What a Rubric Judge Measures: Two Defects in a Shipped Scorer Change What It Measures, Not How Noisily

Anonymous authors
Paper under double-blind review

Abstract

Every number a long-horizon memory benchmark reports passes through a rubric judge. The LLM-as-judge literature treats that judge as a noisy instrument – measure the disagreement, rotate the judges, average the offsets away – which presupposes that it measures one latent quantity up to noise and bias. A judge can instead be perfectly reliable and measure the wrong thing. We audit one deployed benchmark's shipped scorer end to end, as code and as a measurement instrument, re-running it under controlled substitutions to measure what it refuses and what it grants. Two defects are readable in the released code: the judge prompt's question placeholder is never filled, leaving the rubric's only irrelevance clause anchored to a literal placeholder, and nine of the ten dimension evaluators truncate the rubric's own half credit through an integer cast. Both change the construct rather than the noise. Filling the placeholder collapses one dimension's score by an order of magnitude and reverses the sign of the correlation between answer length and score; removing the truncation flattens the largest apparent context-length effect in the audit. Handed the record containing the gold answer – the most favourable case any evidence layer can produce – the pipeline still credits only 45–63% of items, a shortfall traceable to a frontier model's reading step, not the scorer's leniency, while the same judge separately passes a no-model, one-event reader on nearly a fifth of rubric criteria, many of them satisfied by saying nothing. The credit rate then prices an inference auditors want, from a verbatim-containment count to a bound on resolving power: it is licensed only where a scorer never refuses a correct answer, false for every scorer we measure, so tighter rubrics keep the inference valid and make it unaffordable. One preregistered prediction of ours failed in the field's favour. The rule is to audit a judge as code before auditing it as an annotator: both defects here were readable, and fixable, without running a single system.

1 Introduction

A rubric judge sits between every system and every number a memory benchmark reports. The literature that studies this instrument has converged on one diagnosis and a family of remedies: judges disagree, so measure the disagreement, rotate the judges, and average the offsets away (Zhu et al., 2026; Yang et al., 2026; Norman et al., 2026). That programme presupposes that judges measure one latent quantity up to noise and bias. A judge can instead be perfectly reliable and measure the wrong thing, and the wrong thing can be legible in the scoring code before a single system is run.

We audit one shipped scorer end to end, the official evaluator of a deployed long-horizon memory benchmark, reading it as code, re-running it under controlled substitutions, and measuring both what it refuses and what it grants. The object is one benchmark; each mechanism is stated so that it is checkable on any other rubric-judged evaluation for the price of one re-scoring pass.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

What this paper contributes.

1. Two defects in released code, with measured magnitudes (§3). The judge prompt's <question> placeholder is never filled (F1), and nine of ten dimension evaluators truncate the rubric's own half credit through int() (F2). Removing the truncation flattens the largest apparent context-length effect in the audit to 0.000 at every scale.
2. The mechanism, not only the omission (§4). The unbound placeholder disables the rubric's only irrelevance penalty, so supplying the question flips the sign of the length–score correlation for both readers. This is construct substitution, distinguishable from judge variance by a saturation diagnostic an offset model cannot produce, and the standard remedy makes it worse: averaging two constructs yields a stable number that measures neither.
3. The ceiling on perfect provenance, and where it is not (§5.1, §5.2, §6). Handed the record containing the gold, the scored pipeline credits $\iota = 0.614/0.452/0.629$, a ceiling on what any evidence layer can buy. Partitioning the same items by whether the answer carries the gold string puts that shortfall on the reading step rather than the scorer, which credits a demonstrably correct answer 0.972 of the time. A separate leniency belongs to the rubric, and the two act on different items.
4. The step from a count to a bound, priced (§5, §7). The first link of $V \leq A \leq B$ is proved equivalent to $\iota \geq 1$ rather than observed to fail, the repair's looseness is measured on two instruments rather than bounded, and both accountings of the deconfounding price are given, because $1/\iota^2$ and $\Theta(1/\iota)$ answer different questions.
5. A registered prediction that failed in the field's favour (Appendix A). Scorer strictness could inflate item-response difficulty estimates; measured per item against frozen criteria the contamination is not shown, and items that discriminate better are items the scorer is more willing to credit, evidence for IRT-based adaptive selection.

The object, the arms, and one judge. The benchmark under audit ships 20 conversations and 400 probing questions per context split, ten dimensions, and an official evaluator that scores each response with a rubric prompt sent to a judge model (Tavakoli et al., 2025). The arms this paper scores are those of the companion instrument paper (Anonymous, 2026b): two index-only readers returning exactly one event and holding no model (last-occurrence, and BM25-last), a retrieval-plus-model arm (the reproduced ceiling proxy), a full-context arm, and two arms built for this paper, paste (return the gold-bearing turn verbatim) and extract (a model handed only the gold-bearing record). Two prompt modes are compared throughout: official, which replicates the released scoring path verbatim, and fixed, which fills the question placeholder and keeps float scores, changing nothing else. The default judge is pinned (gpt-4o-mini); §4 adds a second (gemini-2.5-flash) where the comparison needs it. Every criterion graded here was frozen in a preregistration file before the corresponding runner existed, and verdicts are reported in the prereg's own wording; the record is Appendix K.

The scorecard is in Appendix P. Every registered claim of this paper, its frozen criterion and its grade are there in one place, generated from a single source so a grade cannot differ between the count and the row.

2 Related work

Judge bias is a property of reading; both defects here are upstream of it. Zheng et al. (2023) name position, verbosity and self-enhancement bias; Wang et al. (2024) move a verdict by swapping candidate order; Panickssery et al. (2024) show a judge favouring its own text; Liu et al. (2023) propose a scoring procedure, form-filling with chain-of-thought and a probability-weighted sum, introduced because integer verdicts tie too many items, which is the collapse F2 produces by casting three rubric levels to two. Each result is a fact

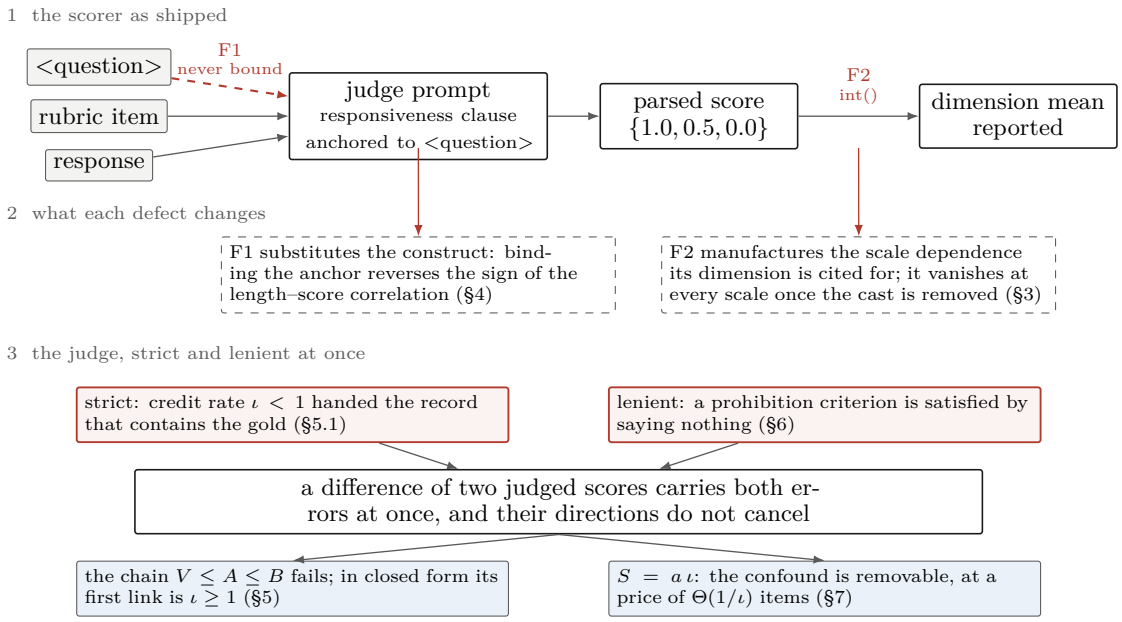

Figure 1. Where each finding sits, and what it feeds. Band 1 is the released scoring path; the two defects are at two specific edges of it (F1: <question> never bound; F2: the parsed score cast through int() on nine of ten evaluators). Band 2 is what each defect does to what is measured. Band 3 is the judge itself, the ceiling on perfect provenance, and the two downstream consequences a difference of judged scores inherits. No magnitude appears in this figure; it carries quantity names and section pointers only.

about what a judge does with the prompt it is handed. F1 is a fact about which prompt it is handed: the <question> placeholder is unfilled before the judge reads anything, invariant to order sensitivity, self-preference and rubric-following alike. Length bias belongs to the same line: a length-controlled estimator (Dubois et al., 2024) debiases a construct with no relevance term in it, since the length–score correlation changes sign once the placeholder is bound and nothing else moves ($+0.038/+0.167 \rightarrow -0.203/-0.194$, Table 2 in §4); no preference for verbosity predicts that, a missing irrelevance term does.

Variance control varies the judge around a fixed prompt; a placeholder never filled is invariant to all of it. Yang et al. (2026) move a score by changing only the evaluator; Norman et al. (2026) separate agreement from validity across 541K judgments; Zhu et al. (2026) cycle judges to cancel the bias term. Each holds the prompt fixed, so no cross-judge control reveals F1, and where two judges measure different constructs, averaging raises precision while lowering validity (§4). Two 2026 reliability instruments (Dev et al., 2026; Peng et al., 2026) vary only judge-side axes, so both return a reliable verdict on the dimension F1 governs, since reliability is agreement and an invariant defect is perfectly agreed; the same blind spot holds for prompt-perturbation work (Bellibatlu et al., 2026). The leniency direction is documented independently (Gonzalez et al., 2026; Weng et al., 2026); we add a mechanism producing leniency and strictness in one rubric at once, and the quantity ι carrying strictness into a downstream bound.

Reading a scorer as code. Biderman et al. (2024) report that scoring choices no paper states decide the number; ours is a defect in a released artifact at a named commit, with a magnitude measured against the arm it changes. Li et al. (2026) name silent-no-op perturbations, found at 100% on two of five benchmarks; that is F1 as a quantity, and their gate would have flagged it, which we state because it is against us. The statistical route to benchmark defects (Truong et al., 2025) cannot reach either defect: F1 is invariant across items, and F2's effect is absorbed as an item fewer systems pass, which a unidimensional

model reads as difficulty rather than a cast. Extended discussion (item-response contamination, generalisability theory, judge non-determinism, neighbouring benchmarks) and the full registered IRT experiment are in Appendix A.

3 The shipped scorer, read as code

Both findings of this section are readable in the benchmark's released code and reproducible from it, with no model and no API call.

What was audited. The scorer is src/evaluation/compute_metrics.py and the judge prompt is unified_llm_judge_base_prompt in src/prompts.py, both read at the repository's head commit 3e12035, dated 2026-02-02, after the benchmark's own publication. The rubric loop's substitution binds <rubric_item> and <llm_response> only; nine int() casts truncate the score on nine of ten evaluators, and the tenth (evaluate_event_ordering) uses float(). Both findings were disclosed to the benchmark's authors on 2026-08-24, naming the commit, the files and the lines, and offering the re-scoring harness before either went anywhere. They replied the next day, undertook a fix and declined the offer to withhold. The repair landed on 2026-08-31 as commit b2da22e. It repairs F2 and leaves F1 in place: the <question> line was deleted from the prompt's input block, but the two clauses that consume the question survive verbatim, and probing_question remains a parameter of every evaluator that never reaches the prompt. The first author's position is that "the rubric and the LLM's response should be sufficient"; §4 is the measurement that bears on it. Every magnitude below is of 3e12035; Appendix J(d) says what b2da22e now makes runnable.

F1: the judge never sees the question. The rubric-scoring loop substitutes only the rubric item and the response into the judge prompt; the <question> placeholder is never filled on any of the ten dimensions, so the prompt's own responsiveness clause is anchored to a literal placeholder. Its magnitude, on the one dimension where the official evaluator's float() lets it be isolated: filling the placeholder takes event_ordering from 0.222 to 0.003 on BM25-last and from 0.280 to 0.000 on last-occurrence (Figure 2, Table 1). evaluate_event_ordering scores this dimension with a normalised Kendall τ_b times an F_1 and attaches the judge's average as a further field; the benchmark reports the rank agreement, so F1 is isolable from F2 only on the one dimension whose reported score it does not set. A judge change alone (official prompt untouched) moves abstention by 0.925, the larger axis, kept apart from the placeholder effect throughout.

F2: half-credit is truncated, and its magnitude is length-coupled, which we registered the opposite of. The prompt defines 1.0, 0.5, 0.0; nine evaluators cast through int(), so 0.5 becomes 0. We froze the prediction that the fixed-minus-official shift per dimension would be stable within ± 0.05 across three scales on at least 8 of 10 dimensions; it holds on 4 of 10 for last-occurrence and 5 of 10 for BM25-last. P-T3-17 is FAILED, so every magnitude in this section is a 100K number. P-T3-19 held, and reading it explains the failure: on event_ordering, where F2 bites hardest, the corrected score is inside the frozen bar of 0.05 at every scale, while the official score moves across 0.204 of range — the defect manufactures the scale-dependence rather than inheriting it. Put beside the companion instrument paper's scale experiment (Anonymous, 2026b), this sharpens rather than adds: the largest apparent length effect in the whole audit, an index-only reader's event_ordering judge score falling $0.280 \rightarrow 0.076$ and rising to 0.167, reads 0.000/0.013/0.000 once the question is supplied. Full scaling detail is in Appendix E.

4 The unbound anchor and the sign flip

The prompt's only downward pressure is its responsiveness clause: "if the response does not address the QUESTION, score 0.0 and stop." With <question> unbound the clause has nothing to evaluate against, so nothing in the score can penalise irrelevance; the score is therefore monotone in how much text a strategy returns under the official protocol, and supplying the question must reverse that. Table 2 gives the rank correlation between the length of the returned turn and its score, over all 400 items, for both readers under both modes.

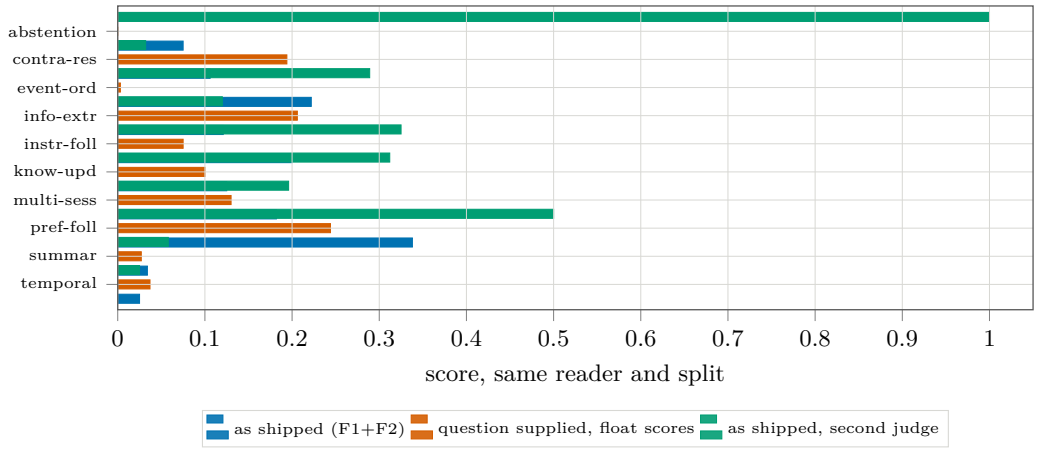

Figure 2. The shipped scorer measures a different construct, not a noisier one. One reader and one split, scored three ways. Blue against orange isolates F1 from F2: supplying the question moves event_ordering by 0.219, the one dimension already scored with float. Blue against green is the same readings under two judges: abstention moves from 0.075 to 1.000, and no additive judge offset can carry a bounded score from near zero to saturation, which distinguishes construct substitution from variance.

Table 1. Two readers × two prompt modes × two judges. official replicates F1 and F2 verbatim; fixed fills the question and keeps float scores.

	last-occurrence		BM25-last		BM25-last
dimension	official	fixed	official	fixed	(2nd judge)
abstention	0.150	0.000	0.075	0.000	1.000
event_ordering	0.280	0.000	0.222	0.003	0.289
knowledge_update	0.100	0.075	0.125	0.100	0.312
preference_following	0.487	0.281	0.338	0.244	0.500
contradiction_resolution	0.094	0.163	0.106	0.194	0.032

Table 2. Rank correlation between the length of the returned turn and its score, over all 400 items, for both readers under both modes.

reader	mode	ρ(length, score)	ρ(correct, score)
last-occurrence	official (anchor unbound)	+0.038	+0.132
last-occurrence	question supplied	**−0.203**	+0.104
BM25-last	official (anchor unbound)	+0.167	+0.066
BM25-last	question supplied	**−0.194**	+0.055

The sign flips for both readers, by 0.24 and 0.36. With the question supplied, returning more text is penalised, as the clause intends; with the anchor unbound it is not penalised at all and for the stronger reader it is mildly rewarded. The official protocol is a one-sided scorer: it detects the presence of rubric-satisfying content and cannot detect that the content is beside the point, the same asymmetry as the audit's own instrument (Anonymous, 2026b). It also explains two numbers elsewhere: our reproduction of the vanilla arm scored above the published vanilla by +0.194 and +0.297, and abstention reaches 1.000 under the second judge, since an abstention rubric needs the responsiveness clause to mean anything and a charitable judge can satisfy it with any text at all.

Why this is construct substitution and not judge variance. Judge-variance results (Zhu et al., 2026; Norman et al., 2026) presuppose that judges measure one latent quantity up to an offset. Two observations say that fails here. First, the mechanism: with <question>

unbound the scorer is not a noisy measurement of "did the response answer the question" but an exact measurement of "does the response contain rubric-satisfying text". Second, the shape: an additive offset on a bounded score cannot carry a dimension from 0.075 to saturation at 1.000 while preserving order, and saturation is the diagnostic. Completing the 2×2 grid (both readers, both judges, Appendix B) shows the deliberately weaker reader winning 6 of 10 dimensions under each judge, what the unbound anchor predicts rather than a judge property, while the judge does reorder two dimensions of ten, both informative rather than marginal (one saturates under the second judge, the other sits at the floor under both) — weak evidence about ranking, strong evidence the scale is not shared. If two judges measure different constructs, averaging produces a stable number that measures neither.

5 From a count to a bound

The companion instrument paper reports V, the share of items whose gold value appears verbatim inside a single logged turn (Anonymous, 2026b): a string property of the corpus, no model, no judge, no reader. An auditor wants to spend that count on a bound: if the answer sits in one turn, an index-only system can have it, so whatever separates two systems on that item is not a state channel, which would turn V into an instrument-free ceiling on the companion theory paper's resolving power (Anonymous, 2026c).

Definition 1 (Containment and blindness). For a dimension $\mathcal{T}$: V = the fraction of items whose gold value appears verbatim inside a single logged turn (a string property, no model or judge); A = the fraction answerable from a single retrieved record; B = the fraction on which $R_{\text{state}} = 0$.

Proposition 1 (The structural chain). If an index-only optimum retrieves the answer-bearing record whenever one exists, and returning it answers the item, then $V \le A \le B \implies R_{\text{state}}(\mathcal{T}) \le (1 - V)\, S_{\mathcal{T}}(\text{ceiling})$.

This is a partial-identification argument (Manski, 1995; 2003) whose first link, "returning that turn verbatim answers the item", is an assertion about a scorer rather than the corpus.

5.1 The first inequality, measured

We measured $V \le A$ on the subset where the question is defined: items whose gold appears verbatim inside some event. Handing the judge that turn scores 0.091 at 100K against a bar of 0.80 frozen beforehand — the turns are long and a rubric asking whether the response states the answer does not credit them. We bought the extraction step: one record, one model, the benchmark's own retrieval-answering prompt, same judge, same items (Table 3).

Table 3. The step from a count to a bound, measured. knowledge_update, fixed-mode judge; the short-gold containment subset where $V = 1$ by construction. Bars frozen in pre-reg/E_ORACLE_PREREG.md.

arm, all index-only unless noted	100K	500K	1M	frozen criterion
n items	22	42	35	$n \ge 10$ to grade
S(paste), return the gold-bearing turn	0.091	0.310	0.257	≥ 0.80: FAILED
S(extract), one record + a model	0.614	0.452	0.629	≥ 0.80: FAILED
S(ceiling), top-8 retrieval + a model	0.364	0.460	0.429	—

Extraction is worth up to +0.523, and it is not enough: on a subset where $V = 1$ by construction, A measures 0.45 to 0.63 even with a model handed the correct record, so $V \le A$ is false, and since $V \le B$ was only ever established through $V \le A \le B$, the chain breaks at its first link. (??) is unsupported, not false: $B \ge A$ may still hold, but the argument for it is gone. On 100K and 1M a single perfect record beats the reproduced ceiling arm (by 0.250 and 0.200, unregistered), so R_{state} cannot even be computed from this ceiling on this subset — consistent with the companion's judge-free retrieval hit rate on the same buried-gold items (Anonymous, 2026b). All of the above lives on $n = 22/42/35$, where one item is worth 0.045/0.024/0.029; full resolution accounting is in Appendix C.

5.2 Whose shortfall is it: the rubric's or the reader's?

ι is named the credit rate under perfect information. The 37–55% it does not grant could be a rubric refusing a correct answer or a reader failing to extract the gold; three controls separate them.

Control 1: change the reader's tier and nothing else. Pooled over 115 items the credit rate is 0.526 against 0.513 (mid- vs. frontier-tier extractor), a difference of 0.013 against a band of 0.15. Invariance under a stronger reader is not evidence the rubric is strict: the two tiers fail on the same items, Jaccard overlap 0.82.

Control 2: a generation budget that cannot bind. Both extract arms shared a max_tokens of 400 with reasoning models; raising it to 2000 confirms the repair took, but ι moves by only $-0.004/0.026$, inside the frozen band — an apparatus defect this paper would have had to report as a confound is measured and immaterial.

Control 3: partition the items by what the answer contains. Split on whether the answer carries the gold value verbatim. P-T3-119 passes on both tiers: the scorer credits a demonstrably correct answer 0.972 (frontier) of the time against a bar of 0.900, so the scorer's own refusal is only 0.028, not the full $1 - \iota$. What is missing is upstream: the answer carries the gold on only 0.461 of items, and on 0.339 of them the frontier reader states that a record containing the value verbatim does not. "Perfect provenance leaves 37–55% of correct answers uncredited" is withdrawn: the ceiling on evidence-tracing is set one step earlier than located. Full decomposition table and example transcripts are in Appendix D.

5.3 The repair and its limit

Proposition 2 (The repair, and what it costs). Since $R_{\text{state}} \leq (1-B)S \leq (1-A)S$, a lower bound on A suffices: multiplying the containment rate by the extract score and counting non-contained items as zero gives $R_{\text{state}}(\mathcal{T}) \leq (1-A)S_{\mathcal{T}}(\text{ceiling})$, $A \geq V \cdot S(\text{extract})$. Unlike Prop. 1, this bound is instrument-dependent.

On knowledge_update, $A \geq 0.374/0.290/0.346$ against $V = 0.61/0.64/0.55$: V exceeds the A it stands in for at every scale, so a bound from V alone is tighter than anything measured, and the repair is looser by 1.45–1.97× — its instrument-dependence is its content, since on this benchmark there is no instrument-free bound.

Definition 2 (Credit rate). $\iota = S(\text{extract})$, the rate at which the scorer credits an answer produced from the record containing the gold value verbatim. It is a property of the (benchmark, reader, scorer) triple, measured at $0.614/0.452/0.629$, so a refusal rate $1-\iota = 0.39/0.55/0.37$.

> $1 - \iota$ is a property of the (benchmark, scorer) pair with no system in it: any resolving power computed as a difference of scores is confounded by it. As rubrics tighten or gold answers lengthen, $\iota \to 0$ while every score remains perfectly computable: a benchmark can grow arbitrarily hard to be credited on while growing no harder to answer. A resolving-power measurement does not become vacuous as $\iota \to 0$; it becomes unaffordable (§7).

6 What the same judge pays for saying nothing

§5.1 measures the judge withholding credit under perfect information; the opposite direction asks what it pays a reader that cannot possibly know the answer. A reader's score carries a floor unrelated to information. On the eight dimensions where the gold is essentially never in one event ($V \leq 0.03$), the one-event no-model reader still passes 17.5% of rubric criteria, and 18.9% of those passes are on negation criteria ("the response does not recommend heavy dependencies") satisfied by saying nothing, concentrated on abstention (100%), contradiction_resolution (87%) and preference_following (31%).

Table 4 gives the pass rate on negation criteria by how much of the record the arm sees, on the eight dimensions where $V \leq 0.03$.

Table 4. Pass rate on negation rubric criteria, by how much of the record the arm saw. The eight dimensions where $V \leq 0.03$, official-mode judge.

arm	what it sees	negation criteria	pass rate
one-event dump, no model	one retrieved event	84	0.119
retrieval + model (ceiling)	top-8 chunks	42	0.333
full context (vanilla)	the whole record	42	**0.548**

The floor is real and it is not information-free. The rate rises monotonically with access, $0.119 \to 0.333 \to 0.548$ (gap 0.429 against a band of 0.15), so P-T3-108 falls in branch (b): the full-context arm earns its passes, negation criteria being simultaneously a floor and a real measurement that shrinks a resolving-power gap by 0.119 rather than 17.5%. A by-product re-judgement prices this judge's own non-determinism: at temperature zero, 29 of 800 shards change, and the ones that change change completely.

One judge, measured in both directions, and neither corrects the other. ι (withholding under perfect provenance) and the 17.5% floor (granting under zero information) act on different items with opposite effects on a resolving power, so a difference of scores carries both errors at once, and both are smaller than they first look: $1-\iota$ is almost all reading failure, and the floor's information-free part is 0.119 of 17.5%. This dependence of credit on wording rather than content is also the mechanism that would break the factorisation of §7 (Weng et al., 2026).

7 The factorisation $S = a\,\iota$

§5.1 reports the structural chain as unsupported because V exceeds the A it stood in for. That is not an empirical shortfall; it is an identity.

Assumption 1 (Separability of access from credit). For a system x, let $E_i(x) \in \{0,1\}$ indicate that x has the gold-bearing evidence on item i, and $C_i \in \{0,1\}$ that the pipeline credits an answer produced from it. Assume $\Pr[\text{scores} \mid E_i = 1] = \iota$, $\Pr[\text{scores} \mid E_i = 0] = 0$ for every system and item, with C_i independent of x given E_i. Then $S(x) = a(x)\iota$ and for an index-only system $a = V$, so $A = V\iota$.

Proposition 3 (The chain's first link, in closed form). Since $A = V\iota$ and $V > 0$, $V \leq A \iff \iota \geq 1$. The bound is false for every scorer with a nonzero refusal rate, overstating by exactly $V - A = V(1-\iota)$.

Proof. $V \leq A \iff V \leq V\iota \iff 1 \leq \iota$ for $V > 0$; and $V - A = V - V\iota = V(1-\iota)$. □

At the measured ι, this gives 0.236/0.350/0.204, matching §5.1's reported gaps to every digit — checked by deconfound.py and drawn in Appendix H (Fig. 8).

Proposition 4 (Deconfounding). Both terms of $R_{\text{state}} = S(\text{access-complete}) - S(\text{index-only})$ carry the same factor ι, so $\widehat{R} = R_{\text{state}}\,\iota$ and $\widetilde{R} = \widehat{R}/\iota$ estimates R_{state} itself, using the extract arm already built for §5.1.

Proof. Under Assumption 1 the two arms score $a_c\iota$ and $a_i\iota$ with the same ι, so $\widehat{R} = (a_c - a_i)\iota = R_{\text{state}}\iota$. □

Proposition 5 (Cost of the correction). $\widetilde{R} = \widehat{R}/\hat{\iota}$ is a ratio of estimators; by the delta method, $\text{se}(\widetilde{R})/\text{se}(\widehat{R}) \geq 1/\iota$. Holding true R_{state} fixed, both variances are $\Theta(\iota/n)$, so items needed for fixed precision grow as $\Theta(1/\iota)$: R_{state} remains identified for every $\iota > 0$ and becomes unaffordable, failing only at $\iota = 0$ exactly.

Two accountings are both real and answer different questions, $1/\iota^2$ pricing $\widetilde{R}$-precision against $\widehat{R}$-precision on two estimands whose scales differ by ι, and $\Theta(1/\iota)$ pricing the same thing under a perfect rubric (3.2× against 1.8× at the measured ι's); neither may be printed alone, both diverge as $\iota \to 0$, and neither is charged to any system. The correction changes what a future audit should report, not this one, since $\widetilde{R}$ needs a nonzero $\widehat{R}$ to divide and this audit established none. Full derivations in Appendix G.

8 Discussion, limitations and conclusion

Three portable checks, each one pass or less: bind the anchor and re-score, using the length–score sign as the tell; measure ι with one extract pass and partition it with none, splitting on whether the answer carries the gold string; count the rubric's prohibitions, a grep. This paper owns ι, its partition, the two code defects, the negation floor, and the priced factorisation; it does not own a resolving power (Anonymous, 2026c) or V (Anonymous, 2026b) (Appendix I).

Limitations. One benchmark, one judge pair; ι's reader-half is measured but not explained; the reproduced ceiling is ours, not the benchmark's best system; the IRT test uses three systems; the oracle subset is small and quantised ($n = 22/42/35$); and the factorisation is tested on one axis and family (Anonymous, 2026c;d). What remains open (generality of $\iota < 1$, a human bound on the reader's share, whether the negation floor is rubric-writing's in general, and what b2da22e newly makes runnable) is priced in Appendix J.

Conclusion. A rubric judge is not a noisy reader of a fixed construct; it is part of the construct. The scorer audited here never sees the question, truncates its rubric's half credit, and rewards saying nothing on absence criteria; under perfect provenance it leaves 37–55% of the score uncollected, almost none of it the scorer's fault. None of that is judge variance or removed by averaging judges, and all of it is visible for one re-scoring pass. The portable results: a diagnostic (bind the anchor, watch the sign), a quantity (ι, pricing a count-to-bound step at $\Theta(1/\iota)$ items), and a partition (split ι on the gold string, since almost none of it belongs to the rubric here).

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 scorer defects (F1, F2) are reproducible from the benchmark's own released code at the named commit with no artifact of ours; every other arm's preregistration, runner and result file is 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 P was frozen in a preregistration file before its data existed; the appendix states the counting rule and the freeze commit for each.

Ethics statement

The disclosure to the audited benchmark's authors (2026-08-24, naming the commit, files and lines, with the re-scoring harness offered before publication) is reported in §3; the authors replied within a day, repaired one defect, and declined the offer to withhold. No human data was collected; all items are the benchmark's own public conversations and probing questions. We are not aware of a further Code of Ethics concern raised by this work.

AI use statement

We report our use of generative AI tools against the categories of the ICLR 2027 AI Policy for Authors. Our use was confined to preparing the document. Tasks in the required-disclosure list: none. Generative AI was not used to generate synthetic data sets, to develop

the theoretical model or conceptual framework, to formulate the mathematical claims, to supply critical ingredients for the proofs or to write them, to propose or refine the hypotheses, to design the methodology or the experiments or to give feedback on them, to set or choose any pre-registered threshold, to implement methods, to clean or reformat any dataset, or to interpret the results. Every result reported here is computed by a script the authors wrote, and every number is checked against the result file it cites. Tasks in the recommended-disclosure list: generative AI was used on the manuscript itself. It condensed a longer version of this work into the present conference format, which involved drafting and rewriting substantial parts of the exposition; it edited for readability, formatted tables and references, and assembled and anonymised the supplementary archive. It was not used to discover the research topic, identify the gap, select the literature, or decide what the paper claims: the claims, the numbers they quote and the qualifications attached to them are the authors', carried over from the longer version.

References

Anonymous. MarginGate: Sparse margin-triggered verification for batch-invariant LLM inference. arXiv preprint arXiv:2605.30218, 2026. States that evaluation pipelines and LLM-as-judge setups assume a fixed prompt and decoding policy define a fixed computation, and that this fails in BF16 inference.

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, 2026a.

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, 2026b.

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, 2026c.

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

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, 2026e.

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.

Rohith Reddy Bellibatlu, Edward Raff, and Wenbin Zhang. JudgeSense: A benchmark for prompt sensitivity in LLM-as-a-judge systems. arXiv preprint arXiv:2604.23478, 2026. Hand-validated prompt-paraphrase pairs measuring verdict stability under semantically equivalent rephrasings; reports that model scale is not a reliable proxy for consistency. The prompt is the variable being perturbed.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. arXiv preprint arXiv:2405.14782, 2024. The

evaluation-harness maintainers on implementation-level irreproducibility: prompt formatting, normalisation and scoring choices that are invisible in a paper and decide the number. The closest prior statement of "read the scorer as code", made from inside the harness rather than against somebody else's released one.

Robert E. Blackwell, Jon Barry, and Anthony G. Cohn. Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores. arXiv preprint arXiv:2410.03492, 2024. States that not all models are deterministic at temperature zero with a fixed seed, and that few benchmark studies quantify the resulting uncertainty because repetition costs money.

Robert L. Brennan. Generalizability Theory. Springer, 2001.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. Science, 380(6641):136–138, 2023.

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.

Lee J. Cronbach, Nageswari Rajaratnam, and Goldine C. Gleser. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16(2):137–163, 1963.

Sunishchal Dev, Andrew Sloan, Joshua Kavner, Nicholas Kong, and Morgan Sandler. Judge reliability harness: Stress testing the reliability of llm judges. arXiv preprint arXiv:2603.05399, 2026. An open-source library that, given a benchmark and a judge configuration, generates reliability tests for binary judgment accuracy and ordinal grading. Four judges over four benchmarks; no judge is uniformly reliable, with consistency lost to text formatting changes, paraphrasing, verbosity and label flips. The perturbations are all on the RESPONSE; the scoring harness is not audited.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In Conference on Language Modeling (COLM), 2024. arXiv:2404.04475. Length bias in an automatic judge, regressed out with a length-controlled estimator that raises the correlation with Chatbot Arena to 0.98. The standing prior result on length and score, and the reason this paper's sign flip has to be reported as a MECHANISM rather than as a discovery of the correlation.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, João Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation. arXiv preprint arXiv:2502.06559, 2025.

Santiago Gonzalez, Alireza Amiri Bavandpour, Peter Ye, Edward Zhang, Ruslans Aleksejevs, and Todor Antić. QEDBENCH: Quantifying the alignment gap in automated evaluation of university-level mathematical proofs. arXiv preprint arXiv:2602.20629, 2026.

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.

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

Wei-Chieh Huang, Weizhi Zhang, Yuchen Wu, Yankai Chen, Eric Hanchen Jiang, Wooseong Yang, Yiwei Yang, Henry Peng Zou, Hanrong Zhang, Ying Nian Wu, Haolun Wu, Kai-Wei Chang, Philip S. Yu, Xue Liu, and Aylin Caliskan. Harness the memory: A holistic

evaluation of memory substrates in memory agents. arXiv preprint arXiv:2608.15008, 2026.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for AI evaluation? arXiv preprint arXiv:2607.15190, 2026.

Joint Committee for Guides in Metrology. Evaluation of measurement data — guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections). Technical Report JCGM 100:2008, BIPM, 2008. Distinguishes a random effect, which enters the uncertainty budget, from a recognised systematic effect, which must be CORRECTED and not reported as uncertainty (§3.2.3–3.2.4).

Erik Larsen. The instability of safety: How random seeds and temperature expose inconsistent LLM refusal behavior. arXiv preprint arXiv:2512.12066, 2025. 876 harmful prompts, four models, twenty sampling configurations: 18–28% of prompts show at least one decision reversal, and the stability index at temperature zero is 0.977.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023. arXiv:2303.16634. Form-filling with chain-of-thought, and probability-weighted summing of the score tokens introduced BECAUSE integer scores tie so many items – the same collapse F2 produces by casting, arrived at from the opposite direction.

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.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. arXiv preprint arXiv:2411.00640, 2024.

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

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.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2404.13076. Self-preference, with self-recognition as the causal handle.

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

Yangda Peng, Yunjia Qi, Hao Peng, Haotian Xia, Guanzhong He, Xintong Shi, Richeng Xuan, Songyuanyi Lu, Yixian Liu, Zhichao Hu, Yuhong Liu, Lei Hou, Bin Xu, and Juanzi Li. Can llm-as-a-judge reliably verify rubrics in agentic scenarios? arXiv preprint arXiv:2606.29920, 2026. RuVerBench: 2,458 instances in research and coding, measuring how prompt design, batching and majority voting change rubric-verification accuracy. Every axis is judge-side; the scorer's own code is not read.

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.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In Advances in Neural Information Processing Systems, Datasets and Benchmarks Track, 2024.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4486–4503, 2021.

Hiroki Tamba. Necessary but not sufficient: Temperature control and reproducibility in LLM-as-judge safety evaluations. arXiv preprint arXiv:2606.26185, 2026. Pinning temperature to zero across providers and model versions still leaves borderline items non-reproducible, and harnesses report single verdicts without accounting for grader variability.

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.

Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Ben Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in AI benchmarks. arXiv preprint arXiv:2511.16842, 2025. Flags invalid benchmark ITEMS from response-pattern statistics, at up to 84% precision for expert review over nine benchmarks. The method assumes a mean score summarises performance, hence a unidimensional latent construct with an expected range per item.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024. arXiv:2305.17926, 2023. Positional bias: swapping the order of two candidates moves the verdict, and the remedy is a calibration over presentation orders. Another judge-side perturbation, and invariant to an unbound placeholder.

Shihao Weng, Yang Feng, and Xiaofei Xie. Beyond accuracy: Policy invariance as a reliability test for LLM safety judges. arXiv preprint arXiv:2605.06161, 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.

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

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, and the sections below hold the extended related work and the full registered IRT experiment, the reordering grid in full, the oracle subset's scope and resolution, the third control in full (the decomposition and example transcripts), the scale experiment at every dimension, the extractor control against its own resolution, the factorisation geometry and cost figure in full, the overstatement identity table, what this paper owns and borrows in full, what remains open in full, the preregistration record, the notation table, and the public scorecard of every graded row. Reviewers are not required to read it.

A Extended Related Work

Table 2 (the length–score correlation) and Table 4 (the negation-floor pass rates) are given in the main text (§4, §6) and are not repeated here.

A judge is not a deterministic instrument, and the exception is not a variance term. Tamba (2026) pin temperature to zero across providers and versions and still find borderline items non-reproducible. Anonymous (2026) locates the violation in bf16 inference. Larsen (2025) report 18–28% of prompts showing at least one decision reversal across twenty sampling configurations, with a stability index of 0.977 at temperature zero. Blackwell et al. (2024) note that few benchmark studies quantify any of this, repetition costing money. We paid for one as a by-product of a re-judge we needed anyway (§6): at temperature zero, 29 of 800 shards change, 11 and 18 on the two arms. What we add is the shape rather than the rate. The shards that change change completely, so the quantity is a rare draw between two nearly disjoint outcomes rather than a variance that shrinks under averaging, and an interval computed as sampling noise prices it wrongly both ways: too wide for the 771 shards that do not move, too narrow for the 29 that do.

The neighbourhood, and a specification defect one level away. Long-horizon memory evaluation is young and crowded. Wu et al. (2025) build 500 questions over timestamped chat histories with five ability labels; Chhikara et al. (2025) and Xu et al. (2025) ship memory architectures and score them on LoCoMo; Huang et al. (2026) compare memory substrates; Microsoft Open Source (2026) releases another agent-memory benchmark this year. Almost all score free-text answers with a model, because a memory answer is a phrase rather than a label. That shared choice is what makes a scorer audit transferable and ι worth measuring on each of them, and §5.2 separates reader from scorer for one pair. We claim nothing about the scorers of these benchmarks, having read none of their code. Panthi & Abdelfattah (2026) observe that a conversational memory system stores several derived forms of the

same information, so the evaluator must choose which form receives credit; switching only that choice moves ndcg on 83.4–94.0% of shared queries and can reverse the ordering of competing architectures. Their credited target is a choice a benchmark author made and could have made otherwise. F1 and F2 are departures from the rubric the scorer itself ships. The two results compound: a benchmark whose ranking already turns on a specification choice is one whose implementation defects cannot be diagnosed by comparing rankings.

Generalisability theory is this paper's home, and F1 is outside what it can express. Cronbach et al. (1963) and Brennan (2001) decompose observed-score variance across facets and report a coefficient for a specified universe of admissible observations, on the premise that the facets are where the disagreement lives while the universe is fixed. F2 sits inside that frame as a rater-side effect with a magnitude. F1 does not. With the anchor unbound there is no responsiveness clause in the rubric being applied, so every observation is an observation of a different construct and the universe itself has moved. A coefficient computed over that universe is a correct number about a universe nobody intended, which is Norman et al. (2026)'s title with a mechanism under it and Bean et al. (2025)'s construct-validity question asked of a scoring script. Metrology says what to do about each: Joint Committee for Guides in Metrology (2008) separates a random effect, which enters an uncertainty budget, from a recognised systematic effect, which must be corrected and may not be reported as uncertainty. Both defects are recognised systematic effects, so the licensed response is a correction rather than a wider interval. That is also the boundary against the error-bar line: Miller (2024) and Blackwell et al. (2024) give the right treatment for a residual that is sampling, and an error bar around a truncated score is a precise interval about the wrong quantity. The checklist and reporting-reform lines sit further out. Reuel et al. (2024) assess benchmarks against documented best practice, Eriksson et al. (2025) review validity issues across the field, and Burnell et al. (2023) ask for per-instance reporting. All three would improve what a benchmark publishes about itself; none prescribes reading the scorer as code, and we know of no checklist item either defect would have failed.

Item response theory could confuse a strict scorer for a hard item. We predicted it would, and on this benchmark it does not. Polo et al. (2024) estimate full-benchmark performance from about 100 items. Hofmann et al. (2025) select items adaptively and report higher validity and lower variance on MMLU with fifty times fewer items. Zhou et al. (2026) rebuild benchmark analysis on IRT across tens of thousands of items. All of it rests on the IRT premise: each item has a difficulty parameter, each system an ability parameter, and the probability of a correct response is a monotone function of their difference, estimated from the joint pattern of scored responses. The premise requires that an item fewer systems pass is an item harder to answer. §5.1 measures a scorer whose credit rate under perfect information is $\iota < 1$. Such a scorer makes an item harder to be credited on while leaving it exactly as hard to answer. The item's difficulty parameter rises, and nothing in the response pattern separates the rise from a genuine one, because the estimator only ever sees scored responses. Separating the terms needs an item answered under conditions where the answer is known to be right, which is the index-only arm of §5.1. Psychometrics names the assumption at risk: an item whose parameters differ across groups holding ability fixed exhibits differential item functioning (Holland & Wainer, 1993), and the requirement that a measurement mean the same thing across the conditions compared is measurement invariance (Millsap, 2011). A scorer with $\iota < 1$ threatens invariance across scorers rather than across examinee groups, a facet the DIF literature does not usually vary.

That is an argument rather than a result. It was written down as P-T3-27–31 in prereg/E_IRT_PREREG.md and the grader committed, and measured it per item on 115 items of the containment subset at no API cost. The difficulty is the three-system $d_i = 1 - \overline{\text{score}}$ over bm25last, lo and the access-complete arm (Table 5). The prediction failed, and both halves of the failure are worth having.

P-T3-28 does not reach its bar, so the claim comes down. The sign is the one we predicted and the magnitude is not: $+0.126$ against a frozen bar of 0.30, at $p = 0.18$. Per the prereg this is recorded as possible in principle, measured here and not shown, and may not be restated as "tends to" or "risks". On this benchmark, with this scorer, we have no evidence that the IRT difficulty parameter carries scorer strictness.

Table 5. E-IRT, 115 items, criteria frozen before the data. d_i is the fraction of the three frozen systems that the scorer did not credit; ι_i is the extract arm's score on the same item, i.e. the scorer's verdict when handed the record containing the gold value. The partial correlation removes $\log(1+n_{\text{holders}})$ and $\log(1+n_{\text{events}})$ from both sides, because "gold verbatim present" does not mean "easy to answer" when many records hold competing values, the prereg forbids reporting the raw figure alone. P-T3-30 was declared weak in the same document that froze it, so it is recorded as untestable and no verdict rests on it.

criterion	quantity	value	verdict
P-T3-27	IQR(d_i) on the containment subset, bar $\geq 1/3$	0.333	supported
P-T3-28	partial $\rho(d_i, 1-\iota_i)$, bar ≥ 0.30 at $p < 0.05$	+0.126	not supported
	permutation p, 10^4 draws	0.18	
	raw ρ, reported and not the criterion	+0.206	
P-T3-29	$\bar{\iota}$(low discrimination) $-$ $\bar{\iota}$(high), bar ≥ 0.15	−0.230	not supported
P-T3-30	judge-swap, declared weak in advance	—	untestable

P-T3-29 came out the other way round. We predicted that items whose difficulty is scorer-induced would present as highly discriminating, so that adaptive selection would prefer them. The high-discrimination third of items has mean $\iota = 0.656$ and the low-discrimination third 0.426: items that separate systems better are items the scorer is more willing to credit, by 0.23, past the bar's magnitude in the opposite direction. That is not a null but evidence for the soundness of IRT-based adaptive selection (Hofmann et al., 2025) on this benchmark, obtained by trying to break it.

What survives. $\iota < 1$ is measured (§5.1) and does open a gap between scored and answering difficulty in principle; we cannot show that gap contaminates a difficulty estimate here. Jiang et al. (2026) reach an adjacent conclusion from inside the method, scalable estimators giving unreliable item-level inferences at benchmark scale, and our result is orthogonal to theirs rather than supporting it. ι is easy to mistake for a difficulty parameter and is not one. Rodriguez et al. (2021) bring item response theory to leaderboards, modelling latent subject skill against latent item difficulty and using the fit to decide what to annotate; their b_i is a property of the item, estimated from responses, while ι is a property of the (family, scorer) pair, measured by handing a responder the records that carry the answer and observing that it is credited only ι of the time. An item a perfect responder is refused credit on is not a difficult item, so the two quantities sit on different axes: IRT places systems and items on one latent scale, and ι says how much of that scale is the rubric rather than the task. A resolving power, being a difference of two scores, is measurable only to within $1-\iota$, and that bound is invisible to a difficulty model however well fitted. What remains is a method: the index-only and extract arms of §5.1 test the scorer-strictness confound on any benchmark, and on one benchmark the answer was no. Three systems give d_i only four values, which is coarse; §8 records it as the first thing a larger system pool would sharpen.

B The reordering grid, in full

The grid's reading has two parts, and the first is the larger effect.

The protocol inverts instrument strength under both judges. The deliberately weaker reader wins 6 of 10 dimensions under each. That is not a judge property but what the unbound anchor predicts, since a scorer with no term for irrelevance cannot reward the retrieval precision that distinguishes the two readers.

The judge does reorder, on two dimensions of ten (Fig. 3a), which refutes a pure offset model. Both flips are informative rather than marginal: abstention flips because both readers saturate under the second judge (0.950 and 1.000), the construct-substitution signature of the unbound anchor rather than a ranking signal, and contradiction_resolution flips at 0.050 against 0.032, between two arms both at the floor. A reordering that happens only at saturation and at the floor is weak evidence about ranking and strong evidence that the scale is not shared, which is the claim this section needs.

Table 6. Which instrument wins, under each judge, both in official mode. The two readers are ordered by construction (BM25-last adds term weighting to the same one-event, no-composition strategy) so a dimension where the weaker one wins is the scorer inverting instrument strength. The gemini column for BM25-last has 397 of 400 items; three judge calls failed after retries and are absent rather than imputed.

	gpt-4o-mini			gemini-2.5-flash		
dimension	LO	BM25L	wins	LO	BM25L	wins
abstention	0.150	0.075	LO	0.950	1.000	BM
contradiction_resolution	0.094	0.106	BM	0.050	0.032	LO
event_ordering	0.280	0.222	LO	0.306	0.289	LO
information_extraction	0.124	0.121	LO	0.207	0.120	LO
instruction_following	0.312	0.200	LO	0.475	0.325	LO
knowledge_update	0.100	0.125	BM	0.188	0.312	BM
multi_session_reasoning	0.140	0.182	BM	0.182	0.196	BM
preference_following	0.487	0.338	LO	0.650	0.500	LO
summarization	0.043	0.034	LO	0.066	0.058	LO
temporal_reasoning	0.013	0.025	BM	0.013	0.025	BM
LO wins on	6/10			6/10		
ranking flips across judges	2/10: abstention, contradiction_resolution					

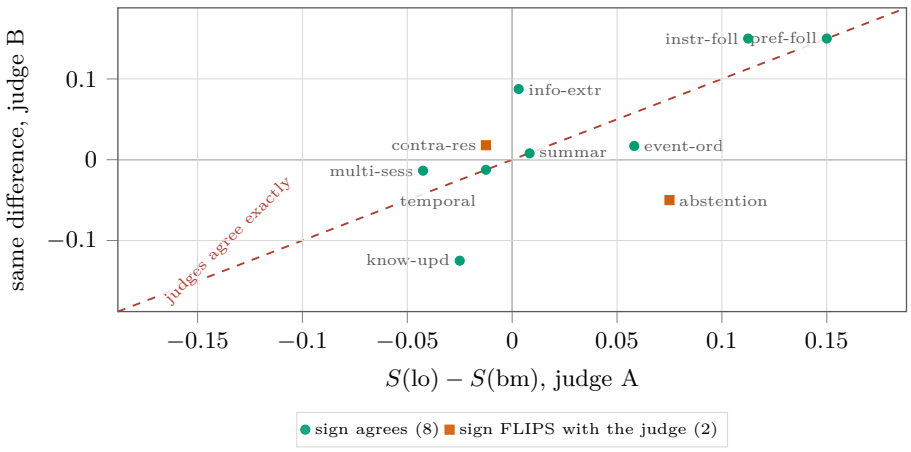

Figure 3. The completed grid of Table 6, and why its two flips are not evidence about the readers. The signed margin LO−BM25L under each judge, one point per dimension; the diagonal is exact agreement between the judges and the two orange points are the dimensions whose sign flips with the judge. The ordered pair of readers inverts under both judges on the same count (6/10), so the inversion is not a property of either judge. Both flipped dimensions place a cell inside the floor band (≤ 0.10) or the saturation band (≥ 0.90) of Table 1, and abstention is at the floor under one judge and saturated under the other, so margins of 0.02–0.05 there are places where the scale has run out rather than rankings.

The consequence for the remedy. If two judges measure different constructs, averaging them produces a stable number that measures neither: precision rises while validity falls, which is Norman et al. (2026)'s title. F1 is a mechanism that produces it, together with a cheap diagnostic: check whether the scoring prompt's anchor is bound, and if not, substitute and re-score. Here that check costs one re-run and moves one dimension by 0.219. It also bites the audit that produced it. Reliability is not a weaker form of resolving power and not a prerequisite we can borrow: a protocol on which one dimension's score swings by 0.925 with the judge cannot settle whether a task is blind, and no amount of variance control reaches the question. The ordering of reliability and validity is textbook and has been applied to LLM judges (Yang et al., 2026); ours is the mechanism, the diagnostic and the two magnitudes.

C The oracle subset: scope and resolution

An unregistered finding, marked as such. On 100K and 1M a single perfect record beats the reproduced ceiling arm, by 0.250 and 0.200 (Figure ??). Then $S(\text{ceiling}) - S(\text{index-only})$ is negative and R_{state} cannot be computed from this ceiling on this subset at all. This is consistent with the judge-free retrieval hit rate the companion instrument paper measures ($h = 0.136/0.262/0.200$, Anonymous, 2026b, row 313b) rather than in tension with it: these are exactly the items whose gold is buried inside a long turn, which is where BM25 over four-turn chunks fails. The audit uses this arm as a reproduced ceiling proxy, and this is the first measurement that locates where the proxy stops being one.

Scope, which the pre-registration made a criterion rather than a caveat. All of the above lives on the short-gold containment subset, $n = 22/42/35$: V's numerator, the items that give V its value, and not the denominator over which the 0.55/0.60/0.50 of (??) is computed, so none of these numbers may be read as that rate. Six further cells on two other dimensions have $n < 10$ and are reported without being graded.

The printed precision on those cells exceeds what they can express. One item is worth $1/n$, which at $n = 22/42/35$ is 0.045, 0.024 and 0.029, so the third decimal of any figure on this subset is not a measurement. However, $1/n$ is an upper bound on the step and not the step: the judge scores each rubric criterion in $\{0, 0.5, 1\}$ and an item's value is the mean over its criteria, so an item can land between 0 and 1 and these cells do, extract at 100K being $13\frac{1}{2}$ items of 22 and ceiling at 500K $19\frac{1}{3}$ of 42. The finest lattice any cell here sits on is a third of $1/n$, still an order of magnitude coarser than the decimal it is printed at. We keep three decimals only where the arithmetic downstream needs them and mark the item rather than rounding the tables, because rounding would hide which differences are one item and which are none. The rule is the criterion companion's own (Anonymous, 2026e), applied inward: the resolution of every cell on this subset is coarser than the precision it is printed at, and a difference should be read as a count of items rather than as a decimal.

C.1 Whose shortfall is it: the rubric's or the reader's?

D Control 3, in full: the decomposition and example transcripts

Control 3: partition the items by what the answer contains. The premise here is a construction invariant rather than a measured arm, which is the only kind of premise gate a manipulation cannot move: the record handed to the reader is the one the containment subset was built from, and P-T3-116 confirms it carries the gold verbatim on 1.000 of 115 items. Split those items on a string property of the answer: does it carry the gold value verbatim? Where it does, the answer is correct on the one operationalisation the benchmark's own rubric uses ("LLM response should state: X"), so a refusal there is the scorer refusing a demonstrably correct answer, which is the event $1 - \iota$ was named for. Table 7 is that partition for all four arms and Figure 4 draws it.

P-T3-119 passes on both tiers: the scorer credits a demonstrably correct answer, so the shortfall is the reading step and not the rubric. Where the answer carries the gold verbatim the scorer credits it 0.967 (mid-tier) and 0.972 (frontier) under the repaired budget, against a frozen bar of 0.900. The scorer's own refusal of a demonstrably correct answer is therefore 0.033 and 0.028, not the $1 - \iota$ of 0.46 to 0.49 that the arm as a whole returns. What is missing is upstream: the answer carries the gold on only 0.391 and 0.461 of items, and on 0.339 of them the frontier reader states in so many words that the record does not contain the information, from a record that contains it verbatim. The attribution "perfect provenance leaves 37–55% of correct answers uncredited" is withdrawn. What ι measures is the credit rate of the whole pipeline under perfect provenance, and §7 needs exactly that, since the factorisation carries whatever the two arms face in common. The ceiling on evidence-tracing stands at the same number and is set one step earlier than this paper had located it.

Table 7. Where the 115 items go, and what the scorer does inside each part. The same items throughout, the containment subset on which the record carries the gold verbatim by construction (P-T3-116, 1.000). The four counts are the items whose answer carries the gold verbatim, reports it absent in the reader's own words, returns nothing at all, and answers otherwise; the second is one regular expression fixed in analyze_iota_decomp.py and applied identically to every arm, and the third is an empty response or the literal [None]. The last two columns are the scorer's credit on the first group and its credit over all 115. The two repaired rows are P-T3-117–120's arms. The partition itself was not registered and is post hoc.

	items, by what the answer does				credit given	
extract arm	gold	absent	nothing	other	the gold	ι
flash, as published	47	5	38	25	0.968	0.526
flash, budget repaired	45	5	37	28	**0.967**	0.522
pro, as published	48	35	0	32	0.979	0.513
pro, budget repaired	53	39	1	22	**0.972**	0.539

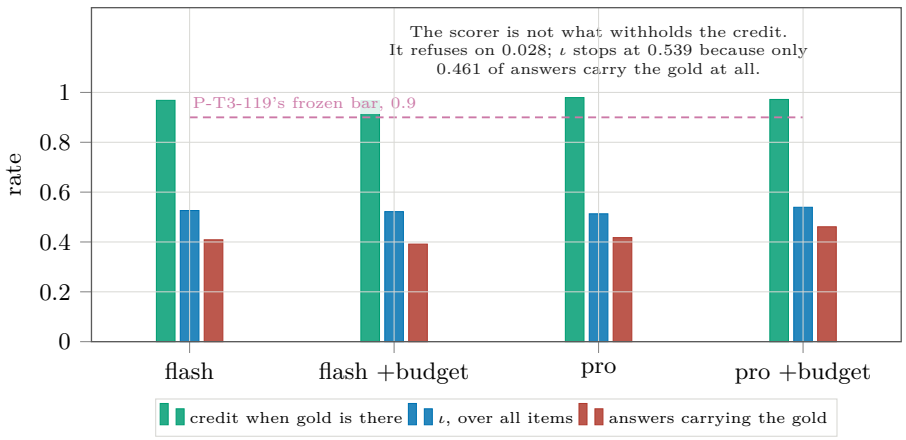

Figure 4. ι is low because the answer rarely carries the gold, not because the scorer refuses it. The three rates on the same 115 items, for the two extract arms as published and as repaired. The credit the scorer gives when the answer carries the gold clears P-T3-119's frozen bar of 0.900 on every arm and sits near saturation; ι over all items is barely half of it; and the third bar says why, because fewer than half the answers carry the gold at all. The gap between the first bar and the second is the reading step, and it is most of $1 - \iota$.

Table 8. One item from each part of Table 7, in the reader's and the judge's own words. The family's own extractor under the repaired budget. The selection rule is mechanical and is the first item in sorted (split, dimension) order within each part, so nothing here is chosen; the generator is refusals_table() and the source is results/extract2/IOTA_REFUSALS.json. Two rows are the scorer being right, which is the point: on the second the record's $50 occurs only inside $500, one of the 3 items of 115 (0.026) whose gold matches with no token boundary, so the reader declining is the reader reading correctly.

gold, and what the answer does	the reader's answer, from a record that contains the gold	the judge's own reason, at the score it gave
3 liters carries the gold 4105-char record	Aim for at least 2-3 liters of water daily, more if you are sweating heavily.	[0.5] The response suggests a daily water intake of 'at least 2-3 liters', which partially meets the rubric criterion of stating '3 liters'. However, it does not explicitly state '3 liters' as required. …
$50 reports it absent 287-char record substring-only match	I cannot answer this question as the information is not provided in the context.	[0.0] The response does not provide the required information of '$50' for the monthly budget for books and subscriptions. Instead, it states that it cannot answer the question due to lack of information, …
april 25 returns nothing 4904-char record	[None]	[0.0] The response is missing entirely, which means it does not provide the required deadline of April 25. Therefore, it fails to comply with the rubric criterion.
12 books by march 1 answers otherwise 5205-char record	12	[0.0] The response does not fully comply with the rubric criterion as it fails to mention the deadline of 'by March 1'. While it correctly states '12 books', it is incomplete without the specified …

Two limits of the partition, in both directions. A gold string absent from the answer does not make the answer wrong: a correct paraphrase lands outside that part, which is why its credit rate is reported and no bound is put on it. Equally, a gold string present does not make the answer right, because the containment test has no token boundary either way, which Table 8's first row shows ("at least 2-3 liters" contains "3 liters" and is graded 0.5). Both slacks widen the same part and neither touches the claim P-T3-119 makes, which is about the part where containment holds and is a statement that the scorer pays there.

E The scale experiment, every dimension and both modes

§3 grades P-T3-17 on the shift's spread and Fig. ?? shows the verdict. Two things it does not show belong here: the shift itself at each split, which is the quantity the criterion was written about, and where on the scale the six cells of the mode grid actually sit. The second matters for a claim §4 makes and cannot support from a five-row table: that the two dimensions whose ranking flips with the judge flip inside a band where the scale has run out (Figure 5).

Table 9. The fixed-minus-official shift at three context lengths, last-occurrence reader. The criterion was ± 0.05 on $\geq 8/10$; it holds on 4. The rightmost columns are the post-hoc reason it failed: the range each mode's own score covers across the three splits. The whole tabular is generated from results/judge/*/, which retires the last hand-typed measurement block in this paper, the one this paper's own audit named as such, and makes the emphasis a rule rather than a hand: the largest official range and the smallest fixed range, both of which fall on event_ordering.

	shift (fixed − official)				range across splits	
dimension	100K	500K	1M	stable?	official	fixed
abstention	−0.150	−0.243	−0.243	no	0.136	0.043
contradiction_resolution	+0.069	+0.057	+0.046	yes	0.026	0.048
event_ordering	−0.280	−0.063	−0.167	no	0.204	0.013
information_extraction	+0.005	−0.034	−0.055	no	0.033	0.066
instruction_following	−0.131	−0.148	−0.076	no	0.117	0.062
knowledge_update	−0.025	−0.057	+0.029	no	0.064	0.046
multi_session_reasoning	−0.016	−0.033	−0.029	yes	0.079	0.092
preference_following	−0.206	−0.124	−0.118	no	0.157	0.069
summarization	−0.024	−0.027	−0.012	yes	0.068	0.064
temporal_reasoning	+0.013	−0.036	+0.007	yes	0.038	0.015
mean				4/10	0.092	0.052

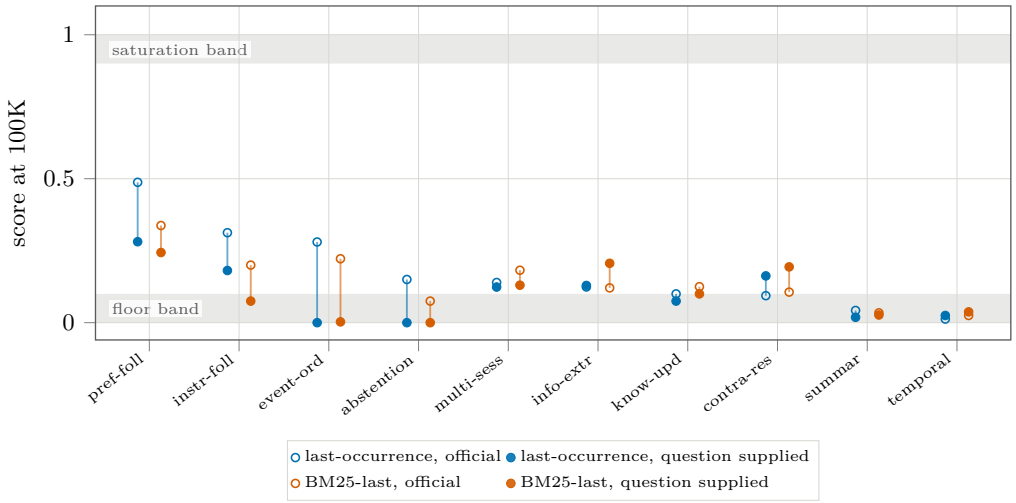

Figure 5. The mode grid at all ten dimensions, both readers, 100K. Open markers are the official protocol, filled markers the same readings with <question> supplied, and the segment between them is what the anchor is worth on that dimension. Table 1 prints five of these ten. The shaded regions are the floor (≤ 0.10) and saturation (≥ 0.90) bands: a margin measured inside either is a place where the scale has stopped resolving, which is the reading Fig. 3 needs. Every reading here is under the official judge, so the saturation band is empty in this figure and abstention appears only in the floor one; its saturated half is the second judge's, and it is Fig. 2 that draws that.

F The extractor control, against the resolution of its own statistic

§5.2 states the result of the one-substitution control: ι does not move when the extractor is replaced by a frontier model, so the credit that is withheld is withheld by the rubric. The control's numbers and the frozen band it was graded against are here (Table 10), together with the thing a table of three-decimal rates cannot print beside itself, the step size of the statistic (Figure 6). One item is worth $1/n$, and at $n = 26$ that is 0.038, larger than any move this control found; the judge's part credit puts the statistic on a finer lattice than that, so the figure prints the item and the caption counts in items rather than claiming the moves fall below a step.

Table 10. ι under two extractor tiers, everything else held. Same items as Table 3 (the short-gold containment subset), same fixed-mode judge, same prompt, 115 items pooled. The rows differ in one field of one request, and the whole tabular is generated from results/extract2/OX.json.

extractor	100K	500K	1M	pooled	
gemini-2.5-flash	0.596	0.412	0.632	0.526	the arm above
gemini-2.5-pro	0.577	0.392	0.632	0.513	this control
difference	−0.019	−0.020	0.000	**−0.013**	band 0.15

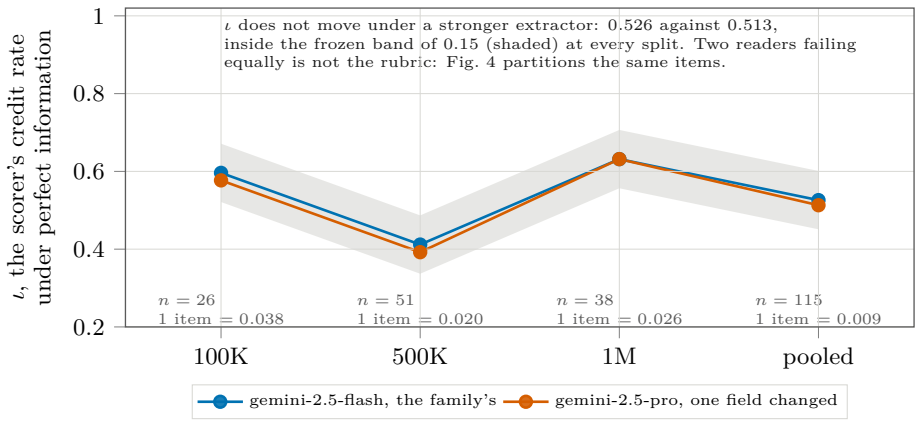

Figure 6. ι under two extractor tiers, everything else held. Same items as Table 3, same fixed-mode judge, same prompt; the rows differ in one field of one request. The shaded region is the frozen band of 0.15 around the family's own extractor. The largest per-split move is 0.020 and on 1M the two tiers agree exactly. Read in the unit the statistic is built from, the whole substitution moves the pooled credit by $1\frac{1}{2}$ of 115 items (half an item at 100K, one at 500K, none at 1M) against a frozen band of 0.15, which is 17 items. The moves are not below the instrument's step: an item is worth $1/n$ but the judge awards part credit, so the lattice is finer than the printed item and these are one to three of its steps. What the control shows is that they are an order of magnitude inside the band, not that they are invisible.

G The overstatement is an identity, not a shortfall

Prop. 3 says the gap §5.1 measures is $V(1-\iota)$ exactly. That is a stronger statement than three matching numbers, because it holds at every ι and not only at the three this benchmark happens to produce, and it is the reason (??) cannot be rescued by a better-behaved corpus. The two effective item counts the correction costs are tabulated here for the same reason §7 refuses to print either alone, and the identity is drawn over the whole range of ι rather than at the three values this benchmark happens to produce (Figure 7).

Table 11. ι read as an axis, and two effective item counts because the question has two answers. Recomputed by deconfound.py from the same numbers analyze_oracle.py prints. $n\iota^2$ is the count at which this subset's precision on $\widetilde{R}$ equals that of $n\iota^2$ items on $\widehat{R}$ under the same rubric, exact here, since $\widehat{R} = 0$ makes the inflation exactly $1/\iota$, but a comparison across two estimands whose scales differ by ι. $n\iota$ is the count a perfect-credit rubric needs to reach the same precision on the same estimand, and it is the accounting Prop. 5's $\Theta(1/\iota)$ rate belongs to; it is a leading-order scaling, not an identity.

split	n	V	ι	$V(1-\iota)$	reported $V-A$	$1/\iota$	$n\iota^2$	$n\iota$
100K	22	0.61	0.614	0.236	0.236	1.63	8.3	13.5
500K	42	0.64	0.452	0.350	0.350	2.21	8.6	19.0
1M	35	0.55	0.629	0.204	0.204	1.59	13.8	22.0
pooled, of 99							30.7 (3.2×)	**54.5** (**1.8**×)

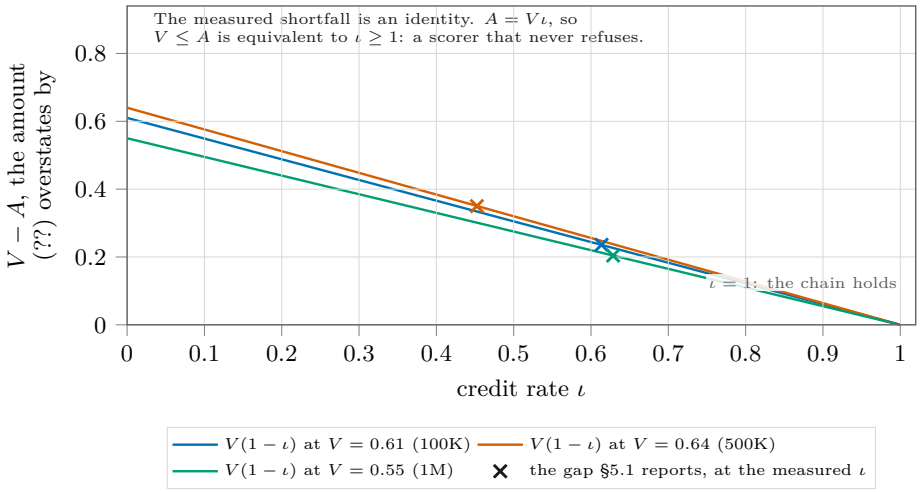

Figure 7. $V - A = V(1-\iota)$, at every credit rate and not only at ours. Each curve is the identity at one split's containment rate V; each cross is the gap §5.1 reports, plotted at that split's measured ι. The crosses land on the curves because the identity is exact, and the curves meet zero only at $\iota = 1$, a scorer that credits every correct answer. The assumption $V \le A$ was never an approximation the data violated: it is equivalent to $\iota \ge 1$, so it fails for every scorer with a nonzero refusal rate and its failure is proportional to the containment rate it was supposed to license.

H The factorisation geometry, and the cost figure, in full

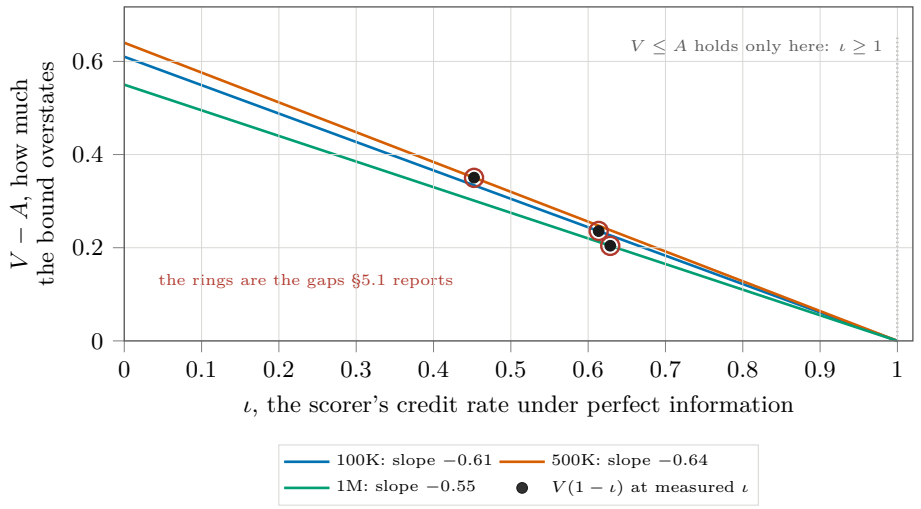

Figure 8. Prop. 3 as geometry, and its identity checked on the page. $V - A = V(1-\iota)$ is linear in ι with slope $-V$, one line per scale, and all three reach zero at $\iota = 1$. So the assumption $V \le A$ that (??) made is not an approximation our data happened to violate: it is the single point where those lines land, and what it asserts is a scorer that never refuses. The filled markers are $V(1-\iota)$ at the measured ι; the rings are the gaps §5.1 reports, which reach the paper by a different route, and they coincide at all three scales. One disclosure: $A = V\iota$ is how (??) was built, so the lines existing is a definition closing. That the reported gaps sit on them is not, and it is the half deconfound.py exists to check.

The main text states Prop. 4 (Deconfounding) and Prop. 5 (Cost of the correction); their full proofs, elided there for space, are given here.

Proof of Prop. 4. Under Assumption 1 the two arms score $a_c\iota$ and $a_i\iota$ with the same ι, because (??) makes credit conditionally independent of the system given the evidence; that

clause, and not Def. 2, is what the step needs. Hence $\widehat{R} = a_c\iota - a_i\iota = (a_c - a_i)\iota = R_{\text{state}}\,\iota$ and $\widehat{R}/\iota = R_{\text{state}}$. The step that does the work is the common factor; were the two arms scored by different rubrics the difference would not factor and no division would recover anything, which is why §7 insists that one scoring rule run across all arms. □

Prop. 5's delta-method variance is, in full,

$$\operatorname{Var}(\widetilde{R}) \;=\; \frac{1}{\iota^2}\Big[\operatorname{Var}(\widehat{R}) \;+\; \widetilde{R}^{\,2}\operatorname{Var}(\hat{\iota})\Big] \;+\; o(n^{-1}), \qquad \frac{\operatorname{se}(\widetilde{R})}{\operatorname{se}(\widehat{R})} = \frac{1}{\iota}\sqrt{1 + \frac{\widetilde{R}^{\,2}\operatorname{Var}(\hat{\iota})}{\operatorname{Var}(\widehat{R})}} \;\geq\; \frac{1}{\iota}, \quad (1)$$

with equality iff $\widetilde{R} = 0$.

The step that holds a fixed is an assumption, and it is the one to attack. Writing the scores as $a\iota$ and "holding the true R_{state} fixed" treats the rubric's strictness and the items' answerability as separate coordinates. A rubric can tighten in a way that also changes which items are answerable at all (adding a criterion that no response to a badly specified item can satisfy removes that item from both terms, moving a_c and ι together), and then the $\Theta(1/\iota)$ rate is a statement about a path through the (a_c, ι) plane rather than about ι alone. The factorisation check on the budget axis (Anonymous, 2026c;d) tests the separation on one axis (coverage) and one family; it does not test it on this one. Prop. 5 is therefore read here as a rate at fixed a_c, which is how it is used, and the joint case is not worked out.

Proof of Prop. 5. $\widetilde{R} = f(\widehat{R}, \hat{\iota})$ with $f(x, y) = x/y$, so $\partial_x f = 1/\iota$ and $\partial_y f = -\widehat{R}/\iota^2 = -\widetilde{R}/\iota$ at the true values. Both arms are independent means over the same item set, so the first-order expansion gives (1) directly; the radicand exceeds 1 unless $\widetilde{R} = 0$, which is the equality case. For the rate, let $S(\text{complete}) = a_c\iota$ and $S(\text{index-only}) = a_i\iota$. Then $\operatorname{Var}(\widehat{R}) = [a_c\iota(1-a_c\iota)+a_i\iota(1-a_i\iota)]/n = (a_c+a_i)\iota/n+O(\iota^2/n)$ and $\operatorname{Var}(\hat{\iota}) = \iota(1-\iota)/n = \iota/n+O(\iota^2/n)$. Substituting, $\operatorname{Var}(\widetilde{R}) = \big[(a_c + a_i) + R^2\big]/(\iota n) + O(1/n)$, so fixing se requires $n = \Theta(1/\iota)$. A binomial variance shrinks with its own mean, and that is the power which cancels: dividing by ι costs two powers and the vanishing scores return one. □

> The proposition prices two different questions, and both columns of Table 11 are real. Writing $\operatorname{se}(\widetilde{R}) = \operatorname{se}(\widehat{R})/\iota$, with a cost of $1/\iota^2$ items, treats ι as known; the extract arm estimates it, and at the values measured on the companion's witness family ($\iota = 55/60 = 0.917$ at $n = 60$, Anonymous, 2026c, row 105) the true inflation is 1.376 where $1/\iota$ gives 1.091, so the effective item count is 31.7 of 60 rather than 50.4. Moreover $1/\iota^2$ prices precision on $\widetilde{R}$ against precision on $\widehat{R}$, two estimands whose scales differ by exactly ι, whereas the cost of measuring the same thing under a perfect rubric is $\Theta(1/\iota)$, because scores shrink with ι and so do their variances. Both accountings are defensible, Table 11 carries both, and neither may be printed as though the question had a single answer. The delta-method variance was checked against $2{\times}10^5$ bootstrap draws and agrees to 1.009× (deconfound.py). One consequence is easy to miss: by (1) the price of deconfounding rises with $\widetilde{R}$ itself, so a benchmark with real resolving power pays more per item for the correction than one with none, while a subset whose $\widehat{R}$ is zero needs no variance adjustment at all and $1/\iota$ is exact there.

The two columns answer different questions, and the less dramatic one is the licensed one. $n\iota^2$ compares precision across two estimands whose scales differ by ι; $n\iota$ is the accounting Prop. 5's rate belongs to, and under it the loss charged to the rubric is 1.8× rather than 3.2× (Figure 9). Printing either column alone makes the table and the proposition look like they disagree. The conclusion needs neither choice: both accountings diverge as $\iota \to 0$, and neither is charged to any system.

> Vacuity is the infinite-sample corner of a finite quantity, not a loss of meaning. At $\iota = 0.05$ a difference of $R = 0.10$ still exists and still means what it meant, and

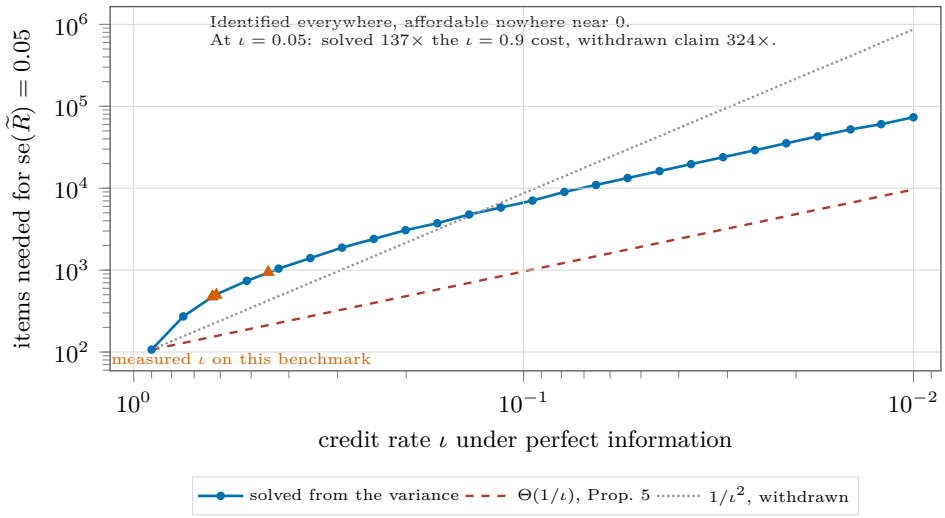

Figure 9. Identified for every $\iota > 0$, and unaffordable as $\iota \to 0$. Items needed for se$(\widetilde{R}) = 0.05$, solved from the delta-method variance of Prop. 5 at $a_c = 0.945$, $a_i = 0$; both axes are logarithmic and ι decreases to the right. This figure carries a retraction. The dotted curve is $1/\iota^2$, which an earlier version of deconfound.py asserted and which would have priced $\iota = 0.05$ at 324× the $\iota = 0.9$ cost; the solved answer is 137×, and its slope settles onto the dashed $\Theta(1/\iota)$'s by $\iota \approx 0.1$ rather than onto the dotted one's. Both reference curves are drawn through the $\iota = 0.9$ cost, which is outside the range where either rate is the whole of the variance, so what they carry is a slope and not a height: the dashed one reads 18× at $\iota = 0.05$ where the solve reads 137×, and the gap between them stops growing while the gap to the dotted one does not. The estimand does not move anywhere on this axis; only the bill does, and the pole at $\iota = 0$ is where it stops being payable.

reaching se = 0.05 on it needs $8{\times}10^4$ items instead of $2.5{\times}10^2$. What the correction does not do is rescue any verdict in the audit. $\widetilde{R}$ needs a $\widehat{R}$ to divide, and the audit of §3's benchmark established a nonzero one on no dimension; dividing 0 by ι gives 0. The correction changes what a future audit should report, not what this one found, and it does not touch the audit's one-sidedness either, because deconfounding a difference does not turn an existential instrument into a universal one (Anonymous, 2026b). The pole at $\iota = 0$ is the same shape as the one a companion paper (Anonymous, 2026a) finds at $N = P$, where a rule's leverage diverges exactly as it stops being testable; that both limits are poles of quantities each programme had been treating as a nuisance constant is the thing we would most like a reader to check.

I What this paper owns and borrows, in full

I.1 What this paper owns and borrows

Three quantities in this programme are easy to confuse, and the confusion would flatter this paper. This paper owns ι, its partition into a reading step and a scorer's refusal, the two code defects and their magnitudes, the negation floor and its informed controls, and the factorisation $S = a\,\iota$ with the price of removing the confound. It does not own a resolving power. R_{state} is defined and bounded in the theory companion; this paper never estimates one, and nothing here is evidence that any task's resolving power is positive or zero. It does not own V: the verbatim-containment count belongs to the instrument companion and appears here only as the left end of the chain $V \le A \le B$, which §5 shows does not hold. What this paper supplies to those two is one inequality and one refusal: a scored gap is measurable only to within $1 - \iota$, and a difference of two judged scores may not be read as a difference of capability until the credit rate beside it is printed. What it takes from

them is the definition of the chain's two endpoints and no graded verdict at all, a constraint papers/cross_grade_audit.py enforces by failing the build.

J What remains open, in full

J.1 What remains open

Neither arm below bears on the two defects this paper reports. F1 and F2 are readable in the released code and their magnitudes are measured here. What the arms below decide is how far $\iota < 1$ generalises beyond the (family, scorer) pair it was measured on.

(a) Is $\iota < 1$ generic, or is it this benchmark's? Every ι here is measured on one benchmark, one reader family and one rubric-based judge. We cannot say whether the shortfall that makes a score difference unmeasurable is a property of judged rubrics in general. We can say where each half must recur. The scorer's half falls below 1 wherever a rubric can withhold credit from a response that carries the answer, which a rubric with several criteria per item does by construction; here that half is 0.028 and it is the small one. The reader's half, the large one, recurs wherever a model is asked to answer from a retrieved record, and has nothing to do with judging. Benchmarks scored by exact match remove the first half and keep the second, which is why the companion instrument paper's counting route is immune to one and not the other (Anonymous, 2026b). The test costs one scoring pass and one string comparison per suite: an extraction arm against a judged arm on the same items, partitioned as in §5.2. Until someone runs it elsewhere, both halves are mechanisms with one data point.

(b) A human upper bound on the reader. The arm this section used to owe was hand extraction on the 115 items, because a tier substitution cannot rule out a shortfall two models share. §5.2 closes the part the paper needed: the scorer pays 0.972 when the gold string is produced, so a rubric strict about a phrasing neither model produces cannot be more than 0.028 of the gap. What a human arm would still buy is the other side, now a question about readers rather than rubrics: how much of the 0.339 on which a model reports a gold-bearing record empty is a hard record and how much is a model that does not look. That is one afternoon of reading 115 short records, it needs no API budget, and it bounds the reading step from above in a way no model substitution does.

(c) Whether the negation floor is this rubric's or rubric-writing's. §6 measures a floor of 0.119 that no information earns and a further 0.429 that access does. Both are one benchmark's rubric set. The mechanism is general and its size is not: any criterion phrased as a prohibition is satisfied by silence, so the floor exists wherever such criteria do, and how much of a suite's score it carries depends on how many there are, 84 across eight dimensions here. Counting them is free for anyone holding a rubric set.

(d) What the landed repair leaves open. F1 and F2 were disclosed on 2026-08-24 and the repair landed on 2026-08-31 as b2da22e. It repairs F2 and leaves F1 in place, and the authors dispute that F1 is one (§3). The repair does not retire §3. Every published BEAM number was produced by the scorer as audited, so the audit describes the record as it stands however quickly the code changes, and what a repair retires is the recommendation rather than the measurement. What b2da22e newly makes possible is the comparison this section could not run. With the int() casts gone, F2 no longer confounds F1 on the nine dimensions whose reported score is the judge's, so supplying the question at b2da22e would isolate F1 on dimensions whose published numbers it sets, which event_ordering is not. That is one scoring pass over the same items, hours and no model training. We have not run it, and the reason is a budget rather than a difficulty: the arms in this paper were bought at 3e12035 and a second scorer is a second purchase of all of them. Until someone does, F1's magnitude on a reported score is unmeasured, and this paper claims it on none. A reader meeting this section should read every magnitude in it as attached to 3e12035, which is why the commit is named wherever a number is.

K Every criterion this paper grades, with its freeze date

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

Table 12. The preregistration record, generated from git.

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

L Every dimension and split, graded and ungraded

P-T3-24's scope clause restricts the graded cells, and the restriction is material: of the 9 dimension-by-split cells in the containment subset, only 3 carry enough items to be graded,

and all 3 are the same dimension. The excluded cells are printed here in full with their values, because a scope clause that hides the numbers it excludes is indistinguishable from a scope clause chosen after seeing them. Ours was frozen before the data; the table lets a reader confirm that excluding them was not doing us a favour.

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

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

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

M The graders' verdicts against the grade the ledger carries

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

Table 14. Every verdict a grader prints, against the grade its criterion carries in the claim ledger. The second column is the grader's own stdout; the third is looked up in claims_all.tex at build time. The vocabularies differ because each preregistration froze its own words, so only the class comparison is meaningful, and a differ in the last column fails the build.

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

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

N What is on disk

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

Table 15. The experimental record, counted at build time.

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

O 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
V	the fraction of items whose gold value appears verbatim inside a single logged turn. A string property of the corpus	Def. 1
A	the fraction answerable from a single retrieved record	Def. 1
B	the fraction on which $R_{\text{state}} = 0$: the state channel separates nothing	Def. 1
$V \le A \le B$	the chain from a count to a bound, and §5 measures where it breaks	§5
ι	credit rate: the rate at which the scorer credits an answer produced from a record containing the gold verbatim. $1-\iota$ is the rate at which it refuses a correct answer, and $\iota = 1$ is the assumption the chain silently needs	Def. 2

Table 16, continued

symbol	what it denotes	defined
$S = a\,\iota$	the factorisation the repair rests on: coverage times credit rate. Constructed for the repaired bound, which is why the theory companion tests it on an axis the construction never used	§7
a	coverage: the share of items whose gold the reader actually retrieved	§7
$\widetilde{R} = \widehat{R}/\hat{\iota}$	the deconfounded resolving power, reached through two scored arms	§7
R_{state}	resolving power, imported: what a scored comparison is a proxy for	(Anonymous, 2026c)

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.

- A, an answerable share here; in Anonymous (2026c) it is a realised comparator arm.
- B, a blindness share here; in Anonymous (2026c) it is a realised comparator arm; in Anonymous (2026e) it is a criterion's band.

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.

P The scorecard, every graded row

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

Table 17. The scorecard in eight lines, counted from the scorecard. Every graded row of Table 19 and Table 20 falls in exactly one class below. Where a family's status names two classes at once, the row is counted in the more serious one, which is the only direction of error a summary of one's own scorecard may have. It is an index, not a substitute, and the rows are all still there: 17 claim families in Table 19, 46 sub-rows in Table 20. What the unflattering rows are for. 37 of the 62 rows carry a positive grade and 22 an unflattering one, and the second number is not a defect count: each of those rows narrows a claim made above it, names a number of ours that moved, or marks a question this apparatus cannot decide. None of them is an experiment left undone: work nobody has run is not graded here at all, and is bounded where it is stated.

grade	rows	what the grade means in this paper
withdrawn / superseded	0	asserted, then taken back or re-assigned; the sequence is in the post-mortems
not supported / failed	9	a frozen criterion was missed, on our own data
self-correction	8	our own arithmetic or our own criterion was wrong and is repaired here
not decidable / not evaluable	1	the apparatus or the band cannot separate the outcomes: not a null
acknowledged limit	4	a boundary of the result, stated where the result is
formal	2	proved from the definitions; no measurement can move it
measured	26	an artifact on disk, and a reader may dispute the artifact

Table 17, continued.

grade	rows	what the grade means in this paper
argued / reported	9	no grade is claimed: an argument, or a number printed without a criterion
other	3	status words outside the eight classes above
total graded rows	62	every numbered row of both tables

Table 19. Claims by evidential status. Grades: formal (proved, no data), code-verifiable (readable in released code, no model and no API call), measured, not supported / failed (frozen criterion missed), not evaluable (criterion mis-specified, not a null), untestable, self-correction, acknowledged limit, argued, reported (not graded). Rows recording grades we issued and then corrected are kept: the sequence is the contribution.

Claim	Status	Evidence
200. F1: the judge prompt's <question> is never bound, on any of the ten dimensions	code-verifiable	head commit 3e12035, lines 347–349: the replace chain substitutes the rubric item and the response only
201. F2: nine of ten dimension evaluators truncate the rubric's half credit through int()	code-verifiable	the tenth, evaluate_event_ordering, uses float()
202. P-T3-17: the fixed-minus-official shift is stable within 0.05 across splits on at least 8 of 10 dimensions	failed; below: 1 not supported	holds on 4 of 10 (last-occurrence) and 5 of 10 (BM25-last)
203. The unbound anchor removes the rubric's only downward pressure	measured	length–score rank correlation $+0.04/+0.17$ official against $-0.20/-0.19$ fixed
204. Construct substitution is distinguishable from judge variance by saturation	argued	abstention 0.075 to 1.000 on a judge change; no additive offset does that
205. Perfect provenance has a ceiling: handed the record containing the gold the pipeline credits $\iota < 1$	measured; below: 1 not supported	$\iota = 0.614/0.452/0.629$; 37–55% of the score uncollected under conditions no evidence layer improves on
206. P-T3-21, e_oracle's main criterion: pasting the gold-bearing turn scores ≥ 0.80	not supported	$S_{\text{oracle}} = 0.091/0.310/0.257$ against a bar of 0.80; P-T3-25 forbids softening it
207. The chain $V \leq A \leq B$ breaks at its first link	measured	in closed form the link is $\iota \geq 1$; it overstates by exactly $V(1-\iota)$
208. P-T3-106, the premise gate for the stronger-extractor arm: the answer contains the gold verbatim on ≥ 0.90 of items	failed as frozen; below: 1 not supported	containment 0.417; a verbatim-containment gate placed in the premise of an arm about answering
209. The judge is lenient: it passes a one-event, no-model reader on rubric criteria	measured	17.5% of criteria; 18.9% of those passes are on negations
210. The prohibition count needs no run at all	measured	84 negation criteria across eight dimensions, by grep over the rubric set
211. A temperature-zero judge is not a deterministic instrument	measured	29 of 800 shards change on re-execution (11 and 18 on the two arms)
212. P-T3-27: dispersion of scorer-induced difficulty on the containment subset	supported; below: 1 not evaluable	$\text{IQR}(d_i) = 0.333$ against a bar of one third
213. Under $S = a\,\iota$ the strictness confound is removable	formal	at a price of $\Theta(1/\iota)$ items; identification survives, affordability does not
215. F1's class is named as a failure mode by a contemporaneous paper, whose gate would have caught it	reported, against us	silent-no-op: the scorer-consumed prompt bit-identical to the unperturbed input, gated at $< 20\%$ (Li et al., 2026)
214. This paper owns ι and the two defects, and no resolving power	acknowledged limit	R_{state} and V are the companions'; nothing here estimates either

Table 19, continued.

Claim	Status	Evidence
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 46 sub-rows are in Table 20 in full, of which 8 self-correction, 5 not supported, 3 acknowledged limit, 1 failed, 1 not evaluable. Nothing is dropped, and nothing is summarised into a kinder word than it had.		

Table 20. Supporting refinements, moved from Table 19 by the stated rule. The direction is deliberate: a selection rule for a scorecard should be biased against its authors, so unflattering rows stay in the main table and supporting detail moves here. claims_all.tex is the source both tables are generated from.

Claim	Status	Evidence
The 46 rows below were moved here from Table 19 by the rule in papers/split_claims.py: a lettered sub-row whose status is supportive moves; every top-level claim and every row carrying a withdrawal, failure or self-correction stays. Nothing is dropped.		
200a. …and the prompt's only irrelevance clause is anchored to that placeholder	code-verifiable	"on-topic with respect to the QUESTION", evaluated against a literal
200b. …so no cross-judge protocol can reveal it	argued	the defect is invariant to the judge, hence to rotation and averaging
200c. F1's magnitude, isolated on the one dimension the official code scores with float	measured	event_ordering 0.222 to 0.003 (BM25-last), 0.280 to 0.000 (last-occurrence)
200d. …and we published one figure for two different manipulations	self-correction	an F1 magnitude of 0.277 matched no arm; 0.925 is a judge change, not a placeholder
200e. …and those magnitudes are now generated from the artifact, not typed	code-verifiable	figs/t6_numbers.tex, 10 macros; the grid's body likewise
200f. F1 manufactures the scale dependence read off this dimension's judge field	measured	0.280 to 0.076 to 0.167 official; 0.000, 0.013, 0.000 with the question supplied
200g. …and the isolating dimension's reported score is a rank agreement, not the judge's	reported, against us	evaluate_event_ordering returns τ_b-normalised $\times F_1$; the judge average is a further field. Named by the authors, 2026-08-31
200h. …so F1's magnitude on a score the benchmark publishes is unmeasured here	acknowledged limit	the nine judge-scored dimensions were confounded by F2 at 3e12035; §J.1(d) prices the run
201a. …and this paper printed the count as eight	self-correction	nine functions contain int(response['score']); "eight, and one uses float" is nine of ten
201b. This paper's own scorecard credited that scale dependence to F2, and to a cast removal	self-correction	it is F1's (row 200f) and the manipulation is the question supplied; event_ordering casts through float, so F2 never touched it
201c. The authors repaired F2 at b2da22e on 2026-08-31 and did not repair F1	code-verifiable	nine int() casts now float(); the <question> line is deleted and both clauses consuming the QUESTION survive
202a. …so every magnitude in the scorer section is a 100K figure	self-correction, P-T3-18	now labelled as one wherever it appears
203a. The protocol inverts instrument strength, under both judges	measured	the deliberately weaker reader wins 6 of 10 dimensions under each; a scorer with no irrelevance term cannot reward precision

Table 20, continued.

Claim	Status	Evidence
203b. The judge reorders on two dimensions of ten, which no pure offset model can do	measured	the completed 2×2 grid; comparing across judges with one reader missing would manufacture it
203c. …but both flips sit at the floor or at saturation, so they are weak evidence about ranking	argued, against us	0.950 and 1.000 on abstention, 0.050 against 0.032 on contradiction_resolution
205a. …and ι belongs to the (benchmark, reader, scorer) triple and transports to nothing	acknowledged limit	one benchmark, one rubric, one default judge
205b. …and the scorer is NOT what sets it (P-T3-119)	measured	where the answer carries the gold verbatim the scorer credits 0.972, so its own refusal of a demonstrably correct answer is 0.028
205c. …the shortfall is the reading step: a frontier model reports a gold-bearing record empty	measured	the answer carries the gold on 0.461 of items; the reader states the value absent on 0.339 of records that contain it verbatim
205d. …and tier-invariance was read as evidence for the rubric, which it is not	self-correction	two readers can fail the same way and these two fail on the same items, Jaccard 0.82; only a partition separates reader from scorer
205da. …and the extractor control was defended with the wrong lattice	self-correction, against us	three sentences called $1/n$ the statistic's step; this judge awards part credit, so an item's value lands on halves and thirds and the moves are one to three steps of the real lattice, not below one
205e. P-T3-118, whether ι is an artifact of this paper's own generation budget	not supported	ι moves -0.004 and $+0.026$ under a budget of 2000 against 400, inside the frozen band of 0.15; the registered prediction was the opposite
205f. …and the budget defect it tested for was real all the same	measured, against us	70 of 230 repaired generations exceed the old 400-token ceiling; the frontier arm's longest answer goes 112 → 223 characters
205g. P-T3-120, whether the reading failure is the budget's	not supported	the share of answers carrying the gold moves -0.017 and $+0.043$, inside the band; the reading failure is real
205h. P-T3-117, the manipulation check, as frozen	failed as frozen	1 of 115 frontier generations returned no content, at finish_reason error; 0 of 230 stopped on length, which is what the gate's own reason asks
205i. The containment test that selects these items has no token boundary	measured, against the count	3 of 115 golds match only inside a longer token ("$50" in "$500 fee"); post hoc, and it is a property of V
206a. …and what it refutes is this paper's PROOF of $V \leq A$, not the inequality	argued	the assumption may hold under another judge, which is a different experiment and not a different wording
206b. P-T3-26, the extract arm at a bar frozen by amendment before its shards existed	not supported	$S_{\text{extract}} = 0.614/0.452/0.629$ against 0.80; the amendment is counted, not overwritten
206c. …and this ledger first recorded P-T3-21 as a premise gate graded not evaluable	self-correction	it is the arm's main criterion and it misses its bar; the premise-gate description belongs to P-T3-106
207a. …and the licensed repair is instrument-dependent	measured	looser by 1.45 to 1.97 times
208a. …so P-T3-107, the criterion it gated, is not evaluable rather than null	not evaluable	the prereg says so in advance and this ledger first attached that grade to the gate instead
208b. …and the failed gate's own statistic is the partition that answers the arm's question	measured	containment of the gold in the ANSWER, 0.391/0.461, is what splits the reader's failure from the scorer's
208b†. …read post hoc, the credit rate is invariant to extractor tier	measured, post hoc	0.526 against 0.513 pooled over 115 items, a move of 0.013

Table 20, continued.

Claim	Status	Evidence
208c. P-T3-116, the premise gate for the partition, built on a construction invariant	supported	the record handed to the reader carries the gold verbatim on 1.000 of 115 items; no manipulation in the arm can move it
208c†. …so the strict half is the rubric's and not the reader's	measured, post hoc	the same conclusion the mis-specified gate was meant to license
209a. P-T3-108 branch (b): the information-free part of that floor	measured	0.119 for a no-model dump against 0.548 for a full-context arm
209b. …so the raw rate overstated what no information earns	self-correction	the floor is 0.119 and the remainder discriminates with accessibility
211a. …and the flips are total rather than partial	argued	mean absolute change 0.008 and 0.014 with a maximum of one whole shard
212a. P-T3-28: scorer strictness contaminates item-response difficulty	not supported	partial $\rho = +0.126$ against a frozen bar of 0.30; permutation $p = 0.18$
212b. P-T3-29: scorer-induced items present as highly discriminating	not supported, opposite sign	-0.230; items that discriminate better are items the scorer credits more
212c. …which is evidence for IRT-based adaptive selection, obtained by trying to break it	measured	past the bar's magnitude, in the direction that favours the field
212d. P-T3-30: the judge-swap arm, declared weak in the document that froze it	untestable	no verdict rests on it
212e. …and three systems give d_i only four values	acknowledged limit	the first thing a larger system pool would sharpen
213a. …tested on an axis the factorisation was not built from	measured	the deconfounding arm; the price is what the arm pays
213b. …and the delta-method price of estimating ι rather than knowing it	formal	inflation 1.376 where $1/\iota$ gives 1.091; agrees with 2×10^5 bootstrap draws to 1.009×
215a. …and what remains ours is the instance rather than the class	argued	their seven realizations were in their own legacy pipeline; ours is released, cited code at a named commit
215b. Two 2026 judge-reliability instruments would return reliable on the dimension audited here	argued, refutable	every axis in both is judge-side and F1 is invariant to all of them (Dev et al., 2026; Peng et al., 2026)