# Do Injected Errors Transport to Organic Errors? A Matched Audit of Corruption Probes for Chain-of-Thought Reasoning

Anonymous submission

Anonymous

**Abstract**

Results from injected-error probes do not transport to the errors language models actually make at the rates a probe implies. We audit the standard corruption-probe protocol on frozen frontier models, holding the item, the prefix position and the trace author fixed. In a pre-registered pilot (three proprietary families, three public math benchmarks, four typed corruptions, 2,976 continuations), injected corruptions were recovered from without any acknowledgment in 37.8-60.2% of continuations per family (recovery of any kind 86.8-97.2%; invisible step-deletions excluded and reported separately), while the models' own reproduced errors were repaired in 0 of 60 error-localised continuations. The author-fixed contrast is i-c: on the same items and positions, a corruption injected into another model's correct trace is repaired far more often than that same model's real error; authorship is fixed, but the error and its preceding derivation both differ. In the exploratory cohort the paired gap is +0.40 to +0.75, every Newcombe 95% interval excludes zero, and exact McNemar tests Holm-adjusted jointly over the six primary contrasts are significant at 0.05 in three and reach adjusted p = 0.0645 in the other three (discordant pairs 6/0, 6/0, 9/1). A separately registered confirmatory cohort on untouched items contains 173 b-c matched pairs and an eligible subset of 122 i-c matched pairs across 1,038 continuations. All six primary point estimates are positive, matching the exploratory signs, with author-fixed point-estimate gaps of +0.227 to +0.415; after Holm correction only the two gpt contrasts reach 0.05, while the sonnet adjusted values are 0.0509 and 0.068 and flash is not separated from zero. The confirmatory point estimates therefore retain the exploratory direction, but this study does not establish that direction separately in every family after multiplicity correction. Two further exploratory arms probe the error distribution itself: transplanting the real error TEXT into a correct derivation gives j-c point estimates of +0.00 to +0.06 with every interval spanning zero, which neither establishes equivalence nor identifies a mechanism, while a formulaic corruption at the same position in the same trace is repaired far more (i-j +0.39 to +0.75), and cutting just before the error restores recovery (f-c +0.39 to +0.70). Exploratory mechanism probes are consistent with a surface-form account: a blind judge flags 66.1% of injected steps as artificially edited versus 20.6% of real error steps, while a small reference LM's token surprisal does not separate the two (AUC 0.47, null). The higher edit-flagging rate is consistent with a surface-form hypothesis, but the reference-LM result is inconclusive and neither probe identifies the cause of the recovery gap. Transplanted real errors are flagged at 31.7% against 20.6% for real errors, an intermediate rate this sample does not resolve; their j-c recovery estimates run from +0.000 to +0.056 with every 95% interval including zero, which establishes neither equivalence nor the absence of a recovery advantage. Recovery estimates from these injected-error probes should therefore not be substituted for organic-error recovery rates without validation on the target error population, and outcome-label calibration needs recovery rates from that population. All corpora, continuations, the validated acknowledgment detector and registration artifacts accompany the submission.

## 1. Introduction

Process supervision, step-level verifiers, and CoT faithfulness studies share a common experimental primitive: perturb an intermediate step and observe the model's downstream behaviour [1,2,8]. The primitive is attractive because it gives the experimenter ground truth about where the error is. It also carries an assumption that has recently come under scrutiny from the training side: that a corruption inserted by the experimenter behaves like an error the model itself would make. Fine-tuning on synthetic error correction fails to transfer to on-policy errors [19], and organic-error corpora document that synthetic taxonomies miss natural failure patterns [20,21]. What has been missing is an instrument-level audit: a within-item, position-matched, authorship-controlled measurement, on frozen frontier models, of how far behaviour under injected corruptions transports to behaviour on real errors. That audit is this paper.

The transport gap is large. Across three model families and three public math benchmarks, injected corruptions are recovered from without any acknowledgment in 37.8% to 60.2% of continuations per family (invisible deletions excluded). Real errors at the same positions on the same items are almost never repaired, whether the model's own (0 of 60 error-localised continuations) or another model's (2.0 to 9.0% across families). Holding the trace author fixed, injecting a corruption into that same author's correct trace restores repair (paired gap +0.40 to +0.75). In this author-fixed subset the two conditions differ in recovery even with the donor author held constant, though error content and the preceding derivation are not varied independently.

Contributions. (i) A matched within-item transport audit: real versus injected errors at a fixed position, with trace authorship controlled by dedicated arms (clean foreign truncation; foreign-authored injection; transplanted real-error text; pre-error truncation) and an instruction ladder (bare / check-first / location-given) connecting to error-location studies [16]. The direction of the finding is anticipated by training-side and corpus studies [19,20,21]; the within-item quantification on frontier models, with authorship identified, is new. (ii) The measurement instruments: a pre-registered surface of unacknowledged recovery across family, benchmark and corruption type, a validated acknowledgment detector (200/200 operator agreement; 0 verdict flips under four null perturbations), organic-error controls, and a detect/localise/recover decomposition of what 'repair' actually is. (iii) Exploratory error-distribution probes with a calibration implication: a blind edit-detection probe separates injected from organic error steps while reference-LM token NLL does not, and an outcome-label identity (standard algebra, not claimed as a contribution) shows the label-noise parameter measured by injected probes does not transport to organic errors. Data availability: corpora, continuations, detector, registrations and analysis code accompany the submission as a runnable package (README).

## 2. Related work

Closest prior work, with deltas. (a) Synthetic error injection for training self-correction fails to transfer [19]: supervised fine-tuning on injected-then-masked errors does not improve self-correction, and the authors attribute the failure to the distribution shift between synthetic and on-policy errors. Our question is complementary and evaluation-side: we do not train anything, and we audit whether the injected-error

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.

*measurement* transports, on frozen frontier models, with item, position and authorship held fixed, which their training-level comparison cannot isolate. (b) Rule-injected versus model-generated mistakes have been compared for detection and correction performance [20], and MWPES-300K [21] documents that natural error patterns diverge from synthetic taxonomies. Both vary the error source across datasets or models; neither pins the same item, the same prefix position and the same trace author. Our arms h/i/c provide that matched comparison, although error provenance and the preceding derivation are not varied independently. (c) Tyen et al. [16] show models often correct errors once the location is disclosed; our location-given arm reproduces this partially and places it on a ladder against undisclosed continuation. (d) Unspoken Hints [22] documents silent adoption of answer hints, acknowledgment and use coming apart for hints; we measure the same dissociation for errors. (e) Reward models have been shown to score consistency rather than causal correctness [23]; our results give the generation-side counterpart: continuations give a complementary measurement of recovery differences across prefix conditions, without isolating coherence or step truth as the driver.

Step attribution and injected perturbations. Thought-anchor style counterfactual resampling attributes importance to steps with on the order of a hundred rollouts per step [1]; black-box causal attribution of CoT steps [2] and DAG-structured views of reasoning [3] rely on experimenter-controlled interventions. Our audit does not replace these tools; it measures the external validity of their shared primitive.

Self-correction. Intrinsic self-correction is contested [4,5], with recent work proposing an accuracy-correction paradox and error-depth effects [6] and spontaneous step-level self-correction as a training target [7]. Our organic arms quantify one end of this debate: with the first erroneous step present in the prefix, unprompted repair of a model's own error was never observed (0/60), and explicit check-first or location-given instructions recover repair of real (foreign) errors only partially (Section 6).

Process reward labels. Monte-Carlo and outcome-derived process labels are known to be noisy, and robust-training remedies exist [9,10,11]. We contribute the measurement side: the noise channel's parameter depends on error provenance, and recovery estimates differ between the sampled injected and organic conditions, with the matched gaps and their uncertainty reported in Sections 6 and 7. ProcessBench [17] and DeltaBench [18] provide human-annotated real-error localisation benchmarks; they measure detection, not spontaneous repair in continuation, and are the natural source of gold labels for instantiating the identity of Section 8. Injected-mistake probes are likewise a standard instrument in the faithfulness literature [8,12,13,15].

What is standard and what is new (labelling required by our review standard; the label-identity algebra is textbook and is not claimed as a contribution).

| component | status |
|---|---|
| Outcome-label precision/recall algebra | standard (textbook) |
| Synthetic-organic error mismatch, direction | known: training-side [19], corpora [20,21] |
| Typed-corruption continuation protocol + validated acknowledgment detector | new instrument |
| Measured unacknowledged-recovery surface (family x benchmark x type, pre-registered) | new measurement |
| Matched within-item transport audit with authorship control (arms h/i) | new design + quantification |
| Mechanism probes (blind edit detection; reference-LM NLL) | new measurement |

## 3. Definitions and measurement

A continuation exhibits *recovery* if its final answer matches the gold answer under deterministic numeric scoring (a single '#### <number>' pattern; no tolerant extraction, no rescues). It exhibits *acknowledgment* if it explicitly flags an error or inconsistency in the given prefix, as judged by an LLM detector. Our headline metric is *gold-answer recovery without acknowledgment*, abbreviated below as silent repair. The name is behavioural, not mechanistic: a continuation can reach the gold answer by correcting the flawed value in place, by re-deriving from the problem statement, or by ignoring the prefix altogether. Section 4 separates these where the data allow, and we never claim in-place correction beyond what the reappearance analysis supports.

Detector validity. The detector agreed with 200/200 stratified operator hand labels (single annotator, an acknowledged limitation), and its verdict flipped on 0 of 272 re-judgements under four null perturbations of the input (whitespace, sentence order, politeness prefix, model paraphrase).

Registrations and serving endpoints. The pilot design, gates and kill conditions were registered on 2026-08-26 before any corruption run (prereg.json); the within-item experiment and its design-B amendment were registered the same day before any continuation ran (prereg_4arm.json). Arms i and e were added on 2026-08-27 in response to an external review and are not pre-registered; their contrasts and the Holm set were fixed in code before any of their results were computed, and the amendment note (amendment_P46) says so. Arms f and j were added in the same session but are NOT covered by that amendment note or by any other: we record the omission here rather than edit the original file, and we therefore treat f-c, j-c and i-j as exploratory throughout. They are outside the six pre-specified primary tests, which are unaffected. Serving endpoints are pinned by exact serving id and access date (2026-08-26/27): databricks-gemini-3-5-flash, databricks-claude-sonnet-5, databricks-gpt-5-5; open-weight replication uses databricks-meta-llama-3-3-70b-instruct and databricks-qwen3-next-80b-a3b-instruct. The adjacency-search protocol (queries, dates, arXiv ids returned) is logged in the release (adjacency_log.json).

Experimental scale (all runs, resumable ledgers, budget-metered).

| Stage | Generations | Purpose |
|---|---|---|
| Base traces (temp 0) | 1,080 | own stepwise traces; 826 correct kept |

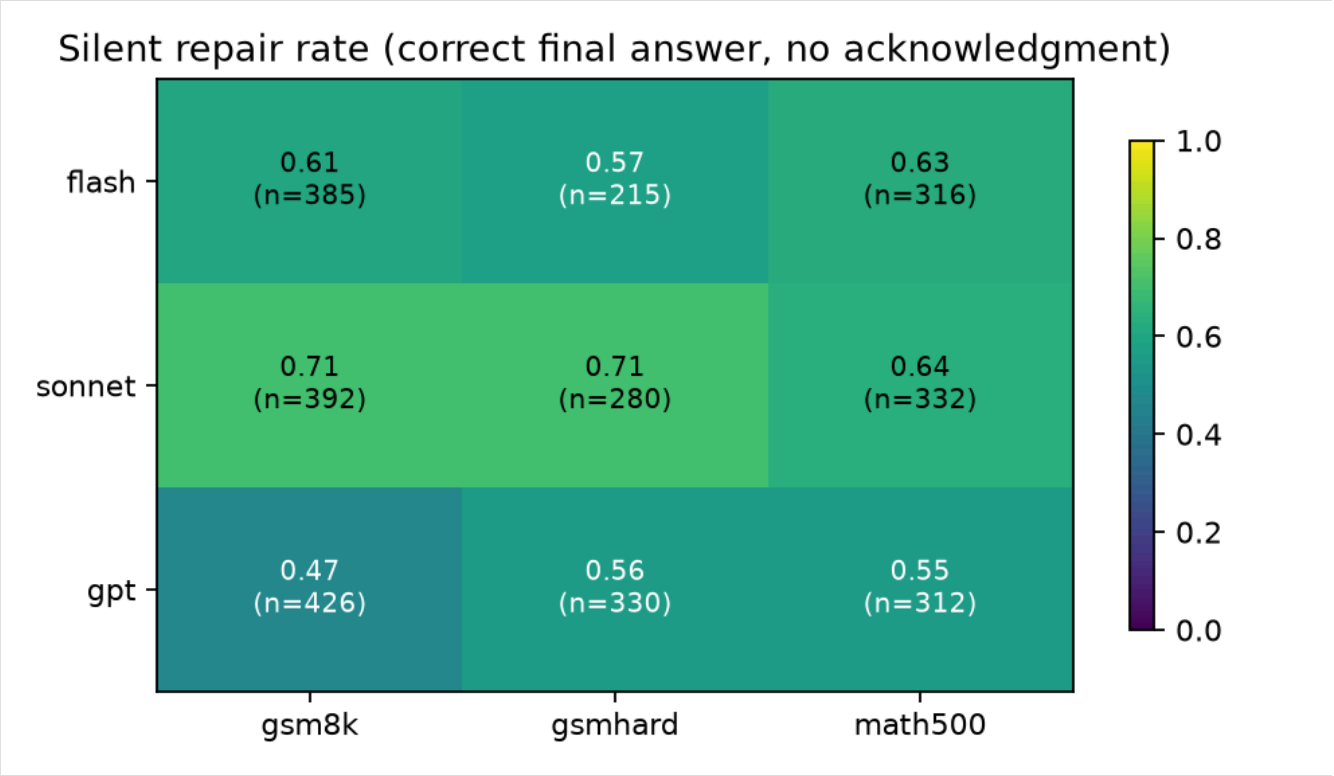
| | | |
|---|---|---|
| Injected-corruption continuations | 2,976 | 4 typed corruptions (12-row smoke batch excluded from all analyses) |
| Acknowledgment judgements | 2,976 | flash detector |
| Organic controls | 490 | regeneration, truncation, localisation |
| Within-item probe (temp 0.8, k=6) | 10,075 | pass rates + trace pools |
| Within-item continuations + judgements | 2 x 996 | arms a-j |
| Mechanism probes | 185 blind judgements + 126 NLL scorings + 329 repair-mode classifications | appendix |
| Confirmatory cohort continuations | 1,038 | Section 7, separately registered, items never used before |

## 4. The injected-corruption surface (pre-registered pilot)

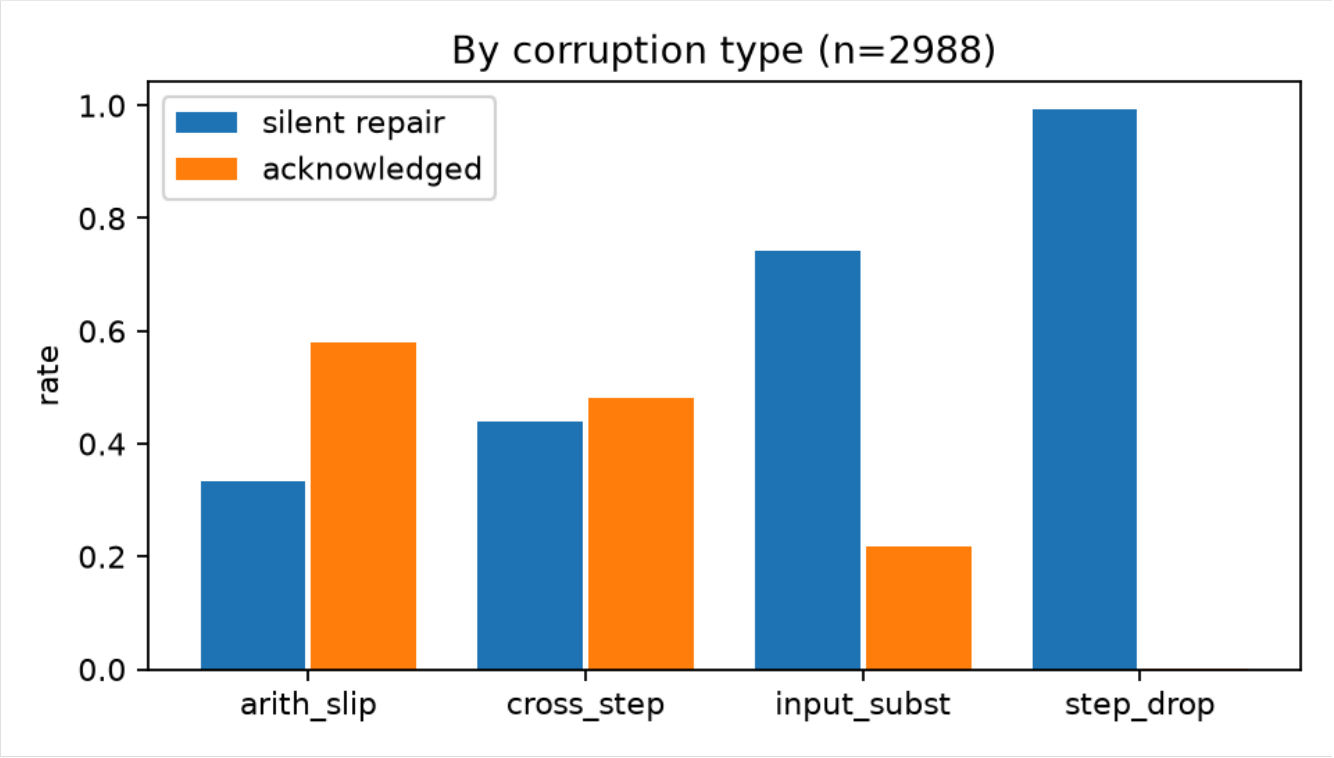
Design (registered before execution): three serving endpoints, three public benchmarks (GSM8K, GSM-hard, MATH-500 numeric subset), four typed corruptions injected into the model's own correct trace prefix, disjoint item pools, one continuation per (item, type) at temperature 0.8. A 12-row smoke batch that validated the pipeline before the registered run is excluded from every number in this paper (n = 2,976).

Per-cell rates (%). Silent repair = gold-answer recovery without acknowledgment.

| family \| benchmark | n | silent repair | recovery (any) | acknowledgment | base accuracy |
|---|---|---|---|---|---|
| flash \| gsm8k | 373 | 60.9 | 91.4 | 32.2 | 0.88 |
| flash \| gsmhard | 215 | 56.7 | 80.5 | 30.2 | 0.50 |
| flash \| math500 | 316 | 62.7 | 93.7 | 33.2 | 0.73 |
| gpt \| gsm8k | 426 | 47.2 | 98.4 | 51.2 | 0.96 |
| gpt \| gsmhard | 330 | 56.4 | 97.9 | 41.8 | 0.73 |
| gpt \| math500 | 312 | 55.1 | 97.1 | 42.0 | 0.72 |
| sonnet \| gsm8k | 392 | 70.7 | 93.9 | 23.7 | 0.91 |
| sonnet \| gsmhard | 280 | 70.7 | 86.1 | 16.8 | 0.65 |
| sonnet \| math500 | 332 | 63.6 | 93.1 | 30.1 | 0.80 |




Silent-repair rate by family and benchmark. All nine cells lie inside the pre-registered non-saturation band [0.10, 0.85] (gate a).




Recovery and acknowledgment by corruption type. Acknowledgment tracks the visibility of the edit: arithmetic slips have the highest acknowledgment rate, step deletions 0.1%.

Per-type rates with Wilson 95% CIs. step_drop is excluded from all pooled headlines: it removes a step rather than presenting an explicit erroneous step, and its recovery rate does not by itself identify how the continuation reached the answer.

| corruption type | n | silent repair (CI) | recovery (any) | acknowledgment |
|---|---|---|---|---|
| arith_slip | 822 | 33.3 [30.2, 36.6] | 88.8 | 58.0 |
| cross_step | 815 | 43.8 [40.4, 47.2] | 90.9 | 48.2 |
| input_subst | 671 | 74.2 [70.8, 77.4] | 95.2 | 21.8 |

| | | | | |
|---|---|---|---|---|
| step_drop | 668 | 99.3 [98.3, 99.7] | 99.3 | 0.1 |

Headline rates per family, step_drop excluded. The recovery-any column is the parameter that enters the label identity of Section 8.

| family | n | silent repair (CI) | recovery any (CI) |
|---|---|---|---|
| flash | 698 | 49.1 [45.4, 52.8] | 86.8 [84.1, 89.1] |
| sonnet | 792 | 60.2 [56.8, 63.6] | 89.5 [87.2, 91.5] |
| gpt | 818 | 37.8 [34.5, 41.1] | 97.2 [95.8, 98.1] |

How recoveries happen (recovered continuations, step_drop excluded). Overt = gold-answer recovery with acknowledgment of an error or inconsistency in the prefix. Inline silent = recovery without acknowledgment, with lexical reappearance of the corrupted value; reappearance is an indicator, not proof of causal prefix use. No detected reappearance = recovery without acknowledgment and without reappearance; the recovery mechanism is unresolved for this category.

| family | recovered n | overt | inline silent | no detected reappearance (mechanism unresolved) |
|---|---|---|---|---|
| flash | 606 | 43.4 (263) | 32.0 (194) | 24.6 (149) |
| sonnet | 709 | 32.7 (232) | 30.7 (218) | 36.5 (259) |
| gpt | 795 | 61.1 (486) | 23.9 (190) | 15.0 (119) |

Registered gate outcomes. Gate (a) non-saturation passed in 9/9 cells. Gate (c) label validity passed. Gate (b), a difficulty-slope criterion, FAILED as registered: the slope of silent repair on (1 - base accuracy) excluded zero in only one of three families. We report this failure as registered and treat the difficulty axis as unresolved; the slope rests on three aggregated benchmark points per family, so these results do not resolve the relationship between difficulty and silent recovery, and the visible structure in the data is carried by corruption type.

## 5. Organic controls: unprompted repair of own errors was not observed

Injected vs organic, same estimand (recovery-any; silent repair and acknowledgment as separate columns). Injected rows are per family, step_drop excluded; organic arms use the models' own wrong traces.

| Arm | n | recovery any | silent repair | acknowledgment |
|---|---|---|---|---|
| Injected, flash | 698 | 86.8% | 49.1% | 41.5% |
| Injected, sonnet | 792 | 89.5% | 60.2% | 30.2% |
| Injected, gpt | 818 | 97.2% | 37.8% | 59.5% |
| Own wrong trace, truncated at 60% of steps | 65 | 1.5% | - | - |
| Own wrong trace, first-error step inside prefix | 60 | 0.0% | - | 0.0% |

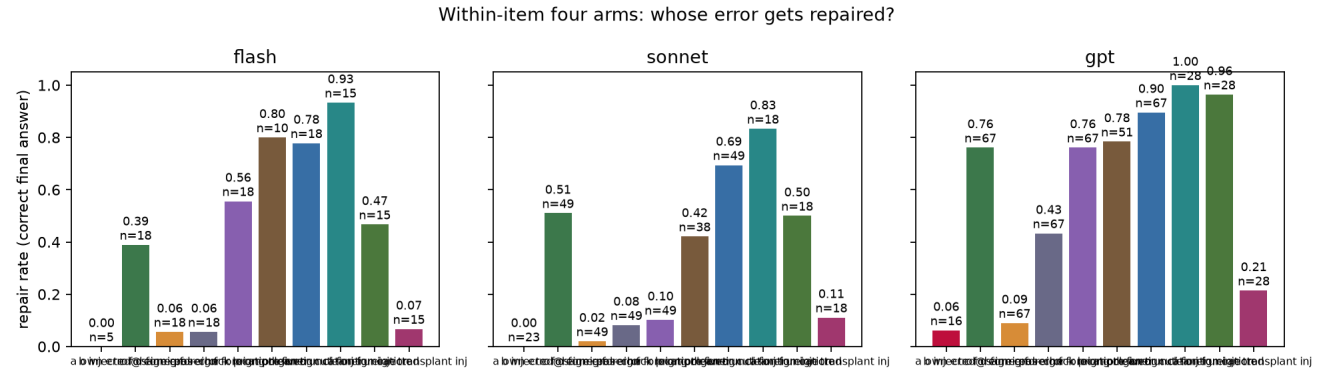
Both controls used the pilot's wrong base traces (temperature-0 reproduction, an acknowledged selection effect addressed in the within-item experiment). Even with the first erroneous step guaranteed inside the prefix by an LLM localiser (83/93 localised), repair and acknowledgment were not observed. Under an independent-binomial approximation for this selected pilot condition, the rule of three puts an approximate 95% upper bound of 5% on the underlying rate; this is not a bound for organic errors in general, and it is consistent with either strong commitment to own errors or effects smaller than this design can detect, and is not evidence of impossibility.

## 6. Within-item transport experiment

The organic controls confound provenance with solvability. Our first within-item design (both a wrong and a correct trace from the same family on the same item) turned out to be infeasible: at temperature 0.8 with k=6 samples, per-item pass rates are strongly bimodal, and only 7-14 mixed items per family exist. We report this as a measured fact and registered an amended design before any continuation ran. Design B fixes the item, the position and the continuing model F, and varies the flawed prefix. Arms: (b) F's own correct trace with a corruption injected at the matched step position; (c) another family G's real wrong trace, first wrong step localised and included; (d) arm-c prefix with an explicit check-the-steps-first instruction; (e) arm-c prefix with the error location disclosed; (g) F's own correct trace cleanly truncated at the same position (upper baseline); (h) G's correct trace cleanly truncated at the same position (authorship control); (i) G's correct trace with the same typed corruption injected at that position; (f) G's wrong trace cut just BEFORE the first error (organic pre-error context, no error present); (j) G's correct trace with G's real error step text transplanted at that position (a fully matched injection: same error text, type, position and length as arm c, different surrounding derivation). Arm (a), F's own wrong trace, exists only on the rare mixed items and is reported descriptively; the pilot's error-localised organic control (0/60) is the own-error anchor.

The author-fixed contrast is i-c: both prefixes are authored by the same other model G and truncated at the same position, and the flaw is G's real mistake inside G's own wrong derivation versus an experimenter's edit inside G's correct derivation. Provenance and the derivation that precedes the flaw therefore move together in i-c, which is why arms j and f are needed: j holds the error text fixed and varies only the surrounding derivation, and f removes the error step from the organic prefix. b-c is the same contrast with F-authored injected prefixes (authorship varies with provenance, as flagged by the external review); h-c varies the presence of a real error under fixed foreign authorship; h-i is the injection cost under fixed foreign authorship. j-c shares the exact error text with c and differs in the derivation that precedes it, so it varies the preceding derivation without isolating coherence from other contextual properties; i-j compares a formulaic corruption against transplanted real-error text in the same

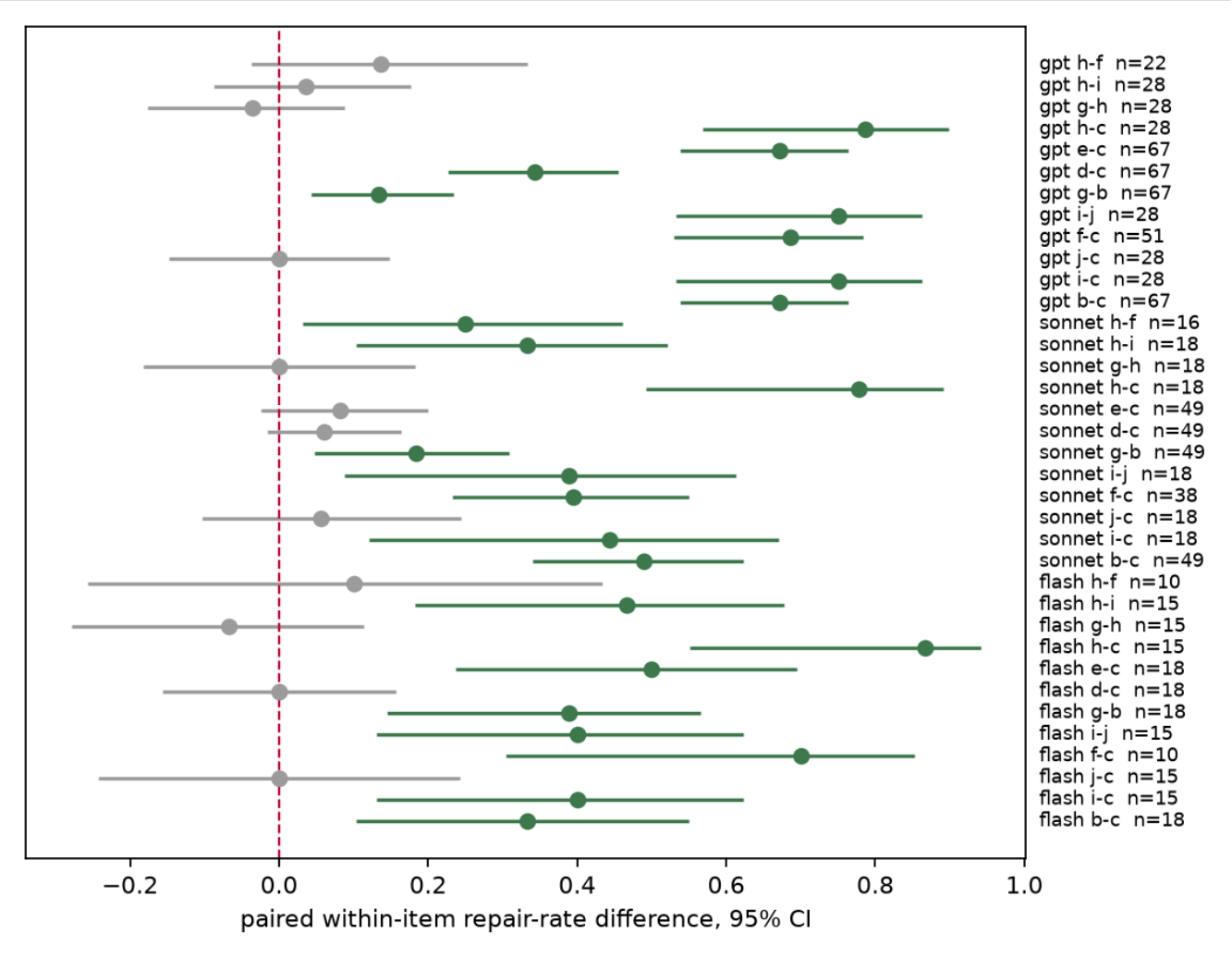
context; f-c removes the error step from the organic prefix. Arm f exists only where the localised first error is not the first step (the pre-error prefix must be non-empty), which is its sole eligibility rule.




Recovery rate by arm and family, within the same qualified items. Per-arm rates and sample sizes are tabulated in the appendix.

Within-item paired recovery differences, provenance and error-distribution contrasts. The instruction and clean-truncation contrasts are in the appendix. CI = Newcombe paired 95% interval; p = exact McNemar (binomial on discordant pairs), reported raw / Holm-adjusted; 6/n = the paired difference produced by the smallest fully one-sided discordant count that reaches two-sided p below 0.05, which is a property of the test and the sample size, not a power-based minimum detectable effect. The Holm family for the primary provenance claim is the six tests {b-c, i-c} x {flash, sonnet, gpt}; the remaining contrasts are adjusted within each contrast's three-family group and are descriptive.

| family | contrast | n | discordant n10/n01 | diff | 95% CI | 6/n | p raw / Holm |
|---|---|---|---|---|---|---|---|
| flash | b-c injected(own) - real | 18 | 6 / 0 | +0.333 | [0.104, 0.549] | 0.33 | 0.03125 / 0.0644531 |
| flash | i-c injected(foreign) - real | 15 | 6 / 0 | +0.400 | [0.131, 0.623] | 0.40 | 0.03125 / 0.0644531 |
| flash | j-c transplanted - real | 15 | 1 / 1 | +0.000 | [-0.242, 0.242] | 0.40 | 1 / 1 |
| flash | f-c pre-error cut - real | 10 | 7 / 0 | +0.700 | [0.304, 0.853] | 0.60 | 0.015625 / 0.015625 |
| flash | i-j formulaic - transplanted | 15 | 6 / 0 | +0.400 | [0.131, 0.623] | 0.40 | 0.03125 / 0.0625 |
| sonnet | b-c injected(own) - real | 49 | 24 / 0 | +0.490 | [0.340, 0.623] | 0.12 | 1.19209e-07 / 5.96046e-07 |
| sonnet | i-c injected(foreign) - real | 18 | 9 / 1 | +0.444 | [0.120, 0.670] | 0.33 | 0.0214844 / 0.0644531 |
| sonnet | j-c transplanted - real | 18 | 1 / 0 | +0.056 | [-0.103, 0.244] | 0.33 | 1 / 1 |
| sonnet | f-c pre-error cut - real | 38 | 15 / 0 | +0.395 | [0.233, 0.549] | 0.16 | 6.10352e-05 / 0.00012207 |
| sonnet | i-j formulaic - transplanted | 18 | 8 / 1 | +0.389 | [0.087, 0.613] | 0.33 | 0.0390625 / 0.0625 |
| gpt | b-c injected(own) - real | 67 | 45 / 0 | +0.672 | [0.538, 0.763] | 0.09 | 5.68434e-14 / 3.41061e-13 |
| gpt | i-c injected(foreign) - real | 28 | 21 / 0 | +0.750 | [0.532, 0.863] | 0.21 | 9.53674e-07 / 3.8147e-06 |
| gpt | j-c transplanted - real | 28 | 2 / 2 | +0.000 | [-0.148, 0.148] | 0.21 | 1 / 1 |
| gpt | f-c pre-error cut - real | 51 | 35 / 0 | +0.686 | [0.530, 0.784] | 0.12 | 5.82077e-11 / 1.74623e-10 |
| gpt | i-j formulaic - transplanted | 28 | 21 / 0 | +0.750 | [0.532, 0.863] | 0.21 | 9.53674e-07 / 2.86102e-06 |




Paired differences with Newcombe 95 percent CIs; filled green markers are intervals excluding zero, grey markers include zero.

Reading. The exploratory i-c estimates are positive with the donor author held fixed and exclude zero in 3/3 families (b-c: 3/3), so an author mismatch between arms is not required for the gap; error text and the preceding derivation still differ, so other prefix differences are not excluded. Real errors are repaired in 2.0-9.0% of continuations against clean-truncation baselines of 69.4-89.6%. Instructions recover repair of real errors only partially: check-first shifts recovery significantly in 1/3 families, location-given in 2/3, consistent with [16]. Arm h exists only where the error donor also produced a correct parsed trace on the same item, so h- and i-contrasts are computed on that common subset; unequal arm denominators reflect eligibility, not attrition (API failures were zero and every continuation was judged). Donor-correct eligibility plausibly skews this subset easier; the i-c contrast is computed entirely within it, so no cross-subset comparison enters the provenance claim and no imbalance was detected within it. A balance check confirms the direction of any selection: arm-c recovery is HIGHER on the eligible subset than on the remainder (7%/6%/21% vs 0%/0%/0% across families), but arm i does not exist outside that subset, so this baseline comparison does not identify whether eligibility shrinks or enlarges the i-c gap relative to a broader item pool. With 15-28 pairs, a fully one-sided 6/0 discordance gives an exact two-sided p of 0.03125 before correction; that is a sample-dependent threshold of the test, not a statement that smaller population differences cannot be detected.

Matched-text contrasts. Transplanting the real error text into the donor's correct derivation (arm j) versus the same text in its own flawed derivation (arm c): j-c = +0.000 [-0.242, 0.242], n=15 (flash), +0.056 [-0.103, 0.244], n=18 (sonnet), +0.000 [-0.148, 0.148], n=28 (gpt). Removing the error step from the organic prefix (arm f) restores recovery relative to c: f-c = +0.700 [0.304, 0.853], n=10 (flash), +0.395 [0.233, 0.549], n=38 (sonnet), +0.686 [0.530, 0.784], n=51 (gpt). Formulaic versus transplanted injection in the same context: i-j = +0.400 [0.131, 0.623], n=15 (flash), +0.389 [0.087, 0.613], n=18 (sonnet), +0.750 [0.532, 0.863], n=28 (gpt).

Reading the matched-text arms. j-c holds the error text fixed while changing the preceding derivation, whereas i-j compares two error texts inside the same correct derivation. j-c point estimates are near zero in every family while i-j point estimates are positive in all three; only gpt passes the three-family Holm correction on i-j. These exploratory results associate recovery with the substituted error text without identifying a unique causal feature. They are consistent with differences between the sampled error distributions, but they do not establish equivalence between j and c, and they do not rule out a contribution from the surrounding derivation. The j-c intervals admit sizable differences in either direction (flash [-0.242, +0.242]), so the j-c result is a failure to detect a difference at this sample size, not evidence of no difference. f-c shows that removing the error step raises recovery, which again does not isolate the role of the pre-error context. Real-error text is therefore a candidate source for future injection probes, but this study does not establish that such probes can substitute for organic-error continuations.

## 7. Confirmatory cohort on untouched items

The within-item cohort of Section 6 is treated as exploratory: its arms were added while the design was still moving. The Section 4 pilot keeps its pre-registered status. We therefore registered a second cohort before any call to it. The item pool is the same three public tiers continued by the identical stride rule at later indices, deduplicated against the exploratory questions, so no item appears in both. The registration fixed the target, the six primary tests, the multiplicity correction and the stopping rules, and forbade inspecting any contrast while collecting. We report the result as registered.

Confirmatory cohort, the two primary contrasts per family. Same estimand, same scoring and same correction as the exploratory table; the two cohorts are never pooled.

| family | contrast | n | discordant n10/n01 | diff | 95% CI | p raw / Holm |
|---|---|---|---|---|---|---|
| flash | b-c | 24 | 4 / 3 | +0.042 | [-0.223, 0.299] | 1 / 1 |
| flash | i-c | 22 | 6 / 1 | +0.227 | [-0.042, 0.458] | 0.125 / 0.25 |
| gpt | b-c | 84 | 36 / 5 | +0.369 | [0.224, 0.493] | 7.84157e-07 / 4.70494e-06 |
| gpt | i-c | 53 | 24 / 2 | +0.415 | [0.245, 0.555] | 1.04904e-05 / 5.24521e-05 |
| sonnet | b-c | 65 | 18 / 6 | +0.185 | [0.015, 0.340] | 0.0226558 / 0.0679675 |
| sonnet | i-c | 47 | 14 / 3 | +0.234 | [0.032, 0.411] | 0.0127258 / 0.0509033 |

Confirmatory cohort by benchmark tier. The second registered extension was restricted to GSM-hard and the MATH-500 numeric subset, which is declared in the registration, so the mix is not the same as the exploratory cohort and the reader can see what each tier contributes. Every matched pair appears here: the tier columns sum to b-c 173, i-c 122, which are the same totals as the family table above. Tier-level tests are omitted when fewer than five matched pairs are available; those rows are kept and marked not tested.

| tier | contrast | n | discordant n10/n01 | diff | p raw |
|---|---|---|---|---|---|
| gsm8k | b-c | 4 | 1 / 0 | +0.250 | not tested |
| | i-c | 4 | 3 / 0 | +0.750 | not tested |
| gsmhard | b-c | 160 | 56 / 12 | +0.275 | 6.2095e-08 |
| | i-c | 112 | 40 / 6 | +0.304 | 3.1028e-07 |
| math500 | b-c | 9 | 1 / 2 | -0.111 | 1 |
| | i-c | 6 | 1 / 0 | +0.167 | 1 |

What replicated and what did not. All six confirmatory point estimates are positive, matching the exploratory signs on items the design had not seen. After Holm correction, evidence against a zero gap is confined to the two gpt contrasts (2 of 6). The sonnet adjusted values are 0.0509033 and 0.0679675 and flash is not separated from zero; the sonnet intervals shown in the table are pointwise, not multiplicity-adjusted, so neither family meets the registered 0.05 threshold. The registered target was 300 matched pairs and the cohort delivered 173 b-c pairs (of which 122 are i-c eligible) across 1038 continuations, because a donor needs a genuinely wrong answer in a short enough trace and these families solve most items. The tier table shows the confirmatory evidence rests mainly on the harder benchmark. We therefore report sign consistency across six point estimates together with confirmatory rejection of a zero gap in gpt only; the other two families remain inconclusive under the registered correction.

## 8. What repair rates do to outcome-derived step labels

**Observation (outcome-label identity; standard algebra).** Let F indicate a truly faulty step, Y = 1{final outcome incorrect} the outcome-derived label, r = P(outcome correct | F=1) the recovery-any rate and e = P(outcome incorrect | F=0). The raw label Y itself, used as a fault predictor, has precision $\pi(1-r) / (\pi(1-r) + (1-\pi)e)$ and recall 1-r against F, where $\pi$ = P(F=1).

This is a property of the labels, not a bound for trained verifiers: a learner with features informative about F can denoise Y and exceed 1-r recall (in the extreme, a feature that reveals F recovers it perfectly from arbitrarily noisy labels). An earlier draft presented this identity as a ceiling for outcome-supervised verifiers; the external review correctly rejected that reading and we withdraw it. What the identity does support is a calibration point: anyone substituting this pilot's injected-condition recovery rates, r in the 86.8-97.2% range (recovery-any, measured above), gives raw-label recall of 2.8-13.2%; substituting the selected own-error control (0/60 error-localised, below roughly 5% by the rule of three) gives a plug-in recall of one, subject to sampling uncertainty and that control's selection conditions. These are illustrations for the sampled conditions, not general ranges for injected-error probes. Recall is all this identity pins down: precision additionally requires the prevalence of wrong steps and the rate at which correct steps are mislabelled, neither of which we measure here, so we make no claim about label cleanliness. Using an injected-condition recovery estimate to calibrate labels for an organic-error target population is not justified without validating that substitution on the target population. A deterministic simulation (theory_check.py; 16 parameter settings) matches the identity to within 0.02. Training verifiers on injected versus organic labels under a fixed budget and cross-evaluating on human-annotated benchmarks [17,18] is the natural downstream test and is registered future work.

What this paper establishes vs what you may want (estimand statement).

| you may want | what we actually establish |
|---|---|

| | |
|---|---|
| 'step verifiers have a hard ceiling' | no such claim; an identity for RAW outcome-derived labels whose parameter r must come from the error population you care about |
| 'injected probes are useless' | these probes measure recovery after typed edits, not edit detectability alone. Exploratory b-c gaps are +0.333 to +0.672; confirmatory b-c gaps are +0.042 to +0.369 and i-c gaps are +0.227 to +0.415, with Holm significance confined to gpt |
| 'models cannot self-correct' | in the exploratory foreign-error arm, unprompted recovery is 2.0-9.0%; the effect of explicit instructions varies by family |

## 9. Attrition and incident disclosure

Per-arm attrition in the within-item experiment is zero after a defect fix: our judgement stage originally truncated the detector's output budget, which silently dropped 29% of verdicts; an attrition audit caught it, the budget was raised, and every continuation was re-judged. All numbers in this paper are from the complete re-judged set. A 12-row smoke batch predating the registered pilot run is excluded from every analysis (Section 4). API failures: 0 rows. Endpoint ids and dates are pinned in Section 3.

## 10. Limitations

Single operator annotated the detector validation set (200 labels); a second independent annotator and an adversarial implicit-acknowledgment set are future work. The localiser is itself an LLM (83/93 coverage on pilot traces). We do not assume a direction for its errors: a systematically late index would change which step the continuation starts from, and mislocalisation could in principle either shrink or inflate the measured gap, and the results presented here do not quantify that sensitivity. The blind edit judge is flash-class and therefore shares a family with one continuing model; the NLL probe uses an independent reference model but not the continuing models themselves. Arms i and e were added post-registration under a dated amendment, and arms f and j were added with no amendment note at all, so every contrast that uses f or j is exploratory (Section 3). The confirmatory cohort's second registered extension was restricted to GSM-hard and the MATH-500 numeric subset, so its item mix is not the exploratory mix and the gsm8k tier contributes four pairs that carry no test. The three main families are proprietary, dated serving endpoints; the open-weight replication is small. Organic pilot controls used temperature-0 reproduced errors (selection toward deterministic failures); the within-item experiment replaces those with k=6 temperature-0.8 sampling but retains eligibility and sampling selection. Benchmarks are numeric math only; scoring is exact-match. Instruction arms (d, e) change the instruction, not only the evidence, so they measure elicitable repair, not disposition alone.

## 11. Conclusion

Across the sampled protocols, injected-error arms have higher recovery point estimates than matched organic-error arms, including comparisons that fix the item, the position and the trace author. In the separately registered confirmatory cohort all six primary point estimates are positive, but only the two gpt contrasts pass Holm correction at 0.05, so the direction is not statistically established separately in every family. A blind edit-detection probe is consistent with a surface-form account: formulaic experimenter edits are flagged at 66.1% against 20.6% for real errors, transplanted real-error text sits between them at 31.7% and is not separated from real errors at this sample size, and a small reference LM's token surprisal does not separate the two conditions (AUC 0.47), which does not exclude separation under other reference models. Corruption-probe results about step-level verification should not be read as statements about organic errors; outcome-derived step labels should be calibrated against organic, not injected, recovery rates; and probes that must inject should draw the injected text from real model failures, whose recovery estimates in our data are +0.000 to +0.056 above organic errors with every interval including zero. Ledgers, registrations, prompts, analysis code and figures accompany the submission as a runnable package (README with pinned endpoint ids and per-stage entry points).

## 12. Use of large language models

Large language models are the OBJECT of study here, and they are also used as instruments inside the pipeline: the first-error localiser, the acknowledgment detector and the blind edit judge are LLM calls, each with its prompt, endpoint id and date pinned in the artifact package. Every primary contrast in this paper is computed from exact-match correctness of the final answer, which is deterministic and involves no model judgement. LLM calls do enter prefix construction and eligibility through the first-error localiser, and they support the acknowledgment decomposition and the mechanism probes, which we report as secondary. The authors also used LLM assistance for drafting and for code review; all reported numbers are produced by the released scripts from the released ledgers, and the authors take responsibility for the content.

## References


1. P. Bogdan et al. Thought Anchors: Which LLM Reasoning Steps Matter? arXiv:2506.19143.
2. Local Causal Attribution of Chain-of-Thought. arXiv:2606.21821.
3. From Chains to DAGs. arXiv:2601.17593.
4. J. Huang et al. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798.
5. Large Language Models Have Intrinsic Self-Correction Ability. arXiv:2406.15673.
6. Decomposing LLM Self-Correction: the Accuracy-Correction Paradox and Error Depth Hypothesis. arXiv:2601.00828.
7. S3cMath: Spontaneous Step-Level Self-Correction. arXiv:2409.01524.
8. Chain-of-Thought Unfaithfulness as Disguised Accuracy. arXiv:2402.14897.
9. SCAN: Self-Denoising Monte Carlo Annotation for Robust Process Reward Learning. arXiv:2509.16548.
10. FreePRM: Training Process Reward Models Without Ground-Truth Process Labels. arXiv:2506.03570.
11. Towards Robust Process Reward Modeling via Noise-Aware Learning. arXiv:2601.12748.
12. Dissociation of Faithful and Unfaithful Reasoning in LLMs. arXiv:2405.15092.

13. Two Regimes of Chain-of-Thought Unfaithfulness. arXiv:2607.23458.
14. Certified Self-Consistency. arXiv:2510.17472.
15. T. Lanham et al. Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.
16. G. Tyen et al. LLMs Cannot Find Reasoning Errors, but Can Correct Them Given the Error Location. arXiv:2311.08516.
17. C. Zheng et al. ProcessBench: Identifying Process Errors in Mathematical Reasoning. arXiv:2412.06559.
18. Can Large Language Models Detect Errors in Long Chain-of-Thought Reasoning? (DeltaBench). arXiv:2502.19361.
19. Synthetic Error Injection Fails to Elicit Self-Correction in Language Models. arXiv:2512.02389.
20. Exposing the Achilles' Heel: Evaluating LLMs' Ability to Handle Mistakes in Mathematical Reasoning. arXiv:2406.10834.
21. Error Classification of Large Language Models on Math Word Problems: a Dynamically Adaptive Framework (MWPES-300K). arXiv:2501.15581.
22. Unspoken Hints: Accuracy Without Acknowledgement in LLM Reasoning. arXiv:2509.26041.
23. Reward Models Identify Consistency, Not Causality. arXiv:2502.14619.

## 13. Appendix: instrument contrasts, mechanism probe and open-weight replication

These analyses support the main result but are not needed for it, so they sit here to keep the main text inside the page limit.

Arms per family (%). Qualified items: flash 18, sonnet 49, gpt 67.

| family | arm | n | recovery | acknowledgment | silent repair |
|---|---|---|---|---|---|
| flash | a own real error | 5 | 0.0 | 0.0 | 0.0 |
| | b injected, own trace | 18 | 38.9 | 22.2 | 22.2 |
| | c foreign real error | 18 | 5.6 | 11.1 | 0.0 |
| | d foreign + check prompt | 18 | 5.6 | 33.3 | 0.0 |
| | e foreign + location given | 18 | 55.6 | 61.1 | 22.2 |
| | f foreign, cut before error | 10 | 80.0 | 0.0 | 80.0 |
| | g clean truncation (own) | 18 | 77.8 | 0.0 | 77.8 |
| | h clean truncation (foreign) | 15 | 93.3 | 0.0 | 93.3 |
| | i injected, foreign trace | 15 | 46.7 | 20.0 | 33.3 |
| | j transplanted error, foreign trace | 15 | 6.7 | 0.0 | 6.7 |
| sonnet | a own real error | 23 | 0.0 | 4.3 | 0.0 |
| | b injected, own trace | 49 | 51.0 | 20.4 | 40.8 |
| | c foreign real error | 49 | 2.0 | 0.0 | 2.0 |
| | d foreign + check prompt | 49 | 8.2 | 22.4 | 2.0 |
| | e foreign + location given | 49 | 10.2 | 24.5 | 6.1 |
| | f foreign, cut before error | 38 | 42.1 | 5.3 | 42.1 |
| | g clean truncation (own) | 49 | 69.4 | 4.1 | 67.3 |
| | h clean truncation (foreign) | 18 | 83.3 | 0.0 | 83.3 |
| | i injected, foreign trace | 18 | 50.0 | 33.3 | 33.3 |
| | j transplanted error, foreign trace | 18 | 11.1 | 5.6 | 11.1 |
| gpt | a own real error | 16 | 6.2 | 12.5 | 6.2 |
| | b injected, own trace | 67 | 76.1 | 61.2 | 26.9 |
| | c foreign real error | 67 | 9.0 | 7.5 | 3.0 |
| | d foreign + check prompt | 67 | 43.3 | 53.7 | 1.5 |
| | e foreign + location given | 67 | 76.1 | 89.6 | 0.0 |
| | f foreign, cut before error | 51 | 78.4 | 2.0 | 78.4 |
| | g clean truncation (own) | 67 | 89.6 | 1.5 | 88.1 |
| | h clean truncation (foreign) | 28 | 100.0 | 7.1 | 92.9 |
| | i injected, foreign trace | 28 | 96.4 | 57.1 | 39.3 |
| | j transplanted error, foreign trace | 28 | 21.4 | 17.9 | 3.6 |

Within-item paired recovery differences, instruction and clean-truncation contrasts. Same estimand, scoring and correction as the main table; these contrasts are adjusted within their own three-family group and are descriptive.

| family | contrast | n | discordant n10/n01 | diff | 95% CI | 6/n | p raw / Holm |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| flash | h-c clean(foreign) - real | 15 | 13 / 0 | +0.867 | [0.551, 0.941] | 0.40 | 0.000244141 / 0.000244141 |
| flash | h-i clean - injected (foreign) | 15 | 7 / 0 | +0.467 | [0.183, 0.678] | 0.40 | 0.015625 / 0.046875 |
| flash | d-c check-first - real | 18 | 0 / 0 | +0.000 | [-0.156, 0.156] | 0.33 | 1 / 1 |
| flash | e-c location-given - real | 18 | 9 / 0 | +0.500 | [0.237, 0.694] | 0.33 | 0.00390625 / 0.0078125 |
| flash | g-b clean(own) - injected(own) | 18 | 7 / 0 | +0.389 | [0.145, 0.565] | 0.33 | 0.015625 / 0.0351562 |
| sonnet | h-c clean(foreign) - real | 18 | 14 / 0 | +0.778 | [0.492, 0.891] | 0.33 | 0.00012207 / 0.000244141 |
| sonnet | h-i clean - injected (foreign) | 18 | 6 / 0 | +0.333 | [0.104, 0.521] | 0.33 | 0.03125 / 0.0625 |
| sonnet | d-c check-first - real | 49 | 3 / 0 | +0.061 | [-0.015, 0.164] | 0.12 | 0.25 / 0.5 |
| sonnet | e-c location-given - real | 49 | 5 / 1 | +0.082 | [-0.025, 0.199] | 0.12 | 0.21875 / 0.21875 |
| sonnet | g-b clean(own) - injected(own) | 49 | 11 / 2 | +0.184 | [0.047, 0.309] | 0.12 | 0.0224609 / 0.0351562 |
| gpt | h-c clean(foreign) - real | 28 | 22 / 0 | +0.786 | [0.568, 0.898] | 0.21 | 4.76837e-07 / 1.43051e-06 |
| gpt | h-i clean - injected (foreign) | 28 | 1 / 0 | +0.036 | [-0.088, 0.177] | 0.21 | 1 / 1 |
| gpt | d-c check-first - real | 67 | 23 / 0 | +0.343 | [0.227, 0.455] | 0.09 | 2.38419e-07 / 7.15256e-07 |
| gpt | e-c location-given - real | 67 | 45 / 0 | +0.672 | [0.538, 0.763] | 0.09 | 5.68434e-14 / 1.7053e-13 |
| gpt | g-b clean(own) - injected(own) | 67 | 10 / 1 | +0.134 | [0.043, 0.234] | 0.09 | 0.0117188 / 0.0351562 |

## 14. Mechanism: injected edits are detectably out of distribution

A candidate explanation for the transport gap is that an experimenter's edit is an out-of-distribution event in an otherwise coherent trace, while a real error is locally coherent with its own derivation; the exploratory probes below do not establish the cause of the gap. Two exploratory measurements probe this on i and c prefixes drawn from the same arms (same author G, same position, same item). The probe sample sizes are reported separately below; correspondence with the primary i-c pairs is not established by these aggregate counts.

Blind edit-detection probe. A judge (flash-class, temperature 0) sees the problem and the partial solution and must say whether exactly one step was artificially edited, without being told which condition it is in.

| prefix condition | n | flagged as edited (Wilson 95% CI) | flagged at the true position |
|---|---|---|---|
| formulaic injection (arm i) | 62 | 66.1% [53.7, 76.7] | 40 |
| transplanted real error (arm j) | 60 | 31.7% [21.3, 44.2] | 9 |
| real error in own trace (arm c) | 63 | 20.6% [12.5, 32.2] | 9 |

Reference-LM surprisal (a null, reported as such). On i/c prefix pairs drawn from the same arms, with pair-level correspondence to the blind-judge sample not established by these aggregate counts, the mean token NLL of the flawed step given its own preceding context, under an independent small open-weight reference model (Qwen/Qwen2.5-0.5B, CPU inference; not one of the continuing families), does NOT separate the conditions: 1.07 nats for injected steps versus 1.14 for real error steps (paired mean difference -0.08, 95% CI [-0.24, 0.09], Wilcoxon p = 0.40, n = 63 pairs; AUC 0.47). These exploratory probes are consistent with greater edit detectability for the sampled formulaic corruptions, but they do not establish that surface form explains the recovery gap, do not exclude surprisal effects under other reference models, and do not identify detection as a causal bottleneck. A transplanted real error (arm j) is arithmetically inconsistent with its correct preceding steps, yet is flagged less often than the formulaic edit (table above), and its j-c recovery estimates include zero in every family (Section 6). The decomposition in the next table reports where recoveries stop, without establishing a causal ordering. The pattern is in the same direction as the verifier-side finding that reward models score consistency rather than causal correctness [23]. The reference model is small; surprisal under the continuing models themselves is not observable to us (no logprob access) and is left open.

What 'repair' is: detect / localise / recover decomposition of continuations (LLM-classified; modes: in_place_fix = corrects the flawed step and continues, fresh_resolve = re-derives from the problem statement, continues_flaw = builds on the flawed value). Recovery of injected corruptions is dominated by detection followed by targeted fixing; real errors are mostly never detected.

| family | arm | n | P(detect) | P(localise \| detect) | P(recover \| localise) | top modes |
|---|---|---|---|---|---|---|
| flash | b injected (own) | 18 | 38.9 | 71.4 | 80.0 | continues_flaw 9, in_place_fix 6 |
| | i injected (foreign) | 15 | 26.7 | 100.0 | 75.0 | continues_flaw 10, in_place_fix 4 |
| | c real (foreign) | 18 | 16.7 | 66.7 | 50.0 | continues_flaw 15, in_place_fix 3 |
| sonnet | b injected (own) | 49 | 34.7 | 100.0 | 58.8 | continues_flaw 29, in_place_fix 17 |
| | i injected (foreign) | 18 | 38.9 | 100.0 | 71.4 | continues_flaw 10, in_place_fix 7 |
| | c real (foreign) | 49 | 0.0 | - | - | continues_flaw 49 |
| gpt | b injected (own) | 67 | 70.1 | 93.6 | 84.1 | in_place_fix 44, continues_flaw 22 |
| | i injected (foreign) | 28 | 57.1 | 93.8 | 100.0 | in_place_fix 15, continues_flaw 13 |
| | c real (foreign) | 67 | 7.5 | 80.0 | 100.0 | continues_flaw 62, in_place_fix 5 |

## 15. Open-weight replication

Open-weight arms (protocol deviation, recorded before running: own correct traces from temperature-0 bases; arm a omitted; donor pools and localisations reused from the main experiment).

| model | arm | n | recovery | acknowledgment |
|---|---|---|---|---|
| llama70b | b injected, own trace | 14 | 50.0 | 28.6 |
| | c foreign real error | 14 | 7.1 | 14.3 |
| | d foreign + check prompt | 14 | 21.4 | 28.6 |
| | e foreign + location given | 14 | 21.4 | 35.7 |
| | g clean truncation (own) | 14 | 64.3 | 7.1 |
| | h clean truncation (foreign) | 8 | 100.0 | 0.0 |
| | i injected, foreign trace | 8 | 87.5 | 37.5 |
| qwen80b | b injected, own trace | 11 | 45.5 | 18.2 |
| | c foreign real error | 11 | 9.1 | 9.1 |
| | d foreign + check prompt | 11 | 18.2 | 27.3 |
| | e foreign + location given | 11 | 18.2 | 18.2 |
| | g clean truncation (own) | 11 | 81.8 | 9.1 |
| | h clean truncation (foreign) | 8 | 87.5 | 0.0 |
| | i injected, foreign trace | 8 | 62.5 | 12.5 |

Open-weight paired contrasts (Newcombe 95% CI, exact McNemar, descriptive; small n).

| model | contrast | n | diff | 95% CI | p raw |
|---|---|---|---|---|---|
| llama70b | b-c | 14 | +0.429 | [0.143, 0.652] | 0.0312 |
| llama70b | i-c | 8 | +0.875 | [0.401, 0.978] | 0.0156 |
| llama70b | h-c | 8 | +1.000 | [0.541, 1.000] | 0.0078 |
| qwen80b | b-c | 11 | +0.364 | [0.059, 0.613] | 0.1250 |
| qwen80b | i-c | 8 | +0.500 | [0.104, 0.730] | 0.1250 |
| qwen80b | h-c | 8 | +0.750 | [0.297, 0.884] | 0.0312 |

The replication is small (items are limited by donor-error availability) and descriptive, but the ordering reproduces: real foreign errors resist repair while injected and clean prefixes recover.

## 16. Appendix: adjacency search log

Direct-adjacency search protocol behind Section 2 (2026-08-27; arXiv API (export.arxiv.org) full-text/title queries). Every hit's abstract was read before citing; deltas are stated in Section 2.

| query | closest hit |
|---|---|
| abs:"error injection" AND abs:"self-correct" | 2512.02389 Synthetic Error Injection Fails to Elicit Self-Correction |
| ti:"Achilles Heel" + direct id 2406.10834 | 2406.10834 Exposing the Achilles' Heel |
| all:"MWPES-300K" | 2501.15581 Error Classification of LLMs on MWPs |
| ti:"Unspoken Hints" | 2509.26041 Accuracy Without Acknowledgement in LLM Reasoning |
| ti:"Consistency, Not Causality" | 2502.14619 Reward Models Identify Consistency, Not Causality |
| (earlier rounds) thought anchors / causal attribution / PRM noise / self-correction / faithfulness sets | refs 1-18, prior_art_probe.json |