POSITIONAL COVERAGE BIAS: LLM COVERAGE JUDGES VERIFY WHERE CONTENT SITS, NOT WHETHER IT IS THERE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges that verify whether required content is *present* in a long-form report do not track content — they track *position*. We name and measure this failure mode of coverage-style LLM-judge metrics, **Positional Coverage Bias**: permuting a report's paragraphs, changing not a single word, makes the judge rescind *more* "present" verdicts than the agent's real content edits do — 0.107 against 0.047 patching and 0.070 regeneration on the deployed metric's own judge, Holm-confirmed, and still $2.6\times$ that judge's repeat noise ($[1.7,\ 4.2]$) when the permutation is restricted to *within* each section, so it is not discourse scrambling. It replicates on a second agent, two datasets and three judges under the full permutation but *not* under the discourse-preserving one: every full permutation decomposes into a cross-section part and a within-section residual, and the residual is largest where reports are weakly sectioned — the regime the audited benchmark's own agent produces, and the scope we claim. Verdicts track position far more than wording, so phantom break rates of the size deployed benchmarks attribute to agents (Chen et al., 2026) arise with no content removed, and in our own measurements the artifact dominates the signal: a generalizability decomposition leaves the revision method only 0.6% of break-rate variance across four near-equivalent edit strategies, and a blind human pass cannot separate content-preserving controls from real loss. It afflicts every judge that *reads the report and checks each item* — three such protocols — and is removable *by construction*: in a content-determined order the judge's input is byte-identical under permutation and the break falls $0.107 \rightarrow 0.028$. The removal is exact, and held to our own threshold our construction is *not* the metric to adopt: it is more wording-fragile than the judge it repairs, so it relocates the nuisance instead of lowering the threshold. An existing off-the-shelf order-agnostic metric has the property already, at a threshold our own interval cannot separate from the others ($+0.027$ $[-0.041,\ +0.095]$). We report the negative result rather than present a dominated design as a fix. What we hand the field is the threshold: an identifiability model makes it estimable from a model-free control battery *before* any agent is run, below which no coverage metric can *certify* that content was lost, and it is non-transportable, so each benchmark must measure its own. We release the control battery, the instruments, judges and ledger.

1 INTRODUCTION

Deep-research (DR) agents browse the web and synthesize long, citation-grounded reports. In practice they are used *interactively*: a user reads a report, gives feedback, and the agent revises. Recent work reports that agents *break* their own reports when they do this—previously-covered content regresses—and quantifies it with an LLM judge as a *break rate*: the fraction of checklist items covered before revision that are no longer covered after (Chen et al., 2026), headlined as the 16–31% "destructive regression" agents supposedly inflict. A low break rate is read as "the agent preserved what it had."

That headline is not separable from the judge that produced it. An LLM coverage judge — asked whether a required item is *present* in a report — rescinds its own "present" verdict when

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.

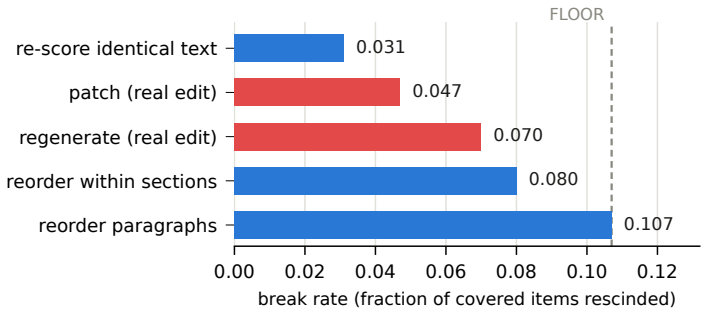

Figure 1: **A word-identical permutation rescinds more covered content than the agent's real edits do — on the judge that produced the number this paper audits.** Break rate over the same 25 ResearchRubrics questions under GPT-4.1-MINI, the deployed MR-DRE judge; blue removes no claim, red is a real revision. The discourse-preserving permutation (paragraphs moved only *within* a section) out-breaks both real arms too, so this is not discourse scrambling. Dashed line is FLOOR, Eq. (1)'s maximum over the content-preserving controls. Paired over the same questions, reorder $-$ regenerate $= +0.037$ $[+0.010, +0.066]$ and reorder $-$ patch $= +0.060$ $[+0.028, +0.092]$; both survive the declared family (§4.1) and the first survives a single family with every instability contrast in the paper.

the identical words are moved to a different position. Permuting a report's paragraphs, changing not one word, breaks *more* coverage than the agent's real patch or regeneration (Fig. 1: 0.107 vs. $0.047/0.070$ on the metric's own original judge, about $4\times$ its repeat noise 0.031), and the break tracks *position* far more than wording. We name this failure mode of coverage-style LLM judges **Positional Coverage Bias** (PCB).

Rather than decompose one published figure, we measure what a judge of that class does to reports of that kind: with no content removed the phantom break reaches 0.107 on the metric's own judge and 0.10–0.23 on a second agent's, and in our measurements it dominates the reported signal. The field gains four things. **(1) A named, measured failure mode of an entire metric class. (2) That a coverage number in this range can be majority artifact:** a generalizability decomposition assigns the revision method only 0.6% of break-rate variance across four near-equivalent edit strategies, and a blind human pass over all 76 flags cannot separate reordered controls from real loss. **(3) A portable, model-free detector with a stopping rule:** an identifiability model makes the floor an estimable *detection threshold*, computable before any agent is run, below which a coverage break rate certifies nothing. **(4) A clean negative result on the obvious remedy, and an existence proof.** The same control induces phantom loss 4.3–$21.3\times$ repeat noise across three read-and-check instruments, so we ask what removes it. Presenting the evidence in a content-determined order removes the position term *exactly* — the judge's input is byte-identical for $664/664$ items under permutation, and the break falls $0.107 \to 0.028$ — which establishes that the term is removable by construction. It is not, however, the metric to adopt: held to our own threshold it scores 0.109 against the 0.107 it repairs, while an *existing* off-the-shelf order-agnostic metric already scores 0.083 and so dominates both. **We report that as the result rather than dress the loss as a win:** the position term is removable, the field already has a better order-invariant option than ours, and no design we tested is clean on both position and wording (Table 2).

The break rate is a property of the (report, rubric, judge) triple, not the agent. Beyond the reorder, five of six judge$\times$dataset cells make benign content-preserving controls break as much as real revisions, and two further legs move the metric without touching a word — a month of judge drift and a meaning-preserving rubric paraphrase. The detection threshold transports only like-for-like: on the three controls every judge ran the RR floor holds ($0.022/0.031/0.021$), but adding a reorder multiplies it 1.6–$4.4\times$ (Table 1), so a published floor is a lower bound and each benchmark must measure its own — and of the four revision arms only the weak open agent's full regeneration, the one that reproduces the published phenomenon, clears it.

Two findings of our own, held to the same standard. Constrained in-place editing does cut raw break rate below full regeneration, significantly once pooled — but corrected by its own floor no arm retains a residual distinguishable from zero, so we report "patch, don't regenerate" as a raw-

break-rate ordering and nothing more. And raising an open agent's context occupancy does raise its break rate, by less than that dataset's own content-preserving controls, so we report it in Appendix D rather than claim it. Both are the standard this paper demands of others, applied to us.

2 Related Work

Evaluating deep-research agents. A cluster of recent benchmarks scores long-form DR reports with weighted expert rubrics and LLM-as-judge grading: ResearchRubrics (Sharma et al., 2026), ResearcherBench (Xu et al., 2025), RigorousBench (Yao et al., 2025), and DeepResearch Bench (Du et al., 2026), the last pairing report-quality scoring (RACE) with citation checking (FACT). Citation faithfulness itself is measured with extract–retrieve–verify pipelines such as VeriScore (Song et al., 2024). These efforts define the metric vocabulary we adopt—coverage, citation faithfulness/groundedness, and presentation—but they all score a *single* report; none models the multi-turn revision loop that we target.

Perturbation sensitivity and measurement validity. That evaluation is sensitive to semantically irrelevant surface form is established: judges show position, verbosity and self-enhancement biases (Zheng et al., 2023), are swayed by answer order (Wang et al., 2024), and favour their own generations (Panickssery et al., 2024); prompt *formatting* alone moves task accuracy by tens of points (Sclar et al., 2024), single-template evaluation is unstable in both scores and model *ordering* (Mizrahi et al., 2024), reordering multiple-choice options moves models several leaderboard positions (Alzahrani et al., 2024), a fixed commercial API name is not a fixed measurement target across months (Chen et al., 2024), and a known confound can be regressed out of a judge-based win rate (Dubois et al., 2024). Jacobs & Wallach (2021) supply the vocabulary this line of work keeps reinventing: a construct is always *operationalised* by a proxy whose reliability and validity must be argued rather than assumed — our identity control is a test–retest check, our modification controls test validity. These works establish that judge scores *shift* under surface and position changes; Positional Coverage Bias is the measurement-level increment they do not state — in *coverage* verification the positional artifact *out-magnitudes* the content effect it is meant to measure, dominates a deployed agent-capability number, and holds across coverage-judge instruments with an estimable detection threshold. The closest priors are the positional processing of Liu et al. (2024) and the judge position bias of Zheng et al. (2023); our contribution is not a new mechanism but its magnitude and consequence in a real metric.

Per-item differentials, and where the noise lives. Differencing per item is *not* new, and we do not claim it. Milani Fard et al. (2016) formalise prediction churn as the expected absolute per-example difference in 0–1 loss and show A/B test power scales with accuracy gain relative to churn; Bhojanapalli et al. (2021) measure 5.7–15.2% churn between deep-net runs whose accuracies differ by ≈ 0.1 points; and Miller (2024) and Bowyer et al. (2025) both prescribe inference on per-item paired differences in preference to aggregate scores. Nor is judge self-disagreement new: Stureborg et al. (2024) find GPT-4 agrees with itself less than humans agree with each other, and Thakur et al. (2025) show high percent agreement can hide large per-item divergence. **Our delta is that break rate reports one side of that churn.** It counts non-target items covered before and uncovered after, never the reverse — the discordant cell of a McNemar table, not the net difference in which the directions cancel. And we measure that they very nearly do: a reword moves the metric 0.022 while mean coverage moves $+0.0005$, a whitespace-only edit 0.021 against $+0.0008$, and a month of judge drift 0.053 while coverage moves *up* by $+0.0125$. The instrument is not biased in one direction; it is noisy in both, and the metric records only the losing half, which is what turns symmetric churn into reported destruction. Remedies for zero-mean sampling error under a reliable grader do not touch that.

Multi-turn revision of DR reports. Chen et al. (2026) (MR-DRE) introduce revision as an evaluation axis and report that five DR agents, while incorporating $>90\%$ of feedback, regress 16–31% of previously-covered content and degrade citations; their prompt-engineering and dedicated reviser sub-agent reduce but do not eliminate the gap. Sabharwal et al. (2026) independently observe, under process-level feedback, that agents "rewrite the full report" and forget up to 24% of satisfied criteria, with gains that do not compound. We build directly on these findings and depart from them at the measurement level: we *audit the metric itself* and find that much of the reported regression is judge

instability under report modification rather than content loss — a bound neither work computes. Our two constructive findings, constrained editing and context occupancy, are held to the same standard and claimed no further (§1).

Edit-based revision, and plan-then-edit. Localized editing appears in adjacent settings: fragment-level patching in STEP-DEEPRESEARCH (Hu et al., 2025), durable claim "ledgers" in PAPERJURY (Wang et al., 2026) and DEEPREVIEWER-2.0 (Weng et al., 2026), and a revision ledger proposed but not implemented by Nogueira et al. (2025); Kruthof (2026) report plan-then-edit propagating factual edits *worse* than direct editing, in a single-turn, single-model synthetic setting whose "plan-then-edit" regenerates the whole manuscript. We implement a diff-constrained patch reviser and a ledger variant whose rule is that *removals of cited claims must be justified*, find the ledger no better than plain patching, and position both as strong near-equivalent baselines rather than a win: once existing content is preserved, the residual metric is dominated by judge instability (§4). Our constrained-diff results agree with theirs in spirit — full regeneration is the destructive regime, and the only mode whose break rate clears our instability floor.

3 METHOD

Multi-turn revision protocol. Each instance is a turn-1 report r_1 for a query q with a checklist $C = \{c_i\}$. A judge J labels each item covered/not; let $S_1 = \{i : J(r_1, c_i) \text{ covered}\}$. A turn-2 message requests one uncovered item $t \notin S_1$; the reviser produces r_2, and S_2 is scored the same way. *Break rate* counts non-target items lost: $\mathrm{BRK} = \frac{|\{i \in S_1 \setminus \{t\} : i \notin S_2\}|}{|S_1 \setminus \{t\}|}$, and *incorporation* is $\mathbb{1}[t \in S_2]$.

Reviser modes (edit strategy isolated). A single reviser LLM with fixed decoding and identical feedback is run in four modes, so differences reflect the strategy, not the model: (i) *regenerate*—rewrite the full report; (ii) *patch*—emit JSON find/replace edits applied programmatically to r_1, so unedited text is preserved verbatim; (iii) *ledger-patch*—patch plus a guard blocking any edit that deletes a cited claim (`<cite>`) without justification; and (iv) *localized minimal edit (LME)*—route the request to one paragraph via a compact digest and edit only it, leaving the rest byte-identical.

Break-rate decomposition. Let $\mathcal{P}$ be a set of *content-preserving* perturbations — transformations π for which r_1 and $\pi(r_1)$ contain the same claims by construction, so a metric tracking content loss must score 0. We use six: *identity* (re-score r_1 unchanged), *reword* (rephrase one paragraph, every fact and citation retained), *neutral-insert* (append text with no epistemic content), *whitespace*, *reorder* (permute whole paragraphs), and *full reword* (rephrase every paragraph, matched to regeneration's edit magnitude).

One definition of the floor. We fix it once and use it everywhere:

$$\mathrm{FLOOR}(D, J) \;=\; \max_{\pi \in \mathcal{P}} \mathrm{BRK}\big(\pi;\, D, J\big), \qquad \mathrm{BRK}_{\mathrm{true}}(m) \;=\; \mathrm{BRK}(m) - \mathrm{FLOOR}(D, J). \tag{1}$$

The floor is per dataset D and judge J, never pooled, and is a *maximum* over $\mathcal{P}$. Two biases run opposite — $\mathcal{P}$ is incomplete (the max understates the true supremum) while a max of noisy means overstates it — so we report FLOOR as an interval and treat every floor, ours included, as a *lower bound*; the incompleteness dominates (App. E). A maximum over an *unnamed* set is not comparable, so we fix $\mathcal{P}$ as the six above and release it: the four model-free controls ship as one generator whose verify mode reproduces every artifact byte-for-byte. Floors over different $\mathcal{P}$ are not comparable — the error we withdraw in §4 — and arms that never permute paragraphs are also compared against the order-preserving floor (App. F). All perturbations are judge-only: any benchmark can measure its floor for one extra scoring pass.

Why order-invariance is the right requirement. A coverage checklist asks whether an item is *addressed*, not whether it is addressed in a particular place, and the benchmarks that deploy it read the output as content presence — so order-invariance is the desideratum implied by the metric's own use, not one we impose. Our own evidence pushes back twice, and we state both here rather than in an appendix: under the reorder net coverage does fall slightly (-0.042/-0.025, App. E), and a blind human rater declines "present" on 23.7% of provably word-preserving flips. Neither licenses a metric that treats position as content: the first is about 40% of that break — the judge reading a permuted report as covering less, which is the phenomenon, not a defence against it — and the second is an error floor we subtract rather than ignore.

When does the floor exceed the signal? Model the judge's chance of re-scoring a covered item as lost under a transformation T as $\Pr[\text{lost} \mid T] = \rho + \nu(T) + \lambda(T)$: repeat noise ρ (flip rate on *identical* text), spurious content-preserving sensitivity $\nu(T) \geq 0$ (order, whitespace, paraphrase), and true loss $\lambda(T) \geq 0$ (only when T removes content). A control c then breaks at $F_\rho + N_c$ and a real arm m at $F_\rho + N_m + L_m$, with $\text{FLOOR} = F_\rho + \max_c N_c$ and L_m the loss the benchmark wants. Two facts follow for *any* checklist-judge metric: a control out-breaks a real arm — the inversion we document — iff $\max_c N_c > N_m + L_m$; and N_m and L_m are not separately identified from a break rate, so any arm at or below FLOOR admits a pure-nuisance reading ($L_m = 0$). FLOOR is therefore a *detection threshold*: loss is certifiable only above it. Both its inputs are measured, not assumed, so it is an operational test rather than a caveat, and our data lie in that regime under all four judges (App. F, which gives the algebra).

Dependability (generalizability theory). The floor asks whether a content-preserving change breaks; the sharper question a ranking metric must answer is whether the break rate can *differentiate the objects a benchmark ranks* at all. We treat the revision method as the object of measurement and the judge and the question as random facets, and estimate the variance components by a balanced random-effects ANOVA — the generalizability-theory design of Cronbach et al. (1972). The generalizability coefficient $E\rho^2$ (relative error) and the dependability $\Phi = \sigma^2_{\text{method}} / (\sigma^2_{\text{method}} + \sigma^2_{\Delta})$ (absolute error) give the share of break-rate variance that is the *method* rather than the judge, the question, or their interaction; a Φ far below the conventional 0.80 means the metric cannot reliably tell arms this closely spaced apart, and a follow-up D-study asks whether averaging more judges would raise it. **A variance component is defined relative to the sampled universe of objects**, so this design bounds the metric's power to separate *near-equivalent edit strategies* and does not license extrapolation to the wider between-agent spread behind the audited range. It connects to the test–retest / inter-rater tradition (intraclass correlation is the single-facet special case); the estimator and $E\rho^2$ are in App. F. Code: `code/`.

4 Experiments

4.1 Setup, statistics, and what sent us looking

We build on MR-DRE's multi-turn revision protocol over three deep-research checklist datasets — ResearchRubrics (RR), ResearcherBench (ResB) and RigorousBench (Rig) — with turn-1 reports from the open DR-Tulu-8B (Shao et al., 2025) and a turn-2 user message asking for one uncovered checklist item to be added. *Break rate* (BRK↓) is the fraction of *non-target* items covered at turn-1 that become uncovered at turn-2, per MR-DRE's own checklist evaluator; *incorporation* (INC↑) is whether the requested item becomes covered. Coverage is scored by DeepSeek-chat with Qwen3-8B as a different-family cross-check, plus Qwen3-32B and GPT-4.1-mini for the floor analyses. One reviser (temperature 0, identical feedback) is run in four modes — full *regenerate*, constrained *patch*, *ledger-patch* and *localized* — so that only the edit *strategy* varies. The reviser is DeepSeek-chat; the modes, licences, decoding and compute are specified in Appendix B, and the verbatim prompts ship with the released code. **Two scope decisions.** MR-DRE's own judge was unavailable for the main sweep, so every claim is *relative* within a fixed judge (our anchor approaches the published range at BRK $= 0.144$); and the primary judge also scores that model family's own revisions, a self-preference risk (Panickssery et al., 2024) the independent checks bound. **Statistics, stated once.** Every contrast is a paired question-level bootstrap (10,000 resamples, seed 0) on matched target-excluded denominators, and every p-value belongs to exactly one pre-stated Holm family: the 4 edit-strategy confirmatory tests, the nine instability contrasts, the two flagship reorder-versus-arm tests, and the twelve replication contrasts *per control*. We never pool families for power. **A family of two containing exactly the tests one wants to pass is the shape of p-hacking, so we report the flagship without that boundary:** in one 11-test family with every instability contrast in the paper, *both* flagship tests survive Holm and nothing else does; in a 23-test family that also pools the second agent's twelve contrasts, reorder-versus-patch survives and reorder-versus-regeneration just misses.

What sent us looking. Before auditing the metric we used it. Constrained in-place editing does cut raw break rate below full regeneration, significantly under both judge families pooled (p=0.003/p=0.030, all 4 surviving Holm), but it inverts on the third dataset and retains no floor-

corrected residual distinguishable from zero (Table 6, App. H): "patch, don't regenerate" is a raw ordering and nothing stronger.

4.2 THE PERMUTATION OUT-BREAKS THE REVISER'S OWN EDITS

If break rate measured content loss, a modification that changes no substantive content should score ≈ 0. It does not. Figure 1 shows it on the judge behind the published number, Table 1 on all four, and App. H the full grid. Re-scoring identical text gives 0 or near it in four of six judge×dataset cells; both exceptions are DeepSeek (0.030 on ResB, 0.069 on Rig, the latter above every control there). A third judge, served locally, settles what those cells cannot: its identity break rate is *exactly* 0.000 (frozen weights, greedy decode), yet under it a paragraph reorder, which changes not one word, breaks 0.091 — above both real arms in its column, and above the 0.064 we measure under the judge that drifts. So the floor is not the drifting judge's artifact: on all four judges the reorder is the largest control and no arm reaches its own floor (Table 1), while a *reword* or an *unrelated neutral insertion* already breaks comparably to real revisions — more on ResB/DeepSeek — so instability is judge- and dataset-dependent and can be severe. Correcting each cell by its own floor (Eq. (1)) leaves one positive residual, ResB regeneration at $+0.082$, which we withdraw: FLOOR is a maximum over four noisy estimates and no interval that resamples it excludes zero, while pooled patch sits *below* its floor (-0.043 $[-0.128, -0.001]$). Against the magnitude-matched control the residual is $+0.071$ $[+0.009, +0.139]$ — carried by the core slice, null on the disjoint 25, short of our Holm threshold (App. F).

4.3 THE THRESHOLD IS LOCAL, AND TOO IMPRECISE TO IMPORT

Within a dataset and judge the floor is agent-independent — the controls perturb the turn-1 report — so any claim that the artifact "share" falls as break rate rises is $\mathrm{FLOOR}/\mathrm{BRK}$ arithmetic, not a result. **Across cells it is neither stable nor precise:** on the one control we ran everywhere, estimates span 0.006 to 0.078 with overlapping intervals ($[0.005, 0.177]$ vs $[0.000, 0.018]$) and the smallest exactly zero in 60% of resamples, so we claim no ratio; and it does not transport across *question samples* either, the ResB regenerate rate moving $2.60\times$ between disjoint 25-question draws of one benchmark ($[1.01, 9.15]$) with a *different* control maximal on each. A published floor is therefore a lower bound: divided into the published 16–31%, the three controls that judge ran give 10–19% and its *complete* battery 35–68%, both judge-matched. Of the four arms only the open agent's full regeneration clears its own floor ($+0.080$, $[+0.018, +0.142]$) — the one arm reproducing the phenomenon we audit (Table 7, App. H).

It is order, not wording. Under all four judges the perturbation that moves the metric most changes no words at all: rewording *every* paragraph while keeping every citation (edit magnitude 0.66, above regeneration's 0.55) breaks $0.040/0.051/0.023/0.048$ against $0.064/0.065/0.091/0.107$ for merely permuting them — a paired $+0.059$ $[+0.029, +0.094]$ on the original judge, for zero lexical change. No revision arm moves paragraphs, so this does not confound our own edit-strategy result (App. E).

A null is not equivalence. On ResearcherBench, where the whole battery ran, the content-preserving edits *are* the largest movements we measure — a whitespace-only edit breaks 0.177, above the real patch at 0.112 and inside the published 16–31% range, though whitespace is content-preserving rather than *positional*. Equivalence tested at a pre-fixed margin holds for three contrasts and none at a margin below our own edit-strategy effect, so ours is a bounding claim, not an equality one (Tables 8, 9, App. H).

4.4 THE NUMBER BARELY SEPARATES THE ARMS IT RANKS

Generalizability theory says this once, as a number. Treating the revision method as the object of measurement and judge and question as random facets (§3), a balanced ANOVA over four methods, two judges and 25 questions leaves the *method* — the quantity a benchmark wants — 1.3% of break-rate variance against 43% for judge×question, with dependability $\Phi = 0.31$ ($[0.00, 0.80]$), $P(\Phi < 0.80)$=0.97; the four-judge design keeps $\Phi \in [0.00\text{–}0.37]$ and puts method at 0.6%, 95% upper bound 7.8%. **Averaging judges cannot repair it:** a D-study reaches only $\Phi = 0.24$ at *fifty* judges (Tables 3, 4). Scoped to the four strategies sampled here, break rate is a dependable measurement of the (judge, question) pair rather than of the reviser.

Table 1: **Under every judge we ran, a word-preserving paragraph reorder breaks more coverage than the agent's real revisions.** Break rate on ResearchRubrics under four judges spanning three model families. The rightmost column is GPT-4.1-MINI, the judge that produced the 16–31% breakage this paper set out to explain; it is measured on 65 questions and the other three on 25, so compare *down* each column rather than across rows. The ordering in that column is paired and Holm-confirmed (§4). Read the FLOOR row — the largest content-preserving control — against the real arms below it: no arm reaches it under any judge, and the control attaining it is in every case the paragraph reorder. The floor is *largest* under the judge that reproduces itself most exactly, so it is not a weak judge's noise. The original judge's column is complete, and the reorder attains its maximum there against a *full* reword at 0.048 — order beats wording by 2.2× on the judge that produced the range. **Both floors are printed, and the second cuts against us:** no arm permutes paragraphs (τ=1.000, App. E), so the *order-preserving* row is the operative comparator, and against it patch clears under QWEN3-8B (+0.006) and regeneration under three of four judges (+0.008/+0.027/+0.011). No positive residual survives resampling the floor inside each bootstrap replicate (App. F) — that is the claim; these point estimates are stated rather than rounded away. Every cell of the original judge's column carries a question-level bootstrap interval, and every ordering this paper claims from it is restated as a paired contrast with its own interval (App. F). ‡ Identity is measured, not assumed.

	content changed?	DEEPSEEK *(primary)*	QWEN3-8B *(cross)*	QWEN3-32B *(stronger)*	GPT-4.1-MINI *(original)*
Content-preserving: break should be ≈ 0 if the metric tracks content					
identity (re-score same text)	no	0.005	0.002	0.000‡	0.031
reword one paragraph	no	0.022	0.024	0.013	0.027
insert neutral text	no	0.014	0.040	0.021	0.023
whitespace only	no	0.021	–	0.018	0.031
paragraph reorder	no	0.064	0.065	0.091	0.107 §
reword every paragraph	no	0.040	0.051	0.023	0.048
FLOOR, all controls	—	0.064	0.065	0.091	0.107
FLOOR, order-preserving only	—	0.040	0.051	0.023	0.048
Real revisions					
patch	minimal	0.024	0.057	0.072^{S}	0.043
ledger-patch	minimal	0.007	0.026	–	–
localized	minimal	0.039	0.047	–	–
regenerate	full	0.038	0.059	0.050	0.058

Superseded value, printed rather than replaced. Re-running the identical reviser and judge two months later gave 0.0093, and the reviser-side evidence (§4.5) attributes the eightfold change to the *reviser* rather than the judge; we keep it visible as the cleanest instance of the drift this paper documents and do not use it as evidence about edit strategy. The GPT-4.1-MINI column is our re-implementation of MR-DRE's verbatim checklist prompt, validated against the DEEPSEEK break rates before use.

4.5 JUDGE, RUBRIC AND REVISER ALL MOVE ON THEIR OWN

Two probes that never touch the report (App. F). Re-scoring identical text with DEEPSEEK-CHAT a month later costs 0.053, an order of magnitude above the same-version identity 0.005 and comparable to a real patch; paraphrasing all 664 criteria with meaning, polarity, entities and numbers preserved (Table 10) moves the break rate 0.0836 on *unmodified* reports. Mean coverage moves +0.013 and −0.010, so a benchmark reporting aggregate coverage sees neither — and the 16–31% came from such a judge at an unspecified date.

A blind human finds no per-item difference either. One rater, blind to which edit produced each of the 76 flips (App. E), does *not* reproduce GPT-4O's separation: 78.7% of control against 79.3% of real flips called present (−0.006 [−0.186, +0.183]). On the 38 whitespace/reorder flips, where every word provably survives, the human still declines "present" 23.7% of the time — 20.0% on the 15 whitespace-only flips, which reorder nothing, so the floor is the rater's and not the text's; corrected by that floor no arm excludes zero (−0.030 [−0.230, +0.171]), and the two adjudicators' divergence is that error floor.

Table 2: **Positional Coverage Bias is a property of the read-and-check protocol; order-invariance is buildable, and what it costs depends on how you build it.** *Top*: the identical model-free reorder — not one word changed — run through three *functionally distinct* read-and-check instruments over the same reports, each against *its own* identity repeat noise. Every one rescinds 4.3–21.3× as many "present" verdicts as its own noise (and a discourse-preserving within-section reorder still clears it, §4). *Bottom*: two designs that never show the judge document order. Ours (D) retrieves per item and orders by relevance with a content-deterministic tie-break, so the judge's input is *byte-identical* under permutation (664/664 items verified); the off-the-shelf alternative (E) is *also* invariant. Applying Eq. (1) to ourselves, the threshold — a *maximum* over controls — is 0.107 (A), 0.109 (D), 0.083 (E): D removes the position term exactly but *relocates* the nuisance into wording rather than lowering the threshold. No design here is clean on both axes; that is the result, not a caveat. **Every cell carries a bootstrap interval** (App. F), and they settle two things: B's and C's reorder intervals ([0.051, 0.231], [0.081, 0.173]) clear their own identity intervals ([0.000, 0.018], [0.008, 0.033]); and D's identity and reorder intervals *coincide* ([0.013, 0.045] vs [0.009, 0.053]) — invariance evidenced, not asserted — with [0.073, 0.146] disjoint from both.

coverage instrument	judge	identity	reorder	reword
read-and-check: the judge reads the report in document order				
A. per-item binary presence *(Mr-Dre)*	4.1-mini	0.031	**0.107**	0.048
B. batch claim/nugget recall	4.1-mini	0.006	**0.128**	—
second judge family	GPT-4O	0.010	**0.074**	—
on GPT-4O *reports*	4.1-mini	0.016	**0.159**	—
C. graded requirement satisfaction *(InFoBench-style)*	4.1-mini	0.020	**0.124**	—
order-invariant: the judge cannot see where the evidence sat				
D. retrieve-then-judge, lexical *(ours)*	4.1-mini	0.028	0.028	**0.109**
semantic retrieval	4.1-mini	0.040	0.039	**0.119**
E. decompose-and-search *(*RAGAS*, verbatim)*	4.1-mini	0.079	0.083	0.076

Bold marks where each design pays: A–C on position, D on wording. **Row E carries no bold, deliberately**: its three cells are mutually indistinguishable (+0.005 [−0.024, +0.033]), so its threshold 0.083 is merely the largest of three values we cannot order. Removal is still detected: deleting the most content-bearing paragraph breaks D at 0.099, far above its 0.028 repeat noise. E's cells pair statements by exact text — its decomposition moves the item set for 6/25 questions even under an identity repeat — and pool over them where A–D average over questions; App. F reports both, which agree.

Not a weak-judge artifact. Figure 1 is the flagship on the *original* MR-DRE judge, paired on matched denominators over the same 25 questions, and both contrasts survive Holm. It is not one agent's either: §4.6 replicates the whole pipeline on a second, different-family turn-1 agent.

4.6 SCOPE: PART OF THE BREAK IS CROSS-SECTION, AND NONE OF IT IS ONE PROMPT'S

A reorder changes no word but does change discourse order, and at n=65 the two come apart. A *within-section* permutation separates them. On the DR-TULU reports under GPT-4.1-MINI — the agent and judge behind the 16–31% — it still breaks 0.080, 2.6× the judge's own repeat noise (+0.049 [+0.025, +0.076]). **Two readings we took from a 25-question version do not survive the larger sample, and we withdraw them:** it does *not* exceed regeneration (+0.010 [−0.014, +0.034], p=.415), and the full permutation *is* distinguishable from it (+0.027 [+0.009, +0.046], p=.002) — so a measurable share of the break is content moving *across* sections. **The second agent does not replicate under it, and we say so:** the within-section reorder is positive in only six of the twelve contrasts and none survive, every interval containing zero in both directions (Table 5). On GPT-4O's sharply sectioned reports the full-reorder break (0.10–0.23 vs within-section 0.03–0.10) is carried by *cross*-section movement, which we attribute to retrievability rather than judge instability. **Both agents now tell one story** the smaller sample had split in two: a cross-section part, plausibly retrievability, and a within-section residual no retrievability account explains, largest where the report is weakly sectioned — the regime the audited benchmark's own agent produces. Nor is it tied to the MR-DRE prompt: the same reorder induces phantom loss 6.2–21.3× each instrument's own repeat noise in two *other* read-and-check judges (Table 2), so Positional Coverage Bias is a property of the protocol, not of a prompt.

4.7 REMOVING THE POSITION TERM, AND WHAT IT COSTS

The position term is removable by construction. Show the judge the same evidence in a *content-determined* order: segment the report into units, retrieve those relevant to each item, and order them by relevance with a content-deterministic tie-break. Under permutation all three are unchanged, so the judge's input is byte-identical — verified for $664/664$ items — and the reorder break falls $0.107 \to 0.028$, exactly this judge's repeat noise, with coverage preserved (0.511) and real deletion still detected (0.099). **That is an existence proof, not a metric we propose:** seeing a retrieved excerpt rather than the whole report, this judge is *more* wording-fragile than the one it repairs (0.109 vs 0.048).

And an existing metric already has the property, which is the result. We take RAGAS's `LLMContextRecall` (Es et al., 2024) — decompose the requirement into claims, verify each against the report — from its shipped prompt file unmodified, under the same judge and statistic. It is position-invariant too (0.083 against its own repeat noise 0.079, p=.76). Applying Eq. (1) to all three designs gives 0.107, 0.109 and 0.083 (Table 2). *So our construction does not lower the detection threshold: it relocates the nuisance from position to wording.* **Held to the paper's own interval standard, that ordering does not survive either:** paired over the same questions, the read-and-check and decompose-and-search thresholds differ by $+0.027$ $[-0.041, +0.095]$ (p=.43). At n=25 the three are not distinguishable, so "the off-the-shelf metric has the lowest" is a point estimate we do not claim as an ordering. Ours is still not the metric to adopt, and the existing recipe has the property already, with a caveat Table 2 quantifies.

The system under test drifts too, enough to reverse an apparent judge-dependence we had read into QWEN3-32B when the same reviser is re-run two months later; we withdraw that reading (App. F, with five further checks that change no headline).

5 CONCLUSION

Under all four judges, the metric's own included, permuting a report's paragraphs breaks more coverage than the reviser's real edits — on 65 questions, every interval clear of zero. A break rate is a property of the (report, rubric, judge) triple, reportable only as an interval. **Limitations.** One reviser and one turn-1 agent; single-run drift; one rater at 23.7%; battery and second agent at 25 questions.

REFERENCES

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Srinadh Bhojanapalli, Kimberly Wilber, Andreas Veit, Ankit Singh Rawat, Seungyeon Kim, Aditya Krishna Menon, and Sanjiv Kumar. On the reproducibility of neural network predictions. *arXiv preprint arXiv:2102.03349*, 2021. URL https://arxiv.org/abs/2102.03349. Preprint; no peer-reviewed version located.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81143–81184. PMLR, 2025.

Bingsen Chen, Boyan Li, Ping Nie, Yuyu Zhang, Xi Ye, and Chen Zhao. Beyond single-shot writing: Deep research agents are unreliable at multi-turn report revision. *arXiv preprint arXiv:2601.13217*, 2026. URL https://arxiv.org/abs/2601.13217.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024. doi: 10.1162/99608f92.5317da47. URL https://hdsr.mitpress.mit.edu/pub/y95zitmz.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2506.11763.

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. URL https://arxiv.org/abs/2404.04475.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025. URL https://research.trychroma.com/context-rot. Web technical report (not peer-reviewed); code at https://github.com/chroma-core/context-rot.

Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, Wei Yuan, Wenjin Deng, Xiaojian Yuan, Xiaoyun Zhang, Xiangyu Liu, Xikai Liu, Yanming Xu, Yicheng Cao, Yifei Zhang, Yongyao Wang, Yubo Shu, Yurong Zhang, Yuxiang Zhang, Zheng Gong, Zhichao Chang, Binyan Li, Dan Ma, Furong Jia, Hongyuan Wang, Jiayu Liu, Jing Bai, Junlan Liu, Manjiao Liu, Na Wang, Qiuping Wu, Qinxin Du, Shiwei Li, Wen Sun, Yifeng Gong, Yonglin Chen, Yuling Zhao, Yuxuan Lin, Ziqi Ren, Zixuan Wang, Aihu Zhang, Brian Li, Buyun Ma, Kang An, Li Xie, Mingliang Li, Pan Li, Shidong Yang, Xi Chen, Xiaojia Liu, Yuchu Luo, Yuan Song, YuanHao Ding, Yuanwei Liang, Zexi Li, Zhaoning Zhang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. Step-deepresearch technical report, 2025. URL https://arxiv.org/abs/2512.20491. StepFun Agent Team technical report.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 375–385. ACM, 2021. doi: 10.1145/3442188.3445901.

Garvin Kruthof. Editpropbench: Measuring factual edit propagation in scientific manuscripts, May 2026. URL https://arxiv.org/abs/2605.02083. arXiv:2605.02083; code https://github.com/kruthof/EditCoherenceBench.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/dc5c768b5dc76a084531934b34601977-Paper.pdf.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. 2024. URL https://arxiv.org/abs/2411.00640. arXiv preprint, 14 pages.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681. URL https://aclanthology.org/2024.tacl-1.52/.

Brenda Nogueira, Werner Geyer, Andrew Anderson, Toby Jia-Jun Li, Dongwhi Kim, Nuno Moniz, and Nitesh V. Chawla. From verification burden to trusted collaboration: Design goals for LLM-assisted literature reviews, December 2025.

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. URL https://arxiv.org/abs/2404.13076.

Rishabh Sabharwal, Hongru Wang, Amos Storkey, and Jeff Z. Pan. Multi-turn evaluation of deep research agents under process-level feedback, 2026. URL https://arxiv.org/abs/2606.09748. Published as a workshop paper at SCALE – ICML 2026 (Oral).

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11324.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. DR Tulu: Reinforcement learning with evolving rubrics for deep research, 2025. URL https://arxiv.org/abs/2511.19399. ICML 2026 (oral).

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2511.07685. arXiv:2511.07685.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724. arXiv preprint, 2 May 2024; no peer-reviewed venue found as of 2026-08.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 404–430, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL https://aclanthology.org/2025.gem-1.33/.

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. URL https://aclanthology.org/2024.acl-long.511.

Yiran Wang, Ruixuan An, Biao Wu, and Wenhao Wang. Paperjury: Due-process review for bounded latex revision, 2026. URL https://arxiv.org/abs/2606.16322.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, Enhao Gu, Qiyao Sun, and Yue Zhang. Deepreviewer 2.0: A traceable agentic system for auditable scientific peer review, 2026. arXiv:2604.09590.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry, 2025. URL https://arxiv.org/abs/2507.16280.

Yang Yao, Yixu Wang, Yuxuan Zhang, Yi Lu, Tianle Gu, Lingyu Li, Dingyi Zhao, Keming Wu, Haozhe Wang, Ping Nie, Yan Teng, and Yingchun Wang. A rigorous benchmark with multidimensional evaluation for deep research agents: From answers to reports, 2025. URL https://arxiv.org/abs/2510.02190. arXiv:2510.02190; v2 (Jan 2026) retitled "Dr. Bench: A Multidimensional Evaluation for Deep Research Agents, from Answers to Reports"; code (MIT): https://github.com/EVIGBYEN/DrBench.

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)*, 2023. URL https://arxiv.org/abs/2306.05685.

A Statements: reproducibility, ethics, and AI use

Reproducibility Statement

All experimental details are in Appendix B: datasets and licenses, the three judges and their decoding, the reviser and its four modes, the content-preserving controls, and compute; the paired-bootstrap and Holm–Bonferroni procedures are specified in §4. Every number in the paper is a macro traced to a run row in `experiments/runs.csv` with its per-run log and output JSON. We release the reviser, the content-preserving controls, the break-flip adjudication set and the evaluation adapters, so any multi-turn revision benchmark can measure its own instability floor. The four controls that need no model call — whitespace-only, paragraph reorder, within-section reorder and neutral insertion — ship as a single generator with a `verify` mode that regenerates every perturbed report behind the numbers in this paper and checks it byte-for-byte: 350 for the first three and a further 75 for the within-section reorder across all three report sets. The released transform is therefore the one we ran rather than a description of it.

Ethics Statement

This work studies the reliability of automated evaluation. It collects no personal data and runs no user study; the one human component is a single-rater adjudication pass in which one rater gave a three-way verdict on 76 items drawn from public research benchmarks (App. E), with no personal or sensitive material involved and no data about the rater recorded or released. All datasets are public research benchmarks used under their stated licenses (Mr-Dre checklists MIT; DR-Tulu-8B Apache-2.0). Our finding—that a widely-used LLM-judge break rate carries a locally-varying instability floor, large enough that the share attributable to real content loss cannot be read off the raw number—has a positive valence: it guards against over-attributing destructive behavior to deployed agents and against optimizing systems toward an artifact-laden metric. We use LLM judges only for benchmark scoring, not for consequential decisions, and report the judges' disagreement and instability explicitly.

AI Use Statement

This research was carried out with substantial assistance from a large-language-model coding and research assistant, under the authors' direction; the authors made every research decision and take full responsibility for the content of this paper. **Required disclosures.** The assistant implemented the experimental code—the four reviser modes, the content-preserving control battery, the judge and adjudicator adapters, the scoring harness, the paired-bootstrap and Holm–Bonferroni analysis, and the figure scripts—and executed the runs on the authors' cluster. It contributed to methodology design and hypothesis refinement: in particular, the pivot that defines this paper—from trying to *fix* revision breakage to *auditing the metric*—was proposed during assisted analysis after two method elaborations returned null results, and the authors chose to pursue it. It also participated in result interpretation and produced first drafts of the prose in all sections. **Recommended disclosures.** The

assistant was additionally used for literature search, related-work drafting, figure creation, reference formatting and readability editing, and a separate LLM was used to give an independent critique of the draft's figures, tables and writing. **Verification.** Every result number in this paper is a LaTeX macro bound to a single row of the released run ledger with its per-run log and output JSON; no number was written from an estimate, a recollection or a model's assertion. We release the code, controls, ledger and logs so that each reported number can be re-derived from its artifact.

Distinctly, and as a matter of method rather than of authorship: LLMs are the *object of study* here. They serve as the deep-research agent under test, as the reviser, as the four judges whose instability we quantify, and as the two independent adjudicators. Those uses are specified in §4 and Appendix B.

B Reproducibility

Data. Core Set = 25 questions each from ResearchRubrics, ResearcherBench, RigorousBench (n=75), via the MR-DRE harness (MIT). Turn-1 reports from DR-TULU-8B (Apache-2.0), served with vLLM. **Judges.** Coverage scored with DEEPSEEK-CHAT (primary) and QWEN3-8B (cross-check), both temperature 0, using MR-DRE's checklist-evaluation prompt; a stronger QWEN3-32B (served locally, tensor-parallel 2) *and* GPT-4.1-MINI—the original MR-DRE judge, run from a non-DGX endpoint since OpenAI is unreachable on the DGX—re-score the RR controls and methods as floor-robustness checks. The GPT scorer is a separate harness that reuses MR-DRE's verbatim checklist prompt; we validated the *port* by pointing it at DeepSeek and confirming it reproduces our DeepSeek reword/patch/regenerate break rates exactly. Its GPT-4.1-mini column is of course different from the DeepSeek column — that difference is the measurement, not an error. Primary DeepSeek/Qwen numbers are reproductions under a consistent judge; every relative claim is cross-judge verified. The primary judge also scores the DeepSeek reviser's own output; the independent Qwen, GPT-4.1-mini, and GPT-4o checks bound this self-preference risk. **Reviser.** DeepSeek-chat, temperature 0, identical per-question feedback across modes; patch/ledger apply model-emitted JSON edit ops programmatically; localized routes via a compact paragraph digest. The tables of §4 use a single feedback assignment and one temp-0 decode per condition; temp-0 decoding is *not* fully deterministic, so reviser-sampling variance is measured separately over five independent decodes per strategy, and additionally across API model versions two months apart. **Controls (C3).** identity (re-score unchanged); reword (rephrase one paragraph, citations verified preserved); *full-report* reword (rephrase every paragraph; edit magnitude 0.66, an edit-magnitude-matched floor for regeneration; $25/25$ reports keep $\geq$ 90% of citations); neutral-insert (unrelated SI-unit note, no epistemic claims); whitespace-only; paragraph reorder — all judge-only, no regeneration. **Adjudication (C3).** 76 break-flips (item covered@turn-1, uncovered@turn-2, non-target) sampled across content-preserving controls and real methods; an independent judge (GPT-4o, temperature 0, run from a non-DGX endpoint since OpenAI is unreachable on the DGX) was given the full after-report and labeled whether the criterion's content is still present; a second adjudicator, which is *not* independent of the pipeline (deepseek-chat is also the primary coverage judge and the reviser, so its leniency bounds rather than corroborates) (DeepSeek) repeats the pass for independence, both in the standard prompt and in a blinded variant that withholds which edit produced the report. The two adjudicators agree on direction but on only about half the individual flips, and the disagreement is one-directional; see §4. We report the re-run values rather than an earlier pass whose verdict file was not retained and which does not reproduce — itself an instance of the drift documented in §4. The 76-flip sample, the prompts, and every model's per-flip verdicts are released. This is an automated proxy; *human* adjudication remains the arbiter and the blind instrument for it is released with the code. **Significance.** Paired bootstrap, 10,000 resamples, seed 0. **Compute.** $\leq$4$\times$A100-40GB; DR-Tulu/Qwen served on 1–2 GPUs; reviser and judges are API/served calls (no training). Env: vLLM 0.8.5.post1, torch 2.6.0+cu124, transformers 4.51.3. **Code and logs.** `code/` (reviser, controls, evaluator adapters); every result traces to a row in `experiments/runs.csv` with its per-run log and JSON.

C Supporting figures

The figures below carry no evidence that is not also reported numerically in §4 and Tables 12–11; they are placed here because the main text is subject to a page limit and this appendix is not. Most are

referenced from the main text; two are read directly from here, and are the two worth reading if you read only two. Figure 4 makes the central claim visually: break rate is decoupled from how much a method actually edits, so the metric is not measuring edit magnitude. Figure 11 in Appendix H is the whole C3 argument in one panel-set — under every judge whose full battery we ran, no real revision reaches the floor, and the control that attains it is always the paragraph reorder.

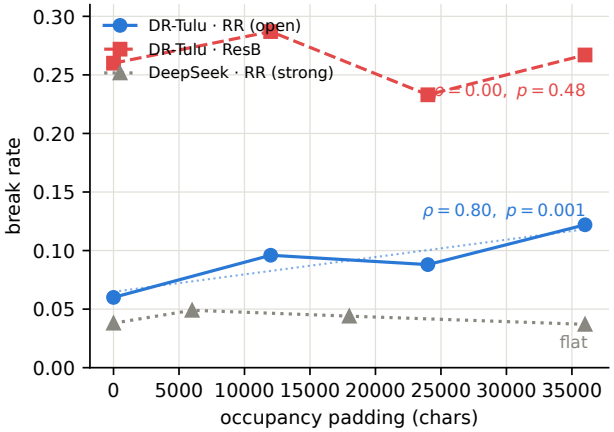

Figure 2: **Occupancy drives breakage for the open agent, not the strong closed reviser** (C1; the numbers are Table 12). Break rate vs. context occupancy—neutral filler of increasing length prepended to the reviser's input, with the edit action held fixed. DR-Tulu-8B self-revising on RR rises with occupancy (Spearman 0.80); the strong DeepSeek reviser is flat, and on ResB the signal is swamped by a high judge floor.

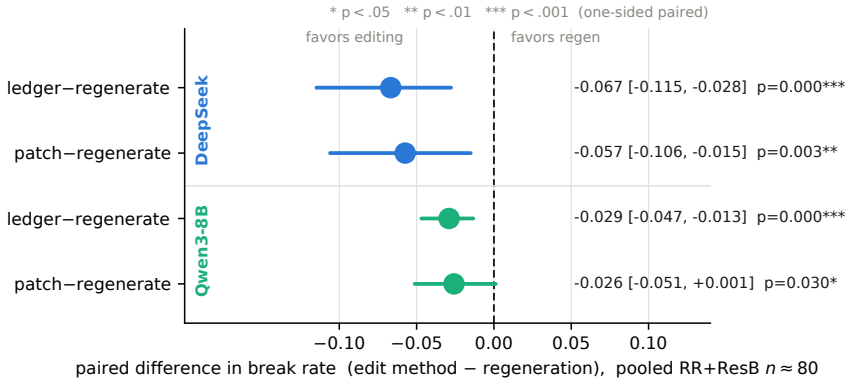

Figure 3: **Constrained editing beats regeneration under both judges** (C2). Paired difference in break rate (edit method − regeneration) with 95% bootstrap CIs, pooled over RR+ResB ($n \approx 80$). All four point estimates favor editing; DeepSeek's intervals and Qwen's ledger interval exclude zero, and Qwen's patch difference is significant by the one-sided paired test ($p = 0.030$).

D THE OCCUPANCY DIAGNOSIS IN FULL

D.1 BREAKAGE IS OCCUPANCY-DRIVEN—IN THE OPEN AGENT

Why occupancy. The mechanistic hypothesis rests on long-context degradation: models under-use information buried in long inputs (*lost-in-the-middle*, Liu et al., 2024) and degrade as raw token count grows even absent a positional confound (Hong et al., 2025). Those results are established for retrieval/QA needles; what follows is suggestive evidence that the effect transfers to multi-turn long-form report revision, and only for the weak open agent, where each regeneration re-exposes the

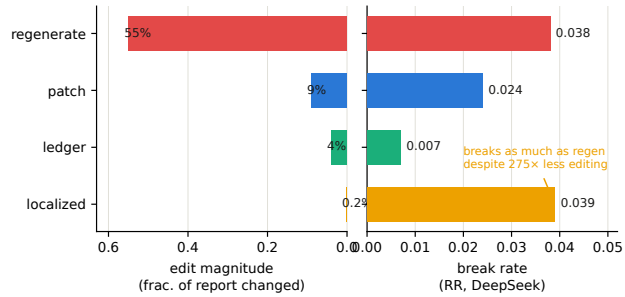

Figure 4: **Break rate is decoupled from how much a method edits.** Edit magnitude (fraction of turn-1 lines changed) vs. break rate (RR, DeepSeek). The localized reviser changes 0.2% of the report yet breaks as much as full regeneration, which rewrites 55%.

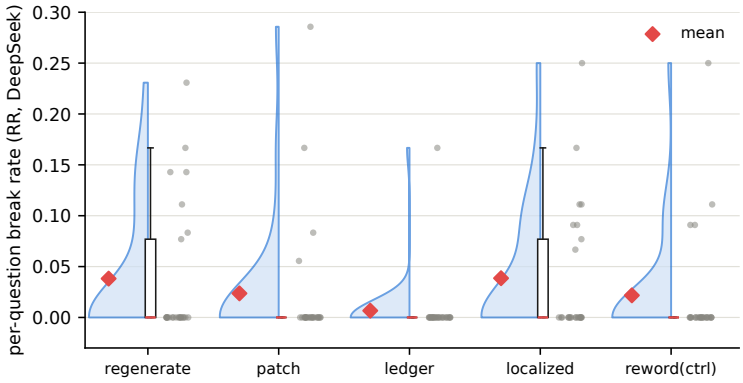

Figure 5: **Break rate is carried by a few unstable items.** Per-question break-rate distribution (RR, DeepSeek). Most questions sit at zero; the mean (red) is pulled up by a small tail of flipped items — the same items the second judge largely does not flip (Fig. 6).

whole report to occupancy-driven loss. Holding the edit action fixed and varying only the reviser's context occupancy, on ResearchRubrics DR-Tulu-8B's own break rate rises with occupancy (Fig. 2, Table 12; Spearman 0.80, paired pad 36k$-$pad $0 = +0.062$, p=0.001), whereas the strong DeepSeek reviser is flat. Breakage is thus a *weak/open-reviser* phenomenon—a context-tracking ("lost-in-the-middle") failure of the deployable agents, not of strong closed models. **Scope.** On ResearcherBench the same sweep is flat (Spearman 0.00, p=0.476; break rate $\sim$0.26 throughout). This is not report length (ResB turn-1 reports are *shorter*, 18K vs. 22K chars—both below the $\sim$33K Setup default, as this self-revision uses a search-capped config); rather, ResB's judge-instability floor is high (§4, Table 11), so its break rate is noise-dominated and any occupancy signal is swamped. The effect is measurable exactly where the judge is stable (RR)—a dependence predicted by C3—but rests on one open agent, four occupancy levels, and one dataset, so we read C1 as suggestive rather than conclusive.

Judge-agreement statistics. Per-item coverage agreement between the two primary judges is moderate (Cohen κ=0.56 over 664 items) while per-report coverage correlates well (r=0.78). The load-bearing contrast is that per-question *break* rate correlates only r=0.27 across judges and the signal sits in a small tail of items (Figs. 5, 6): the judges largely agree about how much a report covers, and disagree about which individual items flipped — exactly the pattern instability predicts and shared content loss does not.

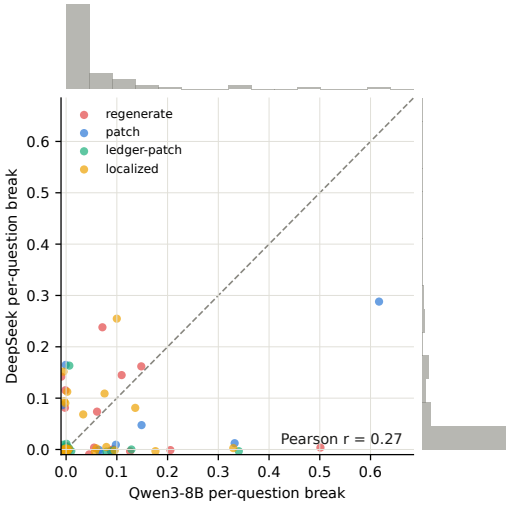

Figure 6: **The two judges disagree on which items break.** Per-question break rate (RR), DeepSeek vs. Qwen3-8B, by method. Mass off the diagonal (r=0.27) means the item a break is attributed to is largely judge-specific.

E INDEPENDENT-MODEL ADJUDICATION

Independent-model adjudication. As a scalable, reproducible complement to the blind human pass below, we asked an independent, different-family judge (GPT-4o), given the *full* after-report, whether each flagged item's content is still present (Fig. 10). On the 47 control flips — where no content *can* be lost — GPT-4o confirms presence for 62%; since content is preserved by construction, every control break is instability whether the error is the primary judge's or GPT-4o's. On the 29 real-revision flips it confirms only 24% (regenerate 19%). The gap is what C3 predicts. Two limits: GPT-4o is itself an LLM judge, so this corroborates rather than proves; and we adjudicated only *flagged* items, with no unflagged negative control, so 62% bounds adjudicator error from one side only.

Adjudication does not converge, and that is the finding. A second adjudicator from a different family agrees on the *direction* and little else (Table 14): the two concur on about half the individual flips, and the disagreement runs one way — the second never returns *lost* where the first returns *present*. Each is wrong in a way the other exposes: the first errs on a third to a half of the controls, where nothing *can* have been lost, so those verdicts are provably mistaken; the second is near-perfect there but credits most *full rewrites* as intact too, which a lenient rater achieves for free, and blinding it to the edit moves it only four points, so the leniency is real rather than a prompt leak. Propagated through the per-item correction the two differ by a factor of five to seven (0.064 vs 0.009, or 0.013 blinded). We therefore print no single per-item number: more model adjudicators cannot settle a systematic disagreement. Only a blind human pass can — the rater sees the criterion and report but not which edit produced them — and we release the tool that performs it.

The blind human pass. One rater saw the criterion and the after-report for each of the 76 flagged breaks, in a randomised order, with the condition that produced each report replaced by an opaque token; a sidecar map joined the verdicts back afterwards and all 76 resolved. The three-way scale is *present* / *partial* / *lost*.

arm	condition	n	present	partial	lost
control	whitespace only	15	12	3	0
control	paragraph reorder	23	17	4	2
control	reword one paragraph	9	8	1	0
real	patch	13	10	2	1
real	regenerate	16	13	2	1

Two readings, and we report both. **Against our earlier corroboration:** the human does not reproduce GPT-4o's separation — 78.7% of control flips and 79.3% of real ones are called present, a difference of -0.006 $[-0.186, +0.183]$. **The stronger reading:** 38 of the 47 control flips are whitespace-only or paragraph reorder, transforms under which every word of the report survives, so any verdict other than *present* there is the rater's own error — and that happens 23.7% of the time. Correcting the real arms by that floor leaves -0.030 $[-0.230, +0.171]$ overall, -0.049 for regeneration and -0.006 for patch: all negative in point estimate, none excluding zero, and unchanged if the scale is cut at *lost* instead of at *present* ($+0.016$, $[-0.097, +0.138]$). Paragraph reorder was called outright *lost* 2 times, against regeneration's 1.

Scope. One rater and one verdict per flip, so inter-rater reliability is unmeasured; at $n{=}47$ control and $n{=}29$ real the intervals are about ± 0.2 wide — a two-proportion test here has 80% power only for a control-vs-real gap of $\gtrsim 0.27$ — so this rules out a large per-item gap rather than establishing exact equality, and a second blind rater with an inter-rater κ is the natural next step. What it does establish is a floor on the instrument: a careful human, blind to condition, declines to confirm content that provably survived in about one flagged item in four. No per-item adjudication of this metric — by a model or by us — can resolve differences smaller than that, which is why two model adjudicators could differ by a factor of five to seven and why we decline to print a single per-item true-loss number.

Two properties of FLOOR as an estimator. First, it is biased *upward*: it is a maximum over noisy control estimates, so it overshoots the quantity it estimates even when the control set is complete. By bootstrap the bias is $+0.0034$ on ResearchRubrics ($+5.3\%$ of the floor), small because a single control wins the maximum in 93% of resamples, and $+8.4\%$ on ResearcherBench, where three controls compete for it. Second, it is biased *downward* whenever the control set is incomplete, and by much more: adding paragraph reorder to the three cheap controls multiplies the floor by 2.9 under DeepSeek, 1.6 under Qwen3-8B and 4.4 under Qwen3-32B — every judge for which we ran it (Table 1). The two biases run in opposite directions and the second dominates, so we do not claim a net direction for $\mathrm{BRK}_{\mathtt{true}}$ and report every floor-corrected residual as an interval.

Reorder is the one control whose churn is not symmetric. The delta we claim over the LLM-judge reliability literature (§2) is that break rate reports one side of a churn whose two sides very nearly cancel: mean coverage moves $+0.0005$ against a break rate of 0.022 for a one-paragraph reword, $+0.0008$ against 0.021 for a whitespace-only edit. Paragraph reorder is the exception, and we say so rather than leaving it out of the list: net coverage falls -0.042 under Qwen3-32B and -0.025 under DeepSeek, roughly 40% of the corresponding break rate. For this one perturbation the judge does not merely churn — it reads a permuted report as genuinely covering less. That is a stronger form of judge dependence than two-sided noise, not a weaker one, and it is why the reorder control carries the floor under all three judges that scored it.

The order confound against C2, withdrawn on measurement. §4 shows the metric tracks paragraph position, which would confound C2 if constrained editing preserved paragraph order and regeneration did not. Neither does. Aligning every turn-2 report's paragraphs to its turn-1 paragraphs — the same blank-line split that builds the reorder control, greedy `difflib` match at ratio 0.60 — the surviving paragraphs reproduce turn-1 order *exactly* in all four arms: Kendall's $\tau{=}1.000$ for patch, ledger-patch, localized and regeneration alike. Mean positional displacement runs 0.019 for patch to 0.048 for regeneration, against 0.344 for the paragraph-reorder control, so the most disruptive arm moves paragraphs seven times less than the control does. Regeneration does not reorder: it rewrites and *adds*, from 27 turn-1 paragraphs to 34, while leaving 0.80 of the originals in place against patch's 0.93. We checked the alignment before believing it, because a τ of exactly 1 across four arms is the shape of an artifact: run against the reorder control itself, where the permutation

is known because a fixed seed generates it, the same matcher returns τ=0.044 and displacement 0.344, and recovers the true position of all 673 paragraphs. The confound is therefore absent rather than untested. One thing the same alignment shows *for* the metric: within patch, the more turn-1 paragraphs survive the lower the break rate ($r = -0.72$ under the primary judge, -0.73 under the cross-check), so break rate does respond to genuine content survival and not only to position. That holds for patch alone — the other three arms show no consistent relation — and we claim it that narrowly.

F FOUR FURTHER CHECKS

The floor > signal model, derived and checked (§3). Write the judge's per-item chance of re-scoring a covered criterion as lost under transformation T as $\rho + \nu(T) + \lambda(T)$: identity repeat-noise ρ, spurious content-preserving sensitivity $\nu(T) \geq 0$, and true loss $\lambda(T) \geq 0$ ($\lambda = 0$ for every content-preserving control by construction). Averaging over the covered items of a dataset gives break rates $B_c = F_\rho + N_c$ for a control and $B_m = F_\rho + N_m + L_m$ for a real arm, with $F_\rho = \mathbb{E}\rho$, $N = \mathbb{E}\nu$, $L_m = \mathbb{E}\lambda$, and $\text{FLOOR} = F_\rho + \max_c N_c$. *(i)* A control out-breaks arm m iff $B_c > B_m$ iff $N_c > N_m + L_m$: subtracting F_ρ from both sides, the common repeat-noise cancels and the comparison is between *sensitivities*, so an arm sitting below a control is not an artifact of the judge being noisy in general. *(ii)* From a single break rate N_m and L_m are not separately identified — only $B_m - F_\rho$ is — so any arm with $B_m \leq \text{FLOOR}$ admits $L_m = 0$; the floor is a *detection threshold*. The regime is instantiated on RR under every judge (identity F_ρ, floor = reorder, and each real arm's excess $B_m - F_\rho$, all as break rate):

judge	F_ρ (identity)	FLOOR (reorder)	regen $B - F_\rho$	patch $B - F_\rho$
DEEPSEEK	0.005	0.064	0.033	0.019
QWEN3-8B	0.002	0.065	0.057	0.055
QWEN3-32B	0.000	0.091	0.050	0.072
GPT-4.1-MINI†	0.031	0.107	(paired $+ +$ 0.037)	(paired $+ +$ 0.060)

In every row $\max_c N_c = \text{FLOOR} - F_\rho$ exceeds each real arm's $B_m - F_\rho$, so the data lie in the $\max_c N_c > N_m + L_m$ regime of (i). † F_ρ under GPT-4.1-MINI is measured, not assumed — a second scoring pass over the byte-identical reports gives 0.031 on the same 65 questions — so this row is an estimate of the model's inputs rather than a bound on them. By (i) it would in any case *cancel* in the paired reorder−regen and reorder−patch contrasts, which are positive and Holm-significant (§4).

Intervals on the flagship column, and every ordering restated as a paired contrast. A referee asked for uncertainty on the cells the paper's claims rest on, and for the orderings to be paired rather than read off marginals. Question-level bootstrap (10,000 resamples, seed 0) over the same 25 ResearchRubrics questions under GPT-4.1-MINI, for every condition whose per-item scores we hold: identity 0.031 [0.020, 0.043]; reword-every 0.048 [0.032, 0.066]; within-section reorder 0.080 [0.056, 0.107]; paragraph reorder 0.107 [0.078, 0.141]; patch 0.047 [0.028, 0.070]; regenerate 0.070 [0.051, 0.090]. Paired on the same questions, the reorder exceeds regeneration by +0.037 [+0.010, +0.066], patch by +0.060 [+0.028, +0.092], identity by +0.083 [+0.048, +0.124] and a full reword by +0.059 [+0.029, +0.094] — all four excluding zero. **Two orderings do not**, and we print them as such: the full reorder against the within-section one (+0.027 [+0.009, +0.046]) and the within-section reorder against patch (+0.045, p=.062). The first is the informative null — on these reports the cross-section component contributes nothing measurable.

Intervals for the instrument battery (Table 2). Same bootstrap, same questions. Read-and-check instruments, identity then reorder: B batch claim-recall 0.006 [0.000, 0.018] then 0.128 [0.051, 0.231]; B under a second judge family 0.010 then 0.074 [0.042, 0.114] (identity [0.000, 0.030]); B on GPT-4O reports 0.016 [0.000, 0.044] then 0.159 [0.097, 0.225]; C graded requirement 0.020 [0.008, 0.033] then 0.124 [0.081, 0.173]. Order-invariant designs, identity / reorder / reword: D lexical 0.028 [0.013, 0.045] / 0.028 [0.009, 0.053] / 0.109 [0.073, 0.146]; D semantic 0.040 [0.011, 0.085] / 0.039 [0.009, 0.084] / 0.119 [0.075, 0.168]; E 0.079 [0.037, 0.129] / 0.083 [0.047, 0.132] / 0.076 [0.033, 0.128]. The two facts worth stating separately: D's identity and reorder intervals coincide, which is what exact position-invariance looks like when it is mea-

sured rather than argued; and E's three intervals overlap one another, so its lower point threshold is not an established ordering.

One estimator note, because row E is not computed like the others. Rows A–D average the per-question break rate over questions; row E pools over statements, which is the convention RAGAS's own decomposition invites because its item set is not fixed across calls. Averaged over questions instead, E reads 0.076 / 0.087 / 0.076 against pooled 0.079 / 0.083 / 0.076, and its intervals are $[0.037, 0.136]$ / $[0.042, 0.135]$ / $[0.033, 0.132]$ pooled against $[0.037, 0.129]$ / $[0.047, 0.132]$ / $[0.033, 0.128]$ averaged. Every conclusion we draw from row E survives either choice: the three cells overlap under both, the reorder-versus-identity gap contains zero under both, and the paired read-and-check-versus-E threshold contrast — which we compute per question for *both* designs, so the estimator is common to the two sides — contains zero under both. We state the convention rather than silently pick the flattering one.

Within-section reorder (retrievability vs. instability). A full paragraph reorder moves content *across* sections, which can reduce a claim's retrievability even though no word changes; a *within-section* reorder — permuting paragraphs only inside each section, discourse order preserved, word-multiset identical — isolates the residual. On the original DR-TULU reports under GPT-4.1-MINI the within-section break is 0.080, barely below the full reorder (0.107) and still above regeneration (paired $+0.010$, $p{=}.415$) and patch (paired $+0.045$, $p{=}.062$). On the more sharply-sectioned GPT-4O reports the full reorder is larger (RR/ResB 0.10–0.23) and the within-section value lower (0.03–0.10), so there the cross-section move does carry real retrievability loss — but the within-section floor still lands at the real-revision level in all six (dataset, judge) cells (full / within / regen / patch, RR then ResB: $0.151/0.071/0.076/0.068$, $0.216/0.065/0.038/0.066$, $0.144/0.066/0.050/0.060$; $0.225/0.104/0.155/0.099$, $0.171/0.031/0.049/0.048$, $0.104/0.053/0.056/0.047$ for DeepSeek/GPT-4.1-mini/GPT-4o). We therefore state the claim as: even a discourse-preserving reorder breaks about as much as the reviser's real revisions, and the extra break of a full reorder we do not claim as instability.

Judge reliability. The two primary judges agree well on per-report coverage ($r{=}0.78$) but only weakly on *which* items break ($r{=}0.27$) — what judge-specific flips rather than shared content loss would look like (Appendix D). **Two elaborations that do not help.** A removal-justifying *ledger* and a low-occupancy *localized* reviser both fail to beat plain patch; at the expanded n the ledger$-$patch difference is bounded within 0.02 under both judges (Table 13), a more informative statement than a large p. C3 explains the null: with content preserved, the residual is instability, which no edit strategy can reduce. **Three revisions instead of one.** Against the *original* report, cumulative damage barely moves over three sequential revisions (regeneration $0.0544 \rightarrow 0.0581$, control $0.0152 \rightarrow 0.0050$) even as reports grow 15–25% (Table 16) — suggestive of a one-time toll rather than a spiral, but these are single runs at $n{=}25$ with no intervals and should not be read more finely than the 0.005 by which the nominally identical first-turn control moved on re-running. **A labelling correction.** An additional fifteen questions per dataset (Table 17) reproduce the ordering on ResearchRubrics while collapsing towards zero on ResearcherBench. We previously called this slice held out; it is not, and we correct that here — it is folded into the $n{\approx}80$ pooled comparison, so it cannot independently replicate a test it is part of. The independent material is the later twenty-five questions per dataset, generated after every analysis here was fixed, which move the pooled comparison to $n{\approx}130$ without changing an ordering.

At a larger sample. Because these are the paper's only positive claims, we generated twenty-five further questions per dataset and re-ran the identical protocol, taking the pooled comparison from $n{\approx}80$ to $n{\approx}130$ (Table 13). Every ordering survives and tightens, and the more useful gain is on the negatives: the failed elaborations are now *bounded* within 0.02 rather than merely non-significant.

But $n{\approx}130$ is an enlargement, not a replication, and we test the difference. Eighty of those 130 are the original pool, so 62% of the "independent" material is the test it would corroborate. The genuinely unused material is the 50 later questions on their own, and on those alone: the direction replicates in *all* 12 cells (three edit arms $\times$ two datasets $\times$ two judges are every one negative), but of the four confirmatory contrasts only one excludes zero — patch$-$regenerate under the cross-check judge, -0.016 $[-0.031, -0.002]$, $p{=}0.028$ — and it does *not* survive Holm over that family of four (0.028 against 0.0125). Under the primary judge the same contrast is -0.028 $[-0.079, +0.012]$. So

C2's significance rests on the $n{\approx}80$ pool; out of sample we can claim a consistent direction and one nominal contrast, and we claim no more than that. It is *not*, however, a single dataset carrying the pooled result. Under the cross-check judge the two datasets agree to within 0.001 (-0.015 on RR, -0.016 on ResB), and for ledger−regenerate under that judge it is ResearchRubrics that carries the effect (-0.014) while ResearcherBench is essentially zero (-0.000); only under the primary judge is ResearcherBench the larger (-0.046 against -0.010). Which dataset carries the effect flips with the judge, so neither does it alone. Filling the set required one cell never run before — the localized reviser on ResearcherBench — with which it now beats regeneration under both judges rather than one.

A magnitude-matched floor absorbs regeneration's residual on one dataset and not the other. The reword floor above changes one paragraph, whereas regeneration rewrites 0.55 of the report; subtracting a one-paragraph floor therefore *under*-corrects regeneration. We built the missing edit-magnitude-matched control: a *full-report* reword that rewrites every paragraph while retaining all facts and citations (edit magnitude 0.66, exceeding regeneration's 0.55; $25/25$ reports keep $\geq 90\%$ of their citations). On RR it breaks at 0.040, indistinguishable from regeneration's 0.038 ($+0.002$, $[-0.041, +0.057]$): there the positive $\mathrm{BRK}_{\mathrm{true}}$ residual is an artifact of correcting with too small a floor, and even full regeneration is largely instability. **It does not carry to ResearcherBench — which we tested precisely because this was the paper's strongest claim, made on the dataset with the smallest residual.** There the matched control reaches 0.125 against regeneration's 0.259, significantly *lower* (-0.133, $[-0.258, -0.030]$, $p{=}0.008$), so we state the instability reading of regeneration for RR only.

The ResearcherBench floor is not an artifact of the filler's wording. The largest ResearcherBench control is a neutral insertion, so its size could in principle be a property of the particular sentence inserted rather than of insertion itself. Swapping the SI-unit note behind 0.140 for a legal disclaimer of similar length gives 0.135 — the same floor to within 0.005. Both variants are produced by the released generator (`code/scripts/make_controls.py`), which reproduces every published control artifact byte-for-byte.

The order-preserving floor, and the one cell it costs us. FLOOR as defined takes the maximum over every content-preserving perturbation, and on ResearchRubrics that maximum is attained by the paragraph reorder under all four judges that scored it. But reorder is also the one control whose churn is not symmetric — net coverage genuinely falls, -0.042 and -0.025 — and *patch*, *ledger* and *localized* never permute paragraphs, so charging them the reorder floor corrects them by a quantity they cannot incur. Restricting the maximum to the order-preserving controls (reword-one, neutral insert, whitespace, full reword) gives 0.040 (DeepSeek), 0.051 (Qwen3-8B) and 0.023 (Qwen3-32B), attained by full reword in every case. Against that stricter comparator the order-preserving arms still sit below the floor — patch -0.016, ledger -0.033, localized -0.001 under DeepSeek; ledger -0.025, localized -0.004 under Qwen3-8B — with one exception we report rather than round away: **patch under Qwen3-8B sits** $+0.006$ **above** its order-preserving floor. That margin is far inside the noise of a 25-question cell and we do not claim it clears. Under Qwen3-32B the printed patch pass clears by $+0.049$ and its re-run sits -0.014 below, which is the drift result of §4 rather than a fact about edit strategy. **Regeneration does not permute paragraphs either** (App. E measures $\tau{=}1.000$ for every arm), so the order-preserving floor is the operative comparator for it too, and against that comparator it sits *above* the floor by point estimate under two of the three judges ($+0.008$ under QWEN3-8B, $+0.027$ under QWEN3-32B; -0.002 under the primary judge). We state that rather than compare regeneration against the looser reorder-inclusive floor: an earlier version of this paragraph justified doing so on the grounds that regeneration permutes paragraphs, which our own alignment measurement contradicts. What the residual survives is settled below by resampling the floor inside every bootstrap replicate, not by the choice of comparator.

The withdrawn residual, and the one that survives. Against the *maximum* floor, ResearcherBench regeneration leaves $+0.081$ $[-0.040, +0.167]$ on the core questions, $+0.011$ on a disjoint 25 and $+0.052$ $[-0.027, +0.095]$ pooled: none excludes zero once the floor is resampled rather than held fixed, which is why §4 withdraws it. Against the *magnitude-matched* control that §3 prescribes where one exists — and one does here, the full reword's edit magnitude 0.709 exceeding regeneration's 0.581 — the same cell gives $+0.133$ $[+0.029, +0.254]$ on the core questions, $+0.011$

$[-0.043, +0.064]$ on the disjoint 25, and $+0.071$ $[+0.009, +0.139]$ (p=0.034) pooled. The entire difference between the two is the max-selection bias: the maximum floor is set by the whitespace control at 0.177, both higher and noisier than the matched control's 0.125. We therefore report one positive residual and refuse to headline it, on two grounds we would accept from a referee: one question slice carries it, and 0.034 does not clear the $0.05/4$ Holm threshold this paper applies to its own confirmatory family.

Does the conclusion depend on which floor we use? It does not, and this is worth stating because on point estimates it looks as though it might. §3 compares arms that never permute paragraphs against the order-preserving floor, and no arm does permute them (App. E), so that floor is the operative one for every arm: 0.040 under the primary judge, 0.051 under the cross-check, 0.023 under QWEN3-32B — all below several arms, where the reorder-inclusive floors are all above them. Recomputing FLOOR as a maximum over the *resampled* controls inside every bootstrap replicate, however, *no* positive residual excludes zero under either definition; the largest, QWEN3-32B regeneration against the order-preserving floor, is $+0.027$ $[-0.001, +0.051]$ and its lower bound is barely negative, which we state rather than round away. One interval does exclude zero, in the other direction: against the reorder-inclusive floor, QWEN3-32B regeneration sits -0.041 $[-0.079, -0.003]$ *below* its floor — under the judge that reproduces itself exactly, a full rewrite provably breaks less than a permutation that changes no word.

The result under the audited system's own definition of "covered". Our break rate calls a checklist item covered when the judge returns a non-zero score, so a *partially satisfies* (0.5) counts. MR-DRE's shipped evaluator does not: its `checklist_eval` counts an item only at a full 1.0. Both read the same three-point scale that MR-DRE's own checklist prompt asks for — we import that prompt module rather than restate it — so the difference is a definition of *covered*, not a second measurement, and it can be applied to the archived per-item judgements without a single new judge call. We re-ran the entire second-agent replication under it. On RESEARCHRUBRICS the phantom break is unmoved: the full reorder costs 0.128–0.222 across the three judges against 0.054–0.074 for patch and 0.063–0.064 for regeneration, and all six reorder-versus-arm contrasts keep intervals excluding zero ($+0.069$ to $+0.148$ against patch, $+0.064$ to $+0.159$ against regeneration; the weakest, GPT-4O, is $[+0.030, +0.111]$). The discourse-preserving control stays where it was (0.065–0.071). **On RESEARCHERBENCH it does not survive intact, and we report that:** under DEEPSEEK the strict rule discards so much of the denominator that the patch rate jumps to 0.310 and reorder-versus-patch collapses to $+0.000$ $[-0.143, +0.143]$ (p=1.00); the same cells under GPT-4.1-MINI and GPT-4O still hold. So the effect is not an artefact of crediting partial coverage, but on the smaller checklists the strict rule costs the power needed to see it. We could not run this check on the flagship cell: its item-level record did not survive, and re-deriving it needs fresh judge calls rather than re-analysis.

G COMPOSITE FIGURES FOR THE POSITIONAL RESULT

Both figures here are drawn from the same cells as Tables 1 and 2; nothing in them is new evidence. They exist because three of the paper's claims are about *ratios* and *scopes* that a table of raw break rates does not make visible: an instrument's phantom loss only means something against its own repeat noise, and the reader has to be able to see at a glance which comparisons we do and do not claim.

H ADDITIONAL TABLES AND FIGURES

The main text is subject to a page limit and this appendix is not, so the tables and figures below live here. The main text references them exactly as if they were inline.

The C2 mitigation in full (Table 6, Table 7, Fig. 3). §4 states the result in one paragraph; this is the treatment behind it. Table 6 gives every arm under both judge families on both datasets: *patch* and *ledger* cut break rate well below full regeneration while *localized* does not (0.039 vs. 0.038 on RR/DeepSeek) — one of the two elaborations whose failure prompted this audit. Pooled over RR+ResB at $n{\approx}80$ the ordering holds under both families (patch$-$regenerate p=0.003/p=0.030;

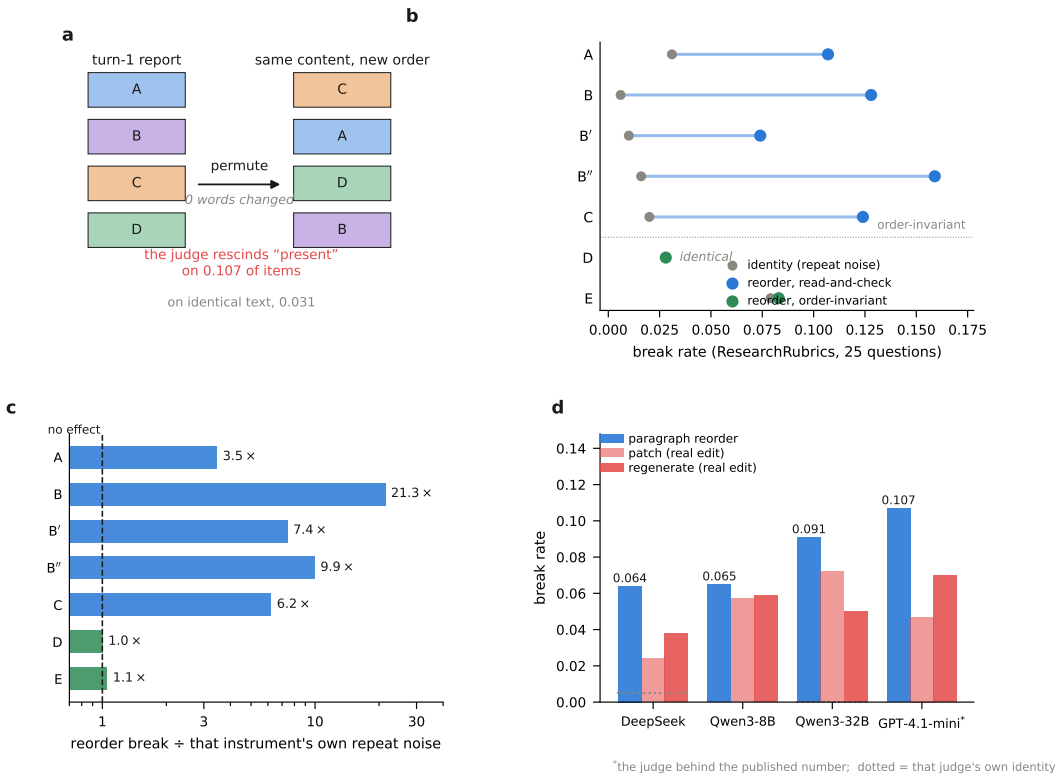

Figure 7: **Positional Coverage Bias is a property of the read-and-check protocol, not of one prompt or one judge.** **(a)** What the control does. The paragraphs of a turn-1 report are permuted with a fixed seed; not one word changes, and every citation, number and claim survives verbatim. The judge nonetheless rescinds "present" on 0.107 of previously covered items, against 0.031 when it re-scores the *identical* text. **(b)** Identity against paragraph reorder for every coverage instrument we ran, same 25 ResearchRubrics questions throughout. The span of each dumbbell is the phantom loss. *A*: per-item binary presence, the MR-DRE protocol. *B*: batch claim/nugget recall. *B′*: the same protocol under a second judge family (GPT-4O). *B″*: the same protocol on a second agent's reports. *C*: graded requirement satisfaction, INFOBENCH-style. *D*: our retrieve-then-judge construction. *E*: decompose-and-search, RAGAS's `LLMContextRecall` run from its shipped prompt file unmodified. The dotted rule separates the read-and-check instruments from the two that cannot see where the evidence sat. **(c)** The same measurements as a ratio to each instrument's *own* repeat noise, on a log axis. This is the comparison that matters and the one raw break rates hide: the instruments' noise floors differ more than twentyfold, so a large break rate is not by itself evidence of anything. Every read-and-check instrument sits far right of 1; both order-invariant ones sit on it. **(d)** The four judges, each against its own two revision arms on the same questions. The permutation out-breaks both real edits under every judge, including GPT-4.1-MINI, the judge that produced the published 16–31%. Dotted rules mark the two judges whose identity re-scoring we measured directly (0.005 and 0.000).

ledger−regenerate $p<0.001$ both), all 4 tests survive Holm–Bonferroni at $\alpha=0.05$, and three of four bootstrap intervals exclude zero (Fig. 3). **The confirmatory family is those four** (two orderings $\times$ two judges); ledger−patch is null, and the nine instability contrasts (Table 9) are a separate family in which nothing survives Holm. **Two scope limits.** The pooling is RR+ResB only — on RigorousBench the ordering *inverts* (patch 0.100 vs. regenerate 0.090), so we exclude it and do not claim the mitigation there — and the gain trades incorporation for break rate under DeepSeek (patch 0.80 vs. 0.92), a favourable trade rather than dominance, roughly even under Qwen. Table 7 then applies Eq. (1) arm by arm: only the weak open agent's full regeneration clears its own floor, which is why the main text reports this as a raw-break-rate ordering and nothing stronger.

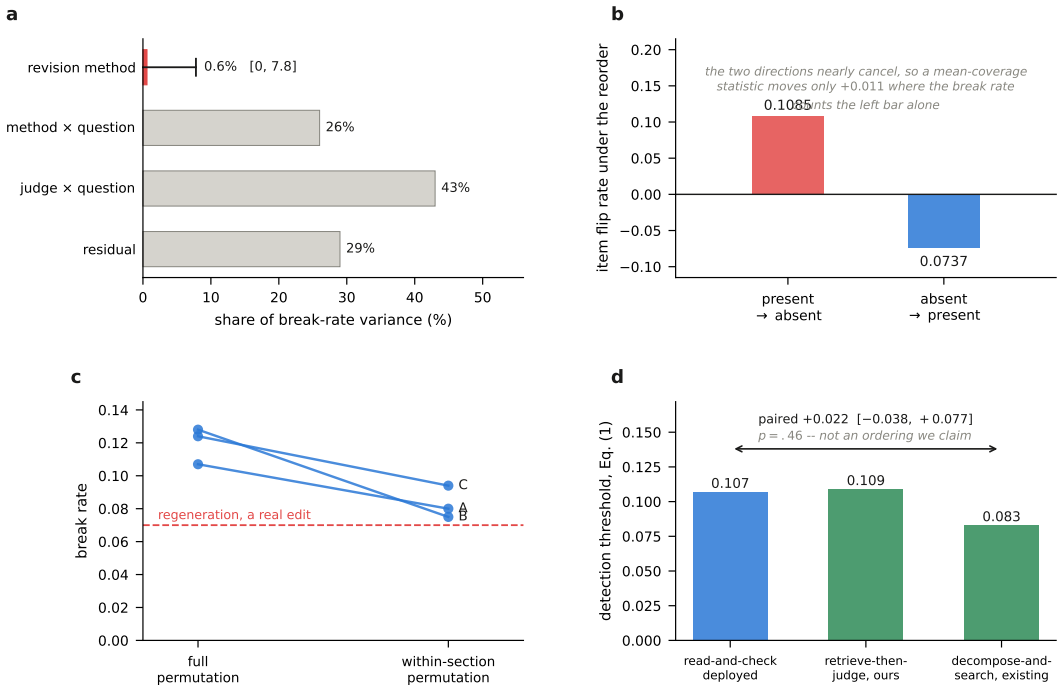

Figure 8: **What a coverage break rate is made of, and what we decline to claim from it. (a)** Generalizability decomposition over four judges and four edit strategies. The revision method—the thing the metric is deployed to measure—takes 0.6% of the variance, with the interval shown; judge $\times$ question takes 43%. The scope matters and we state it rather than let the figure imply more: the four strategies are deliberately near-equivalent, so this bounds the metric's power to separate *those*, and does not license extrapolation to the between-agent spread behind the audited range. **(b)** The permutation moves items in both directions. Counting only the left bar is what an item-level break rate does; because the two nearly cancel, a mean-coverage statistic moves only $+0.011$ on the same data and sees almost nothing. This is the measurement that refuted an earlier two-property design principle of our own, which we withdrew rather than write up. **(c)** Restricting the permutation to *within* each section, so discourse order is preserved, on the three read-and-check instruments. The effect shrinks and does not vanish: all three stay above the same judge's real regeneration. On a second agent's sharply sectioned reports it does not replicate, which is the scope we claim and no more (Table 5). **(d)** The detection threshold of Eq. (1) for the deployed read-and-check protocol, for our order-invariant construction, and for the existing off-the-shelf one. Our construction removes the position term exactly and does *not* lower the threshold; and paired over the same questions the three are not distinguishable at $n{=}25$, so we report the negative result and claim no ordering.

Decode noise, within a session. Every revision in Table 6 is a *single* temperature-0 decode, which is not deterministic, so we re-ran the protocol five times varying only the decode, judge fixed (Table 15). *Our own* reviser keeps the ordering by a comfortable margin — the regenerate−patch gap is $5.6\times$ its own decode spread under the zero-floor judge, which is the judge Table 15 holds fixed and therefore the ratio that table exhibits, and $5.4\times$ when the same five decodes are re-scored under the paper's primary judge, which is not tabulated here; it favours patch in all five repeats under both (two-sided paired t over the five paired repeats, $p < 0.001$) — but a *weaker* reviser does not: its "constrained" edits break as often as its rewrites (0.0420 vs. 0.0444, $p{=}0.379$), so the mitigation is conditional on a reviser competent enough to emit real edit operations. The ledger variant has the largest decode spread of any strategy (0.0166), the mechanism behind a null we could previously only describe. This bounds *session* variance only; replication variance is far larger.

Table 3: **Generalizability-theory decomposition: the break rate barely measures the arm it is used to rank.** Variance components with the revision *method* as the object of measurement and judge and question as random facets, by balanced random-effects ANOVA (§3). Across three designs the method explains only 0–6% of break-rate variance and the dependability is $\Phi \in [0.00\text{–}0.37]$, far below the 0.80 a measurement needs to rank on. In the headline RR design the components are: method 1.3%, judge×question 43%, method×question 26%, residual 29% — the judge×question interaction, not the method, is the largest term. A D-study on that design (questions fixed) reaches only $\Phi = 0.24$ at *fifty* judges: averaging raters cannot make the metric dependable, because the method×question variance is irreducible in the judge count. Per-component bootstrap intervals for the four-judge design are in Table 4: the method share is 0.6% with a 95% upper bound of 7.8%, so the object of measurement is the smallest term under every resample. We report the relative-error coefficient $E\rho^2$ — the standard choice when the use is *ranking* the objects — beside the absolute-error Φ: on the multi-arm designs it is no higher (≤ 0.45 at four RR methods, 0.00 once GPT-4.1-mini joins), and rises toward a usable value only on the two-arm ResB design (0.72), where a single contrast rather than four must be separated — so the ranking coefficient does not rescue the metric on the designs that actually rank several arms.

design	method variance	$E\rho^2$	dependability Φ
RR, 4 methods × 2 judges × 25 q	1.3%	0.45	**0.31**
RR, 4 methods × 3 judges (incl. GPT-4.1-mini)	0%	0.01	0.01
RR, 4 methods × 4 judges (adds GPT-4o)	0.6%	0.24	0.18
RR, 2 methods × 3 judges × 25 q	≈ 0%	0.00	0.00
ResB, 2 methods × 2 judges × 24 q	6.3%	0.72	0.37
conventional bar for a reliable measurement	—	—	0.80

Table 4: **Per-component variance decomposition with bootstrap intervals: the object of measurement is the smallest term, under every resample.** Full seven-component breakdown of the headline 4 methods × 4 judges × 25 questions G-study (DeepSeek, Qwen3-8B, GPT-4.1-mini, GPT-4o), each component as a percentage of total variance with a 95% question-level bootstrap CI (B=5000, seed 0). The *method* — the quantity a break-rate benchmark exists to measure — is 0.6% of variance with an upper CI of only 7.8%: even under the most method-favourable resample it explains under a tenth of the variance. The three noise terms it must be read against (method×question, judge×question, residual) together carry 89%. This is why the dependability CI is wide (Table 3): Φ is a ratio of a near-zero numerator to that noise, so the data cannot pin it, yet $P(\Phi < 0.80) = 0.98$ — the metric is very unlikely to be dependable regardless of where in the interval the truth lies.

variance component	% of total	95% bootstrap CI
method *(object of measurement)*	0.6	[0.0, 7.8]
judge	0.5	[0.0, 5.5]
question	9.3	[1.2, 16.8]
method × judge	0.3	[0.0, 5.8]
method × question	35.7	[3.4, 58.4]
judge × question	22.6	[2.4, 45.5]
residual	31.0	[20.0, 59.3]

Table 5: **The result replicates on a second, different-family agent under the full permutation — and does not under the discourse-preserving one.** We regenerate turn-1 reports with GPT-4O (much stronger and a different family from the original open DR-TULU-8B; knowledge-grounded, no web search — a stated limitation), then run the identical pipeline on ResearchRubrics and ResearcherBench under three judges. Paired diffs are control *minus* real arm on matched (target-excluded) denominators over the same questions, raw two-sided bootstrap p in parentheses ($10{,}000$ resamples, seed 0); † marks survivors of a *single* Holm family over all twelve contrasts, applied separately within each control. *Top:* the full paragraph reorder breaks more than *both* real arms in all twelve contrasts, seven of twelve surviving Holm. *Bottom:* restricting the permutation to within each section leaves only six positive by point estimate and none surviving, with every interval containing zero in both directions. **We report this against our own interest:** on these sharply sectioned reports the full-reorder replication is carried by *cross*-section movement, which we attribute to retrievability rather than to judge instability, and we therefore do not claim the "exceeds" result replicates here. It does survive the discourse-preserving control on the DR-TULU reports that produced the published number (§4), which is the scope we claim. **We tested the obvious escape and it does not hold.** A section containing a single paragraph has nothing to permute, so on sharply sectioned reports the within-section control degenerates to the identity control — inert for $9/25$ of the GPT-4O RR reports and $7/25$ on ResB, against $0/25$ on DR-TULU. Restricting to the questions it actually perturbs does raise the raw within-section break (e.g. $0.066 \to 0.093$ on RR/GPT-4O) but leaves the family seven of twelve positive and one of twelve surviving, so the null is not manufactured by an inert control. Non-significance at $n{=}25$ is not equivalence, and we do not read the bottom panel as one.

dataset	judge	control	regen	patch	control−regen	control−patch
full paragraph reorder (cross-section movement allowed)						
RR	DEEPSEEK	0.151	0.076	0.068	+0.075 (.049)	+0.083 (.004)†
RR	GPT-4.1-MINI	0.216	0.038	0.066	+0.178 (.000)†	+0.150 (.000)†
RR	GPT-4O	0.144	0.050	0.060	+0.094 (.000)†	+0.084 (.000)†
ResB	DEEPSEEK	0.225	0.155	0.099	+0.070 (.335)	+0.126 (.014)
ResB	GPT-4.1-MINI	0.171	0.049	0.048	+0.123 (.000)†	+0.123 (.000)†
ResB	GPT-4O	0.104	0.056	0.047	+0.048 (.188)	+0.058 (.019)
within-section reorder (discourse-preserving; same reports, same judges)						
RR	DEEPSEEK	0.071	0.076	0.068	−0.005 (.877)	+0.003 (.891)
RR	GPT-4.1-MINI	0.065	0.038	0.066	+0.027 (.096)	−0.002 (.911)
RR	GPT-4O	0.066	0.050	0.060	+0.016 (.283)	+0.006 (.616)
ResB	DEEPSEEK	0.104	0.155	0.099	−0.051 (.118)	+0.006 (.964)
ResB	GPT-4.1-MINI	0.031	0.049	0.048	−0.017 (.247)	−0.017 (.238)
ResB	GPT-4O	0.053	0.056	0.047	−0.003 (.947)	+0.006 (.732)

Table 6: Multi-turn revision on MR-DRE/ResearchRubrics (turn-2 content feedback, $n{=}25$, single revision; the reviser is temperature-0 deterministic). Break rate (BRK↓) and incorporation (INC↑), under two judge families. Same reviser (DeepSeek) and identical feedback across methods; only the edit strategy differs. *Patch* and *ledger* reduce break rate below full regeneration — clearly under DEEPSEEK, marginally under QWEN3-8B (0.057 against 0.059) — *localized* does not, sitting marginally *above* regeneration under DeepSeek (0.039 vs. 0.038), so read C2 as covering patch and ledger only. Neither elaboration beats plain patch robustly (see text). Anchor: DR-Tulu-8B self-revision (vanilla regenerate) BRK $= 0.144$.

	DeepSeek judge		Qwen3-8B judge	
Method	BRK ↓	INC ↑	BRK ↓	INC ↑
regenerate (full rewrite)	0.038	0.92	0.059	0.96
patch (constrained edit)	0.024	0.80	0.057	0.96
localized / LME (ours, low-occ.)	0.039	0.72	0.047	0.96
ledger-patch (ours)	0.007	0.80	0.026	0.96

Pooled significance (RR+ResB, $n{=}80$, paired bootstrap; DeepSeek / Qwen): patch−regenerate $p{=}0.003$ / $p{=}0.030$; ledger−regenerate $p{=}0.000$ / $p{=}0.000$; ledger−patch $p{=}0.213$ / $p{=}0.390$ (not significant — ledger ≈ patch).

Table 7: **One floor, four revision arms: only the arm that comes closest to the published range clears it.** ResearchRubrics, primary judge, the same questions throughout, so every row is paired. FLOOR is Eq. (1)'s maximum over five controls — paragraph reorder at 0.064, 95% CI $[0.038, 0.102]$ — and identical for every row, the controls perturbing the *turn-1 report* rather than the revision. Floor and arm are recomputed inside each resample, so the interval carries the floor's own uncertainty, which is most of it. "ctrl $<$ arm" counts how many of the five controls sit significantly below that arm. **Only the first row — the arm whose break rate comes closest to the range this paper audits — demonstrably clears its floor**; for the other three the comparison is indeterminate here. The ratio column is a descriptor: within a cell FLOOR is fixed by construction, so it falling as BRK rises is arithmetic. The last row is judge-matched, dividing the original judge's own floor (0.031) into the range that judge produced rather than ours, because Eq. (1) forbids transporting a floor across judges — but it is a *lower* bound: that floor is a maximum over three controls where ours is over five, and adding the reorder alone multiplies the floor by 1.6 to 4.4 under the three judges for which we bootstrapped it (§4, Table 1). It also matches the judge *only*: the reports, the agents and the question draws behind 0.16–0.31 are not ours, and §4 measures the regenerate rate itself moving $2.60\times$ between two question samples of one benchmark, so this row bounds the artifact share rather than decomposing it.

revision arm	reviser	BRK	$\frac{\text{FLOOR}}{\text{BRK}}$	BRK − FLOOR	ctrl $<$ arm
DR-TULU-8B, full regenerate	weak, open	0.144	44%	$+0.080$ $[+0.018, +0.142]$	5/5
DR-TULU-8B, constrained revise	weak, open	0.060	107%	-0.004 $[-0.052, +0.040]$	2/5
DEEPSEEK, full regenerate	strong, closed	0.038	167%	-0.026 $[-0.069, +0.017]$	0/5
DEEPSEEK, constrained patch	strong, closed	0.024	270%	-0.040 $[-0.084, +0.004]$	0/5
published range, judge-matched only	—	0.16–0.31	10–19%	—	n/a

A bold difference excludes zero: the arm breaks more than its own floor, so a real-loss residual survives. The column is BRK_{true} exactly as Eq. (1) defines it, so its sign matches Table 11 rather than opposing it. The first two rows are the *same* agent under two revision actions, which is why its break rate moves $2.4\times$ while the floor does not; the anchor quoted elsewhere (0.144) is the regenerate row. FLOOR is also biased *upward* as an estimator, being a maximum of noisy terms, which runs against the incomplete-set argument above; Appendix E quantifies both.

Table 8: **Noise budget: every source of movement in the metric, on one scale.** All entries are break rates, so any row reads against any other. The top panel is near-zero only because the pipeline is frozen and decoding greedy — close to guaranteed, not measured. Perturbing any *one* leg of the (report, rubric, judge) triple moves the metric: the middle panel the report, the one-month row the judge, the rubric paraphrase (Table 10) the standard. DS = DEEPSEEK-CHAT, Qw = QWEN3-8B; dashes are conditions not run, not zeros.

	RR/DS	RR/Qw	ResB/DS	ResB/Qw	Rig/DS	Rig/Qw
The judge, on text that did not change at all						
identity, same version	0.005	0.002	0.030	0.000	0.069	0.000
identity, **one month apart**	0.053	–	–	–	–	–
*Content-preserving **modification** (nothing substantive changed)*						
reword one paragraph	0.022	0.024	0.078	0.037	0.020	0.006
reword **every** paragraph	0.040	0.051	0.125	–	–	–
insert neutral text	0.014	0.040	0.140	0.016	0.020	0.020
whitespace only	0.021	–	0.177	–	–	–
paragraph reorder	0.064	0.065	0.155	–	–	–
Real revisions (content genuinely changed)						
patch	0.024	0.057	0.112	0.009	0.100	0.007
ledger-patch	0.007	0.026	0.074	0.024	0.149	0.014
localized	0.039	0.047	–	–	–	–
regenerate	0.038	0.059	0.259	0.064	0.090	0.061

Identity rows are the pairwise break rate between two *independent* scorings of the same unchanged turn-1 reports, with nothing excluded; the same-version row averages three repeat pairs. QWEN3-32B gives 0.000 on ResearchRubrics, the one dataset on which we measured it, and mechanically so: the binary coverage decisions break rate is computed from are identical on all 25 questions (24 of 25 for the continuous coverage score). We did not measure that judge's identity on the other two datasets and do not assume it. The one-month row re-scores identical text with the same model *name* a month later; the reverse direction gives 0.071, and aggregate coverage moves only $+0.013$ over that month — the churn is invisible in the headline number and is absorbed entirely by this per-item differential metric.

Table 9: **Every floor comparison, as a paired difference, tested both ways.** Control and method are scored on the same questions, so the defendable form of "this control reaches that method" is a paired difference. Positive = the *control*, which destroyed no content, broke more. 10,000-resample bootstraps over questions, primary judge, a second disjoint question slice pooled in where we have one. **Bold** = excludes zero above; † = excludes zero below, i.e. the method exceeds its own magnitude-matched floor, which is evidence of real loss. Verdict column: EQUIV = the 90% interval lies within a margin δ=0.05 fixed in advance (two one-sided tests), DIFF = the 95% interval excludes zero, INCON = neither, which we report as inconclusive rather than as support. §4 gives the multiplicity and margin caveats, both of which cut against us.

data	contrast	n	mean	95% CI	p	verdict
RR	paragraph reorder − regenerate	50	+0.0353	[+0.008, +0.062]	.011	DIFF
RR	reword *every* paragraph − regenerate	25	+0.0018	[−0.041, +0.057]	.997	EQUIV
RR	reword one paragraph − patch	25	−0.0019	[−0.039, +0.034]	.940	EQUIV
RR	whitespace only − patch	50	−0.0001	[−0.025, +0.024]	.991	EQUIV
ResB	reword *every* paragraph − regenerate	24	−0.1334	[−0.258, −0.030]†	.008	DIFF
ResB	whitespace only − patch	49	+0.0430	[−0.023, +0.123]	.244	INCON
ResB	paragraph reorder − patch	49	+0.0392	[−0.013, +0.099]	.157	INCON
ResB	insert neutral text − patch	24	+0.0278	[−0.097, +0.177]	.736	INCON
ResB	reword one paragraph − patch	24	−0.0347	[−0.128, +0.083]	.526	INCON

Multiplicity. One family of nine: Holm's first threshold is 0.05/9=0.0056 and the smallest p here is 0.008, so no row survives family-wise correction — including the reorder row, our only positive instability result, which we therefore do not call confirmed. **Margin.** 3 rows are EQUIV at δ=0.05, all on RR; at δ=0.02, below our own pooled C2 effect, 0 are. The two *reword every paragraph − regenerate* rows disagree by dataset: on RR that control is indistinguishable from a full rewrite, on ResearcherBench it breaks significantly *less*, so regeneration there destroys something it does not.

Table 10: **The rubric's wording moves the metric more than the agent does.** We paraphrase every criterion with its meaning, polarity, entities and numbers preserved (item identifiers and weights untouched) and re-score reports that were *not modified at all*. Compared *within the judge that produced it* — QWEN3-32B, whose own re-scoring is deterministic by construction — the effect is roughly twice that judge's break rate for a full rewrite on this dataset (0.050). We do not set it against a one-month drift for this judge: served from frozen local weights it has none, and borrowing our API judge's 0.053 would be the cross-judge comparison we decline elsewhere. Mean coverage barely moves, so a benchmark reporting only aggregate coverage cannot see any of it. An independent stronger model confirms the paraphrases preserve meaning, and restricting to that verified subset leaves the effect intact.

unchanged reports, perturbed rubric	value	
break rate, original rubric → paraphrased	0.0996	Judge QWEN3-32B,
break rate, paraphrased → original	0.0661	
pooled over items rather than questions	0.0858	
restricted to verified meaning-preserving criteria	0.0836	
criteria judged meaning-preserving by an independent stronger model	659/664	
mean coverage, original rubric	0.5616	
mean coverage, paraphrased rubric	0.5514	
change in mean coverage	-0.0102	

decoding greedily on frozen weights, so byte-identical re-scoring returns identical labels by construction (0.000). That rules out sampling noise, not the judge responding to how a criterion is worded, which is what this table measures. The first two rows use the paper's per-question averaging; the third pools criteria directly, which is why it differs slightly. Criterion *order* cannot matter here and was not tested: the evaluator issues one judge call per criterion.

Table 11: **How much of measured "revision breakage" is judge instability, and why no single number answers it.** Break rate (fraction of covered@turn-1 items that drop@turn-2) for *content-preserving* controls vs. real revision methods, on three datasets, under two judge families (DeepSeek / Qwen3-8B). Identity re-scores the same text. *Reword* rephrases a paragraph (all citations retained); *neutral-insert* appends unrelated text with no epistemic claims — neither changes the report's substantive content. Yet both yield break rates comparable to, or above, those of real edits; under this judge only full *regenerate* clears the instability floor on ResB (on RR, whose floor is higher relative to its effects, even regeneration stays within it). Bottom panel: floor-corrected $\mathrm{BRK}_{\mathrm{true}}$ exactly as Eq. (1) defines it — each cell minus its *own* dataset's largest content-preserving control, with no pooling across datasets and no averaging across judges. Every residual here is the difference of two numbers printed above it in this same table, so it can be checked by subtraction. One of the four is positive in point estimate — ResearcherBench regeneration at $+0.082$ — but **it does not survive an interval that resamples the floor** rather than holding it fixed: $+0.081$ $[-0.040, +0.167]$ on these questions, $+0.011$ $[-0.069, +0.047]$ on a disjoint 25, $+0.052$ $[-0.027, +0.095]$ pooled. We therefore report no positive residual on this dataset. On ResearchRubrics regeneration sits -0.026 *below* a floor set by paragraph reorder, and both patch cells are below theirs.

	content changed?	RR	ResB	Rig
Content-preserving controls (break ≈ 0 if the metric tracks content loss)				
reword para. (cites kept)	no	0.022 / 0.024	0.078 / 0.037	0.020 / 0.006
neutral insert	no	0.014 / 0.040	0.140 / 0.016	0.020 / 0.020
whitespace-only	no	0.021	0.177 †	–
paragraph reorder	no	0.064 / 0.065	0.155	–
Real revision methods (RR shown; see Table 6)				
patch (constrained edit)	minimal	0.024 / 0.057	0.112 / –	–
regenerate (full rewrite)	full	0.038 / 0.059	0.259 / –	–
Floor-corrected $\mathrm{BRK}_{\mathrm{true}}$*, Eq.* (1)*: break − that cell's own* FLOOR*, never pooled*				
FLOOR (largest control)	—	0.064	0.177	–
regenerate	full	-0.026	+0.082	–
patch	minimal	-0.040	-0.065	–

Identity re-scoring is reported once only, in the noise budget of §4, where it is the mean of three independent repeat pairs per judge; this table omits it rather than printing a second estimator of the same quantity under the same name. The whitespace-only row is DeepSeek-scored (single value). Paragraph reorder was scored under three judges on ResearchRubrics — 0.064, 0.065 and 0.091, the last under the judge whose identity re-scoring is exactly 0.000 (Table 1) — and under DeepSeek on ResearcherBench. It changes no words yet breaks more than any real edit in every column where both were run. †: whitespace-only on ResB (0.177) *exceeds* the real-patch break rate there (0.112). ‡: DeepSeek is non-deterministic on RigorousBench (identity noise floor 0.069), so its Rig control values sit within self-noise and should be read via the deterministic Qwen judge; Qwen is deterministic on all three datasets. Edit magnitude (fraction of turn-1 lines changed) is decoupled from break: regenerate 0.55, patch 0.09, ledger 0.04, localized 0.002 — localized changes 0.2% of turn-1 lines yet breaks like a full rewrite. **Cells are not equally supplied, and FLOOR is a maximum, hence monotone in the control set.** Counting the rows above: RR/DeepSeek and ResB/DeepSeek carry four controls each; every other cell carries two, the two cheapest. So a floor is only comparable to another floor over the same controls, and an arm that exceeds its own cell's largest control may be exceeding an incomplete floor rather than a real one. The size of that gap is measured, not hypothetical: adding paragraph reorder alone multiplies the RR floor by 1.6 to 4.4 across the three judges for which we bootstrapped it (§4). Every floor here from two controls should therefore be read as a lower bound, and no arm read as clearing it.

Table 12: **Breakage is occupancy-driven for the open agent, flat for the strong closed reviser.** Break rate (RR, n=25) as a function of context occupancy: the edit action is held fixed (revise the turn-1 report given feedback) and only neutral filler prepended to the reviser context is varied. For DR-Tulu-8B revising its own report, break rate roughly doubles with occupancy (Spearman $= 0.80$; paired pad 36k$-$pad 0 $= +0.062$, p=0.001); the strong DeepSeek reviser is flat.

Reviser	pad 0	pad 12k	pad 24k	pad 36k
DR-Tulu-8B, RR (open, self-revise)	0.060	0.096	0.088	0.122
DR-Tulu-8B, ResB (self-revise)[‡]	0.260	0.287	0.233	0.267
DeepSeek-chat, RR (strong, closed)[†]	0.038	0.049	0.044	0.037

[†]DeepSeek pads are 0/6k/18k/36k; the point is the flat trend (Spearman ≈ 0). [‡]On ResB the trend is also flat (Spearman 0.00, pad 36k$-$pad 0 p=0.476), but for a different reason: ResB's judge-instability floor is high (Table 11: reword 0.078, neutral 0.140under DeepSeek), so break rate is noise-dominated and any occupancy signal is swamped. The occupancy effect is detectable on RR precisely because RR's instability floor is low.

Table 13: **The mitigation at a larger sample.** We generated 25 further questions per dataset and put them through the identical protocol, taking the pooled comparison from $n\approx$80 to $n\approx$130. Every ordering the paper claims survives and tightens; more usefully, the two comparisons the paper reports as NULL are now *bounded* rather than merely non-significant — ledger-patch differs from plain patching by less than 0.02 under both judges.

	DEEPSEEK		QWEN3-8B	
revision strategy	n	break rate	n	break rate
regenerate	129	0.0903	130	0.0433
patch	129	0.0444	130	0.0214
ledger-patch	129	0.0438	130	0.0228
localized	129	0.0630	130	0.0275
Paired differences (each question compared with itself), 95% bootstrap CI and one-sided p				
patch − regenerate	−0.0460	[−.081, −.015] .001	−0.0219	[−.039, −.005] .007
ledger-patch − regenerate	−0.0465	[−.077, −.020] .000	−0.0205	[−.033, −.009] .000
localized − regenerate	−0.0273	[−.057, +.003] .040	−0.0158	[−.030, −.003] .010
ledger-patch − patch	−0.0005	[−.022, +.023] .465	+0.0014	[−.013, +.015] .597

Pooled at the question level over ResearchRubrics and ResearcherBench, 25 core + 15 held-out + 25 new per dataset. Each question is tagged with the pool it came from so a strategy is never compared against a different question set. The *localized* row required filling a cell that had never been run (that strategy had never been applied to ResearcherBench); before it was filled the comparison sat at n=105 and was significant under only one judge.

Table 14: **Deciding whether a break was real is itself adjudicator-dependent.** For each flagged break we ask an independent model whether content satisfying the criterion is still present. On the content-preserving controls it *must* be: anything below 100% there is the adjudicator's own error. GPT-4O errs on roughly a third to a half of them; DEEPSEEK is near-perfect on the controls but then also credits most *full rewrites* as intact, which is leniency rather than accuracy. Removing the prompt's hint about which edit was applied (third column) moves it by about four points, so the gap is not a prompt artifact. Consequently the per-item true-loss estimate in the lower panel spans a factor of five, and no additional model adjudicator can settle it.

		"content still present" (higher = flip was a false break)		
	condition	GPT-4O	DEEPSEEK	DEEPSEEK, blind
control	reword	56%	100%	100%
control	whitespace	53%	100%	100%
control	paragraph reorder	70%	100%	91%
real	patch	31%	85%	85%
real	regenerate	19%	94%	88%
Per-item true loss it implies, BRK×*(control present − this condition's present)*				
	patch	0.0210	0.0105	0.0076
	regenerate	0.0637	0.0093	0.0122

n=76 flagged breaks (47 control, 29 real); per-condition samples are 9–23, so these are wide. Neither column is independent of the break rate it corrects — both are computed from the same pooled flip set — so agreement with any aggregate estimate is arithmetic, not corroboration, and we no longer quote one as confirming the other. The blind human pass (Appendix E) supersedes both: it separates the arms not at all. The two adjudicators agree on only about half the individual flips and disagree in one direction only — DEEPSEEK never says lost where GPT-4O says present. That is a systematic strictness gap, not noise. The blind human pass agrees with neither of them (Appendix E).

Table 15: **Reviser decode variance, against the effect it has to beat.** Repeating the *same* revision protocol moves the break rate by as much as the differences between revision strategies — but how much depends on *which* reviser, and that is the finding. Each cell is one independent decode of all 25 questions with the judge held fixed at QWEN3-32B; re-scoring byte-identical text under that judge gives a break rate of exactly 0.000 (§4), so the spread below is the reviser's, measured rather than assumed. For the *weak* reviser at temperature 0 the regenerate−patch gap (0.0024) is smaller than either strategy's own spread (0.0057, 0.0040) and the mitigation does not separate at all; at temperature 0.7 it is only 1.4× that spread. For the paper's own DEEPSEEK reviser the same gap is 0.0381, or 5.6× regenerate's spread, and patch wins in all five repeats. The ordering is therefore a property of the reviser's competence rather than of the decoding temperature, which is the form C2 takes in §4.

reviser	strategy	independent decodes					mean	sd
		1	2	3	4	5		
QWEN3-8B *reviser, temperature 0 (repeat the protocol)*								
	regenerate	0.0420	0.0376	0.0420	0.0516	0.0488	0.0444	0.0057
	patch	0.0468	0.0373	0.0384	0.0428	0.0446	0.0420	0.0040
	ledger-patch	0.0311	0.0559	0.0559	0.0311	0.0355	0.0419	0.0129
	localized	0.0299	0.0266	0.0139	0.0164	0.0266	0.0227	0.0071
QWEN3-8B *reviser, temperature 0.7 (stochastic decoding)*								
	regenerate	0.0490	0.0518	0.0709	0.0441	0.0748	0.0581	0.0138
	patch	0.0389	0.0342	0.0335	0.0307	0.0591	0.0393	0.0115
	ledger-patch	0.0260	0.0577	0.0188	0.0159	0.0270	0.0291	0.0167
	localized	0.0231	0.0324	0.0152	0.0326	0.0217	0.0250	0.0075
DEEPSEEK *reviser, temperature 0 — the paper's own protocol, re-run*								
	regenerate	0.0493	0.0422	0.0583	0.0420	0.0451	0.0474	0.0068
	patch	0.0088	0.0084	0.0088	0.0088	0.0116	0.0093	0.0013
	ledger-patch	0.0414	0.0107	0.0084	0.0399	0.0126	0.0226	0.0166
	localized	0.0241	0.0153	0.0192	0.0258	0.0331	0.0235	0.0068

The DEEPSEEK block is the same-source arm: the API ignores a seed argument, so its repeats are literally "run the published protocol again", which is what a replicator would do. Its spread bounds re-running the protocol *within* one session; across API model versions two months apart the same protocol moves far more, which §4 reports separately.

Table 16: **Breakage does not compound over turns.** Three revisions applied in sequence, with break rate measured after each one against the *original* turn-1 report, so each column is cumulative damage rather than per-turn damage. The real arms grow only slightly even though the reports themselves grow by 15–25% in length, and the content-preserving control does not grow at all. Repeated revision is close to a one-time cost, not a spiral.

		break rate vs. the original report		
	revision applied	after 1	after 2	after 3
real	regenerate	0.0544	0.0562	0.0581
real	patch	0.0398	0.0302	0.0407
control	reword control (a different paragraph each turn)	0.0152	0.0134	0.0050

Reviser QWEN3-8B, judge QWEN3-32B (measured self-noise 0.000), n=25; later turns exclude every checklist item requested so far, so the denominator stays comparable across columns. The control had to be run twice: rewording the *largest* paragraph is idempotent at temperature 0 — the second pass reproduced the first byte-for-byte on $19/25$ reports — so the reported control rewords a *different* paragraph each turn, verified by a per-turn byte-identity check ($0/25$ identical between turns 1 and 2).

Table 17: **Extension slice: 15 further questions per dataset. Not held out** — we called it that in an earlier draft and retract the label here as well as in the text: these questions are folded into the pooled $n{\approx}80$ comparison, so this slice cannot independently replicate a test it is part of. The independent material is the later 25 questions per dataset, generated after every analysis here was fixed. On ResearchRubrics the ordering reproduces; on ResearcherBench every condition, control and real alike, collapses towards zero, so this slice has no resolving power there — reported because it bounds how far these effects generalise.

	ResearchRubrics$_{\text{ext}}$		ResearcherBench$_{\text{ext}}$	
condition	DEEPSEEK	QWEN3-8B	DEEPSEEK	QWEN3-8B
reword (control)	0.0248	0.0564	0.0111	0.0000
patch	0.0121	0.0458	0.0000	0.0095
ledger-patch	0.0356	0.0650	0.0167	0.0000
localized	0.0095	0.0558	0.0333	0.0095
regenerate	0.0449	0.0981	0.0111	0.0000

$n{=}15$ per dataset (question ids 26–40), scored under both primary judges with the same checklist evaluator and the same targeted-item exclusion.

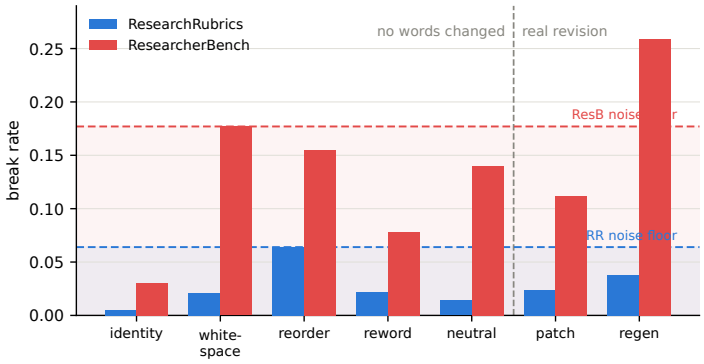

Figure 9: **Break rate tracks perturbation of the scoring pipeline, not content loss.** Break rate (DeepSeek judge) across modifications ordered by how much content changes; left of the divider nothing but formatting/wording/order changes. Content-preserving perturbations are comparable to or above real revisions; on ResearcherBench a whitespace-only edit (0.177) exceeds a real patch (0.112). Shaded bands mark each dataset's floor under *this* judge (its largest content-preserving control, Eq. 1): here every RR method including regeneration stays within the floor while on ResB only regeneration clears it. An earlier version of this caption said regeneration sits above the largest control under QWEN3-8B and QWEN3-32B; that was true only while those judges' maxima were taken over the three cheap controls. With the reorder and full reword since scored under both (Table 1) their floors are 0.065 and 0.091, and regeneration (0.059, 0.050) sits below each. This figure shows the primary judge only, so those two columns are read from that table rather than from these bars. Identity (no modification) is the only condition at or near zero throughout, though not exactly zero in every cell (Table 8).

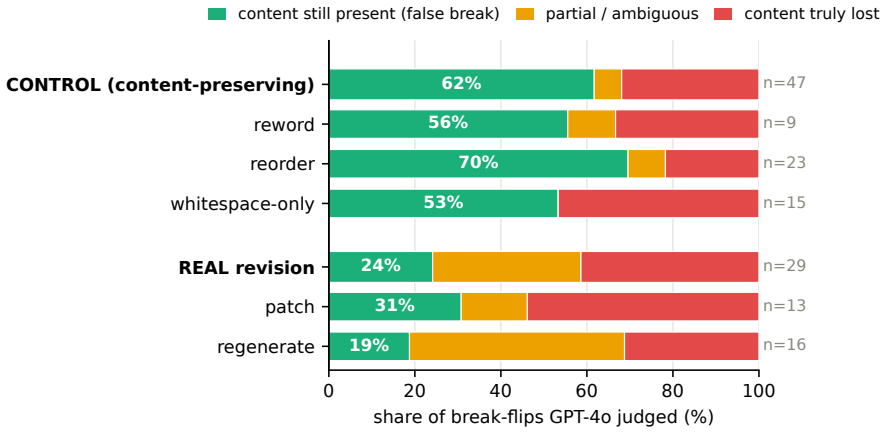

Figure 10: **Model adjudicators disagree with each other and with a human.** For each break-flip, an adjudicator given the *full* after-report rules whether the item's content is still present. GPT-4O separates the arms sharply (62% of control flips present against 24% of real ones); a second model adjudicator barely separates them at all; and the blind human pass separates them *not at all* (78.7% vs 79.3%, difference -0.006 $[-0.186, +0.183]$; Appendix E). We therefore no longer read the GPT-4O gap as confirmation of C3 — the figure is retained to show the spread between adjudicators, which is the finding.

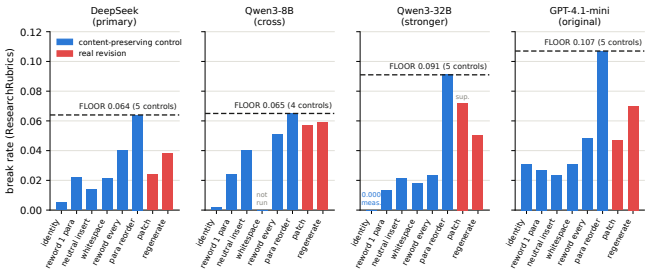

Figure 11: **The judge-instability floor holds across four judges and three families—including the original benchmark judge.** RR break rate by condition (rows) and judge (columns), where GPT-4.1-MINI is MR-DRE's own judge. One panel per judge, same y-scale throughout; blue is a content-preserving control, red a real revision, and the dashed line is FLOOR, Eq. (1)'s maximum over the controls actually scored under that judge (identity excluded, as in Eq. (1)). The figure is the argument: under all four judges *no* red bar reaches the line, and the control that attains it is in every case the paragraph reorder, which changes not one word. Under GPT-4.1-MINI — the judge that produced the range this paper audits — all five controls were scored, so that line is a maximum over the complete set; the only "not run" cell in the figure is whitespace under QWEN3-8B. Cells are annotated rather than left ambiguous: identity under QWEN3-32B is a *measured* 0.000 ("meas."), not an unrun cell, and that judge's patch bar is the pass the paper supersedes ("sup."; re-run 0.0093, §4) — the floor exceeds it either way. Generated by `paper/figures/src/make_fig_judges.py`, which fails the build if any value in it has drifted from the macro in `tables/numbers.tex` it must equal.