conclusion no longer depends on either
115c. …and the floor it drew was not a line	self-correction	the axis is ybar, so the addplot written as a dashed rule at $y = 0.6$ rendered as two dashed bars standing from 0; the caption pointed at "the dashed line" and there was none on the page
115d. …and the same cap defect sat at a second site, latent	self-correction	fig:widthsat took the notebook cap to be max(observed length) over the first family's shards. unbound_arm.py imports CAP from state_arm.py, so the cap is 240 there too and both bind; the figure printed the right number for the wrong reason, and would have printed the widest notebook the first time one ran short of its ceiling