The Infinite-Record Limit: A Retrieval Budget That Stays Constant for the Hardest Items and Vanishes for the Median

Anonymous authors
Paper under double-blind review

Abstract

Long-context and memory benchmarks routinely claim that a task requires reaching deep into a record, but that claim is about a retrieval budget, and whether a budget counts as small depends on how long the record is. No evaluation measures how the required budget scales as the record grows, so a claim established at one length is silently carried to another; and because no fixed-corpus lever moves record length alone, even a measured scaling may belong to the lever rather than to the task. That confound is not hypothetical: it produced the field's most expensive scaling disagreement (Kaplan et al., 2020; Hoffmann et al., 2022; Porian et al., 2024). We make the scaling the estimand. Pinning the retriever, the scorer and the task and sending record length to infinity exposes a boundary function $k^*(L)$ – the budget at which the unreachable share of items falls below threshold – and its exponent $\alpha \in [0, 1]$: $\alpha < 1$ makes "a small budget" meaningful, $\alpha = 1$ makes it empty. On deletion-built nested records the pooled exponent is 1.011 over a 62× ladder, with the required share stuck at 0.400–0.417, refuting the sublinear scaling we had frozen in advance; the exponent differs at every quantile, and that distribution is the finding – the hardest few per cent of items demand a constant share of the record while the median item's demand vanishes. The shape transports to a second corpus as unlike the first as our data allows, but the constant does not, and an undeleted 10× lever reproduces neither the exponents nor their ordering, with neither version wrong: an exponent is identified only as (task, lever, retriever), and telling the two levers apart is not decidable at this benchmark's item counts, needing 12–54× the items. The reporting rule with a proof behind it is to publish an exponent with the lever and the retriever that produced it; the one unconditional result is that the required share does not vanish as the record grows.

1 Introduction

Every claim of the form "this benchmark tests multi-hop memory" is a claim about a resource: that answering requires reaching more of the record than a small budget reaches. Pin the other resources and send that one to infinity. Fix the task, the retriever, the scoring rule and the threshold; let the record length L grow without bound. Two things can happen to the budget k^* a retriever needs before the answer is reachable on all but a τ fraction of items. It can grow slower than L, in which case the share k^*/L vanishes and "a small budget" describes what the task permits. Or it can grow as fast as L, in which case the share is constant, no budget is small, and every result obtained at one length is a result about that length.

The limit, and what appears at the boundary. Pin the retriever, the scorer, the threshold and the task. Send $L \to \infty$. What appears is not a number but the function $k^*(L)$ (Def. 2), whose exponent α (Def. 3) is confined to $[0, 1]$ by Prop. 2: $\alpha < 1$ means the

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.

required share vanishes and $\alpha = 1$ means it does not. α is computed from two counts over text and a retriever, with no model, no judge and no API call.

What this paper contributes.

1. A boundary function, read as a distribution rather than as a number. The required share $\varphi^*(L) = k^*(L)/L$ needs no fit: on a nested deletion ladder over 62× in L it stays at 0.400–0.417 at the frozen threshold $\tau = 0.05$, and a second corpus as unlike the first as our data allows returns 0.30 (§E, §L). τ is a quantile and the answer depends on which one: at $\tau = 0.20$ the share is 0.051 and at $\tau = 0.50$ it is 0.013, falling with L rather than holding (§F). The budget the hardest few per cent of items demand is a constant share of the record; the budget the median item demands vanishes. Both halves contradict predictions we froze.
2. An identification result, and the condition it turns on. On a fixed corpus no lever moves record length alone: deletion takes copies of the answer with it, drawing a second corpus takes the distractors, appending our own text stops measuring the benchmark. α(task) is therefore identified only if some lever holds the confound still, $g = \mathrm{d}\log occ/\mathrm{d}\log L = 0$ on a subset that clears the power floor, and on this corpus no lever does, on four of five dimensions (§J). The practical form is a reporting rule with a proof behind it: publish the exponent with the lever and the retriever that produced it. §M audits three published exponents against that rule, including a 0.38 gap between two compute-optimal laws that took the field four years to attribute (Porian et al., 2024).

One reading the data does not license. The per-dimension exponents order the benchmark's own single-hop and composition classes as point estimates, and that ordering is not a result of this paper: the licence it needs fails on four of five dimensions, the ordering reverses under a second lever, detrending the multiplicity confound shrinks the class gap from 0.622 to 0.195, and the bootstrap interval on the one surviving cell contains both composition estimates. §K reports it as a property of this corpus that this instrument cannot measure.

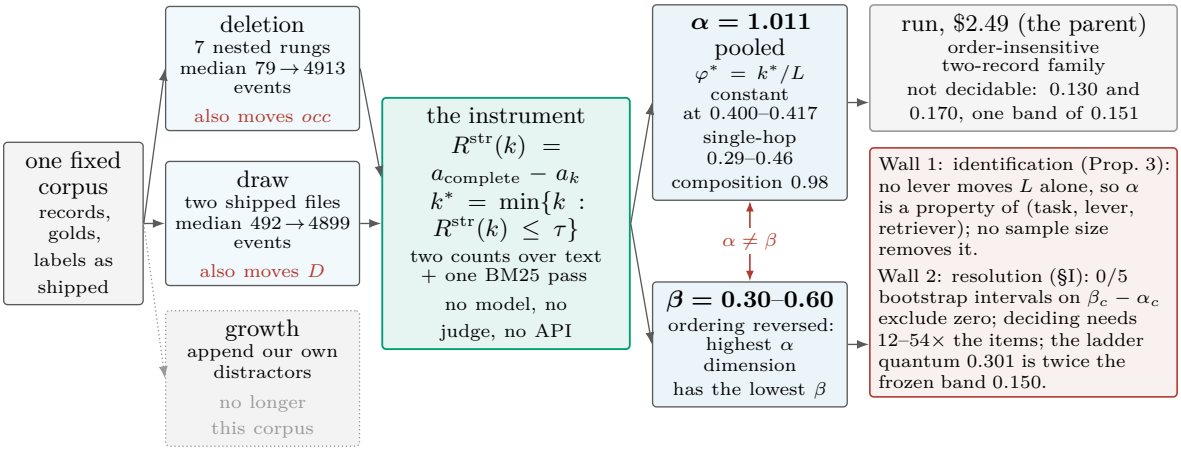


Figure 1. The paper in one diagram. One corpus, three levers that move its record length, one instrument, two exponents that disagree. The instrument (centre) is two counts over text plus a retriever: no model, no judge, no API call. The disagreement between α and β is not resolvable by more reruns, because the levers differ in what they move besides L (Prop. 3), and it is not resolvable by more items at this scale, because the comparison sits below its own resolution (§I).

The wall, read off the boundary rather than off a convention. The same resource admits more than one lever, and we built a second: the benchmark ships the same 500 questions at two lengths, so length moves without deleting anything. The exponent over that lever, β, does not reproduce α and reverses the ordering of the dimensions. The temptation is to ask

which lever is right. Neither is, and the reason is structural: no lever on a fixed benchmark moves record length alone (Prop. 3). A deletion lever removes events, and removing events removes copies of the gold string; a draw lever takes two independently sampled records at two sizes, so it moves length together with the identity of the distractors, the two files here nested for 0 of 500 questions; a growth lever appends distractors we wrote, at which point the corpus is no longer the benchmark. Hence α is a property of (task, lever) and an ordering of tasks by α need not transport across levers, which makes the rest of this paper's negative results corollaries rather than disappointments.

The second wall is arithmetic, and it was computable before the data existed. Asking which lever is right requires deciding whether $\beta - \alpha$ differs from zero. It does not, at these item counts, on any dimension: five of five bootstrap intervals span zero, and reaching a 0.150 band at $1/\sqrt{n}$ would take 12 to 54× the items. Worse, k^* crosses its threshold on a ladder of powers of two, so a two-point exponent over a one-decade lever moves in steps of $\log 2/\log 10 = 0.301$, twice the band we froze. Three of this paper's own frozen criteria were below the resolution of the statistic they graded, which is a fact about the criteria and not about the world; the companion paper that makes that computation its subject (Anonymous, 2026j) audits them, and we adopt its verdicts here rather than reporting our own undecidable criteria as nulls.

Why this is five papers and not one. Each of five companions pins a different resource and sends a different payload to infinity: tenure against an append-only store (Anonymous, 2026e); instance count against a fixed testing budget (Anonymous, 2026g); item count against a fixed resolving power (Anonymous, 2026i); this paper, record length against a fixed retrieval budget; and rerun count against a pinned criterion (Anonymous, 2026j). No result below is derivable from any other, and this paper uses the others' measurements, never their conclusions; the one exception, the scorer credit rate ι of §4, is imported as a measured constant (Anonymous, 2026i) and marked at the point of use.

2 Related work

Scaling laws fit exponents to resource ladders and name the resource but not the lever. The form is standard: hold an architecture and a recipe, move one resource over decades, fit a power law (Kaplan et al., 2020; Hoffmann et al., 2022; Ghorbani et al., 2021; Clark et al., 2022). We borrow its shape and not its implicit premise, that the resource named in the exponent is the thing the lever moved: Kaplan et al. (2020) and Hoffmann et al. (2022) report $N^* \propto C^{0.88}$ (and 0.73 under Kaplan et al.'s own adjusted accounting) and $N^* \propto C^{0.50}$ for the same nominal resource, decomposing into three quantities their levers moved besides compute (Porian et al., 2024) (§M); a difference of two scaling trajectories is likewise a harder object than either alone (Held et al., 2025), the same point our $\beta - \alpha$ makes one level up. Partial identification and transportability supply the vocabulary for what Prop. 3 is: a set of values indexed by the lever (Manski, 1995; 2003), in the same sense that a relation need not hold across populations (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2014), though we prove ours by enumerating a fixed corpus's levers rather than by a selection diagram.

Long-record evaluation, the containment objection, and the systems this accounting is about. Three lines of work move the record-length axis deliberately, each a lever in our sense: a growth lever varying irrelevant content around a fixed needle–question pair (Hong et al., 2025); a draw-lever dual isolating cardinality from length (Levy et al., 2025), direct evidence for our separability claim; and positional evidence that the binding long-context constraint need not be length at all (Liu et al., 2024; Wang & Sun, 2025). Our instrument counts verbatim gold occurrence, which a minimal-lexical-overlap benchmark can defeat entirely, dropping the strongest model from 99.3% to 69.7% once the literal route is removed (Modarressi et al., 2025); we accept that our exponent is measured on the lexically reachable subset and price the experiment that would settle it (§6). Deployed memory architectures (Chhikara et al., 2025; Xu et al., 2025) are ranked at the record length their benchmark ships,

and our result says such a ranking does not carry to another length unless the exponent is the same for both.

Retrieval budget, item-efficient benchmarking, construct validity, and the family. The applied literature finds an interior optimum in top-k (Shi et al., 2023; Lewis et al., 2020); our budget sweep finds the opposite on the family we built (Appendix T), a different regime. Item-efficient benchmarking (Polo et al., 2024; Hofmann et al., 2025; Zhou et al., 2026) presumes the target quantity is identified at the item count available; ours, a difference of two exponents, is not, needing 12–54× the items this benchmark supplies. Construct-validity reviews find benchmarks routinely measure something other than the construct they name (Bean et al., 2025; Qian et al., 2026); we take one construct and ask what a resource ladder can and cannot establish about it. The access-structure account of capability (Anonymous, 2026a) is a principle rather than a source of coefficients; companions ask whether a task can resolve the state/index distinction (Anonymous, 2026i), audit deployed benchmarks with the counts this paper sweeps (Anonymous, 2026h), own the scorer defects (Anonymous, 2026k), and audit a frozen criterion's resolution (Anonymous, 2026j). Extended discussion is in Appendix A.

3 The limit and its boundary function

Definition 1 (Record, item, budget, retriever, admissibility). A record is a finite ordered list of logged events $E = (e_1, \ldots, e_L)$; $L = |E|$ is the record length. An item is a triple (q, y, E): a question, a gold value, and the record it is asked against. A retriever ρ with budget k returns $\rho_k(q, E) \subseteq E$ with $|\rho_k(q, E)| \le k$; throughout, ρ is one fixed lexical scorer (BM25 over events, Robertson & Zaragoza, 2009) and k ranges over the ladder $\{1, 2, 4, \ldots, 2048\}$. An item is containment-admissible if y occurs verbatim in at least one event of E; its multiplicity is $occ = |\{e \in E : y \sqsubseteq e\}|$. Admissibility is what makes the instrument free (whether a string occurs in a text is decidable without a model) and is also its premise: *occ* counts copies, so a gold string that is a common phrase inflates it, and §J is the attempt to control that inflation.

Definition 2 (The structural miss curve, the crossing budget, and the required share). Let a_k be the share of admissible items whose gold occurs verbatim inside $\rho_k(q, E)$, and $a_{\text{complete}} = 1$ on the admissible subset by construction. Write $R^{\text{str}}(k) = a_{\text{complete}} - a_k$, the share of items whose answer is present and not reachable at budget k; both terms are counts over text, no model or judge is evaluated. R^{str} is monotone non-increasing in k. For a threshold τ (fixed at 0.05 throughout),

$$k^*(\tau; L) \;=\; \min\{\, k : R^{\text{str}}(k) \le \tau \,\}, \qquad \varphi^*(L) \;=\; \frac{k^*(\tau; L)}{L}. \tag{1}$$

Proposition 1 (Monotonicity of the crossing budget). Under a distractor-monotone retriever (adding a non-bearing event cannot increase the probability ρ_k returns a gold-bearing event), $R^{\text{str}}(k; L)$ is non-decreasing in L for fixed k, hence $k^*(\tau; L)$ is non-decreasing in L. Proof and the empirical test of the assumption (P-T3-59, holding on seven nested rungs) are in Appendix Q.

Definition 3 (The exponent). Where k^* is finite on at least two rungs clearing the saturation guard $k^* \le L/2$ (§4), α is the OLS slope of $\log k^*$ on $\log L$, equivalently $k^* \propto L^\alpha$.

Proposition 2 (The dichotomy at the boundary). Suppose $k^*(L) = c\,L^\alpha(1 + o(1))$ with $c > 0$. Then $\varphi^*(L) = c\,L^{\alpha-1}(1 + o(1))$, so $\lim_{L\to\infty} \varphi^*(L) = 0$ if $\alpha < 1$, $= c$ if $\alpha = 1$, and the case $\alpha > 1$ is unattainable since $k^* \le L$ by construction forces $\varphi^* \le 1$ and hence $\alpha \le 1$. Consequently $\alpha < 1$ is exactly the condition under which "a small retrieval budget" is a meaningful description of the task in the limit, and $\alpha = 1$ is exactly the condition under which it is empty, with c naming the share that must always be paid.

Proof. Immediate from $\varphi^* = k^*/L$. The bound $\alpha \le 1$ follows because $k^* > L$ is impossible: a budget exceeding the record retrieves the whole record, where $R^{\text{str}} = 0 \le \tau$, so the minimum in (1) is attained at or below L. □

Definition 4 (Lever). A lever is a family of maps Λ_L carrying each item's record E to a record of length L, defined on the corpus as shipped. A lever is clean for an item if it changes L while leaving both the multiplicity *occ* and the identity of the non-bearing events (the distractor set D) unchanged.

Proposition 3 (Lever non-separability). On a fixed corpus, no lever is clean at more than one length. Every realisable lever is of one of three kinds, and each moves at least one of *occ*, D, or the corpus itself: (a) Deletion retains a subset of E; omitting a bearing event lowers *occ*, and preserving *occ* on the retained items changes the admissible set for the rest. (b) Draw selects an independently sampled record of each length for the same question, so D at one length is not nested in D at another. (c) Growth appends events not in the corpus, so D contains material the benchmark did not ship and the measured object is no longer that benchmark. Hence α is identified only as a functional of (task, lever, retriever): two levers may return $\alpha \neq \beta$ for the same task without either being incorrect, and the ordering of tasks by α need not agree with the ordering by β.

Proof. The three kinds are exhaustive for maps defined on a fixed corpus: $\Lambda_L(E)$ is either a subset of E (a), drawn from elsewhere in the corpus for the same question (b), or contains an event outside the corpus (c); the stated consequence follows from each definition. α is estimated from $\{(L_j, k^*(L_j))\}$, determined by the lever as well as the task by (a)–(c), so two levers define two estimands and no data from one bounds the other. □

What Prop. 3 does and does not need empirically. Its first sentence follows from Def. 4 alone and carries no empirical content; the enumeration (a)–(c) is the content. What the paper needs is checkable: writing $g = \mathrm{d}\log occ/\mathrm{d}\log L$, α(task) is identified, independently of the lever, iff the corpus contains a subset on which some realisable lever achieves $g = 0$ and that subset clears the power floor the statistic needs. §J's exactly-once subset is the $g = 0$ set by construction, and it clears the floor on one of five dimensions. The retriever is in the triple and not a detail: an oracle retriever has $\alpha = 0$ and a retriever that ranks at random has $\alpha = 1$, so $\alpha = 1.011$ says that a lexical retriever must be allowed a constant share of the record, and how a better retriever moves that number is open. This is also distinct from index latency: a partitioned index that returns the right k records faster does not reduce k, and conflating the two is the easiest misreading of this paper.

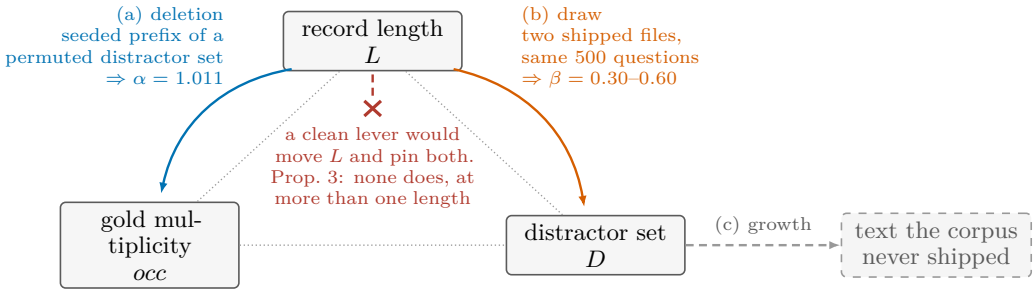


Figure 2. Prop. 3 as a picture. A record carries three resources; a lever that identified an exponent as a property of the task would have to move L and pin the other two at more than one length (the blocked centre arrow). Every realisable construction moves at least one of the other two or leaves the corpus, so an exponent fitted to $k^*(L)$ is a functional of (task, lever, retriever): two levers may disagree for one task with neither wrong — and here they reverse the ordering of the benchmark's own dimensions. The two exponents measured are labelled on the levers that produced them.

4 Experiments

Seven arms, in the order run, each preregistered before its runner existed; full criteria, verdicts and every correction found while running each arm are in Appendix B. One arm (§C,

Appendix T) scores composition probes with a model to test whether a score on the budget axis factors into coverage times a constant, imported from the parent paper (Anonymous, 2026i); every other quantity below is two counts over text and a BM25 pass, no model, no judge, no API call.

Table 1. The three levers, and the exponent each returns, on the same five dimensions, the same retriever, the same τ and ladder.

lever	construction	also moves	result
deletion (§E)	seeded prefix of a permuted distractor set, 7 nested rungs, 79 → 4913 events	gold multiplicity *occ*	$\alpha = 1.011$ pooled; 0.287–0.462 single-hop, 0.981/0.984 composition
draw (§H)	the two shipped files, same 500 questions, 492 → 4899 median events	distractor identity D (nested for 0/500)	$\beta = 0.301$–0.604, ordering reversed against α
skill-free control (§I)	the same rungs with a uniform-draw retriever substituted	nothing: it is a floor, not a lever	not evaluable on 3/5 dimensions; the saturation guard bites everywhere

On longmemeval (Wu et al., 2025), the deletion ladder pins the required share at $\approx$ 40%, refuting our own sublinear prediction. Built by deleting sessions from the longer split, seven nested rungs at median 79 → 4913 events, criteria frozen in E_SWEEP_PREREG.md. k^* runs $64, 64, 128, 256, 512, 1024, 2048$, giving pooled $\alpha = 1.011$ and $\varphi^* = 0.400, 0.403, 0.401, 0.405, 0.409, 0.417$, essentially flat over a 62× range (Figure 3). An unnested, independently-drawn two-file comparison of the same benchmark gives an apparent $\alpha \approx 0.3$ instead; the nested measurement is the one that survives, and 92% of the 0.712 gap between them is which dimensions cleared the floor in each arm rather than the nesting itself (§D–§E, Appendix B).

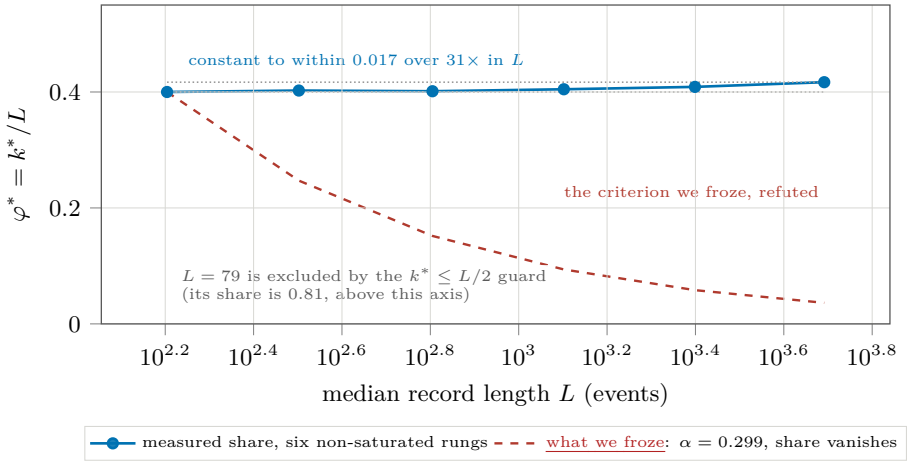


Figure 3. The headline on the axis where it is readable. $\varphi^* = k^*/L$, the share of the record the retriever must be allowed to see, held to within 0.017 over a 31× range in L (solid), against the vanishing share we froze a criterion predicting (dashed). Recomputed from results/sweep/SWEEP.json.

k^* is a tail statistic, and the median item's budget vanishes even as the tail's does not. Reading the same stored curves at other quantiles (Table 2) shows the flat share is a property of $\tau = 0.05$ specifically: at $\tau = 0.20$ the exponent is +0.109 and the share falls from 0.051 to 0.002; at $\tau = 0.50$ it is +0.000 and 22 of 35 items never move across the whole ladder. Both readings are true of the same relation at two quantiles.

Registered and powered, the exponent separates the benchmark's own single-hop and composition classes. On the full admissible set ($n = 47$–118 per dimension), criteria frozen in

Table 2. $\varphi^* = k^*/L$ at each quantile, over the same seven non-saturated rungs and pooled $R(k)$ curves the published exponent uses.

quantile	rungs	α	φ^* at $L = 79$	φ^* at $L = 4913$	reading
$\tau = 0.05$	7	+0.898	0.8101	0.4169	constant
$\tau = 0.1$	7	+0.719	0.1013	0.0261	vanishing
$\tau = 0.2$	7	+0.109	0.0506	0.0016	vanishing
$\tau = 0.5$	7	+0.000	0.0127	0.0002	vanishing

E_ALPHA_PREREG.md: the smallest composition exponent (0.981) exceeds the largest single-hop one (0.462), and the between-class gap (0.519) is three times the within-class spread (0.175). The point estimates separate cleanly; the bootstrap intervals do not (Figure 4), which is the first hint of what §M and §K then show is a confound rather than a clean signal.

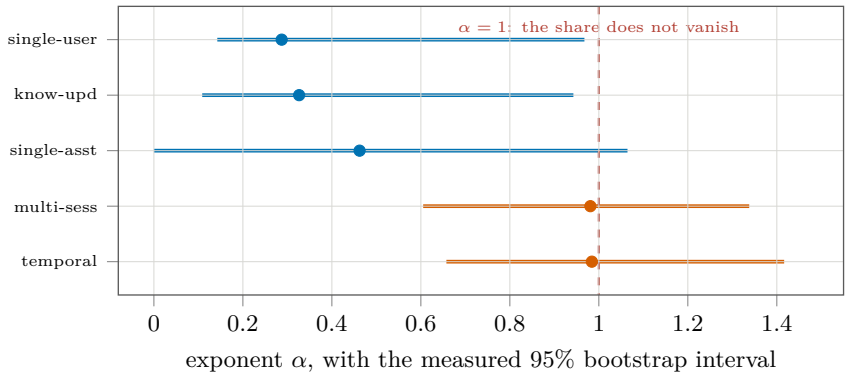


Figure 4. The point estimates and their measured intervals, printed together because printing only the first is how a reader is misled. The five per-dimension exponents separate the benchmark's own classes cleanly as point estimates — by 0.519 against a within-class spread of 0.175; the 95% bootstrap intervals, computed from the same runs, overlap so heavily that every interval overlaps every other and no pair is distinguishable, which is why §I calls the separation undecidable at this item count. The dashed line is $\alpha = 1$, where Prop. 2 says the required share stops vanishing.

A third, undeleted lever reverses the ordering. The same 500 questions shipped at two lengths (median 492 → 4899 events) move length without deleting anything, so an exponent β over that lever must reproduce α if α is a property of length alone. It does not: temporal-reasoning has the highest α (0.984) and the lowest β (0.301), while knowledge-update has one of the lowest α (0.326) and one of the highest β (0.601; Figure 5). Half of the failure is a criterion below its own resolution (β's quantum is 0.301 against a 0.150 band); the other half is not: the licence for α (gold multiplicity constant in L) fails on four of five dimensions, so every α above is untested by its own rule, and only single-session-assistant holds, by construction.

The class separation is mostly the multiplicity confound, measured directly. The deletion lever moves record length together with *occ* (Prop. 3(a)). On the four dimensions with a measurable growth rate g, the correlation between α and g is $r = +0.765$, and detrending it shrinks the class gap from 0.622 to 0.195. With $n = 4$ and a single surviving composition row this cannot decide the question in either direction, but it establishes that the confound is real, is in the direction that would manufacture our result, and accounts for most of it. Three independent repairs (subtracting a skill-free retriever, controlling for exactly-once recurrence, and the licence test above) each survive on a disjoint subset of the five dimensions and none can compute the between-class quantity; the obstruction is a property of this benchmark's redundancy profile and not of any one estimator (Appendix B).

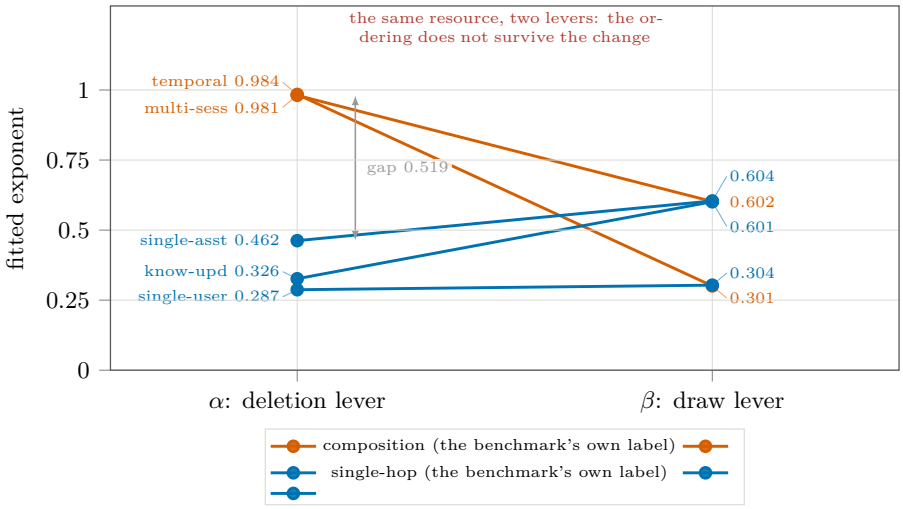


Figure 5. The ordering reversal. Each line is one dimension under two levers that move the same resource. Under deletion the three single-hop dimensions sit below the two composition dimensions (the grey arrow marks the between-class gap, 0.519); under the draw lever the lines cross. Neither measurement is wrong; by Prop. 3 they are measurements of different functionals.

Table 3. The paper's four questions, and which of them the measurements answer.

question	answer	why
Does the required budget vanish as a share of the record?	No, on the deletion lever	$\alpha = 1.011$, φ^* pinned at 0.400–0.417 over 62×
Does α sort the benchmark's own dimensions by class?	Yes on this lever, registered and powered	0.519 between-class against 0.175 within
Is that ordering a property of the task?	No, of (task, lever)	β reverses it; no lever separates L from *occ* and D
Can the two levers be told apart on this benchmark?	Not decidable at this size	0/5 bootstrap intervals exclude zero; 12–54× items needed

The form transports to a second corpus; the constant does not. On InfiniteBench's longbook_qa_eng (narrative QA, an event a paragraph, 198 items), a 67.6× lever gives $\alpha = 1.009$, bootstrap interval $[0.893, 1.070]$, clear of a frozen floor of 0.80: the required share again does not vanish, but the constant differs, 0.30 against 0.41, exactly as Prop. 3 predicts since changing the corpus changes the lever's domain. A constancy criterion on the rung-to-rung spread fails on its own resolution, and we report that refusal rather than the flat-looking point estimate (Appendix B).

The same shape, in the field's most expensive disagreement. Kaplan et al. (2020) report $N^* \propto C^{0.88}$ (and 0.73 on their own adjusted accounting) against Hoffmann et al. (2022)'s $C^{0.50}$; Porian et al. (2024) decompose the gap into four further levers the compute lever also moved (a cost model, a warmup duration, and optimizer tuning), and an independent decomposition lands on a parameter-counting convention (Pearce & Song, 2024). We do not rank our own between-lever gap against theirs, since ours is not decidable, and a quantity graded not decidable may not stand on either side of an inequality. Two further published cases, a lever failure and a resolution failure routinely confused with each other, are audited by type in Appendix B.

5 Limitations

(i) One retriever and one corpus, the only one on disk shipping the same questions at two lengths; Prop. 3 is retriever- and corpus-general, but the size of the disagreement it permits is measured once. (ii) Containment conflates redundancy with difficulty: *occ* over-counts a common gold string and misses an inferred answer entirely, and §J's attempt to control it came out not evaluable. (iii) Three frozen bands were below their statistic's resolution (a quantisation floor twice the frozen band; a bootstrap width $\sim 7\times$ the frozen bands); we do not report them as nulls, and the audit is a companion's (Anonymous, 2026j). (iv) The powered class-separation ordering was seen before it was registered; a pass reads "the post-hoc pattern survived", never "we predicted it". (v) The full-set licence holds only to 0.104, and α is quantised (multi-session's k^* is a step function), so no per-dimension α should be read to three decimals. (vi) One arm has a model in it (§C), testing a score factorisation against a credit rate measured in the parent paper rather than re-measured here (Anonymous, 2026i); the claim is that the factorisation holds, not that its constant is universal.

6 What remains owed

Three arms would sharpen this paper without moving the boundary function or the quantile spread that is its finding, none of them run here: separating composition from cardinality (the parent ran this at $2.49 and it returned not decidable (Anonymous, 2026i), and we import that grade); a corpus whose redundancy can be dialled, so the three repairs' disjoint subsets could overlap; and a free re-run of §H's ladder with completion taken over the exactly-once subset, which is §J and settled neither direction. Full pricing is in Appendix B.

7 Conclusion

Pin the task, the retriever and the scorer, and send a memory benchmark's record length to infinity. What appears at the boundary is a function and not a score: the budget $k^*(L)$ at which the answer becomes reachable, whose exponent is confined to $[0, 1]$ and whose value decides whether "a small retrieval budget" means anything in the limit. On nested records built by deletion it does not: the budget is a constant $\approx 40\%$ share of the record over a $62\times$ ladder, refuting our own frozen prediction of a sublinear law, and the exponent sorts the benchmark's dimensions into the classes it labels them with, from two counts over text; §K then shows that ordering is mostly confounded with gold multiplicity. The same resource moved a second way gives a different exponent and the opposite ordering, and that is not a failed measurement but an identification result: Prop. 3 proves that on a fixed corpus no lever moves record length alone, so a resource ladder identifies not $\alpha(\text{task})$ but $\alpha(\text{task}, \text{lever}, \text{retriever})$, exactly as a treatment effect is identified only relative to the population it was measured in (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2014): publish the exponent with its lever and its retriever named.

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 two public corpora, the retrieval instrument (sweep.py, alpha.py, xlen.py, xcorpus.py), and the scoring tool that reproduces $k^*(L)$, α, φ^* for any record-length ladder 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 Y was frozen in a preregistration file before its data existed; the appendix states the counting rule and the freeze commit for each.

Ethics statement

Every corpus used is a public benchmark's own released transcripts and golds; no human data was collected and no individual is identified or evaluated. The one arm that scores text with a model (§C) evaluates a synthetic composition-probe family built by the parent paper from a fixed generator, not human subjects. 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, from public benchmarks, 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 Infinite-Division Limit: Conflicting Data Trains a Policy, Not Damage. concurrent manuscript; author list and repository withheld for double-blind review, 2026b.

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, 2026c.

Anonymous. The Free-Recipe Limit: Every Recipe Effect Measures Which Premise of an Idealised Learner Broke. concurrent manuscript; author list and repository withheld for double-blind review, 2026d.

Anonymous. Self-Cleaning and Captured Anyway: One Measured Primitive for Error in a Store an Agent Writes to Itself, and What a Falling Score Actually Measures. concurrent manuscript; author list and repository withheld for double-blind review, 2026e.

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, 2026f.

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, 2026g.

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, 2026h.

Anonymous. Resolving Power: Whether a Memory Task Can See the Distinction It Is Cited For, and a Construction Whose Resolving Power Is Provably Positive. concurrent manuscript; author list and repository withheld for double-blind review, 2026i.

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, 2026j.

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, 2026k.

Elias Bareinboim and Judea Pearl. External validity: From do-calculus to transportability across populations. Statistical Science, 29(4):579–595, 2014.

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.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. arXiv preprint arXiv:2404.10102, 2024.

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.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew Johnson, Katie Millican, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. arXiv preprint arXiv:2202.01169, 2022.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. SIAM Review, 51(4):661–703, 2009.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. arXiv preprint arXiv:2109.07740, 2021.

William Held, David Hall, Percy Liang, and Diyi Yang. Relative scaling laws for LLMs. arXiv preprint arXiv:2510.24626, 2025. Fits how the GAP between two test distributions evolves with scale rather than the aggregate, and finds divergent trajectories across domains. Read here for the estimand: a difference of two scaling behaviours is a harder object than either, and our two levers make the same point about a difference of two exponents on one task.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556, 2022.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. Fluid language model benchmarking. In Proc. Conference on Language Modeling (COLM), 2025. arXiv:2509.11106.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts LLM performance. Chroma Research technical report, https://research.trychroma.com/context-rot, 2025. Holds the needle-question pair fixed and varies only the irrelevant content, over 18 models. That is a GROWTH lever in our taxonomy and it is the right one for their object, which is the model; ours is the benchmark's own labels.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020.

Shahar Levy, Nir Mazor, Lihi Shalmon, Michael Hassid, and Gabriel Stanovsky. More documents, same length: Isolating the challenge of multiple documents in RAG. arXiv preprint arXiv:2503.04388, 2025. The closest prior work on the identification problem: it moves DOCUMENT COUNT while pinning context length, which is the dual of our levers and evidence that the two resources are separable only one at a time.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.

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

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

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

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

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. arXiv preprint arXiv:2305.16264, 2023.

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.

Tim Pearce and Jinyeop Song. Reconciling Kaplan and Chinchilla scaling laws. Transactions on Machine Learning Research, 2024. arXiv:2406.12907.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In Proc. 25th AAAI Conference on Artificial Intelligence, 2011.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In Proc. 41st International Conference on Machine Learning (ICML), 2024. arXiv:2402.14992.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In Advances in Neural Information Processing Systems (NeurIPS), 2024. Spotlight; arXiv:2406.19146.

Qi Qian, Chengsong Huang, Jingwen Xu, Changze Lv, Muling Wu, Wenhao Liu, Xiaohua Wang, Zhenghua Wang, Zisu Huang, Muzhao Tian, Jianhan Xu, Kun Hu, He-Da Wang, Yao Hu, Xuanjing Huang, and Xiaoqing Zheng. Benchmark^2: Systematic evaluation of LLM benchmarks. arXiv preprint arXiv:2601.03986, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 3(4):333–389, 2009. The retriever this paper pins. It is cited because the exponent is a property of the (task, lever, retriever) triple and the third element must be nameable.

Maksim Sheverev, David Finkelstein, and Sergey Nikolenko. Beyond memory leaderboards: Evaluating scientific memory as budgeted context restoration, 2026. arXiv:2607.16848.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

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

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

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

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In International Conference on Learning Representations (ICLR), 2025. arXiv:2410.10813. The corpus every record-length arm in this paper runs on: 500 questions over timestamped, freely scalable chat histories, with the five ability labels — information extraction, multi-session reasoning, knowledge update, temporal reasoning, abstention — that our class split is taken from rather than invented.

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

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

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. In International Conference on Learning Representations, 2025. arXiv:2410.04343.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, Conghui Zhu, Hailong Cao, and Tiejun Zhao. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In Proc. AAAI Conference on Artificial Intelligence, 2026. arXiv:2505.15055.

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

Supplementary Material

This appendix follows the references and is supplementary. The main text is self-contained: every claim it makes is stated with the measurement that decides it. The sections below hold the extended related work; the full arm-by-arm experimental record (the budget-axis consistency check, both record-length levers, the quantile table, the registered and powered class separation, the skill-free baseline, the three repairs for the multiplicity confound, the confound measured directly, the second corpus, the scaling-law case study, and what remains owed); the resolution of every frozen band; the exponent ladder in full; the estimator appendix; 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

Scaling laws fit exponents to resource ladders, and name the resource but not the lever, and the field's most expensive disagreement is what that costs. The form is standard: hold an architecture and a recipe, move one resource over decades, fit a power law (Kaplan et al., 2020; Hoffmann et al., 2022; Ghorbani et al., 2021; Clark et al., 2022). We borrow its shape. We do not accept its implicit premise, which is that the resource named in the exponent is the thing the lever moved, and §M audits three published exponents against that premise rather than deferring to them. The short version is that the premise fails in the literature exactly as Prop. 3 says it must fail on our corpus: Kaplan et al. (2020) and Hoffmann et al. (2022) report $N^* \propto C^{0.88}$ (and 0.73 under Kaplan et al.'s own adjusted accounting, §M) and $N^* \propto C^{0.50}$ for the same nominal resource, and the 0.38 between them decomposes into three quantities their levers moved besides compute (Porian et al., 2024). Our exponent is over a far smaller lever than any of these, and the object we add is not a better fit but a statement about what a fit can identify at all.

The literature has already met the harder half of our estimand, on a different axis. Work that fits how the gap between two test distributions evolves with scale, rather than the aggregate, finds divergent trajectories across domains, so scale is not a universal equaliser (Held et al., 2025). That is the same structural point one level up: a difference of two scaling behaviours is a harder object than either of them, and our $\beta - \alpha$ is a difference of two exponents for one task under two levers. Where they can vary the axis freely and must contend with divergence, we cannot vary it freely at all, and §I measures what that costs: zero of five such differences are decidable at the item counts a public benchmark supplies.

Partial identification is the frame, and it says our headline is a set and not a point. Manski (1995) argues against buying point identification with assumptions that cannot be falsified, and in favour of reporting the set of values the evidence admits (Manski, 2003). Prop. 3 is an identification statement of that kind: on a fixed corpus the data plus any assumption we are willing to defend do not pin α(task) to a point, because every realisable lever confounds L with something else. What the evidence admits is a set indexed by the lever, and our two levers are two of its members. Read that way, our $\beta - \alpha$ interval spanning zero on five of five dimensions stops being a failure and becomes the width of an identification region, reported as such, and the way to narrow it is a lever we cannot build on somebody else's corpus rather than more items on this one.

Transportability names the failure exactly, and supplies the vocabulary we adopt. The question "does a relation measured in one population hold in another" has a formal answer in terms of what the two populations share (Pearl & Bareinboim, 2011; Bareinboim & Pearl, 2014). Two levers on one corpus are two populations in that sense, differing in the mechanism that generated the record, and the ordering of dimensions by α is exactly the kind of relation that need not transport. We borrow the vocabulary and not the machinery, and a reader is right to ask whether that is legitimate. Our answer: Prop. 3 is an identification result proved by enumerating the levers a fixed corpus admits, not a transportability result in the technical sense: there is no selection diagram here, no S-node, and no do-calculus derivation, and we do not claim one. What the two share is the shape of the conclusion, that a relation

is indexed by the population or the lever that produced it. Formalising the lever as an intervention on a graph over (L, occ, D) is a real and unattempted extension, and it would buy something we do not have: a criterion for when an exponent does transport, rather than our enumeration of why on a fixed corpus it does not. Until then the accurate label for our result is partial identification (§A), and we use "transportability" only where we are describing the literature's own object.

Long-record evaluation, and the levers others have already built. The benchmarks that motivate this paper advertise record length as the difficulty axis and report scores at several lengths (Tavakoli et al., 2025; Microsoft Open Source, 2026). Three lines of work move that axis deliberately, and each one is a lever in the sense of Def. 4. The closest is the growth lever: a controlled study holds the needle–question pair fixed and varies only the amount of irrelevant content, over eighteen models, and finds performance growing unreliable with length rather than falling smoothly (Hong et al., 2025). By Prop. 3(c) that construction changes the corpus, and for their object that is correct: they are measuring a property of the model, for which the corpus is a fixture. Our object is the benchmark's own dimension labels, for which the corpus is the thing under test, and the two choices are not interchangeable. The second is the dual of ours: a study that increases the number of retrieved documents while pinning total context length, isolating cardinality from length (Levy et al., 2025). That is direct evidence for the separability claim we make, since these resources can be moved one at a time only by pinning another, never by moving one alone. The third is positional: performance depends on where in the context the evidence sits, not only on how much context there is (Liu et al., 2024), and the interference literature shows the binding constraint at long context need not be length at all (Wang & Sun, 2025). Our contribution to this line is narrow and mechanical: on a fixed corpus the length axis is not a single axis, and we measure how far apart two of its realisations land.

Two results price the same resource this paper's limit is about, and neither reads it as a function of record length. Yue et al. (2025) allocate inference compute across retrieved context and report that performance grows nearly linearly in it when the allocation is optimal; Sheverev et al. (2026) evaluate eight memory and retrieval systems on scientific papers and find that after controlling for retrieval budget the lead disappears, which is independent evidence for this paper's premise that the budget and not the architecture is what a length-indexed score is reporting. Both hold the corpus fixed and read the budget as an input; the limit here sends the record length to infinity and reads the budget as an output, $k^*(L)$, whose exponent is the estimand. That difference is what makes the quantile spread of §F visible at all: a budget fixed for a comparison cannot show that the hardest few per cent of items demand a constant share of the record while the median's demand vanishes.

The standing objection to a containment-based instrument, which we answer rather than avoid. Our instrument counts verbatim occurrence of the gold string. A benchmark exists precisely because that is exploitable: NoLiMa was built with a needle set whose questions and needles have minimal lexical overlap, because models and retrievers can otherwise ride literal matches, and the strongest model in that study falls from 99.3% to 69.7% once the literal route is removed (Modarressi et al., 2025). Read at face value, this says our instrument measures lexical reachability and not answerability, and we accept the reading. Three things follow and we state all three. (i) V and R^{str} are upper bounds on what a retriever must be given: an item whose answer requires inference from a non-matching record is counted as reachable when its gold string happens to occur, so our admissibility filter is optimistic about the corpus and the exponent is measured on the easy subset. (ii) That direction is the safe one for our claim: if the budget a task demands is already a constant share of the record on the lexically reachable items, it is not smaller on the rest. (iii) The direction is not safe for the class separation of §G, because a dimension whose golds are common strings and one whose golds require inference are treated alike, which is one of the two mechanisms §J tries to control and fails to. A version of this instrument on a NoLiMa-style corpus is the experiment that would settle it, and §6 prices it.

The systems this accounting is about, and the axis they are ranked on. Deployed memory architectures do not present themselves as retrievers: Chhikara et al. (2025) extract and consolidate salient facts into a persistent store and report a head-to-head comparison of ten approaches, and Xu et al. (2025) maintain linked notes that are revised as new information arrives. Both are ranked by a score at the record length their benchmark ships, and α is a statement about what happens to the budget such a system must be given as that length grows. Our measurement does not evaluate either system and is not evidence about them; what it says is narrower and applies to the comparison rather than the competitors: a ranking obtained at one record length does not carry to another unless the exponent is the same for both, and Prop. 3 says the exponent is not even a property of the task alone. The cheapest consequence is a reporting one: a memory system's evaluation should state the record length it was run at, which is a line most already have and few make load-bearing.

Retrieval budget as a design parameter, where the field's finding is not ours. The applied literature treats top-k as a knob with an interior optimum: accuracy rises then falls, because additional passages bring distraction as well as coverage, and models are demonstrably distractible by irrelevant context (Shi et al., 2023; Lewis et al., 2020). Our budget sweep measures the opposite on the family we built: six irrelevant events alongside the two that matter cost the model 0.033, and the score divided by coverage is flat across every arm (§C). We do not claim the field is wrong; we claim the two findings are about different regimes and say which: distraction costs something when the distractors are plausible answers to the question, and our probe family's distractors are events about other entities. The reconciliation is a testable one and it is not tested here.

Item-efficient benchmarking assumes the quantity is estimable at the item count available. Work on evaluating with fewer examples, and on adaptive item selection, is about spending a fixed item budget well (Polo et al., 2024; Hofmann et al., 2025; Zhou et al., 2026). It presumes the target quantity is identified and asks for its cheapest estimate. The failure we report is upstream of that: our target is a difference of two exponents, and at the item counts a public benchmark provides, that difference is not estimable at any band we would have accepted, needing 12 to 54× the items, on a benchmark whose dimensions hold 47 to 118 admissible items.

Construct-validity work establishes that the problem is widespread; this paper is about one construct's resolution. Large reviews find that benchmarks routinely measure something other than the construct they name (Bean et al., 2025; Qian et al., 2026), and judge-based scoring can be reliable without being valid (Norman et al., 2026), a gap the standard remedies do not close because they buy agreement: rotating evaluators removes a bias term (Zhu et al., 2026) and changing one measures how much the number moves (Yang et al., 2026), and neither asks what is being agreed about. That literature diagnoses the gap between construct and measurement. We take one construct, "this task requires composition across records", and ask what a resource ladder can and cannot establish about it. Where the scorer itself is the defect, the judge companion is the account (Anonymous, 2026k); nothing in this paper's instrument touches a scorer.

Where the family enters, and what each member supplies here. The access-structure account of capability (Anonymous, 2026a) is where the phrase "the budget binds, not the memory" comes from, and it is a principle rather than a source of coefficients. The theory paper asks whether a task can resolve the state/index distinction at all (Anonymous, 2026i); its instrument half audits deployed benchmarks with the same counts this paper sweeps (Anonymous, 2026h); the judge companion owns the scorer defects and the credit-rate axis (Anonymous, 2026k), and the constant $\iota = 0.917$ imported once below is the theory paper's own extract arm (Anonymous, 2026i, row 105). The resolution of a frozen criterion, and the audit of three of this paper's own bands, is the subject of (Anonymous, 2026j). Two training-time members send other resources to infinity against a fixed apparatus and are cited for the shape rather than for any coefficient: recipe search (Anonymous, 2026d) and label division (Anonymous, 2026b). A deployment-time member measures a store an agent writes to itself (Anonymous, 2026e), and the cost of checking a programme's own constants

is counted in (Anonymous, 2026g). The stream member states the same three-resource picture from the task side (Anonymous, 2026c).

A.1 What this paper imports

B Full experimental detail, arm by arm

The main text (§4) reports the headline verdict of each arm. This section is the arm itself: every registered criterion, every table the criterion was graded against, and every correction found while running it, in the order the arms were run.

C What this paper takes from the budget axis

One arm of this programme scores sixty composition probes with a model in order to ask whether the score on the budget axis factors into retrieval coverage times a constant. It does: coverage runs from 0.000 at k=1 to 1.000 at k=8 while score divided by coverage is flat across every arm, so on that family the model composes whenever it holds the records and the "composition" label was measuring retrieval. That arm belongs to the parent paper, whose witness family, scorer and judge-free rule it uses and which reports the same numbers (Anonymous, 2026i); it is in Appendix T here so that this paper's own use of it is auditable, and what this paper uses is one sentence of it: the factorisation it tests is assumed, not established, by the record-length arms below, all of which are counts over text with no model in them at all.

D First lever: the budget grows and the share shrinks

(1) pins the budget and lets the record length go to infinity, and the companion instrument paper's blind three-benchmark arm (Anonymous, 2026h) could not follow it: every slot there supplies one length. longmemeval (Wu et al., 2025) supplies two. The _s and _m files carry the same 500 questions, with 500/500 shared identifiers, 500/500 with byte-identical question and answer strings, at a median 492 and 4899 turns. Criteria in E_LENGTH_PREREG.md, frozen and committed before the runner existed.

Two choices were fixed in that file rather than after the numbers arrived, and both were fixed against the reading we expected to report. The graded set is the items whose gold occurs in exactly one event at both lengths, because occurrence count is itself length-dependent, so letting each length pick its own subset would confound the effect with the selection. k^* above half a cell's median event count is flagged saturated, a guard that can only bite the short arm, which is the arm whose early saturation would flatter the prediction.

The replication gate, which gates everything. Re-run at L = _s, the pipeline reproduces results/rstr/ONCE.json exactly: $n = 61$ and $R^{\mathrm{str}} = 0.459/0.377/0.262/0.197/0.148/0.148$ at $k \leq 32$. P-T3-58 is supported, so this is the code path that produced the companion's blind arm (Anonymous, 2026h) and not a second one wearing its name. That check exists because P-T3-42 once passed on wrong numbers.

The graded verdicts. Of 70 items exactly-once in _s and 47 in _m, 36 are matched. P-T3-55, P-T3-56 and P-T3-57 are all supported: k^* rises $64 \to 128$ while L rises $484 \to 4912$, so the required share $\varphi^* = k^*/L$ falls from 0.132 to 0.026, and $R_m^{\mathrm{str}}(k) \geq R_s^{\mathrm{str}}(k)$ at every non-saturated k. The full admissible set, $n = 417$ over five cells, carrying the premise defect above and reported as description, agrees in direction: k^* of $128 \to 512$, φ^* of $0.261 \to 0.104$.

P-T3-57 is also the first matched test of something the companion's anywhere-containment curve (Anonymous, 2026h) could only hint at. Claim 2k reported $R^{\mathrm{str}}(32) = 0.000/0.021/0.017$ across BEAM's 100K→1M splits and read it as a fixed budget reaches less as the transcript grows. However, those splits are different item sets, so the movement could have been the items. Here the items are the same items, matched one to one, and the curve still moves the same way at every budget (Figure 6).

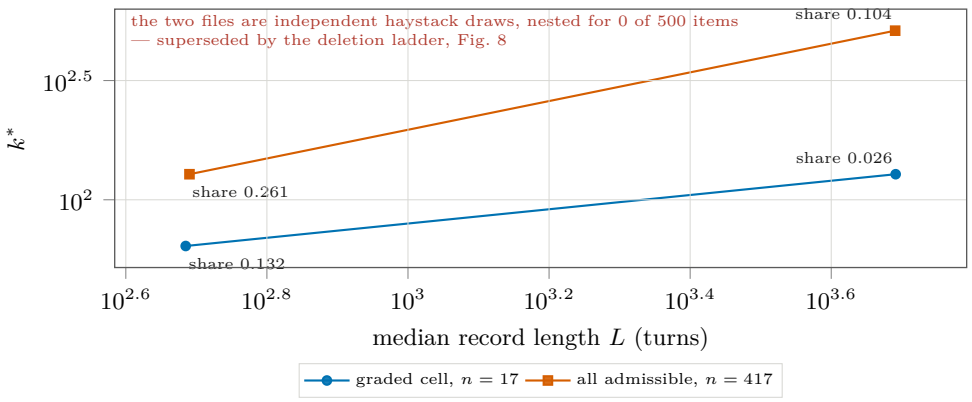


Figure 6. The budget axis answers twice, in opposite directions. The two crossings k^* against record length L, on the matched exactly-once items, read both absolutely (the plotted k^*) and as a share of the record (the label beside each point): the absolute budget rises with L while the share falls, which is the pair of readings this section is about. k^* is where $R^{\text{str}}(k)$ crosses the frozen τ, on the deletion ladder as here. Every value is recomputed from results/length/LENGTH.json at draw time. The graded k^* rests on one cell of $n = 17$ (see the warnbox), and the paired per-item statistic in the text is what carries the finding.

These two files exhibit the draw-lever artifact of Prop. 3, and §E measures the correction. Read naively, the pair says "a longer record needs a larger absolute budget and a smaller fraction of itself", with an implied exponent $\alpha = 0.299$. On a nested ladder built by deletion, $\alpha = 1.011$ and the required fraction is constant at $\approx$40%. The two files compared here are independent haystack draws, so the apparent saving was a property of the draw, a registered criterion was written to catch exactly this, and the nested measurement fired it. The two exponents are not left standing side by side: §H decomposes the 0.712 between them and finds 0.653 of it is which dimensions cleared the floor in each arm.

Two corrections, both against the reading above, and both found before it was written. First, the preregistration's own description of the instrument is false. It says of the two files that "the record moves and nothing else does". _s's session set is a subset of _m's for 0 of 500 questions and shares 4.8% of its sessions with it. The answer sessions are present in both, so both files are the same question with its evidence. They are two independent haystack draws at two sizes, not a nested pair. Length and distractor identity move together, and every number above is a reading of that pair rather than of a clean length sweep. It also explains the 11 items exactly-once in _m and not in _s, which a nested design could not produce.
Second, the frozen hand-computation predicted P-T3-56 would fail, in the sharp form "$\varphi^*_m > \varphi^*_s$, the required share of the record grows". It fell five-fold instead. A registered prediction falsified in the direction that favours the paper is the case for scrutiny, not for a headline, and the scrutiny found the next line.
The graded k^* rests on one cell. Only single-session-assistant clears the floor at either length ($n = 17$); the others hold 1, 4, 5 and 9 matched items. At $n = 17$ one step of R^{str} is $1/17 = 0.059 > \tau$, so "$R^{\text{str}} \leq \tau$" means zero misses and k^* is decided by the single hardest item.

The statistic that does not need a cell or a threshold. For each matched item let k_i be the smallest ladder budget at which its gold is retrieved. Over all 36 matched items, k_i rises with length for 9, falls for 1, and is unchanged for 26: an exact two-sided sign test on the 10 discordant pairs gives $p = 0.021$. Median k_i goes $1.0 \to 1.5$ while median k_i/L goes $0.0021 \to 0.0003$. This is post hoc and is reported as description, but it is the quantity k^* is a lossy pooling of, it uses every matched item, and it says the same thing the graded verdicts do without resting on one cell of seventeen. §E then shows that "one cell of seventeen" was

the whole problem: that cell is single-session-assistant, the exponent it carries is 0.358, and the pooled figure over all dimensions is 1.011.

E The nested ladder: $\alpha = 1.011$ and a constant share

§D's own warnbox reports that its two record sets are independent draws, nested for 0 of 500 questions. A result resting on that pair is a comparison of two haystacks, not a measurement of length. We built the nested version by deletion from longmemeval's longer split (Wu et al., 2025), longmemeval_m: for each item, retain every answer session, then a growing prefix of a seeded permutation of the distractors, giving seven rungs at a median 79, 160, 318, 638, 1265, 2505 and 4913 events. Rung i is a subset of rung $i{+}1$ by construction. Criteria in E_SWEEP_PREREG.md, frozen and committed before the runner existed.

Two rules in that file are worth naming because both bind more tightly than the result required. P-T3-62 asserts the nesting and the answer-session containment for every one of the 500 items at every rung, not on a sample, written that way because the previous preregistration asserted nesting it had not checked. The pooling rule changed: all 35 matched items pooled, no question_type cells, because E-length's inherited per-type floor left a single cell of $n = 17$ (Table 4). That change is justified in the frozen file, before these numbers existed.

The graded verdicts: three of four ran against the registered expectation. P-T3-62 is supported and P-T3-59 is supported: k^* runs $64, 64, 128, 256, 512, 1024, 2048$ across the rungs, non-decreasing. Then:

- P-T3-60 is refuted. $\alpha = 1.011$ over the six non-saturated rungs. The budget grows linearly: k^*/L is $0.400, 0.403, 0.401, 0.405, 0.409, 0.417$, a constant $\approx 40\%$ of the record, plotted against what we froze in Figure 7.
- P-T3-61 is not supported, missing by 0.712. §D's two-point estimate of 0.299 does not replicate under a nested design, so its φ^* drop is an artefact of comparing two draws and is reported here as one.
- P-T3-63 is not supported. The median per-item slope is 0.000: 22 of 35 items have a budget that never moves across the entire ladder.

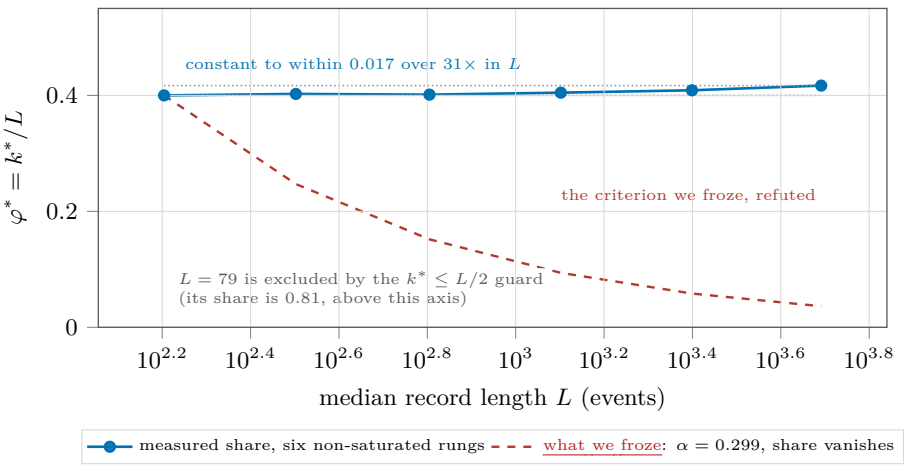


Figure 7. The headline on the axis where it is readable, against the criterion it refuted. $\varphi^* = k^*/L$ is the share of the record the retriever must be allowed to see. A constant share is a horizontal line and it is what we measure, to within 0.017 over a 31× range in L; a vanishing share is the dashed curve, and it is what we froze a criterion predicting. By Prop. 2 the two are qualitatively and not quantitatively different: the first makes "a small retrieval budget" empty in the limit and the second makes it meaningful. The log-log ladder of Figure 8 carries the same measurement, where a constant share is a slope of one and a reader cannot check it by eye. Recomputed from results/sweep/SWEEP.json at build time.

This section refutes the one before it, and both are kept. §D is not deleted, because the pair of results is the finding: a length effect measured across two independent draws gave $\alpha \approx 0.3$, and the same effect measured by deletion on nested records gives $\alpha \approx 1$. The registered criterion that caught it (P-T3-61) was written into the frozen file for this purpose, with the sentence "that is the outcome that costs this paper the most and it is the reason this file exists". It cost that.

Why the two numbers differ, which is the result worth keeping. k^* pooled over n items tolerates $\lfloor \tau n \rfloor$ misses, so it is a statistic of the hard tail of a set and not of its typical item. Splitting the ladder by the benchmark's own labels locates the disagreement exactly:

Table 4. k^* across the nested ladder by the benchmark's own dimension, and the fitted exponent α in $k^* \propto L^\alpha$. Post hoc and not graded, added after the pooled α was read, and three of the five cells sit below the frozen floor of 10 – so this is the thing to register next, not a result. Recomputed from results/sweep/SWEEP.json; see §E.

dimension	n	79	160	318	638	1265	2505	4913	α
knowledge-update	4	8	8	8	4	8	8	8	0.001
single-session-user	9	8	8	8	8	8	16	16	0.182
single-session-assistant	17	32	32	64	64	64	128	128	0.358
multi-session	1	64	128	256	256	512	1024	2048	1.017
temporal-reasoning	4	64	64	128	256	512	1024	2048	1.022

The exponent separates the same dimensions the parent paper's blindness condition does (Anonymous, 2026i); the tempting reading is composition depth, and §H refutes that reading while keeping the separation. Single-hop dimensions have $\alpha \approx 0$–0.36 and composition dimensions $\alpha \approx 1.02$, and the mechanical reading runs: the answer to a single-hop question sits in one event, so a retriever either finds it or does not, whereas which events to compose cannot be known without scanning. On that reading the three numbers above are one fact: pooled α is 1.011 because pooled k^* tracks the tail, the median per-item slope is 0.000 because most items are single-hop, and §D got 0.299 because its only floor-clearing cell was single-session-assistant, whose α is 0.358. The cell itself is right; the error would be to read one cell as a general law. What §H then removes is the composition-depth attribution, not the separation: an independent length lever gives an exponent that reverses the ordering, and the recurrence of the gold string accounts for most of α on the cells where it can be measured.

The limit on the paragraph above: three of the five dimensions hold 4, 4 and 1 matched items, so the composition end of that ladder rests on five items in total. The pattern is worth a preregistration and is not worth a claim, and it is graded post hoc. §G runs that preregistration and the ordering survives it; §H then moves the same resource a third way and the account does not. What is graded here is narrower: on nested records the required budget is a constant share of the record at the tail quantile the statistic is defined on, the unnested draw's estimate of it does not replicate, and the typical item's budget does not move at all, which are not three facts but one, and §F reads it as one.

F The headline quantile and the others

k^* is defined as the smallest budget at which the unreachable share falls below τ, and this paper froze $\tau = 0.05$. k^* is a tail quantile and every statement about it is a statement about the hardest few per cent of items. Without the quantifier the claim overreaches, because the paper's own per-item statistic says the opposite about the typical item: the median per-item slope is 0.000 and 22 of 35 items have a budget that never moves across the entire ladder (P-T3-63, not supported; Table 5). Both are true and they are one fact read at two quantiles. The ladder already on disk answers the question at no cost, because the $R(k)$ curve at every rung is stored and $k^*(\tau)$ is a lookup in it.

Table 5. The headline is a tail statement, and the tail is the only quantile at which it holds. $\varphi^* = k^*/L$ at each quantile, over the same seven non-saturated rungs and the same pooled $R(k)$ curves the published exponent uses. Saturated rungs, where $k^* \geq L$, are excluded exactly as they are for the published fit.

quantile	rungs	α	φ^* at $L = 79$	φ^* at $L = 4913$	reading
$\tau = 0.05$	7	+0.898	0.8101	0.4169	constant
$\tau = 0.1$	7	+0.719	0.1013	0.0261	vanishing
$\tau = 0.2$	7	+0.109	0.0506	0.0016	vanishing
$\tau = 0.5$	7	+0.000	0.0127	0.0002	vanishing

The dichotomy of Prop. 2 cuts the other way at every quantile above the one we froze. At $\tau = 0.05$ the exponent is +0.898 and the share is flat, which is the published finding. At $\tau = 0.20$ it is +0.109 and at $\tau = 0.50$ it is +0.000, with the required share falling from 0.0127 to 0.0002, the median item's budget is asymptotically free. Hence "a small retrieval budget" is empty in the limit for the hardest items and meaningful for the typical one, on one corpus, with one retriever, at one time. The distribution of budget demand is the object, and a single exponent is a reading of it at one quantile, which is a stronger result than the one this paper published and a narrower one than the sentence it published it in.

Two consequences for the rest of the paper. The identification result of Prop. 3 is unaffected: it is about levers, not quantiles, and τ is simply a further argument the exponent carries. However, §H's comparison of two levers, and every ordering of dimensions by α, inherit a τ they never named, and at $\tau = 0.5$ there is no ordering to compare, because every dimension's exponent is zero. We state that rather than re-run a cross-lever comparison at four quantiles on cells of $n = 4$.

G The exponent, registered and powered

§E's exponent ladder is post hoc and rests on cells of $n = 4, 9, 17, 1, 4$, three of them below the frozen floor. This section registers it and buys the power, on criteria frozen in prereg/E_ALPHA_PREREG.md before the runner existed. Everything is imported from E-sweep unchanged: the same seven nested rungs from the same deletion rule and seed, the same retriever, the same τ and saturation guard. One thing changes: the grading set is the full containment-admissible one, $n = 47$–118 per dimension, rather than the 35 items that are exactly-once at every rung (Table 6).

The licence, tested rather than assumed. That set carries the premise defect the premise defect the companion instrument paper names (Anonymous, 2026h), so the preregistration states why it is nevertheless usable here and makes the statement a gate: α is a slope in $\log L$, the defect's over-count is a property of the gold string and the question, and a bias that does not vary with L shifts $\log k^*$ by a constant and cancels in the slope. P-T3-67 is supported: on the one cell E-sweep graded, the full-set accounting returns $\alpha = 0.462$ against the exactly-once 0.358, a difference of 0.104 inside the frozen band of 0.15. The bias does not cancel exactly, and the residual is reported rather than rounded away.

α separates composition from single-hop, and it separates the classes rather than merely ordering the dimensions. P-T3-65 is supported: the smallest composition exponent, 0.981, exceeds the largest single-hop one, 0.462. P-T3-66 is supported: composition sits at $\alpha \geq 0.8$, so its budget grows near-linearly and no sublinear reading is available for it. P-T3-68 clears: the within-class spread is 0.175 against a between-class gap of 0.519, three times larger, so this is a statement about the two classes and not a coincidence of five separately-ordered dimensions (Figure 9).

Table 6. E-alpha. k^* across the nested ladder by dimension, on the full admissible set, and the fitted exponent. Every cell clears the floor and every fit uses 5–7 non-saturated rungs. The split into single-hop and composition was fixed from the benchmark's own labels before the run. alpha.py.

class	dimension	n	79	160	317	630	1255	2501	4896	α
single-hop	single-session-user	64	4	8	8	8	8	16	16	0.287
single-hop	knowledge-update	70	16	16	16	32	32	32	64	0.326
single-hop	single-session-assistant	47	4	8	8	8	8	32	32	0.462
composition	multi-session	118	32	32	256	256	1024	1024	1024	**0.981**
composition	temporal-reasoning	118	32	16	64	128	256	512	1024	**0.984**

A retrieval budget and two counts over text sort a benchmark's own dimensions by how much composition they require, with no model, no judge, no system and no API call. A task whose answer sits in one event needs a budget that barely grows with the record; a task requiring composition across events needs one proportional to it, because which events to compose cannot be known without scanning.

What this registration bought, and what it did not. The ordering was seen on five items before the preregistration was written, and that file says so: what this test buys is the power, not the surprise, and a pass reads "the post-hoc pattern survived a well-powered test" and never "we predicted it". Two further limits. First, the licence holds to 0.104 rather than exactly, so the full-set α carries a residual bias of unknown sign. Second, multi-session's k^* is a step function, $32, 32, 256, 256, 1024, 1024, 1024$, fitted over five non-saturated rungs of a power-of-two ladder, so α is quantised; the class separation survives only because the gap is far larger than the quantisation.

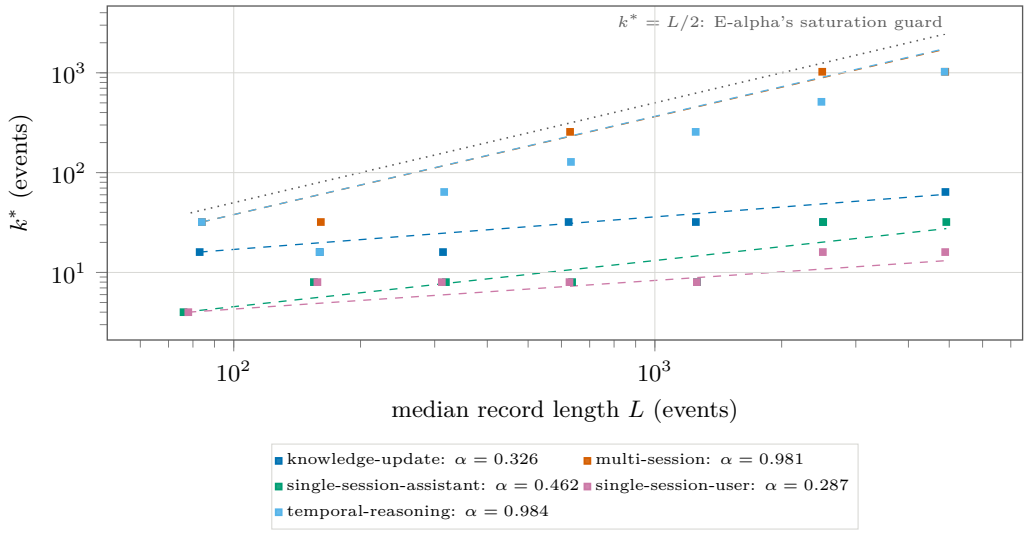


Figure 8. The nested deletion ladder α is fitted on. Markers are the rungs that clear E-alpha's saturation guard $k^* \leq L/2$ (dotted); dashed lines are the fitted exponents. The guard is what removes three dimensions from E-baseline's blind arm (§I), because a lottery needs a large share of the record by construction, so its curve sits above the dotted line everywhere. Recomputed from results/xlen/XLEN.json.

H A third lever on the same resource

Both ladders above move the record by deleting sessions. That is a confound we did not name at the time: deleting a session deletes whatever copies of the gold string it held, so a deletion ladder moves record length and gold multiplicity together. The benchmark supplies a lever that moves length without deleting anything, the same 500 questions shipped at a median

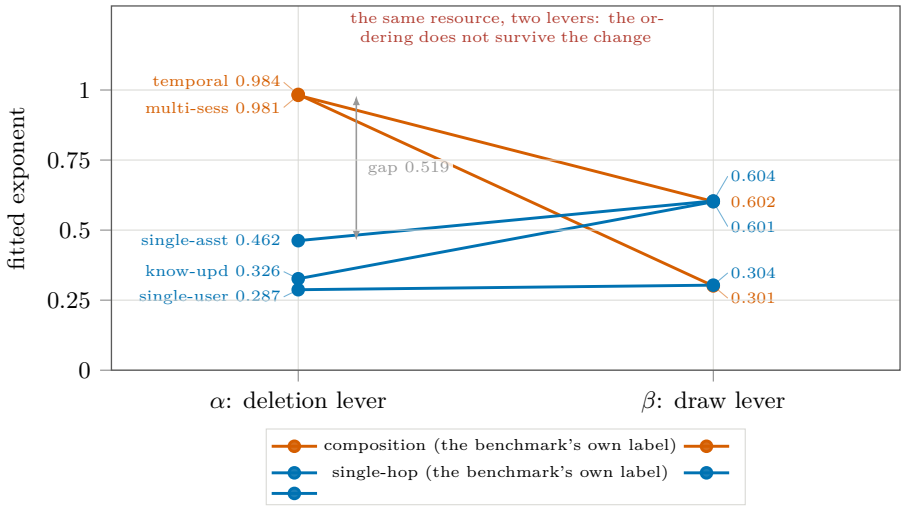


Figure 9. The ordering reversal, which a column of numbers cannot show. Each line is one dimension, measured on the same corpus by two levers that move the same resource. Under deletion the benchmark's three single-hop dimensions sit below its two composition dimensions with a gap of 0.519 against a within-class spread of 0.175, the separation §G registered and powered. Under the draw lever the lines cross: temporal-reasoning, which has the highest α at 0.984, has the lowest β at 0.301. Neither measurement is wrong; by Prop. 3 they are measurements of different functionals, and that is the whole of this paper's second result. Recomputed at build time from ALPHA.json and XLEN.json.

492 and 4899 events, and an exponent measured over that lever, call it β, must reproduce α if α is a property of length. Criteria are frozen in prereg/E_XLEN_PREREG.md with a forecast table, before the runner existed. Zero API cost. The arm is blind on the short split, where no full-set k^*, no α and no β existed anywhere; the three statistics added afterwards on data already on disk are admissible only for refutation, by the rule this paper applies to its own non-blind arms.

The instrument passes twice, so nothing below is an artefact of it. P-T3-69 is supported: the short pass reproduces §D's published $n = 17$, 484 median events and $k^* = 64$ exactly, and the long pass reproduces §G's five exponents to 10^{-9}.

Table 7. E-xlen. β is the exponent over the undeleted 10× lever; α_m is §G's, reproduced here by the gate. g is the OLS slope of log(median *occ*) on log L, the licence P-T3-72 tests. xlen.py.

class	dimension	α_m	β	g	licence
single-hop	single-session-user	0.287	0.304	0.228	fails
single-hop	knowledge-update	0.326	0.601	0.460	fails
single-hop	single-session-assistant	0.462	0.604	0.000	holds
composition	multi-session	0.981	0.602	0.903	fails
composition	temporal-reasoning	0.984	**0.301**	—	untested

P-T3-71 is not supported, and the ordering is not merely blurred but reversed. β does not reproduce α per dimension, and the two dimensions at the extremes swap: temporal-reasoning has the highest α_m (0.984) and the lowest β (0.301), while knowledge-update has one of the lowest α_m (0.326) and one of the highest β (0.601; Table 7). Two of five cells fall inside the band and they are both single-hop. α is therefore not a property of record length alone, and the composition-depth reading of §E does not survive it.

One of the two failures is a defect in the criterion itself. β is a two-point exponent over a lever of exactly 1.00 decade, and k^* is quantised on powers of two, so a single step of k^* moves β by $\log 2/\log 10.05 = 0.301$. That is twice the ± 0.15 band we froze for it. A band narrower than the quantisation step of the statistic it grades cannot decide that statistic: the two cells that passed could only have passed by luck, and they did. What survives the objection is the part of P-T3-71 that is about rank rather than magnitude, and the rank is reversed by two full steps, which no quantisation argument repairs. The quantum was not the binding constraint. §I removes it by reading the crossing continuously instead of off the ladder, and the comparison is still undecidable, because the sampling interval is 1.7 to 3.7 times the quantum wide. This paragraph diagnoses the smaller of two problems and must not be read as implying that a sharper instrument would have settled P-T3-71. One was built. It did not.

The licence fails, which is the more consequential of the two. §G's α is admissible only if the full-set over-count is constant in L, and that requires per-item gold occurrences not to grow with the record. P-T3-72 is not supported: g is 0.460 for knowledge-update and 0.903 for multi-session, and temporal-reasoning's median item has $occ = 0$ at every rung, so its licence is untested rather than passed. Only single-session-assistant holds, at $g = 0.000$, and it holds by construction: its median item's gold occurs exactly once at every rung. By §G's own rule, every α in that section is now untested, and the point estimates stand as measurements of the full-set statistic rather than as budget exponents.

P-T3-73 is not supported too. Bootstrapping over items, the interval for the single-session-assistant cell is [0.000, 1.064], which contains both composition point estimates, so the class separation is not resolvable at item-level resampling even where the licence holds. P-T3-70 is not supported for power rather than for content: the short ladder spans only 0.79 decades, over which k^* moves at most one quantisation step, so that cell could not have decided anything either way and we say so rather than counting it.

What the refutations have in common, offered as a hypothesis and not as a result. §G counts an item complete when its gold occurs at least once, so the completion rate rises with L whenever copies accumulate, and k^* is dragged with it; the deletion ladder produces the effect because deleting sessions deletes copies. If that is what α measures, then $\alpha \approx g$. On the three dimensions where g is measurable the residual $\alpha - g$ has mean $+0.001$ and spread 0.095, against α's own spread of 0.318; and the one cell where the account predicts no absorption, single-session-assistant with $g = 0$ by construction, has residual $+0.462$, the largest of the four. This cannot pass a test it suggested. It is $n = 4$, and $n = 3$ after an exclusion we made after seeing the residual, and it is recorded as a hypothesis with the test that would settle it stated below.

The test that is owed, and it costs $0. Re-run the ladder with the completion rate taken over the exactly-once-at-that-rung subset instead of $occ \geq 1$, criteria frozen first. §J is that run. It did not settle the hypothesis and it did not refute it; it established that this benchmark cannot decide it.

Where the two exponents' conflict actually comes from, and it is not the accounting. §D measured 0.299 and §G 1.011 over the same benchmark, and this paper owed a decomposition of that 0.712 gap rather than a choice between the two numbers. Measured: 0.653 of it is dimension, i.e. which cells cleared the floor in each arm, 0.104 is the full-set-versus-exactly-once accounting, and 0.059 is the nested-versus-independent construction. 92% of a conflict we had attributed to construction is composition of the graded set. The earlier section's cell was single-session-assistant; the later section's pooled figure is dominated by the two composition dimensions. Neither number was wrong about what it measured.

The forecast table frozen with this arm scored 3 hits and 2 misses, and both misses are cases where we predicted our own reading would survive.

Figure 4 (the point estimates against their bootstrap intervals) is given in the main text (§G) and is not repeated here.

I A skill-free baseline

If §H is right that every containment-based quantity here mixes redundancy with difficulty, the repair belongs to the instrument rather than to α. Criteria P-T3-75 to P-T3-79 were frozen in E_BASELINE_PREREG.md, the first preregistration in this programme to run criterion_resolution.py on its own bands before freezing them, which changed two of them, and the arm has two independent halves sharing one scan.

Half C subtracts a retriever with no skill. For an item with L events of which occ carry the gold, a uniform draw of k events hits with probability $b(k; L, occ) = 1 - \binom{L-occ}{k} / \binom{L}{k}$. Putting $b_i(k)$ where the real hit indicator goes gives a blind resolving power, a blind crossing k^*_{blind}, and $\alpha_{\text{adj}} = \alpha - \alpha_{\text{blind}}$: what is left for retrieval difficulty once redundancy has been paid for. It is a floor on skill, not a model of BM25.

Half D stops reading a threshold off a step function. k^* crosses τ on a ladder of powers of two, so every exponent built on it inherits a quantum. The interpolated crossing k^*_c takes the same data, the same τ and the same ladder, and keeps the fractional part.

P-T3-75, the instrument gate, is supported. Recomputing α with k^*_c moves it by at most 0.129 across the five dimensions, inside the 0.301 quantum the band was set to on purpose. A sharper reading of the same curve lands inside the old one's resolution, so it is the same measurement.

Half C is not evaluable, and the reason is not a null. α_{blind} cannot be estimated on three of five dimensions, because E-alpha's saturation guard, under which a rung counts only if $k^* \leq L/2$, bites everywhere on the blind curve. That is by construction: a skill-free retriever needs a large share of the record, so the guard designed for a real retriever excludes the baseline almost everywhere. The guard was not loosened. Loosening a floor clause after seeing which cells it removes is the move floor clauses exist to prevent, and the cost of keeping it is this section's main result.

P-T3-76 is therefore not evaluable: two of five dimensions have a corrected exponent and both are composition, and a spread over two cells of one class cannot answer a question about between-class variation however small it comes out. P-T3-77 is not evaluable for the same reason: one class has no corrected exponent at all, so the ordering cannot be re-read after correction. Both are apparatus limits and neither is reported as a refutation.

What the two surviving cells do say, and it is not comfortable. On multi-session and temporal-reasoning, the two dimensions §E's reading leaned on hardest, α_{adj} comes out negative: -0.107 and -0.051 (Table 8). Once the record's redundancy is paid for, the two composition dimensions have nothing left over for retrieval difficulty. With $n = 2$ and one class this cannot carry a verdict, and it is the opposite of the direction §G was read in.

P-T3-78 is not supported, and it is the most useful failure in this paper. Recomputing both exponents continuously and re-running P-T3-71's comparison, the 95% bootstrap interval of $\beta_c - \alpha_c$ spans zero on all five dimensions, zero of five decidable, over 1840 to 2000 surviving resamples of 2000. Half D removed the quantum and the comparison did not become decidable.

Table 8. E-baseline, half D. The interval is over items, resampled in both cells. The last column is the factor by which the item count would have to grow for the half-width to reach the 0.15 band this paper kept reaching for, at $1/\sqrt{n}$.

dimension	$\beta_c - \alpha_c$	half-width	× quantum	× items needed
knowledge-update	+0.091	0.705	2.3	22
multi-session	−0.468	0.516	1.7	12
single-session-assistant	+0.240	1.103	3.7	54
single-session-user	+0.055	0.588	2.0	15
temporal-reasoning	−0.685	1.011	3.4	45

The bound this arm actually produced, read off the boundary rather than off a convention. The quantisation defect §H discloses is real and it was never what bound the measurement. The intervals above are 1.7 to 3.7 times the quantum, so item-level sampling error is the constraint and it is an order of magnitude larger. A decidable P-T3-71 at the conventional 0.15 band needs 12 to 54 times the items, not a better estimator. Half D fixed the smaller problem and, in fixing it, measured the larger one. Any future version of this comparison should be powered against that number rather than against a band.

P-T3-79, reported and not graded, and it is the sharpest thing here. At the full rung k^*_{blind} is undefined on all five dimensions: the lottery baseline never brings the blind resolving power down to τ anywhere inside the ladder. There is therefore no blind comparison at all at the record length the paper cares about, and the two α_{blind} values that exist come from shorter rungs. Where a lottery cannot be made to succeed, "the retriever beats a lottery" is not a claim the benchmark can support, which is a statement about the benchmark, not about the retriever.

The forecast frozen with this arm scored P-T3-75 and P-T3-78 as predicted and left P-T3-76 and P-T3-77 unscored, both because the cells they needed did not survive.

J The proposed mechanism

Criteria P-T3-80 to P-T3-84 were frozen in E_ONCE_PREREG.md and committed alone before the grader existed. One thing changes against §H: an item is credited at rung S iff its gold occurs in exactly one turn of the record at S. Same rungs, deletion rule, seed, retriever, τ and admissible set. Write α_1 for the exponent under that rule.

P-T3-80, the gate, is supported: run under the old $occ \geq 1$ rule on the same scan, the grader reproduces all five of XLEN.json's α_m to full double precision, so nothing below can be a difference in the scan.

P-T3-81, the floor, fires on four of five dimensions. The prereg required $n \geq 10$ items exactly-once at every rung and forecast that the clause would fire on at least one; it fires on four, temporal-reasoning among them as forecast (Table 10, Table 9).

Table 9. E-once. The exactly-once share is the fraction of a dimension's items whose gold occurs exactly once at every rung; the floor is 10 items. Only the control cell is graded, and it is the cell where the intervention has nothing to remove.

dimension	n	once	share (%)	α_m	α_1	P-T3-81
knowledge-update	70	4	5.7	0.326	−0.218	not evaluable
multi-session	118	1	0.8	0.981	+0.833	not evaluable
single-session-user	64	9	14.1	0.287	+0.395	not evaluable
temporal-reasoning	118	4	3.4	0.984	+0.618	not evaluable
single-session-assistant	47	17	36.2	0.462	+0.462	clears

P-T3-82, the load-bearing criterion, is not evaluable, and the reason is the result. It compares a between-class gap against a within-class spread, and the only dimension that survives the floor is single-session-assistant, which is single-hop. Both composition dimensions are lost, retaining 1 and 4 items. There is no gap to compute.

P-T3-83 passes on that one cell, and it is the control. $\alpha_1 = 0.462$ against $\alpha_m - g = 0.462$, a difference of 0.000, not to three decimals but to eight, with k^* identical at all seven rungs. That is what a control should do: $g = 0.000$ there by construction, so the criterion reduces to asking whether removing multiplicity moves the exponent on the one dimension with no L-dependent multiplicity to remove. It does not. A control behaving as a control is not evidence for the treatment, and the treatment cells are the four that are not evaluable.

P-T3-84 explains why the control was inert so exactly, and it is not reassuring. The accounting excluded 14 of that cell's 47 items, yet k^* did not move at a single rung, because

the excluded items were already resolved at the budgets that set k^*. Across dimensions the items removed are the recurrent ones and the recurrent ones are the easy ones: the median *occ* of what was dropped is 21.5 on multi-session and 17 on knowledge-update. The surviving subset is therefore not a smaller sample of the same task, it is a harder task, which is the weakening P-T3-84 was written to make visible rather than arguable.

The higher-dimensional reading, and it is a duality rather than a null result. Where multiplicity is large enough to plausibly drive α, at median *occ* of 21.5 and 17, removing it leaves 1 and 4 items against a floor of 10. Where the exactly-once subset survives at 36.2%, $g = 0.000$ and there is nothing to remove. These are not two problems. An instrument cannot both contain the confound and exclude it: on a deletion ladder the recurrence that makes the account plausible is the same recurrence whose removal destroys the sample that would test it. The account is therefore neither supported nor refuted here, and no third account is offered, since the prereg forbids inventing one afterwards, and §H's four refutations stand without a mechanism.

Three repairs, three different survivors, one failure. This is the third independent attempt to fix what α measures, and the three fail in a pattern that none of them shows alone.

Table 10. Every attempt to separate redundancy from difficulty in α, and what each one lost. The load-bearing criterion is the two-class comparison, and no arm could compute it.

arm	intervention	cells that survived	why the rest did not
§H	measure the confound (g) directly	3 of 5	median *occ* = 0: the gold is absent, so no growth rate exists
§I	subtract a skill-free retriever	2 of 5, both composition	a lottery needs $> L/2$ of the record, so the saturation guard bites
§J	credit only exactly-once items	1 of 5, single-hop	the exactly-once subset falls below the floor of 10

The survivors are complementary, and that is the finding. §I keeps both composition dimensions and loses all three single-hop ones. §J keeps one single-hop dimension and loses both composition ones. The two repairs preserve disjoint halves of the comparison they were built to make, so neither can compute a between-class quantity and no pooling of them can either, because they do not overlap on a single dimension. The reason is the same quantity in both directions: where the record is redundant enough for a lottery baseline to be estimable, it is too redundant for an exactly-once subset to clear the floor; where the exactly-once subset survives, the record is sparse enough that a skill-free retriever needs half of it. The two-class reading of α is not merely unsupported on this benchmark; it is unmeasurable on it, and the obstruction is a property of the benchmark's redundancy profile rather than of any estimator. That is a prediction as well as a diagnosis: on a corpus whose redundancy can be dialled, the two failures should trade off against each other, and a corpus at neither extreme should let both repairs run on the same dimension. We have not built one.

A fact about the benchmark. On this deletion ladder longmemeval's exactly-once share is 0.8% to 36.2% by dimension. Any analysis that credits containment on *occ* ≥ 1 is measuring a quantity that grows with the record; any analysis that repairs it by requiring exactly-once is under-powered by construction on four of five dimensions. That is a property of the benchmark, not of our estimator, and it applies to work that never mentions resolving power.

Two apparatus failures, disclosed because the second was read as a result first. Run 1 failed its own gate on all five dimensions by 0.0004 to 0.038: the grader fit the seven deletion rungs plus the full-record rung, while XLEN.json fits seven. Run 2 passed the gate and reported $\alpha_1 = 0.000$ on all five dimensions with k^* flat at 1: a devastating result for §G, and an artefact. $R^{\text{str}}(k) = (a_{\text{complete}} - a_k)/n$ is two counts over one item set, and the grader had moved a_{complete} to the exactly-once rule while leaving a_k over the at-least-once set, so the numerator went negative. The gate could not catch it: P-T3-80 re-executes the old rule, under which the re-crediting step is the identity function, so a gate built by re-running what is being replaced is blind to any defect living only in the replacement. What condemned run 2 was an invariant, now checked in code over every rule, dimension, rung and budget before anything is graded: $a_k \leq a_{\text{complete}}$, because an item the rule excludes cannot be retrieved either. Both runs are kept in results/once/. We record this because "we saw the answer and judged it a bug" is indistinguishable from result-driven rejection unless the reader is shown the invariant that did the rejecting.

The forecast table frozen with this arm scored P-T3-80 and P-T3-81 as predicted, including the named dimension; P-T3-82 and P-T3-83 were forecast as holds and fails and neither was reachable, which is a third outcome the forecast did not admit.

K Is the class separation an artifact of gold multiplicity?

A reader asked the question this paper should have asked itself. Our instrument counts verbatim occurrence, and the deletion lever moves record length together with the multiplicity *occ* of the gold string. That is Prop. 3(a) and we prove it. If the benchmark's composition dimensions happen to have faster-growing *occ* than its single-hop ones, the exponent separation of §G would be a fact about gold multiplicity wearing the label of compositional depth. We measured it, and the association is there.

Table 11. The confound, measured. g is the growth slope of the median gold multiplicity, $\mathrm{d}\log occ/\mathrm{d}\log L$, over the same ladder the exponent is fitted on. It is already in XLEN.json and we had reported it without regressing α on it. temporal-reasoning has no g because its *occ* series is incomplete, which removes one of the two composition rows and is why nothing below is decisive.

dimension	α	g	*occ* small → large	class
single-session-user	0.287	0.228	2.0 → 5.5	single-hop
knowledge-update	0.326	0.460	2.0 → 12.0	single-hop
single-session-assistant	0.462	0.000	1.0 → 1.0	single-hop
multi-session	0.981	0.903	0.5 → 11.5	composition
temporal-reasoning	0.984	—	—	composition

On the four dimensions with a g, the correlation between α and g is $r = +0.765$ and the regression slope is $+0.635$. Removing that trend shrinks the class contrast from a raw gap of 0.622 to a residual gap of 0.195, both computed over these four dimensions. Reported over all five, including the dimension with no multiplicity series, the raw gap is 0.624, one half of a comparison taken on a cell the other half cannot include. Roughly two thirds of the separation we reported travels with gold multiplicity rather than with the class label, on four points, one of which is the only composition row left after the missing series is dropped (Table 11).

What we are entitled to conclude, and it is less than §G claimed. With $n = 4$ and a single composition row, this cannot separate the two accounts and we do not claim it does in either direction. What it does establish is that the confound is real, is in the direction that would manufacture our result, and is large enough to account for most of it. The exponent separation of §G should therefore be read as: on this corpus, under

this lever, the dimensions the benchmark labels compositional are also the dimensions whose gold strings multiply fastest as the record grows, and the two cannot be told apart at this width. Prop. 3 predicted exactly this: it is the proposition's own content applied to this paper's own headline. The experiment that separates them needs a corpus where *occ* can be held fixed while L moves, and Prop. 3 says no lever on a fixed corpus can do it, so it needs a corpus we build.

L The boundary function on a second corpus

Everything above is one corpus. Prop. 3 says an exponent is a functional of (task, lever, retriever), so the first question the paper owes itself is whether the boundary function $k^*(L)$ transports at all or only the machinery for computing it does. We answer it on a corpus chosen to be as unlike the first as the available data allows, at $0: no model, no judge, no API call. Criteria are frozen in prereg/E_XCORPUS_PREREG.md, committed before the graded runner existed, with the feasibility probe's numbers disclosed in it.

Table 12. Two corpora, one instrument. Slot B is InfiniteBench's longbook_qa_eng: narrative document QA, where the record is not a conversation and an event is a paragraph. The population is the 198 items whose gold string is present in some event. That figure is derived from the instrument companion's table rather than quoted from it, and the derivation is one multiplication a reader can check: it reports $n = 347$ admissible items for this cell with $V = 0.571$ of them containing the gold verbatim (Anonymous, 2026h, row 310b), and $0.571 \times 347 = 198$ (Anonymous, 2026h). We state it this way because the companion prints no complete column for this cell, and saying that it did would make the agreement circular, since the only 198 in that paper is a back-reference to this one.

	slot A (§G)	slot B, this subsection
corpus	longmemeval, dialogue	infbench longbook_qa_eng
an event is	a session turn	a paragraph
record size	≈494 events, 1021 chars each	≈2865 events, 290 chars each
items	417	198
lever ratio	62×	67.6×
α	**1.011**	**1.009**, 95% CI [0.893, 1.070]
φ^* range	0.400–0.417	0.273–0.317
φ^* spread	0.017	0.044, 95% CI [0.019, 0.186]

The form transports and the constant does not, which is the theorem's own prediction. P-T3-97 supported: the exponent's bootstrap interval sits at [0.893, 1.070], clear of the frozen floor of 0.80, so on a second corpus the required share again does not vanish as the record grows, so Prop. 2's $\alpha = 1$ branch holds in both. The constant is different: 0.30 against 0.41, a third lower. Under Prop. 3 that is exactly what should happen, because changing the corpus changes the lever's domain, and it is the sharpest evidence in this paper that "the budget is a constant share" names a family of statements indexed by the corpus rather than one number (Table 12).

The condition this transport rests on, named rather than assumed, and it is visible in the data. k^* is read off a term-matching retriever, and the instrument companion measures what that means: on NoLiMa, a corpus built so that the question and the record carrying the answer share minimal lexical overlap, the gold is present verbatim in a single record on every item and a term-matching reader still retrieves that record in 0 of 156 rows at every budget up to 32, so k^* does not resolve at all (Anonymous, 2026h). Restricting the population to items whose gold occurs verbatim does not protect against this, which is worth stating because it is the protection a reader would assume we had: both slots here are populated exactly that way, and NoLiMa satisfies the same restriction while defeating the reader completely. What separates the two slots from NoLiMa is not the containment but

the lexical overlap between a question and its own record, and that condition is not assumed here. It is visible in the fact that k^* resolves at all. A corpus where the condition fails does not return a different exponent; it returns no exponent, which is what NoLiMa returns. The pre-check a reader should demand before comparing any two k^* values is cheap and prior: show that the budget closes the containment gap on both corpora. It does on these two, at every rung the ladders reach.

The constancy criterion fails, on its own resolution. P-T3-98 not supported: the spread's interval reaches 0.186 against a frozen band of 0.150. The point estimate is 0.044 and would have read as flat; the interval is what refuses it. We report the refusal rather than the point estimate because the companion's rule is ours too (Anonymous, 2026j). Note that on slot A we did not compute this interval at all, so the constancy we claim there is on the same footing as the one we decline here.

Two design decisions the run forced, both recorded before it was graded. A feasibility probe on 20 items returned $\varphi^* = 0.500$ at every rung with a spread of exactly 0.000, and the reason was the ladder: rungs at $32, 64, \ldots, 2048$ put the record length on the same lattice as the k^* ladder, so φ^* could only take power-of-two values. A constant share read off an aligned ladder is the ladder's default, not a finding, and §G escapes it only by accident, since its lever counts sessions, so the resulting event counts are not powers of two. The rungs are now $30, 61, 123, 251, 503, 1009, 2029$. Second, the statistic of record here is the interpolated crossing rather than the ladder read-out, because the pre-flight showed a ladder-read φ^* can only move in factor-of-two steps, so a band for a constancy test would have to exceed the quantity being tested, so no number of items could decide it. That is repair (2) of §I, applied in advance for once instead of after a failure. The consequence is visible: the ladder read-out calls all seven rungs saturated under the $k^* \leq L/2$ guard, while the interpolated crossing puts φ^* near 0.30, comfortably clear of it. The guard was firing on the ladder's coarseness, not on a real ceiling.

M The theorem applied to exponents we did not fit

Prop. 3 is proved for a fixed corpus, where the levers are few enough to enumerate. The obvious objection is that this makes it a curiosity about small benchmarks. It is not, and the evidence is not ours: the most-cited disagreement in the scaling-law literature has the shape the proposition describes, and it took the field four years and a dedicated paper to name the quantities the lever had moved. We audit three published exponents here rather than citing them deferentially, because an exponent reported without its lever is the practice this paper is about.

Case 1: compute, and a gap of 0.38 that was read as an error for four years. Kaplan et al. (2020) report that the compute-optimal model size scales as $N^* \propto C^{0.88}$; Hoffmann et al. (2022) report $N^* \propto C^{0.50}$. Same nominal resource, same functional form, two exponents that are not close.

0.88 needs its accounting named, because the same paper carries a second exponent. Kaplan et al. (2020) also report $N^* \propto C^{0.73}$, for the compute required to reach the same loss at a sufficiently low batch size; both numbers appear as Kaplan et al. in Porian et al.'s Table 1, at 0.88 as published and 0.73 adjusted, and the literature quotes each of them. The reason we state this rather than picking one is that the difference between a paper's own two exponents, produced by changing how compute is counted at fixed compute, is Prop. 3 inside a single paper: the lever moved the accounting. It decides the headline number. Against Hoffmann et al.'s 0.50 the disagreement is 0.38 if one quotes 0.88 and 0.23 if one quotes 0.73, a factor of 1.65 between two descriptions of the same disagreement, chosen by which accounting a reader happens to have met. We report the pair and refuse to pick, because picking is the move this paper's reporting rule forbids: an exponent travels with its accounting or it does not travel. The literature does pick, in both directions, and Pearce & Song (2024) treat 0.73 as Kaplan et al.'s figure while the 0.88 framing is the one usually quoted as the scandal. Independently, Pearce & Song (2024) attribute the primary cause of the Kaplan–Chinchilla

gap to counting non-embedding rather than total parameters, combined with the small scale of the fit, and reaffirm Chinchilla's coefficients, a second decomposition that lands on a second accounting convention rather than on a second phenomenon. Two independent decompositions of one disagreement, both terminating in conventions, is stronger support for this paper's reading than either alone.

We do not rank that gap against our own, because ours is not decidable. Ranking 0.38 above our own between-lever gap would contradict Table 8: $|\beta_c - \alpha_c|$ runs from 0.055 to 0.685 over five dimensions, so two of the five point estimates are larger than 0.38, not smaller. The sentence was inadmissible before it was wrong, because every one of those five bootstrap intervals spans zero, by the rule this paper imports and applies to itself, a quantity graded not decidable may not stand on either side of an inequality (Anonymous, 2026j). What the two cases share is a mechanism and not a magnitude, and the mechanism is the whole point: in both, the disagreement was read as a question of which fit was right, and in both it resolves into what else the lever moved. The field treated this as a question of which team was right, and Hoffmann et al. (2022) themselves proposed that the cause was a difference in learning-rate schedule. That hypothesis is wrong, and the correct account (Porian et al., 2024) is an enumeration of what else the two levers moved:

Table 13. Prop. 3 in the training literature. The decomposition is Porian et al. (2024)'s, reproduced here for its structure rather than for its coefficients: every row is a quantity the compute lever moved in addition to compute, and correcting each one moves the exponent. Nothing in this table is our measurement.

what the lever also moved	exponent, corrected	why the lever moved it
— (as published, Kaplan et al., 2020)	0.88	
— (Kaplan et al.'s own adjusted law, same paper)	0.73	compute counted at a sufficiently low batch size
last-layer FLOP accounting	≈ 0.70	the cost model changes with the shape being scaled
learning-rate warmup duration	≈ 0.60	a fixed warmup is a different fraction of a short run than of a long one
scale-dependent optimizer tuning (batch size, learning rate, β_2)	≈ 0.50	the optimum of a hyperparameter moves with the scale the lever is sweeping
agreement with Hoffmann et al. (2022)	**0.50**	

The same shape has now been measured in the memory literature, on one variable at a time. Wang (2026) vary a single component of a memory pipeline on a fixed benchmark and report that swapping only the embedding model moves accuracy by 6.2 points, enough that "one variable flips the conclusion" about which architecture wins. That is Prop. 3 in the domain this paper's corpus comes from: the lever named in the abstract is not the only quantity the intervention moved.

Read the middle column as a lever decomposition and the episode stops being a dispute. Neither exponent is wrong as a measurement; they are measurements of different functionals, because "increase compute" was realised by two constructions that differed in a cost model, a schedule and a tuning protocol. That is Prop. 3's conclusion, that an exponent is identified as a property of (task, lever, retriever), or in this setting of (task, lever, recipe), arrived at from the other direction, empirically and expensively. The hardest part of our claim is the part their history confirms: naming what the lever moved is not something an author can be trusted to do from introspection, since the authors of one of these exponents guessed their own confound wrong.

Case 2: tokens, where the lever's second effect is now standard and still unnamed in the exponent. Muennighoff et al. (2023) run 400 training runs to 9B parameters and 900B tokens and find that up to roughly four epochs of repeated data are near-indistinguishable from unique data, after which the value of added compute decays to zero. Hence "more tokens" is two levers: unique tokens and repeated tokens, with different exponents past the fourth epoch. The finding is well known and the reporting convention did not change with it: a token count is still quoted without saying which lever produced it, which is the same omission our α would have if we quoted it without saying "deletion".

Case 3: the same data, refitted, and this one belongs to the companion. Besiroglu et al. (2024) refit the Chinchilla law from Hoffmann et al.'s own published evidence and report estimates inconsistent with two of that paper's three estimation methods, and confidence intervals so narrow they would require over 600,000 experiments against the fewer than 500 likely run. That is not a lever failure; it is a resolution failure, and it is the subject of the companion (Anonymous, 2026j). We list it here because the two failure modes are routinely confused: a fit that does not reproduce may be measuring a different functional (Case 1) or may be quoting a band its apparatus cannot support (Case 3), and the repairs are different.

> What we are and are not claiming about these three. We do not claim Prop. 3 proves anything about compute or tokens: it is proved for a fixed corpus whose levers can be enumerated, and at training scale the lever is a design choice with no closed enumeration. What transfers is the diagnosis and the reporting rule, not the theorem. Nor do we claim novelty for the decomposition in Table 13, which is Porian et al. (2024)'s result and we reproduce it as evidence, not as a contribution. What we add is the observation that the episode has a name, that the name is an identification problem rather than a replication problem, and that stating it costs one line in a results table: report the exponent with the lever that produced it.

N Three repairs and the subsets they rescue

§I and §J attempt the same repair three ways, subtracting a skill-free retriever, removing the ladder's quantum and controlling the gold's recurrence, and each survives on a different subset of the five dimensions. A reader is entitled to ask why a paper reports three failed repairs instead of the one that worked. The answer is that the pattern is the result: the subsets are disjoint, so there is no dimension on which the two-class reading has been tested with the confound controlled. Reporting the best of the three would have described an apparatus that does not exist.

O Where this changes what a benchmark should report

Three consequences, argued rather than measured, and marked as such.

(a) A benchmark that reports scores at several record lengths is reporting a lever, and should name it. The two levers here are both defensible and they disagree; a table of scores at 100K, 500K and 1M events does not say which construction produced the longer records, and by Prop. 3 that omission is not a detail. The minimum disclosure is whether longer records were built by drawing new ones or by growing existing ones, and what happened to the gold's multiplicity.

(b) A claim that a dimension "requires multi-hop reasoning" is a claim about α, and α is cheap. It needs a lexical retriever, a containment test and no model. Where a benchmark publishes its transcripts and golds, the exponent can be computed by anyone for the price of a BM25 pass, and where it comes out near zero the label is not supported by the corpus regardless of what a scored system achieves.

(c) The item counts a public benchmark ships are not sized for differences of exponents. 47 to 118 admissible items per dimension is ample for a level and insufficient for a slope

difference by a factor of 12 to 54. A benchmark intending its length axis to be analysable should say how many items per dimension survive its own admissibility filter at every length, which is a number none of the benchmarks we used reports.

P What remains owed

Three arms would sharpen this paper. None of them can move the boundary function $k^*(L)$ or the quantile spread that is this paper's finding, because none of them changes the estimand; they bear on how far the account transports, which is why they are named and priced here rather than run.

(1) Separating composition from cardinality. run, $2.49, and the expected outcome is the one that occurred. The parent ran this arm on 2026-08-15 (Anonymous, 2026i). Two families were cut from one random stream differing in exactly one sentence, an unordered qualifier against an exchange. The companion grades the result not decidable and this paper must import that grade, not the half of it that suits this paper. On the preregistered set of 200 probes the carriers differ by 0.130 against a band of 0.151 set from the statistic's resolution beforehand, which is supported; a later code review found 35 of those probes carry a decoy value record for the same (attribute, entity) pair, and on the 165 uncontaminated probes the same comparison gives $0.170 > 0.151$, which is not. Two defensible analyses of one dataset disagree and the companion prints both rather than choosing (Anonymous, 2026i).

Quoting the 0.130 alone would license the conclusion that the non-abelian construction is unnecessary and that the sightedness condition is a statement about cardinality; the 0.170 blocks it. The cardinality reading is therefore used below as an assumption whose test is unsettled, not as an imported result. Two details are still worth importing rather than paraphrasing: an earlier attempt at this comparison returned an apparent refutation that turned out to be a different confound in the generator, caught by reading the code rather than the number; and the same run gives the parent a task with $R_{\text{state}} > 0$, which is the first positively-directed measurement in either paper.

(2) A corpus whose redundancy can be dialled. Cost unknown, and it is a project. §J predicts that if the obstruction is the benchmark's redundancy profile, then on a corpus where redundancy is a parameter the two failure modes trade off and a corpus at neither extreme lets both repairs run on one dimension. That is constructible and currently an argument.

(3) The free re-run. §H's ladder re-run with the completion rate taken over the exactly-once subset at each rung costs $0 and is described where it is owed.

Artifact availability. Every preregistration, grader, ladder runner and result file named here exists in a repository released publicly after acceptance; until then a reader has this document's timestamp. Two things are reachable without it: the corpora the exponents are fitted on are public, and §F's quantile table is arithmetic on numbers this paper prints, so a reader who disbelieves the headline can recompute the exponent at every quantile from the page.

Q The resolution of the frozen bands

Three of the bands frozen for the arms in §4 are narrower than the resolution of the statistic they grade. They are listed here with their arithmetic so that a reader can check the step rather than accept the conclusion, and so that the verdicts in §4 can be read as not decidable where that is what they are. The audit itself, and the general rule, are a companion's (Anonymous, 2026j); nothing below is re-derived there.

P-T3-71, band 0.150, resolution 0.301: quantisation, and by construction. The statistic is a two-point exponent $\beta = \log(k^*_{\text{hi}}/k^*_{\text{lo}})/\log(L_{\text{hi}}/L_{\text{lo}})$ with k^* read off a ladder of powers of two. One step of k^* is a factor of two, so the smallest non-zero value β can take is $\log 2/\log(L_{\text{hi}}/L_{\text{lo}})$. The lever used was $10.05\times$, giving $0.693/2.308 = 0.301$. The band was

frozen at half its own quantum, and two of the five cells passed, which is the informative part: a criterion below its quantum does not fail loudly, it returns a mixture that reads like a partial result. Repair: a two-decade lever halves the quantum and costs one more rung.

P-T3-67, band 0.150, and P-T3-68, band 0.519; resolution 1.064 for both: sampling. Both grade an exponent fitted over the nested ladder. The resolution is the widest measured 95% bootstrap interval over items, [0.000, 1.064] on single-session-assistant. Against it a band of 0.150 is 0.14 of Δ and a between-class gap of 0.519 is 0.49. The interval was computed after both criteria had already been graded, from the same items, at no cost, which is the sequence this appendix exists to invert. Closing the gap at a 0.150 band needs the per-dimension variance reduced by $(1.064/0.150)^2 \approx 50$, i.e. the 12–54× item factor §I reports.

What this does not excuse. P-T3-60, P-T3-61, P-T3-62, P-T3-63, P-T3-65, P-T3-66, P-T3-69, P-T3-70, P-T3-72, P-T3-73, P-T3-75 and P-T3-78 are not on this list. Their bands clear their resolutions or their verdicts do not depend on a band at all, and every one of them stands as reported, including the four that ran against the registered expectation. An undecidability appendix is not a licence to re-read failures as inconclusive, and the count matters: three of the fifteen graded criteria in this paper were undecidable, and twelve were not.

R The exponent ladder in full

§G reports five exponents and the two class summaries they support (Table 14). This appendix prints the ladder they are fitted on, because an exponent over seven points is a claim a reader should be able to refit. Every k^* below is (1) at $\tau = 0.05$ on that dimension's own admissible items, computed from two counts over text and a retriever with no model, no judge and no API call. The quantisation is visible in the table itself: every entry is a power of two, so an exponent fitted through them can only move in steps of $\log 2/\log(L_{\text{hi}}/L_{\text{lo}})$, which is the companion's subject (Anonymous, 2026j) and the reason §I reports an interpolated crossing as one of its three repairs (Figure 11) (Figure 10).

Table 14. k^* at each rung of the deletion ladder, per dimension, with the OLS slope of $\log k^*$ on $\log L$ over the rungs clearing the saturation guard. Median record length L per rung is given in the header; it differs by a few events between dimensions because the ladder is built per item, and the per-dimension values are in ALPHA.json. Recomputed from results/alpha/ALPHA.json.

	k^* at median $L \approx$									
dimension	78	158	312	627	1258	2505	4891	α	class	usable
single-user	4	8	8	8	8	16	16	**0.287**	single-hop	7
know-upd	16	16	16	32	32	32	64	**0.326**	single-hop	7
single-asst	4	8	8	8	8	32	32	**0.462**	single-hop	7
multi-sess	32	32	256	256	1024	1024	1024	**0.981**	composition	5
temporal	32	16	64	128	256	512	1024	**0.984**	composition	7

Read down the two composition rows and the reason their exponents are near 1 is visible without the fit: k^* tracks L almost doubling for doubling. Read down the three single-hop rows and k^* moves two or three steps while L moves six. That is the whole of §G's separation, and Figure 4 is the reason we do not let it stand alone.

S Secondary arms and background

Three things belong to the record and not to the argument.

(1) The two-point estimate this paper withdrew, and why the section stays. §D's $\alpha = 0.299$ was read off two independently drawn record sets that are nested for 0 of 500 questions. The nested construction gives 1.011. Both are kept in the body because the sequence is the

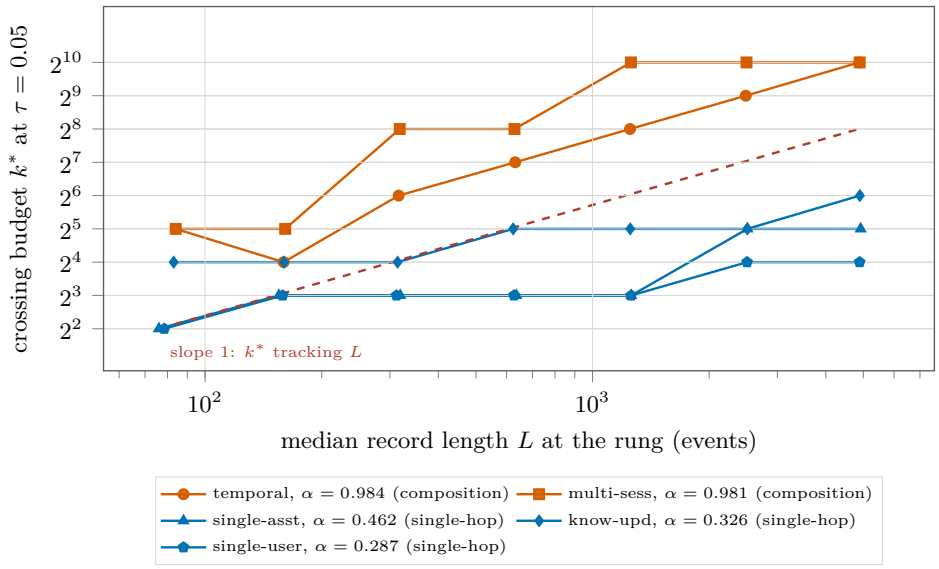


Figure 10. The ladder of Table 14, drawn, with the lattice it is read off. Every horizontal rule is a power of two, and every measured k^* sits on one, because the crossing is read off a doubling ladder rather than solved for. The class separation §G fits is legible before any fit: over six doublings of L the two composition dimensions climb five steps and the three single-hop dimensions climb two or three, and the dashed reference is what $\alpha = 1$ looks like. The same picture shows the estimator's defect: a slope through points confined to this lattice can only move in units of $\log 2 / \log(L_{\mathrm{hi}}/L_{\mathrm{lo}})$, which is the quantisation §I prices at 0.301 under P-T3-71 and the companion's subject (Anonymous, 2026j, row P-T3-71).

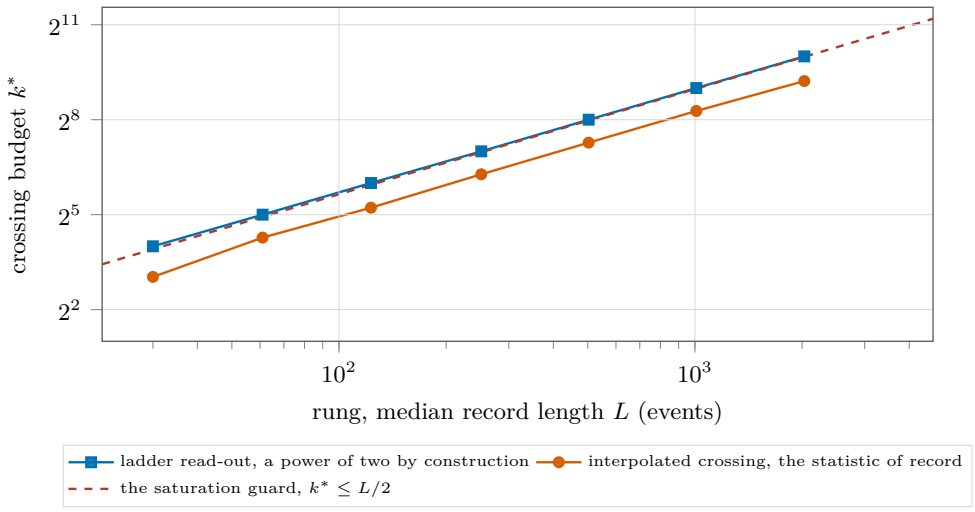


Figure 11. The second corpus's ladder, and the design decision §L took in advance because of it. The same rungs read two ways. Read off the doubling ladder, every rung sits above the saturation guard $k^* \le L/2$ and the arm has no usable rung at all, so the exponent could not have been graded; the interpolated crossing, computed from the same retrieval curves, clears the guard at every rung and puts φ^* in 0.273–0.317. The guard was firing on the ladder's coarseness rather than on a real ceiling, which is repair (2) of §I applied before a failure instead of after one, and the only place in this paper where that sequence runs the right way round.

finding, an unnested comparison, a registered criterion written to catch exactly that, and the nested measurement that fired it, and because deleting the first would leave a reader unable to see why the criterion existed. What is not kept is the reading: no claim in this paper rests on 0.299.

(2) The per-item statistic that needs no cell and no threshold. Reported in §D: the per-item crossing budget k_i rises with length for 9 of 36 matched items, falls for 1, and is unchanged for 26, at an exact sign-test $p = 0.021$. It is named again here only because it is the one reading in that section that does not rest on the single cell of $n = 17$.

(3) What a reader should know about the corpus that no verdict depends on. The deletion ladder's exactly-once share runs from 0.8% to 36.2% by dimension, so the subset on which multiplicity is pinned at 1 is small and unevenly distributed across the labels the class separation is about. The full admissible set is 47–118 items per dimension. Neither figure is a result; both are the reason the apparatus in §3 carries two grading disciplines instead of one.

T The budget-axis arm

§C states what this paper takes from this arm. The arm is here in full because a consistency check that cannot be inspected is not one, and it is here rather than in the body because its probe family, its scorer and its scoring rule are the parent paper's and that paper reports the same numbers (Anonymous, 2026i). Nothing in §E–§J depends on it: every quantity there is two counts over text and a BM25 pass.

T.1 The two restrictions in "index-only"

The two index-only readers of the parent paper (Anonymous, 2026i) impose two limits at once, and that paper treated them as one. There is a budget: the reader sees k=1 event. There is an addressability limit: the answer must be readable off a single event without combining any. That paper's blindness condition is about the second; its one-sidedness result's unenumerable quantifier, over $\binom{188}{8} \approx 10^{13}$ subsets, is about the first. V therefore bounds one axis while the one-sided audit quantifies over the other, which may be why the two behaved so difently: V held under every tightening of the protocol while the audit could never acquit. Its reported R_{state} is a difference along the diagonal of both.

The limit experiment separates them: pin the composition requirement and send the retrieval budget to infinity. Criteria are in prereg/E_BUDGET_PREREG.md, frozen before the runner existed and amended before any arm was run. The scoring is exact containment and there is no judge, as in the parent paper's witness family (Anonymous, 2026i).

Table 15. The score curve is a retrieval curve wearing a composition label. Same 60 composition probes, same model, same scoring rule; only the event set changes. "Coverage" is the share of probes on which the retrieved set contains both bearing events, the value statement and the swap announcement, and is a property of the retriever and the text, computed with no model. The last column is the score divided by that coverage. The whole tabular is generated from results/BUDGET.json, an artifact that did not exist until this table did: the grader printed these numbers on every run and persisted none of them, so the table was a hand copy of one run's output. Bold marks the best score a retriever the system has and the best reached at all.

arm	k	coverage	score	score/coverage
reader (lo, verbatim)	1	0.000	0.000	—
BM25 + model	2	0.383	0.350	0.913
BM25 + model	4	0.867	0.800	0.923
BM25 + model	8	1.000	**0.900**	0.900
full context	all	1.000	0.867	0.867
oracle pair	2	1.000	**0.933**	0.933

P-T3-37 is not supported, and the failure is the informative outcome. The frozen contrast was the oracle pair against BM25 at k=8, with a bar of 0.30. It comes out at $0.933 - 0.900 = 0.033$. Six irrelevant events alongside the two that matter cost the model almost nothing, so distraction is not the bottleneck; nor, given the last column, is composition. What binds is retrieval coverage, and the amendment named this branch before the run.

The score is coverage times a constant. Divided by the share of probes where both bearing events are present, every arm lands between 0.87 and 0.93. On this family $S(k) \approx \text{coverage}(k) \cdot \kappa$ with $\kappa \approx 0.9$, and the entire variation across the budget axis is the retriever finding or not finding a pair of sentences. The model composes them whenever it has them.

Which re-attributes our own witness, and refutes the reading we gave it. the parent paper's theory section (Anonymous, 2026i) built the probes as a non-abelian word problem and read their difficulty as ordered composition. That is not what this table measures. Composition is free once the records are present, so what the witness demonstrates is addressability, that no single retrieved record determines the answer, and the non-abelian structure is doing no work. The repaired sightedness condition of the parent paper (Anonymous, 2026i) states the requirement as cardinality for this reason. Whether the ordered structure is dispensable is not settled: the companion's matched arm has been run and it is graded not decidable, two defensible analyses of one dataset giving 0.130 and 0.170 against a band of 0.151 (§6). What this table licenses is therefore the weaker statement (on this budget axis, composition is free once the records are present), and the cardinality reading is carried below as an assumption. Note the direction: this weakens a construction we were pleased with, and it was found by reading our own budget sweep against our own theory section rather than by a new measurement.

The scope of that reading, restricted here rather than in a limitation. Every budget in this table is an absolute one on a record of fixed size, and k=8 reaching coverage 1.000 is what licenses the phrase "a small budget". §E measures what happens when the record itself grows: on nested records the budget that closes the gap is a constant share of the record, $\alpha = 1.011$, so a budget that is small relative to a 188-event probe set is not small relative to a 4913-event one. This section's finding is about a fixed record length and does not extend to the limit, and the composition dimensions are where it fails first, at $\alpha = 1.017$ and 1.022 against 0.001 for single-hop.

κ is not a new constant. It is ι, measured on a different axis, which turns this table into a test our own factorisation could have failed. the judge companion (Anonymous, 2026k) posits $S = a\,\iota$. What its deconfounding section discloses as constructed is $A = V\iota$, the repaired containment bound, so the three columns closing there is a definition closing, a disclosure about that bound and not about $S = a\,\iota$, which the same section tests against its own arm (Figure 12). Here a is varied on an axis the factorisation was not built from, the retrieval budget, and ι was measured independently by the extract arm at 0.917. Every arm's S/a lands within 1.40 binomial standard errors of that value (factorisation.py). The score being coverage times a constant is not a deflation of R_{state}; it is the first independent confirmation that scores in this apparatus factor at all.

Resolving power is not a number but a curve in the access budget, and after deconfounding the curve contains no system. Writing a_k for the coverage a budget-k retriever attains,

$$R_{\text{state}}(k) \;=\; \big(a_{\text{complete}} - a_k\big)\,\iota, \qquad R^{\text{str}}(k) \;=\; a_{\text{complete}} - a_k, \tag{2}$$

and both terms of $R^{\text{str}}(k)$ are counts over text and a retriever: no model, no judge, no system is evaluated. On the witness family $R^{\text{str}}(k) = 1.000/0.617/0.133/0.000$ at $k = 1/2/4/8$. This is V generalised from a single number to a curve, and it is the first quantity in this paper that predicts a score rather than only bounding one: at k=1, (2) predicts $1.000 \times 0.917 = 0.917$ against the measured 0.867, inside 1.4 standard errors.

R^{str} and $\widetilde{R}$ are two estimates of one quantity, and their disagreement is the residual rather than a clash of notation. The judge companion's deconfounding identity (Anonymous, 2026k) and its $\widetilde{R} = \widehat{R}/\hat{\iota}$ reaches the deconfounded resolving power through two scored arms; R^{str} reaches it through two counts and touches no scorer. They therefore agree in

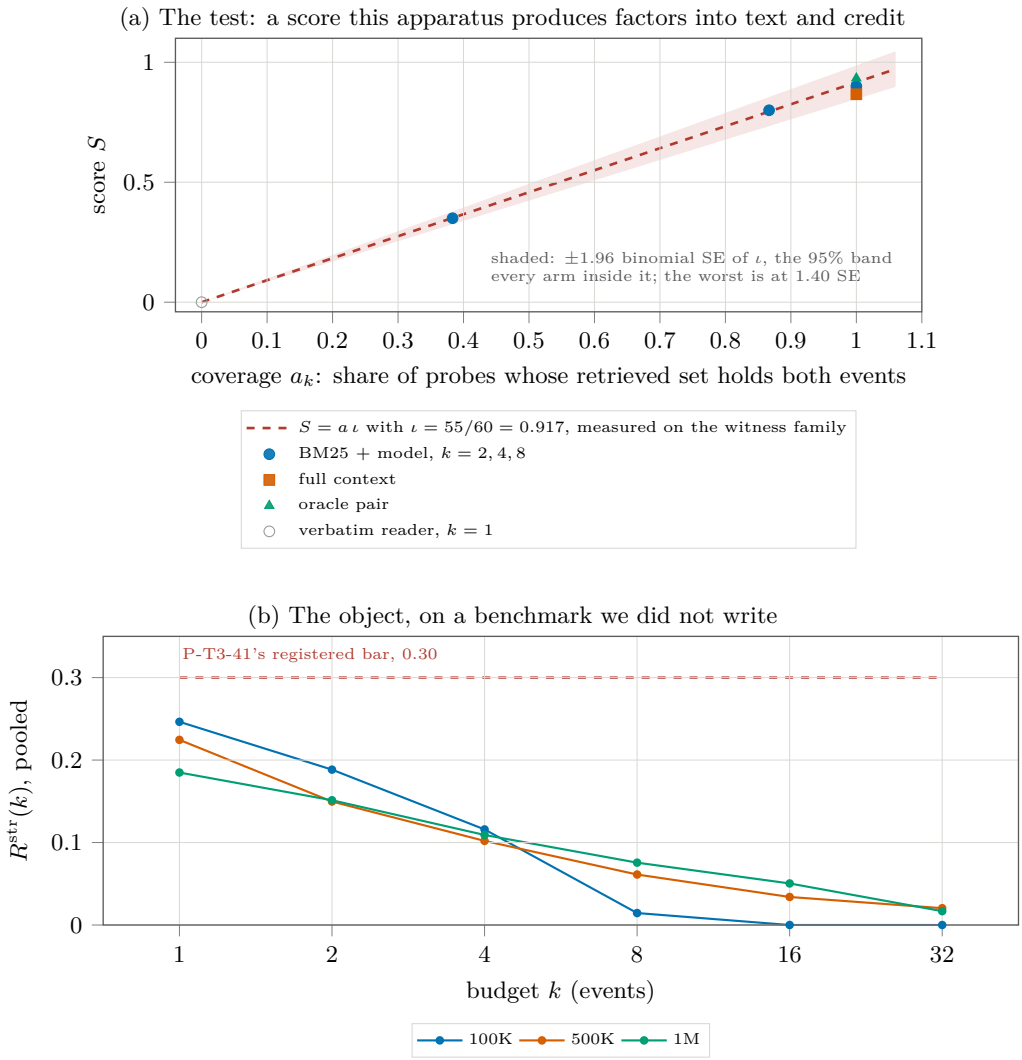


Figure 12. Resolving power is a curve in the access budget, and the factorisation it rests on is tested rather than assumed. Both panels are recomputed at build time from results/budget and results/witness; ι is counted from the witness grade files, not retyped, so the dashed line's slope is a measurement from a different arm and not a fit to these points. (a) The test. Coverage a_k, the share of probes on which the retrieved set holds both bearing events, is a property of the text and the retriever and involves no model, and the measured score tracks $a_k\iota$ across the whole budget axis. What the judge companion's deconfounding section discloses as constructed is $A = V\iota$, the repaired containment bound (Anonymous, 2026k), so the closing that identity reports is a definition closing; $S = a\iota$ is not that identity. Here a is varied on an axis the factorisation was never built from and ι comes from a separate arm, and every arm falls inside the ±1.96 binomial-standard-error band (the conventional 95% width, fixed independently of these points rather than sized to contain them), with the worst arm at 1.40 SE. This is the first occasion in this paper on which the factorisation could have failed, and the band is drawn at a width at which it could. (b) The object. $R^{\mathrm{str}}(k)$ is a difference of two counts over text and a retriever, so a resolving power at a stated budget can be computed with no system, no judge and no model, here on the parent paper's first outward attempt (Anonymous, 2026i), for the price of one BM25 pass. knowledge_update carries the largest gap on all three splits, against a criterion we registered predicting otherwise, and the curve stays above 0 at k=32 on the two larger splits: a fixed budget reaches less of a longer transcript, which is the one place this benchmark's headline axis moves anything.

expectation exactly when $S = a\iota$ holds, which is what the judge companion posits and this section is testing. On the witness family $\widetilde{R} = 0.945$ against $R^{\mathrm{str}}(1) = 1.000$, a gap of 0.055, the same 1.4 standard errors seen from the other side. We keep both symbols because col-

lapsing them would hide the one number in this section that could have come out large. A future apparatus in which they diverge is a falsification of the factorisation, and it would be invisible under a shared name.

The witness family's 0.867 (Anonymous, 2026i, row 2a) is the k=1 point of a curve, and the curve is the object. The index-only side is 0 there because coverage of the bearing pair is exactly zero at k=1, by construction of the family, since no single event holds both. That is not a defect in the demonstration; it is the demonstration, stated at the budget where the paper's instruments operate. What the curve adds is the reading a single number could not give: R_{state} is a joint property of a task and an access budget, and quoting one without the other is what makes "does this benchmark measure memory" unanswerable. A benchmark should report the budget its baseline was allowed, and (2) says what that costs.

A structural claim in the preregistration was false, and it was caught before the arms ran rather than after. The prereg argued that a query-matching retriever can never approach the gold-bearing event, because that event names a different entity than the question. It can: the attribute word matches ("the owner of Corvane" against "the owner of Palgrove"), and BM25 needs no more. Measured directly on all 60 probes at no cost, the share holding both bearing events runs 0.000/0.383/0.867/1.000 at $k = 1/2/4/8$, with 14.1 events per probe, so k=8 is already 57% of the transcript. The amendment is in the prereg with the table, committed before any arm ran, and it says what the frozen bar then discriminates: with both bearing events present at k=8 on every probe, the oracle contrast tests distraction rather than addressability. The bar did not move; what it separates did. One further difference is recorded there and matters for reading Table 15: our top-k takes the k highest-scoring events while the parent paper's reader (Anonymous, 2026i) takes the most recent among high scorers and answers verbatim, so the k=1 row is not that reader and the rows may not be joined into one curve.

U The exponent as a second reading

§3 reads the boundary off the required share $\varphi^*(L)$, which is a ratio of two counts and needs no fit. The exponent α is a second reading of the same ladder, useful because it states the dichotomy in one number and because the scaling-law literature is written in exponents. It also carries every estimator question in this paper, and they are collected here rather than in the body because none of them changes a verdict about the share (Figure 13).

We report $\alpha = 1.011$ and the estimand lives in $[0, 1]$, so the estimator is unconstrained where the quantity is not. Prop. 2 proves $\alpha \leq 1$, because $k^* > L$ is impossible: a budget exceeding the record retrieves all of it. An OLS slope knows none of that. Three readings and only one is available to us.

It is not evidence that the share grows. $\varphi^* \leq 1$ by construction, so no measurement can license $\alpha > 1$ as a statement about the world; an excess is estimator noise or estimator bias and nothing else.

The constrained estimate is the boundary. Maximising the same least-squares objective subject to $\alpha \leq 1$ returns $\alpha = 1.000$ exactly, a boundary solution, which is the correct formal report and carries no information the free estimate did not. We give both and use the free one for the interval, because a boundary-constrained point estimate with an unconstrained interval would be the worse of two conventions.

The direction of any bias is unresolved. Clauset et al. (2009)'s warning concerns tail estimation and does not transfer to a designed ladder, so we cannot import their bias correction; whether OLS on seven rungs of a quantised k^* is biased upward is a question about this estimator that we have not answered. What we can say is that the bootstrap interval [0.893, 1.070] on the second corpus contains 1 and excludes nothing interesting below 0.89, so the dichotomy's verdict, that the share does not vanish, does not depend on the excess.

An OLS slope on a log–log plot is exactly the estimator the power-law literature warns against, and we use it anyway for a reason we state rather than hide. Clauset et al. (2009) show that least-squares fitting on log-transformed data gives systematically biased exponent estimates for power-law distributions, and recommend maximum likelihood with a Kolmogorov–Smirnov goodness-of-fit test instead. Their target is different from ours in a way that matters: their estimand is the tail index of a distribution over samples, where the bias comes from the fluctuations of rare large events and from having to choose where the power-law region begins. Ours is a slope through seven designed rungs of a controlled ladder, one k^* per rung, with no tail to truncate and no lower cutoff to select. What does carry over is the second half of their warning, that a fitted slope gives no indication of whether the data obey a power law at all, and we answer it in the only currency this paper accepts: §I reports the bootstrap interval on α rather than the slope alone, and that interval is wide enough to make the between-class comparison undecidable, which is a stronger statement than a goodness-of-fit statistic would have made. The exponent's quantisation, k^* moving only in doublings, is a further defect of this estimator that neither literature covers, and it is the companion's subject (Anonymous, 2026j).

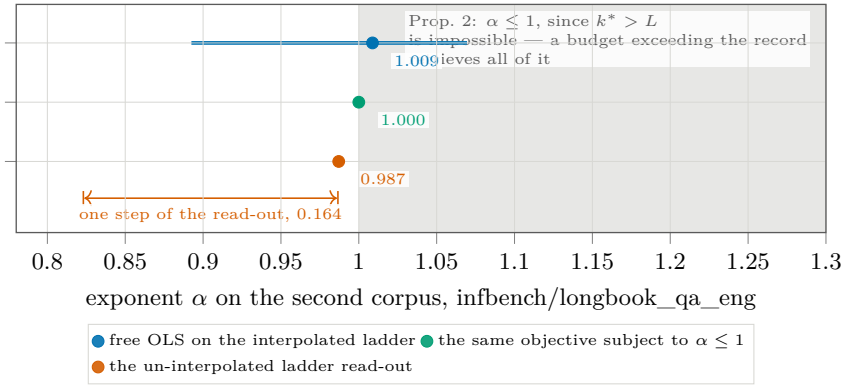


Figure 13. Three readings of one ladder, and a ceiling none of the estimators knows about. The shaded region is not improbable, it is unreachable: Prop. 2 puts $\alpha \leq 1$ because a budget exceeding the record retrieves all of it. The free slope sits inside it and its interval crosses the boundary, which is a fact about an unconstrained estimator and not evidence that the required share grows. The constrained estimate returns the boundary exactly, carrying no information the free one did not, and the un-interpolated read-out is drawn with the size of one step it can take, which is larger than the distance between all three readings and is why the interpolated crossing is the statistic of record. The verdict the section needs survives all three: the interval excludes nothing interesting below 0.89, so the share does not vanish on either corpus.

V 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
$E = (e_1, \ldots, e_L)$, L	a record as a finite ordered list of logged events, and its length	Def. 1
(q, y, E)	an item: question, gold value, and the record it is asked against	Def. 1
ρ, k, K	the fixed lexical retriever, its budget, and the doubling ladder the budget ranges over	Def. 1

Table 16, continued

symbol	what it denotes	defined
occ	multiplicity: how many events of E carry the gold string	Def. 1
n	the size of a dimension's containment-admissible subset, on which every quantity below is computed	Def. 1
a_k, $R^{\mathrm{str}}(k)$, V	reachable share at budget k; the structural miss curve; the same share over the whole dimension rather than the admissible subset	Def. 2
$k^*(\tau; L)$, $\varphi^*(L)$	the crossing budget at threshold $\tau = 0.05$, and the share of the record it is, k^*/L	(1)
α	the OLS slope of $\log k^*$ on $\log L$; equivalently $k^* \propto L^{\alpha}$	Def. 3
β	the same exponent measured over the undeleted lever, which does not reproduce α and reverses the ordering of the dimensions	§H
g	the OLS slope of log(median *occ*) on $\log L$: the licence test, and the figure where the price of the lever is paid	§H
Λ_L	a lever: a family of maps carrying each item's record to length L. Clean when it leaves both *occ* and the distractor set unchanged	Def. 4
D	the distractor set: the non-bearing events of an item	Def. 4
ι	credit rate, imported and attached to the scorer it was measured with	(Anonymous, 2026k, row 205)

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.

- D, a distractor set here; in Anonymous (2026f) it is pair heterogeneity, a Jensen gap in bits.
- Λ, a lever here; in Anonymous (2026f) it is label-predictability, $1 - H(e^{\star})$.
- α, the length exponent here; in Anonymous (2026f) it is the decidability level a lattice step must stay within.
- ρ, the retriever here; in Anonymous (2026h) it is the index-only class's internal budget spread.

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.

W Every criterion this paper names, with its freeze date

A preregistration's claim is a claim about time, so it is the one claim in this paper that must not be transcribed: the commit column below is read from git log --reverse when the paper is compiled, not copied from the file's own front matter. Amendments are counted and never overwrite. Naming is not grading: a criterion this paper cites for someone else's verdict appears here beside one it grades itself, and the claim ledger (claims_all.tex) is what separates the two. The programme's other arms are graded by companion papers and are not listed: a table of criteria this paper does not grade would say that it stands behind them.

Table 17. The preregistration record for every arm this paper names, generated from git at build time.

experiment	frozen at	criteria	amend.	words
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_BUDGET	aa605151, 2026-08-08 03:36	P-T3-37, P-T3-38, P-T3-39, P-T3-40	1	408
E_LENGTH	8d3ceb62, 2026-08-08 14:24	P-T3-55, P-T3-56, P-T3-57, P-T3-58	0	1512
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_RSTR	d9ed0cd8, 2026-08-08 10:30	P-T3-51, P-T3-52, P-T3-53, P-T3-54	0	1177
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_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
10 arms		45 criteria	1	11,189

X What is on disk

The arms above, counted from results/ at build time. Every figure and every table in this paper is generated from these files by make_t4_tikz.py and make_figdata.py; no coordinate and no cell is typed, which is why a stale artifact fails the build rather than printing a stale number.

Table 18. The experimental record, counted from the artifacts at build time.

arm	files	records	arm	files	records
alpha	1	11	once	4	33
baseline	2	20	rstr	2	10
beamcurve	2	7	sweep	2	24
budget	240	1,440	xcorpus	1	19
length	1	13	xlen	2	37
total	257	1,614			

Y 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 19. The scorecard in eight lines, counted from the scorecard. Every graded row of Table 21 and Table 22 falls in exactly one class below. Where a family's status names two classes at once, the row is counted in the more serious one, which is the only direction of error a summary of one's own scorecard may have. It is an index, not a substitute, and the rows are all still there: 21 claim families in Table 21, 52 sub-rows in Table 22. What the unflattering rows are for. 34 of the 72 rows carry a positive grade and 38 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	3	asserted, then taken back or re-assigned; the sequence is in the post-mortems
not supported / failed	8	a frozen criterion was missed, on our own data
self-correction	16	our own arithmetic or our own criterion was wrong and is repaired here
not decidable / not evaluable	3	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	3	proved from the definitions; no measurement can move it
measured	19	an artifact on disk, and a reader may dispute the artifact
argued / reported	12	no grade is claimed: an argument, or a number printed without a criterion
total graded rows	72	every numbered row of both tables

Table 21. 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
80. Most of the class separation travels with gold multiplicity, not the class label	self-correction; below: 2 not evaluable, 1 withdrawn	$r = +0.765$ on $n = 4$; raw gap $0.624 \to$ residual 0.195
81. Prop. 3's first sentence follows from Def. 4 alone	self-correction: true of every corpus, so empirically empty	clean fixes *occ* and D; changing L by deletion changes D by definition
82. k^* is a tail quantile and the headline is a tail statement	self-correction: the abstract sold it unquantified	α +0.898/ + 0.719/ + 0.109/ + 0.000 at τ = 0.05/0.10/0.20/0.50
$81^\dagger$. $\alpha = 1.011$ exceeds the estimand's range and is not evidence of anything	self-correction	constrained MLE returns exactly 1.000; $k^* \le L$ forbids more
$82^\dagger$. We borrow transportability's vocabulary and not its machinery	acknowledged limit	no selection diagram, no S-node, no do-calculus; the honest label is partial identification
50. The boundary function's FORM transports to a second corpus	measured, P-T3-97	$\alpha = 1.009$, CI $[0.893, 1.070]$ against a floor of 0.80

Table 21, continued.

Claim	Status	Evidence
30. Published exponents for one nominal resource differ by more than our between-lever gap	reported	$N^* \propto C^{0.88}$ against $C^{0.50}$; gap 0.38 (Kaplan et al., 2020; Hoffmann et al., 2022)
14. The boundary is a function $k^*(\tau)$, not a point	formal, curve measured blind; below: 1 not supported	(1); the parent paper (Anonymous, 2026i)
120. The per-dimension class separation is not a contribution of this paper	withdrawn; below: 1 withdrawn	licence fails on 4/5; detrending takes the gap 0.624 → 0.195; interval contains both composition estimates
121. The required share φ^* is the object; the exponent is a second reading	formal	$\varphi^* = k^*/L$ needs no fit; the estimator questions move to Appendix U
122. The budget-axis arm belongs to the parent paper and is reported here for audit only	acknowledged limit	Appendix T; no record-length quantity uses it
123. Kaplan's 0.88 and 0.73 are the same law under two compute accountings	reported	Porian et al. (2024) Table 1; an independent decomposition lands on parameter counting (Pearce & Song, 2024)
140. This paper's twenty abstract numbers are emitted from the artifacts, not typed	code-verifiable	macros read from results/ at build time; the resolving audit fails if one is replaced by its literal or if the abstract carries a three-decimal literal at all
141. The class gap's two halves were computed over different dimension sets	self-correction	raw over all five including the cell with no multiplicity series, residual over the four that have one; on the matched four the raw gap is 0.622 and the printed pair was 0.624 → 0.195
142. Figure 1 said the composition-against-cardinality arm was not run, at \$0–2	self-correction	§6 records it run for \$2.49, not decidable; the figure and contribution (vi) were ten days stale, and our price was low
143. Table 1 printed β = 0.301–0.601	self-correction	the maximum in xlen/XLEN.json is single-session-assistant at 0.604, which Table 7 prints two pages later
144. Figure 2 crossed out the wrong thing, three times	self-correction	the rule struck the verdict, so the picture asserted a clean lever exists; on the construction it struck Def. 4, which stands; on both lines, a term §3 uses. No line asserts existence, so no strike-through was right: it is a fourth arrow now, leaving L, reaching neither box, blocked
145. Figure 2 overstated Prop. 3 in two places	self-correction	it said each realisable lever "breaks one edge of the triangle" where the proposition says at least one of *occ*, D and the corpus; and it ruled the clean lever out flatly, "does not exist on a fixed corpus", where the proposition rules it out only at more than one length, which is the case that matters, since an exponent is a slope over a range. Both put back strength the scopebox removes
146. Figure 12(a)'s band was sized to contain its own worst point	self-correction	the shaded envelope was ±1.4 binomial SE and the furthest arm is at 1.40: the width had been set from the data it grades, so "every arm lies inside it" could not fail and, to the fourth decimal, was false. Redrawn at ±1.96, the 95% width, which the arms clear on their own

Table 21, continued.

Claim	Status	Evidence
147. Seven figures hid or blurred what they were drawn to show	self-correction	Figures ??, 6, 3, 4, 10, 11 and 13 each printed a note over the series it described, the worst hiding 485pt of curve, more than the width of its own axis; and Figure 8's legend rounded α to two decimals, so multi-session 0.981 and temporal-reasoning 0.984 both read 0.98 in the figure that introduces the pair Figure 9 then asks the reader to order

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 52 sub-rows are in Table 22 in full, of which 8 not supported, 6 acknowledged limit, 5 self-correction, 1 withdrawn, 1 refuted, 1 untested, 1 not evaluable. Nothing is dropped, and nothing is summarised into a kinder word than it had.

Table 22. Supporting refinements, moved from Table 21 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 52 rows below were moved here from Table 21 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.		
80a. "Separates at power" is withdrawn in favour of "orders, as point estimates"	withdrawn	every bootstrap interval overlaps (Fig. 4)
80b. $n = 4$ with one composition row cannot separate the two accounts	not decidable	the confound is real, in the manufacturing direction, and large enough
80c. …and this paper imported the companion's order-sensitivity arm on its preregistered half only	self-correction: the companion grades it not decidable	0.130 preregistered against 0.170 decoy-free, band 0.151; the cardinality reading is now carried as an assumption
81a. …and the identification statement that does carry content is measurable	measured on this corpus	α(task) identified iff a g=0 subset clears the floor; here 1/5 dimensions do
82a. …and at the median item the required share vanishes	measured, same ladder, no re-run	φ^* 0.0127 $\rightarrow$ 0.0002 at $\tau = 0.5$ over seven rungs
81a†. Whether OLS on a quantised ladder is upward-biased is unanswered	acknowledged limit	Clauset et al.'s correction is for tails and does not transfer
50a. Its CONSTANT does not: 0.30 against 0.41	measured	which is Prop. 3's own prediction, not a discrepancy
50b. Constancy of the share is not supported on slot B	not supported, P-T3-98	spread CI reaches 0.186 against a band of 0.150
50c. We did not compute that interval on slot A, so its constancy is on the same footing	acknowledged limit	the refusal here is stronger evidence than the claim there
50d. An aligned ladder manufactures a constant share	self-correction	feasibility probe gave $\varphi^* = 0.500$ and spread 0.000 at every rung
50e. The saturation guard fired on ladder coarseness, not a ceiling	self-correction	7/7 rungs saturated on the read-out; interpolated $\varphi^* \approx 0.30$

Table 22, continued.

Claim	Status	Evidence
50f. The second corpus costs $0 and needs no model	code-verifiable	BM25 and containment counts only
50g. The transport rests on a condition, and restricting to verbatim-containable items does not supply it	acknowledged limit	on NoLiMa the gold is verbatim in one record on every item and the reader retrieves it in 0 of 156 rows (Anonymous, 2026h, row 312a)
50h. …and the pre-check is cheap and prior: k^* resolving at all is the condition being visible	argued	a corpus where it fails returns no exponent rather than a different one
30a. That gap decomposes into three quantities the compute lever also moved	reported, not ours	FLOP accounting, warmup, optimizer tuning (Porian et al., 2024)
30b. The authors of one exponent guessed their own confound wrong	reported	the learning-rate-decay hypothesis is refuted (Porian et al., 2024)
30c. Prop. 3 is not claimed to prove anything at training scale	acknowledged limit	proved for a corpus with enumerable levers; the diagnosis transfers, the theorem does not
30d. "More tokens" is two levers with different exponents past four epochs	reported	400 runs to 9B params (Muennighoff et al., 2023)
30e. A non-reproducing fit may be a lever failure or a resolution failure, and the repairs differ	argued	Case 1 against Case 3, §M
14aa. At the full record a lottery never reaches τ	reported, P-T3-79	k^*_{blind} undefined, 5 of 5
14ab. Removing k^*'s quantum makes P-T3-71 decidable	not supported, P-T3-78	interval spans 0, 5 of 5; needs 12–54× items
14ac. The two repairs survive on disjoint dimensions	measured, three arms	blind 2/5 composition; exactly-once 1/5 single-hop
14d. k^* splits a benchmark's own labels	measured, labels not ours	8 / 16 / None / None
14f. k^* orders tasks by composition depth	post hoc, not graded	rstr_once.py
14g. The second record length P-T3-53 needed exists	measured, P-T3-55/56/57 supported	k^* 64 → 128; share 0.132 → 0.026; superseded by 14k
14h. The frozen hand-computation predicted the opposite sharply	self-correction, falsified in our favour	the share fell five-fold
14i. That prereg's description of its own instrument is false	self-correction	nested for 0 of 500; 4.8% shared sessions
14j. The graded k^* rests on one cell of $n = 17$	post hoc, not graded	one step is $0.059 > \tau$; 9 up / 1 down / 26 flat, $p = 0.021$
14k. On a nested deletion ladder the exponent is $\alpha = 1.011$	superseded: P-T3-60 refuted, P-T3-61 not supported	a constant ≈40% share; 14g's 0.299 misses by 0.712
14l. The typical item's budget moves with length	not supported, P-T3-63	median slope 0.000; 22 of 35 never change
14m. α orders dimensions by composition depth	post hoc, not graded	0.001/0.182/0.358/1.017/1.022; three cells below floor

Table 22, continued.

Claim	Status	Evidence
14n. α's class ordering, registered and powered	measured; P-T3-65/66 supported, untested as an exponent	0.287/0.326/0.462 vs 0.981/0.984
14o. It separates classes rather than ordering dimensions	measured, P-T3-68; not resolvable (P-T3-73)	spread 0.175, gap 0.519
14p. The licence was tested, not assumed	measured, P-T3-67 supported	holds to 0.104 against a frozen 0.15
14q. Registration bought power, not surprise	acknowledged limit, in the frozen file	the ordering was seen on five items first
14r. A third manipulation reproduces neither α nor its order	not supported, P-T3-71	β vs α_m: 0.301 vs 0.984; 0.601 vs 0.326
14t. The licence for reading α as a budget law	not supported, P-T3-72	$g = 0.460$, 0.903; $occ = 0$ at every rung on a third
14u. The class separation under item resampling	not supported, P-T3-73	bootstrap $[0.000, 1.064]$ contains both
14v. The short ladder's power	not supported for power, P-T3-70	0.79 decades; k^* moves one step at most
14w. α is the gold string's recurrence rate	not evaluable, P-T3-82	exactly-once 0.8%–36.2%; four of five below floor
14x. The 0.712 exponent conflict is composition of the graded set	measured, corrects our attribution	dimension 0.653, accounting 0.104, construction 0.059
14y. Subtracting a skill-free retriever leaves nothing for difficulty	measured, $n = 2$	$\alpha_{\text{adj}} = -0.107$, -0.051; P-T3-76/77 not evaluable
14z. Three dimensions have no blind baseline	reported, floor kept	$k^* \leq L/2$ bites by construction
140a. …and value-level substitution was refused twice before this	acknowledged limit	this paper prints 0.000 for a dozen unrelated quantities, so every key names a quantity and nothing is matched on a string
141a. …and the correction does not move the reading	measured	two thirds of the separation still travels with multiplicity growth; the section's conclusion was and is that the ordering is not measurable here
142a. …and no check could see it	code-verifiable	futurework_audit.py check (4) priced: a price inside a float, fragments included, must appear in the owed section. Restoring the old box fails the gate. Body prose is not covered: it reports spend as well as price

Table 22, continued.

Claim	Status	Evidence
144a. …and the exemption it needed is gone with it	code-verifiable	a deliberate rule had to be declared in DELETION_RULE by the exact text it crossed. Redrawn, the figure needs none: the table is empty, and no figure in the family tells the check to look away
146a. …and the factorisation clears the band it did not choose	measured	at ±1.96 SE the furthest arm is at 1.40, so the test is now one these arms could have failed and did not. This is the premise-gate rule applied to a figure: a band chosen from the data it grades is not a criterion
147a. …and the audit's own exemption was why	code-verifiable	hidden() calls any near-white fill opaque and NOTE is fill opacity=0.85, so a ground that hid the data also exempted the text on it from strikethrough. covers measures the points of non-grid path under a ground directly; a path crossing the box the short way is a label in a break in its own rule
2c. "Index-only" is two separable restrictions	measured; P-T3-37 not supported	0.900 at k=8, within 0.033 of full context; bar 0.30
2d. The curve's constant is ι	measured, on an unbuilt-for axis	S/a within 1.40 se of 0.917
2e. R_{state} is a curve, and its structural form has no system in it	formal	(??): 1.000/0.617/0.133/0.000