LOCA-RCA: A LABEL-FREE COUNTERFACTUAL BENCHMARK FOR DEPENDENCE ON CITED EVIDENCE IN RCA AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents increasingly produce root-cause diagnoses accompanied by citations to observability telemetry, yet nothing checks whether the diagnosis actually *depends* on the evidence it cites. We introduce LOCA-RCA (Label-free Observability Citation Ablation for Root Cause Analysis), a training-free and *label-free* benchmark that removes exactly the evidence an agent itself cited, re-runs it, and compares the outcome against removing an equal number of non-cited items. Across 7 models and 3 deployments (335 OpenRCA incidents, of which Market and Bank are two schemas of one testbed), citations are counterfactually load-bearing well beyond a size-matched removal of evidence the agent did *not* cite, for most models, and the dependence runs through two channels with different operational meanings: switching to a different root cause (*flip*) or refusing to answer (*abstain*). Reporting them separately reorders the naive ranking. **Whole-pool citation makes specificity structurally unmeasurable** rather than low, which we report as a benchmark output rather than a missing row; a $k = 3$ citation budget removes the behaviour on the one deployment and two models where we tested it. A salience-matched control leaves the contrast intact in 7 of 9 model-deployment cells. **Item-level human reliability is established under none of four annotation rounds**, so the metric stands on its construction; and base accuracy against gold is 0–11% on two deployments, so this is an audit signal about grounding, not a filter for right answers. We release the construction scripts, probes, metrics, per-incident traces and annotation material.

1 INTRODUCTION

Site-reliability (SRE) and AIOps teams increasingly delegate incident triage to LLM agents that read the telemetry — logs, metrics and traces — and return a root cause together with the evidence they consulted. The citation is what makes such an answer auditable: an on-call engineer accepts or rejects a diagnosis by checking the lines it points at. Yet nothing in current practice verifies that the diagnosis *depends* on those lines. A model can name the right component while its citation is decorative, attached after the conclusion was already formed, and the resulting artifact is indistinguishable, on the page, from one whose conclusion genuinely rests on those lines. The question is **citation attribution without a reference answer**, and it is narrower than chain-of-thought *faithfulness*: whether the conclusion moves when its stated grounds are removed, not whether the reasoning describes the computation.

Existing evaluations do not close this gap. RCA benchmarks score the final root cause against ground truth (Xu et al., 2025; Cai et al., 2026; Cui et al., 2024); CoT-faithfulness work intervenes on the model's own reasoning trace or on answer correctness (Shen et al., 2026; Han et al., 2026; Lanham et al., 2023). None asks the operational question directly: *if the evidence the agent cited were not there, would the diagnosis change?* That question is causal, it is about the input rather than the reasoning text, and, crucially for deployment, it can be asked without any gold root cause.

We therefore build LOCA-RCA, a training-free and label-free probe. For each incident we take the evidence ids the agent itself cited, remove exactly those, and re-run it; a matched control removes an equal number of *non*-cited items, so the signal is the contrast rather than raw sensitivity to a smaller

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.

pool. A sufficiency check keeps only the cited items. Everything is behavioural: no gold label, no judge, no training.

Three findings shape the paper. First, citations *are* counterfactually load-bearing well beyond a size-matched removal of evidence the agent did not cite, on three deployments and seven models, with two exceptions we report rather than round away (Mistral-7B on Telecom, whose interval includes zero, and Claude-Opus-5, unmeasurable on all three). Second, and this is what a single aggregate number hides, dependence is expressed through two channels, switching the root cause (*flip*) or refusing to answer (*abstain*), and reporting them separately reorders the naive ranking: on Market and Telecom small models express dependence by naming a different cause, frontier models by declining to answer, and the two call for different responses from whoever reads the diagnosis. We rest that split on what the channels *mean* rather than on human recognition of them, which our annotation rounds did not establish. Third, we find *no evidence* that dependence is ordered by model tier on the cells where a fair matched comparison exists, and for one closed model the contrast is not estimable at all, because it cites the whole evidence pool, which makes specificity structurally unmeasurable for it.

Contributions:

- **A label-free counterfactual probe** for evidence dependence in RCA: ablate exactly the evidence the agent cited, with a matched random-removal control and a sufficiency check. No gold root cause is required (Section 4).
- **A two-channel readout rather than one score**: *flip* and *abstain* are co-primary because they call for different responses from whoever reads the diagnosis, and reporting them separately reorders the ranking a single score gives (Section 5).
- **Measurements on 7 models over 335 incidents and 3 deployments**, including a structural one: citing the whole pool makes specificity *unmeasurable* rather than low, which a single aggregate scores as perfect dependence. Every claim ships with its status and the population it is estimated on (Table 1; Section 7).
- **Three controls, and the numbers they move**, each catching a way a naive version of this metric misleads (Appendix B).

2 Related Work

What the neighbouring benchmarks intervene on. Counterfactual reasoning about microservice failures predates this line (Gan et al., 2021), LLM-based incident RCA is deployed in production (Chen et al., 2024), and stated-confidence reliability there has been studied directly (Zhang et al., 2024), none of them removing the evidence the agent itself cited. Four recent benchmarks come closer, each intervening on a different object: JustDiag ablates its own *mechanisms* and rescores against silver references (Bi et al., 2026); FORGE withholds whole telemetry *modalities* and validates its taxonomy against *ground-truth* correctness (Riddell et al., 2026); AgentHallu attributes which step broke the *final answer* ($\mathbf{1}[y \neq y_{gt}]$) and FaithCoT-Bench perturbs *reasoning steps* (Liu et al., 2026; Shen et al., 2026); RFEval injects a flawed reasoning chain with the *input held fixed*, the inverse of ours (Han et al., 2026). None removes a *specific cited item*, and none has a label-free headline metric.

vs. prior and concurrent work on the same intervention. Removing gold does not just move this intervention to a new setting, it changes what the intervention can measure: without a reference answer only *change* is observable, and change cannot separate abandoning a bad conclusion from losing hold of a good one. The flip/abstain split is what that forces, the only structure left for recovering operational meaning once correctness is unavailable, rather than an addition to the primitive. The primitive itself is not ours and its lineage is older than a "concurrent work" framing would suggest: DeYoung et al. (2020) defined *comprehensiveness* and *sufficiency*, the two arms we reuse and name after them, Carton et al. (2020) normalised them against the *random*-rationale baseline that is our size-matched control, and Terrenzi et al. (2026) used this exact three-arm design for agentic GraphRAG concurrently with this work. Relative to Terrenzi et al. (2026): they score against reference answers on 30 multi-hop text questions where we are label-free on 335 telemetry incidents, and

they report one aggregate accuracy drop where we decompose it. We run no comparator pipeline on RCA telemetry.

vs. contemporary RCA benchmarks. Gong et al. (2026) and Wang et al. (2026) are both gold-dependent and neither probes evidence dependence; their failure statistics are what motivates measuring it.

vs. abstention research, and positioning. Treating refusal as a first-class outcome follows Wagner (2026), and Afolabi et al. (2026) flag masking artefacts in single-step ablation and leave neutral-placeholder controls to future work, a control our size-matched random removal does not supply. Our claim is therefore the conjunction, not any element of it: *label-free* evidence ablation $\times$ *RCA/telemetry*, where no gold cause exists $\times$ release as a reproducible instrument, plus the flip/abstain decomposition, tested against a pre-registered human study, that one aggregate number hides.

3 BENCHMARK

From raw telemetry to a citable evidence pool. Production telemetry spans days and millions of rows, so an agent cannot cite it directly. For each incident we parse the time window *from the report text*, never from ground truth. An incident whose window cannot be parsed is dropped rather than given the gold time. We then assemble a bounded pool from the three modalities inside it. Because the pool defines what *can* be cited, and therefore what the ablation can remove, we state its construction in full and fix it before looking at any outcome. Each modality contributes its most anomalous items, capped at 8 in rank order: logs at or above WARN or carrying an HTTP status of 400 or more, traces that failed or ran in the top decile of their own operation, and metrics ranked by $|x - \mathrm{median}|$ against their own series median. The pool is therefore at most 24 items and in practice 16 where a modality carries no signal. Every item gets a stable id (`L#`, `M#`, `T#`) and an index back to its source row, so a citation can be resolved, verified and removed.

What the ablation therefore removes. It removes *the items the agent cited*, which is not the same as every trace of the component it named: after removing the cited ones the diagnosed component is still named by some remaining item in 68%, 29% and 35% of probed cases on Market, Bank and Telecom.

Eliciting machine-readable citations. Agents answer under a five-section contract. The sections are reasoning, root-cause entity, root-cause fault type, cited evidence ids, and confidence or abstention. That makes the citation a set of ids rather than prose. Section 7 quantifies how much the elicitation format itself matters.

Data. Three OpenRCA (Xu et al., 2025) deployments: Market (cloudbed-1+2, 148 incidents, log+metric+trace), Bank (136) and Telecom (51). Bank's access logs are almost entirely HTTP-200 traffic with no diagnosable signal and Telecom ships no log modality at all, so both have metric+trace pools only; we record this per incident (`log_salience`) rather than padding the pool with uninformative lines, and it turns out to matter (Section 7). Telecom is the one genuinely different *system* rather than a second schema of the same testbed, so it is the only place a claim can be checked against a second *source* at all. **We therefore avoid the phrase cross-system**: two independently sourced systems, one of them contributing 51 incidents, support a comparison between two sources and not a claim about systems in general. All 335 incidents obtained their window from the report text alone, and ground-truth fault labels are carried along but enter no metric, feeding only the fault-type grouping of Appendix L. We release construction scripts and indices rather than redistributing the upstream telemetry, in line with its licence.

4 PROBE

An incident is a pair (q, E) with q the natural-language report and $E = \{e_1, \ldots, e_m\}$ the evidence pool. An agent $\mathcal{A}$ maps (q, E) to $(r,\, c,\, C,\, \alpha)$: reasoning r, a root-cause conclusion $c = (\text{entity}, \text{fault})$, cited ids $C \subseteq E$, and an abstention flag α. The intervened runs below write

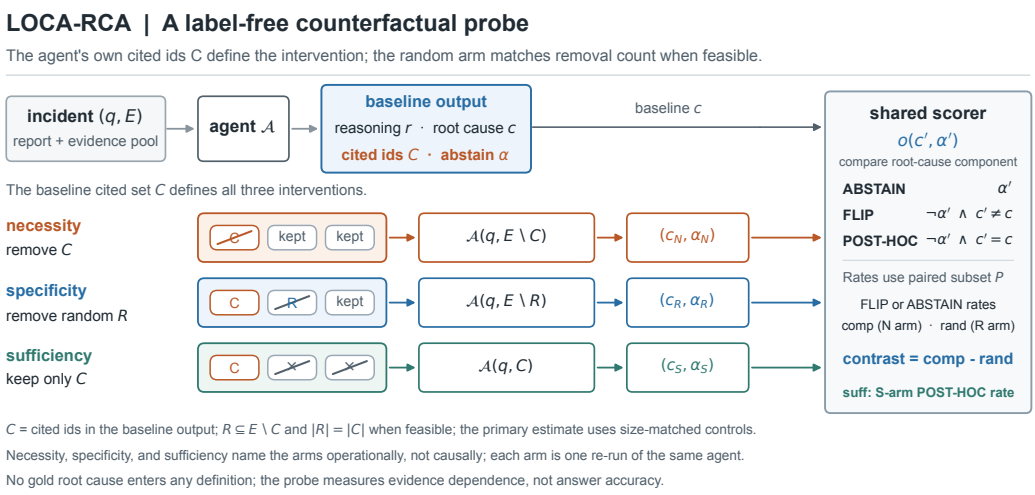

Figure 1: The probe. The reported signal is the *contrast* between the necessity and specificity arms, so a model that is merely sensitive to a smaller pool scores zero.

· in the citation slot because the re-run's own citations are recorded but enter no metric. We run $\mathcal{A}$ greedily so every comparison is between deterministic outputs, verified with a null control (Appendix B).

For each incident with a concrete cited diagnosis ($\neg\alpha$, $C \neq \emptyset$, entity present) we perform the three interventions of Figure 1. The arm names are operational, not causal: they say which items were withheld, not that the cited set is necessary or sufficient beyond this intervention.

$$\text{necessity:} \quad (r_N, c_N, \cdot, \alpha_N) = \mathcal{A}(q,\, E \setminus C) \tag{1}$$

$$\text{specificity:} \quad (r_R, c_R, \cdot, \alpha_R) = \mathcal{A}(q,\, E \setminus R), \quad R \sim \mathrm{Unif}\big(\{S \subseteq E \setminus C : |S| = m\}\big) \tag{2}$$

$$\text{sufficiency:} \quad (r_S, c_S, \cdot, \alpha_S) = \mathcal{A}(q,\, C) \tag{3}$$

Here m in (2) is $\min(|C|, |E \setminus C|)$, a minimum rather than $|C|$ because **a same-size R exists only when $|C| \leq |E \setminus C|$**: the seeded sampler degrades rather than failing, so on a majority-cited pool the arms remove different numbers, the one quantity the control exists to fix. On **Market, which carries the headline results, at most 7.7% of any measurable model's incidents are affected and restricting to size-matched controls moves the contrast by at most** 0.017; on Bank and Telecom the share reaches 64%.

$c' \neq c$ below means the canonical root-cause *component* differs, not the fault wording, after canonicalisation that merges runtime/port variants of the same service (`adservice.ts:8088` ≡ `adservice`).

One incident, concretely. On `openrca-Telecom-021-task_7` Qwen-7B cites two metric items showing `redis_002` memory above median and concludes `redis_002 / memory_pressure`. **Remove those two** and it answers `db_003 / failure`: a FLIP. **Remove a size-matched pair it did not cite** and it stays on `redis_002`, so the change was not a reaction to a smaller pool. The incident contributes $1 - 0 = +1$.

5 METRICS

What the number is, and what we do not call it. The quantity below is a *behavioural, prompt-conditional counterfactual sensitivity to self-cited evidence*, measured on single-shot answers to a fixed prompt over a fixed, pre-built evidence pool. We call it *evidence dependence* for brevity. A high contrast says the answer is causally sensitive to those specific items under this elicitation, not that the model's stated reasoning describes its computation, and not that its citation is the explanation. It is therefore a *comparative* instrument: absolute levels are not portable across elicitation formats, so every claim in Section 7 is a comparison within one format.

Write P for the *paired* subset of probed incidents on which both the necessity and the specificity intervention ran, $S \supseteq P$ for those on which the sufficiency intervention ran (it needs no control), and $\Pr_P$ for a proportion over P. **The primary estimate uses the $m = |C|$ subset of P**, the pooled one all of it. Every contrast is computed on P: differencing a rate measured on all probed incidents against a rate measured on P mixes denominators and produces spurious values (Appendix B). For an intervened output (c', α') define the outcome

$$o(c', \alpha') = \begin{cases} \text{ABSTAIN} & \alpha' \\ \text{FLIP} & \neg\alpha' \wedge c' \neq c \\ \text{POST-HOC} & \neg\alpha' \wedge c' = c \end{cases} \tag{4}$$

so (4) makes ABSTAIN and FLIP partition the cases in which the diagnosis did not survive, and POST-HOC is the case in which the conclusion is unmoved by the removal of its own support. Write o^i_N, o^i_R and o^i_S for $o(\cdot,\cdot)$ applied to incident i's three arms (1)–(3),

$$\text{comp} = \tfrac{1}{|P|}\textstyle\sum_{i\in P} \mathbf{1}[o^i_N \neq \text{POST-HOC}], \qquad \text{rand} = \tfrac{1}{|P|}\textstyle\sum_{i\in P} \mathbf{1}[o^i_R \neq \text{POST-HOC}] \tag{5}$$

$$\text{contrast} = \text{comp} - \text{rand}, \qquad \text{suff} = \tfrac{1}{|S|}\textstyle\sum_{i\in S} \mathbf{1}[o^i_S = \text{POST-HOC}] \tag{6}$$

Here (6) differences the two rates of (5), and sufficiency is the mirror image: keeping only the cited evidence should *preserve* the conclusion. Decomposing the contrast by channel,

$$\text{flip-c} = \Pr_P[o_N = \text{FLIP}] - \Pr_P[o_R = \text{FLIP}], \; \text{abstain-c} = \Pr_P[o_N = \text{ABSTAIN}] - \Pr_P[o_R = \text{ABSTAIN}] \tag{7}$$

with $\text{contrast} = \text{flip-c} + \text{abstain-c}$ in (7). All quantities are label-free: no gold root cause enters any definition. Confidence intervals are percentile bootstrap over incidents ($B = 2000$; 20000 in the appendices). **The resampling unit differs by result**: per-model intervals resample probed *rows*, of which Market has 148 sharing 14 incident ids, while every cross-model claim resamples those 14 *clusters* (Appendix F). Recomputed on clusters, every measurable model's per-model interval still excludes zero, the closed models widening most (Appendix F.3). flip-c **and** abstain-c **are co-primary; their sum is a *behavioural disruption contrast* and nothing more**, since a single total mixes two mechanisms an operator would respond to differently.

6 METRIC VALIDITY

Four annotation rounds, three of them pre-registered, asked whether humans read the ablation the way the metric codes it. **Item-level reliability is established under none of the four**, and this section sets out what that does and does not cost the paper. The round-by-round detail, the instruments and the per-rater numbers are in Appendices J, K and M.

What the rounds do establish. Round 2 (pre-registered, $n = 39$, three annotators) finds almost no item-level agreement, Fleiss $\kappa = 0.018$, so **we make no per-item human-ground-truth claim**. Against a null that shuffles each rater's own label vector across items, preserving their marginals exactly, ABSTAIN ($+0.179$, $p = 0.004$) and POST-HOC ($+0.145$, $p = 0.006$) clear rater habit and FLIP does not ($+0.060$, $p = 0.15$; Table 33). **That asymmetry, and not a reliability result, is the empirical basis for reporting the channels separately.** Round 3 then asked the question the construction actually needs, whether removing the cited evidence destroys the support a human reads in it. It is **met on an amended criterion and failed on the originally committed one** (0.444 against a pre-registered ≥ 0.70), on **18 judgments**, the smallest base in the paper. Both numbers appear and we report it as a pilot result rather than an established property.

What they do not. Round 3's item-level gate fired, mean pairwise $\kappa = 0.000$ against a gate of 0.40, and stopped the round before its 72-item main set ran. Round 4, which we built after a pre-registration that said *no third instrument* and disclose rather than let that document stand, cleared its pilot gate ($\kappa = 0.643$ on 12 items) and then **failed its pre-registered main set**: $\kappa = 0.369$ [0.200, 0.521], Fleiss 0.323. A pilot that clears a gate does not predict the set it gates. **The metric therefore stands on its construction and not on human agreement**, which is the branch PREREG_ROUND4 §8 wrote in advance.

Table 1: **What this paper establishes, before the evidence for it.** Read each claim only at the strength its status allows. SUPPORTED: an interval excluding zero on the stated cells after the corrections of Appendix F. EXPLORATORY: a direction we observe but cannot defend at the available n or without an untested control. REFUTED: measured and absent, not merely unproven. NOT ESTABLISHED: measured and inconclusive, the instrument having missed its own pre-registered gate. Table 12 lists every claim; Table 13 the five the measurements did not support.

Claim	Where it holds	Status
Removing cited evidence changes the conclusion more than a size-matched non-cited removal	3 deployments, 6 measurable models, Mistral-7B on Telecom excepted	✓ SUPPORTED
Dependence runs through two channels whose relative size reverses the naive ranking	Market and Telecom; *not* Bank, where both frontier models lead the flip channel	✓ SUPPORTED
Citing the whole pool makes specificity structurally unmeasurable	Claude-Opus-5, all 3 deployments	✓ SUPPORTED
The contrast survives matching on the measured salience proxies	7 of 9 model×deployment cells; modality, trace status and rank are matched, entity and semantic relevance are not	✓ SUPPORTED *where tested*
Frontier models depend on their citations more than open ones	Market, only against the weakest open model, at Holm 0.040/0.046 under joint clustering; only the Gemini-Flash pair also survives selection bounds	∼ EXPLORATORY
The metric is usable as a screen for correctness	21 cells; it keeps 92% of rows	× REFUTED
Item-level human validity of the metric	4 rounds, 9 annotators; the 72-item main set ran and missed its gate at $\kappa = 0.369$	? NOT ESTABLISHED

Two pre-registered directions the data reversed. Each rater's agreement with the metric now *excludes* zero ($\kappa = 0.429, 0.292, 0.256$), where at $n = 12$ each had contained it; we do not upgrade the claim, because agreement against a judgment the raters cannot reliably make between them is bounded by that unreliability. And PREREG_ROUND4 §7 fixed *flip*'s direction in advance: we expected it **not** to clear its own permutation baseline, and said that clearing it would weaken the reading of flips as confabulation. It cleared (0.269 against 0.144, $p = 0.0006$), so **we report the weakening**: a flip is at least partly readable from the evidence rather than purely a model artefact.

Scope of the flip result. Round 2's items came from Market, whose pools carry a diagnosable log signal. Where evidence is sparse, as on Bank and in any regime in which the cited items carry most of the information, removing them can legitimately leave a *different* best answer, so a flip need not indicate confabulation. We treat "flip is largely confabulation" as established nowhere.

7 RESULTS

Table 1 states the headline claims with their status before the evidence for them. Three more were withdrawn after measurement and one weakened by a pre-registered direction the data contradicted (Table 13).

How to read the three per-model tables. Each row uses the incidents that model was eligible on, so differencing two rows compares two populations; every cross-model claim below uses the matched recomputation of Table 10 instead. **The raw column is not a ranking either**: `contrast` has ceiling $1 -$ control, and normalising by it reorders the models on all three deployments (Appendix F.3).

Citations are counterfactually load-bearing beyond a size-matched control. Table 2: every model's contrast excludes zero, so removing what the agent cited moves the diagnosis strictly more often than removing an equal number of non-cited items, the exception being Claude-Opus-5, whose control exists on only 13 of 115 probed incidents.

Capability is not *shown* to order dependence, on the cells where a fair comparison exists. We state this narrowly because the cells that would license the strong version do not exist: Claude-Opus-5's contrast is not estimable anywhere. Neither comparison that could order the rest separates: matched Mistral-7B − Qwen-7B on Market and GPT-5.5 − Qwen-32B on Telecom both contain zero (Tables 10 and 16), which are **nulls, not equivalences**. We equally do not claim frontier

Table 2: **Evidence dependence on Market (148 incidents).** The channels are reported separately because they mean different things to whoever reads the diagnosis (a changed root cause and a refusal call for different next actions), not because one is better validated than the other (Section 6). **contrast** is the sum of the two unrounded channel estimates, so it can differ from the sum of the rounded ones shown by 0.001. CIs are 95% percentile bootstrap over probed rows (Market has 148, sharing 14 incident ids); the cross-model tests of Appendix F resample those 14 clusters instead. †Too few incidents admit a specificity control for a contrast to be estimable (structurally unmeasurable, not zero). ‡Estimated on a subset the model selected for itself by abstaining at the base step (n_{prb}/n below one half), so it is not measured on the same incidents as the unmarked rows. **Rows are per-model descriptives on each model's own eligible incidents**; cross-model claims use Table 10 instead.

model	**abstain-c**	**flip-c**	**contrast** (= sum)	**suff.**	n	n_{prb}	n_{pair}
GPT-5.5‡	0.671 [0.55,0.78]	0.114 [0.01,0.22]	0.786 [0.68,0.88]	0.80	148	70	70
Gemini-Flash	0.630 [0.53,0.73]	0.000 [-0.09,0.09]	0.630 [0.54,0.72]	0.54	148	100	100
Qwen-32B	0.459 [0.36,0.56]	0.061 [-0.02,0.14]	0.520 [0.42,0.62]	0.67	148	98	98
Qwen-14B	0.407 [0.31,0.51]	0.044 [-0.06,0.14]	0.451 [0.35,0.56]	0.70	148	91	91
Qwen-7B	0.058 [0.01,0.11]	0.245 [0.13,0.35]	0.302 [0.20,0.41]	0.57	148	139	139
Mistral-7B	0.000 [0.00,0.00]	0.399 [0.30,0.49]	0.399 [0.30,0.49]	0.43	148	148	148
Claude-Opus-5	n/a†	n/a†	n/a†	0.29	148	115	13

systems span the whole range: one of the three was never measured (Appendix N probes an open reasoning model instead).

Small models carry their contrast in the flip channel. Figure 4 (Appendix A) splits the contrast by channel. Mistral-7B and Qwen-7B carry essentially all of their signal in flips (0.399, 0.245) with abstention contrasts of 0.000 and 0.058; Qwen-14B/32B and the frontier models invert this (abstention 0.407–0.671, flips near zero). **We do not describe the flip channel as confabulation**, and the reason has changed: round 2 could not recover the flip class above rater habit at $n = 39$, but round 4's 72-item main set *does* (0.269 against a baseline of 0.144, $p = 0.0006$), against the direction we pre-registered. Whether the re-diagnosis is invention or a legitimate reading of what is left stays unresolved. The channels do separate mechanically: on Telecom FLIP is modal where the diagnosed component survives the ablation (66.7%) and ABSTAIN where it is erased (63.1%), with POST-HOC appearing once in 168 erased cases (Appendix F.8). **That mechanism conditions on survival within a model**; whether the split also tracks a model's own disposition to decline is a cross-model question seven models cannot settle, and we report it unsettled (Appendix F.3).

Negative flip contrasts are censoring, not behaviour. Qwen-14B's flip contrast is -0.222 on Bank and -0.128 on Telecom. FLIP and ABSTAIN are mutually exclusive, so an arm abstaining on a share a can flip on at most $1 - a$: a model that answers cited-evidence removal by declining has its own flip channel bounded, and a control arm that shifts components at any rate then drives the difference below zero. Conditioning both arms on not abstaining removes it, and every cell turns positive (Appendix F.10).

Abstention: two mechanisms we cannot separate. Reading abstention as evidence-sensitivity has a competing explanation. A model already heading for a wrong answer gives up when evidence thins. The gold component, entering no metric, would be the check on it, and **it cannot serve as one**: the correct cell is tiny wherever accuracy is, with 19 of 21 model-deployment cells carrying fewer than 8 correct incidents.

Citing the whole pool leaves no control, and small pools make it easy. Claude-Opus-5 cites 97.8% of the pool on average and the *entire* pool in 88.7% of incidents (Figure 3 and Table 8, Appendix A). No non-cited item then remains to form a control, so the contrast is undefined rather than low. Its comprehensiveness is 1.000, measured rather than forced. Specificity is not merely unmeasured but structurally unmeasurable, which the schema of Appendix F.4 reports as a field rather than a missing row. A benchmark reporting only an aggregate score would rank this as perfect dependence. **The rise across deployments is a pool-size artifact, not evidence sparsity**: citing

Table 3: **Telecom head-to-heads, recomputed where *both* models were eligible.** GPT-5.5 − Mistral-7B's lower bound sits on a resampling-grid boundary and is left unstarred (Appendix F.1). *excludes zero.

comparison	n common	own subsets	common subset [95% CI]
Gemini-Flash − Mistral-7B	18	+0.731	+0.778 [+0.556, +1.000]*
Qwen-32B − Qwen-7B	36	+0.388	+0.306 [+0.028, +0.556]*
Gemini-Flash − Qwen-32B	16	+0.266	+0.125 [-0.125, +0.375]
GPT-5.5 − Mistral-7B	18	+0.477	+0.389 [+0.000, +0.722]

everything is mechanically easier from a smaller pool, and conditioned on pools of 16 the rate is flat at 96, 95, 98%. It belongs to one model and to small pools, not to evidence density.

The evasion is promptable away. Capping the budget ("cite at most $k = 3$") restores measurement entirely on Telecom, the worst case: Claude-Opus-5's whole-pool rate goes $0.975 \rightarrow 0.000$, n_{paired} 1→42 and the contrast becomes estimable at +0.429 [0.262, 0.595]. Capped and uncapped contrasts are different-sized interventions and must not be pooled; we report the cap as a separate arm, where it gives the paper's only removal-size-matched comparison (Claude-Opus-5 +0.429 vs Gemini-Flash +0.455, not separated; Appendix D).

The model spread is a deployment effect, not a sparsity effect. Bank has no diagnosable log signal, and rather than blurring the models together it spreads them (Table 6). Dropping the log modality from Market's own pools holds deployment, faults and entities fixed and lands on Telecom's exact composition. The within-family separation does not widen but disappears, $+0.218$ to 0.000 $[-0.218, +0.218]$, while Telecom with the same composition shows $+0.388$ for that pair (Appendix F). **We had read the Bank spread as sparse evidence being where dependent and decorative citation come apart; this within-system test does not support that and we withdraw the reading.**

A third, genuinely different system. Telecom is another vendor's telemetry rather than a second schema of the same testbed (Table 7). Each model is probed only where it produced a concrete cited diagnosis, so per-model contrasts sit on different self-selected sets and every head-to-head is recomputed where *both* models were eligible. Two of the four separations this deployment appears to show do not survive (Table 3).

On a matched population Gemini-Flash separates from Mistral-7B, Qwen-7B and Qwen-14B but **not** from Qwen-32B, and GPT-5.5 from no open model at all, so the defensible statement is conditioned on both the deployment and the control. **That survives neither correction it should**: against the *salience-matched* control the frontier/open pair is +0.107 [-0.04, +0.26], and once the between-model tests resample the 14 incident clusters rather than the 148 rows *and* are Holm-corrected on those clustered p-values, **the only frontier/open separations left on Market are the two against Qwen-7B**, the weakest open model. Applying both corrections jointly rather than in sequence leaves those two intact but moves them from $p < 0.001$ to Holm 0.040 and 0.046, inside the threshold and not comfortably (Appendix F). Bounding those pairs against worst-case eligibility selection rather than sampling noise leaves **one** of the two (Appendix F.1). On Telecom the frontier/open pair is not distinguishable under either control, though within-family scaling holds there too (Qwen-32B − Qwen-7B +0.306 [+0.028, +0.556] on Telecom, +0.235 [+0.071, +0.398] on Market). A frontier/open split survives only in the weakest of these forms.

Does the contrast survive matching on salience? We add a control removing non-cited items matched on modality, trace error-status and nearest within-modality rank. **Those are proxies for salience, not a complete account of it**: the arms are not matched on entity, temporal position, semantic relevance to the question, redundancy with other items, or causal proximity to the fault, so what survives is matching on *measured* salience and not every alternative to authorship. Run on Market and Telecom the contrast survives in 7 of 9 model-deployment cells (Table 23), and the 2 that do not are not the smallest models. What orders the nine cells is each one's random-removal rate ($\rho = -0.750$, exact $p = 0.0127$): where random removal already changes most conclusions there is little margin left for the control to attribute. **We offer that as a candidate diagnostic, not**

a validated rule: nine cells with the same models repeated are not nine independent observations, the cells either side of the cut differ by 0.021 with overlapping intervals, and within Market the association is absent (Appendix F.5).

Where the selection comes from. Only 6 of 148 Market incidents and *no* Telecom incident are eligible for all seven models, so **we abandon the seven-model ranking rather than report it as unsupported**: no population exists on which it could be estimated.

What carries over, and what does not. Within-family scaling reproduces on two of the three deployments and fails on the one with the weakest pools (Figure 2, Appendix A). *Channel assignment, by contrast, is not stable*, and not only for weak models. What reproduces without exception is the *negative* claim motivating the decomposition: an aggregate contrast does not tell you which channel produced it, and the channel is not recoverable from the model's identity alone.

The contrast is not just the component going unmentioned. Split by whether the diagnosed component survives the ablation, the contrast excludes zero in *both* strata on Market and Bank, and Gemini-Flash is *higher* where the component survives, which a mentions-only account cannot produce.

8 LIMITATIONS

This paper runs the *diagnostic* configuration and recommends the *budgeted* one, which it has not validated at scale: $k = 3$, two models, one deployment. **No per-item human standard is claimed anywhere**: four rounds did not settle item-level reliability, and the metric stands on its construction (Section 6). All three deployments come from OpenRCA, of which only Telecom is a genuinely distinct system; Pham et al. (2025) is the independent-source alternative we did not use. **Every interval here is a lower bound**: they resample incidents, not the random control's *draw* and not API nondeterminism, and on an API model the eligible set is not stable (Appendix F.6).

Table 4: **Dependence against correctness, Telecom, pooled over 7 models on paired rows ($n = 213$).** Cells are shares of all rows; the right pair is the quantity the metric is, disrupted by the cited removal *and* not by the random one.

	Single-arm necessity		Net contrast	
Removal disrupts the answer	**Correct**	**Wrong**	**Correct**	**Wrong**
yes	32 (15.0%)	163 (76.5%)	23 (10.8%)	58 (27.2%)
no	1 (0.5%)	17 (8.0%)	10 (4.7%)	122 (57.3%)

The modal case is a wrong diagnosis whose citation is load-bearing on the necessity arm alone, 76.5% of rows. **That arm is not the metric**: subtracting the random control leaves 27.2%, and the modal case becomes a wrong diagnosis the metric does *not* flag, 57.3% (Table 4). Either way **the cell most worth flagging is the one dependence cannot find**. The net partition does lift precision against gold, 0.155 to 0.284 [+0.040, +0.218], but **the lift is the control arm's**: given a row that survived random removal, cited-removal disruption adds +0.017 [-0.009, +0.043] (Appendix F.10). The pool names the gold component on 43% of Telecom incidents; where it does not, no diagnosis can be correct (Appendix F.3). **Base accuracy against gold is 0–11% on Market and Bank** (Appendix G). Finally the scope: every result is a single-shot answer to one prompt over a flat, pre-built pool, not an iterative agent choosing what to retrieve (Kuvshinova & Jin, 2026; Tian et al., 2026).

9 CONCLUSION

LOCA-RCA asks whether a diagnosis changes when exactly the cited evidence is removed, against removing as many uncited items, with no gold root cause anywhere. The contrast excludes zero in every measurable cell but one, and splitting it by channel reorders the models. The probe, the traces and a schema that refuses to print a contrast without a denominator are released with it.

AI USE STATEMENT

This work was carried out with a general-purpose AI coding assistant, and we describe that use in full because the paper's subject matter makes the distinction between assisted and unassisted work especially load-bearing.

The assistant was used for tasks in the *required-disclosure* category: it participated in refining the research question; in designing the probe and every control (the size-matched random-removal control, the null re-ask control, the salience-matched control, the multi-draw control and the top-k citation cap were proposed in that dialogue); in designing the round-3 *and round-4* annotation instruments and drafting both pre-registrations, including the numeric gates that later stopped round 3; in implementing the dataset construction, probe, metric and analysis code; and in *interpreting* the results, proposing readings of the measurements and the wording in which they appear. Every interpretation that survives here was checked against the released numbers by an author, and the readings that did not survive that check are recorded as withdrawn in Table 13 rather than removed. It was also used for *recommended-disclosure* tasks: literature search and gap identification, drafting and editing the manuscript, and generating figures and tables from the persisted statistics.

The assistant *did* implement the analysis code, so it sits on the path from raw traces to every number here; what it did not do is put a number on that path from outside it. No experimental result, no reported number and no human annotation was fabricated, hand-written or generated by a model. All numbers are computed by the released scripts from persisted per-incident traces of real model runs, and re-running those scripts reproduces them. Every annotation in rounds 1–4 was produced by a human rater working on blind items, with no AI involvement in any response, and an LLM judge was explicitly rejected in favour of human annotation. **One place needs stating separately, because it is the only point where assistant output touched the human validity chain**: round 4 teaches its raters what sufficiency means through eight calibration items, and those eight intended answers and their one-line rationales were *drafted by the assistant* from a judgement line fixed in advance, then confirmed or overridden item by item by an author before any rater saw them. The generator refuses to emit an annotator workspace until every item carries an explicit author confirmation, and it records for each item whether a draft was AI-prepared; all eight are recorded as drafted and all eight as author-confirmed. Drafts were derived from the stated sufficiency line and never from the metric's outcome, since drafting from the outcome would manufacture the agreement the study exists to test. Where an analysis was designed by the assistant, it was committed, with its thresholds, before the data it would be applied to existed, so that the design could not be tuned to the result; the round-3 stopping rule fired against the outcome we hoped for and is reported as it fell. That general statement has one exception we flag rather than smooth over: round 4 exists *because* round 3 failed, so its design, though itself pre-registered before its own data, was chosen with knowledge of the previous outcome. The code was verified by a unit-test suite and by re-running the full pipeline. Every claim about a cited paper was checked against that paper's text, and the reference list was compiled from bibliographic-database metadata rather than generated. The author has reviewed all AI-assisted work and takes responsibility for the final content of this paper, including its text, its claims and its released artifacts.

ETHICS STATEMENT

The human validation involved nine annotators across an exploratory round and three pre-registered rounds: one in round 1, three in round 2 (one of whom is an author of this paper), three in round 3 and three further annotators in round 4, from both of which authors were excluded. The annotation is a judgment task over machine-generated text with no personal data and no third-party subjects. We did not seek an IRB determination. We state that as a fact rather than asserting none was required, because the raters are themselves the human subjects here and we publish per-rater performance data. All eight independent annotators volunteered and were unpaid: the two round-2 raters contributed roughly two hours each, the round-3 raters 20, 24 and 51 minutes, and the round-4 raters between 49 and 84 minutes of active time across calibration, pilot and the 72-item main set, as recorded by the instrument itself. Timings are given unlabelled here and in Section 6: with three raters a median and a range are close to the full set of values, so what this actually withholds is the mapping from a duration to a pseudonym that appears elsewhere in the paper. That is a modest protection and we describe it as one. All eight declined authorship and consented to the release of

their per-item answers under the pseudonyms A2, A3, B1–B3 and C1–C3; they are identified by pseudonym only and no acknowledgement names them in any version of this paper.

We have changed what we release about them. An earlier version published per-rater response times and calibration scores alongside the answers. Pseudonymity is weak in a pool of three, and timing traces make re-identification easier for anyone who knows who was asked. Nothing in the paper's analysis needs per-rater timing: it is a pre-registered *descriptive* endpoint, so we now release it aggregated (median and range across raters, which is what the paper quotes) and withhold the per-rater series and the per-rater calibration scores. The per-item answers stay, because inter-rater agreement cannot be recomputed without them, and that is the trade we think is right. Raters were told at recruitment that their answers would be released pseudonymously; they were *not* specifically told that timing would be, which is why that part is withdrawn rather than kept with a note. Any rater may ask for their data to be removed at any time, with no reason given, and we will re-release the artifact without it.

Recruitment and relationships bear on how freely consent was given, so: all volunteers were recruited from an author's academic and professional acquaintances, none is in a supervisory or evaluative relationship with an author in either direction, none is a co-author on any other work with an author, and none was compensated. Unpaid volunteer labour of this kind is common in small annotation studies and is still worth naming as a limitation of the arrangement rather than a neutral fact: a paid, independently recruited pool would be better evidence and we did not have one. We state plainly that author-performed annotation is a conflict of interest. That is why round 3 excluded authors by design, the generator refusing to produce a workspace under an author's rater id in rounds 3 and 4 alike, and why every rater's numbers are reported separately rather than only in aggregate (Section 6 and Limitations). The remaining mitigations are procedural rather than statistical: all raters saw byte-identical items presented blind, without the automated verdicts and without each other's answers; the round-2 option→outcome mapping and hypotheses, and the round-3 instrument, gates and analysis, and the round-4 instrument, gates and analysis, were committed publicly before any annotation existed, so no reported direction could be selected after the fact. We report the resulting near-chance inter-rater agreement in full rather than discarding the raters who disagreed with the first, and round 3 is reported including the pre-registered stopping rule that ended it short of its main set and the originally committed endpoint that failed before the amended one passed. Round 4 is reported including the pilot that cleared its gate, the pre-registered main set that then failed it, and the two pre-registered readings the larger sample reversed. One of them, the direction fixed for *flip*, cuts against a reading this paper relies on. The annotation material is released so that others can redo it. The telemetry is synthetic-fault data from research testbeds and contains no personal information; we redistribute construction scripts and indices only, in line with its licence. A faithfulness probe can be gamed. Section 7 documents one behaviour (citing the entire pool) that already defeats it, and we report such evasions rather than suppressing them.

Reproducibility statement

Section 3 specifies incident construction, Section 4 the three interventions and the seeding of the control, and Section 5 the estimator and its confidence intervals; the supplementary material contains the code and a checklist mapping every reported number to the command that produces it. For every run we persist *per-incident* outcome records including the **raw model outputs** of each intervention, so any future revision of the metric can be recomputed offline without re-querying a model. Every number here was produced by that path, and the records for all three deployments and every control suite are in the release rather than described by it (9.6 MB). We say this precisely because the claim was briefly untrue: the auxiliary suites were released and the two deployments carrying the main results were not, so the headline numbers rested on data a reader could not obtain. Two smaller instances of the same gap are recorded in Appendix F: a loader whitelist that silently dropped the salience-matched arm, so the control separating authorship from salience could not be recomputed from the released traces at all, and a statistics table that omitted the one column a reader would go looking for. Runs record the model id, SDK version and UTC timestamp; local models are loaded 4-bit and decoded greedily, and the pipeline reproduced identical aggregates on three independent re-runs. API nondeterminism is not assumed away but measured by the null control (Appendix B). The annotation material and the round-2, round-3 and round-4 pre-registrations are released with commit timestamps predating the annotations, the round-3 gates are asserted by unit tests that fail if a threshold is edited without editing its pre-registration, and every amendment to that pre-registration

is dated and recorded rather than silently applied. Tables and figures are generated from the persisted statistics by scripts, so no number in the paper is transcribed by hand.

REFERENCES

Halimat Afolabi, Zainab Afolabi, Elizabeth Friel, Jude Roberts, Antonio Ji-Xu, Lloyd Chen, Egheosa Ogbomo, Emiliomo Imevbore, Phil Eneje, Wissal El Ouahidi, Aaron Sohal, Alisa Kennan, Shreya Srivastava, Anirudh Vairavan, Laura Napitu, and Katie McClure. Faithful or just plausible? evaluating the faithfulness of closed-source LLMs in medical reasoning. In *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 1562–1591. PMLR, 2026.

Tingzhu Bi, Xinrui Jiang, Xun Zhang, Pengcheng Su, Congjie He, Jinglin Li, Ping Wang, and Meng Ma. Justdiag!: A diagnostic justification engine for accountable root cause analysis. *arXiv preprint arXiv:2606.19407*, 2026.

Yuanhong Cai, Xiaohui Nie, Kanglin Yin, Changhua Pei, Yongqian Sun, Shenglin Zhang, Haibin Liu, Guiyang Liu, Xidao Wen, Fang Situ, and Dan Pei. A multi-dataset benchmark for evaluating LLM agents in microservice failure diagnosis. *arXiv preprint arXiv:2606.29193*, 2026.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. Evaluating and characterizing human rationales. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9294–9307, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.747. URL https://aclanthology.org/2020.emnlp-main.747/.

Yinfang Chen, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, Jun Zeng, Supriyo Ghosh, Xuchao Zhang, Chaoyun Zhang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Tianyin Xu. Automatic root cause analysis via large language models for cloud incidents. In *Proceedings of the Nineteenth European Conference on Computer Systems, EuroSys 2024, Athens, Greece, April 22-25, 2024*, pp. 674–688. ACM, 2024. doi: 10.1145/3627703.3629553. URL https://doi.org/10.1145/3627703.3629553.

Tianyu Cui, Shiyu Ma, Ziang Chen, Tong Xiao, Shimin Tao, Yilun Liu, Shenglin Zhang, Duoming Lin, Changchang Liu, Yuzhe Cai, Weibin Meng, Yongqian Sun, and Dan Pei. Logeval: A comprehensive benchmark suite for large language models in log analysis. *arXiv preprint arXiv:2407.01896*, 2024.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Yu Gan, Mingyu Liang, Sundar Dev, David Lo, and Christina Delimitrou. Sage: practical and scalable ml-driven performance debugging in microservices. In Tim Sherwood, Emery D. Berger, and Christos Kozyrakis (eds.), *ASPLOS '21: 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Virtual Event, USA, April 19-23, 2021*, pp. 135–151. ACM, 2021. doi: 10.1145/3445814.3446700. URL https://doi.org/10.1145/3445814.3446700.

Albert Gong, Kyuseong Choi, Abhineet Agarwal, Jason Schechner, Ryan Huang, Raj Agrawal, Anish Agarwal, and Raaz Dwivedi. Orca-bench: How ready are language model agents for oncall? *arXiv preprint arXiv:2607.28545*, 2026.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. Rfeval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.17053.

Anastasiia Kuvshinova and Seungmin Jin. Auditable graph-guided root cause analysis for Kubernetes incidents. *arXiv preprint arXiv:2606.08590*, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. Agenthallu: Benchmarking automated hallucination attribution of LLM-based agents. *arXiv preprint arXiv:2601.06818*, 2026.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. Rcaeval: A benchmark for root cause analysis of microservice systems with telemetry data. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (eds.), *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, pp. 777–780. ACM, 2025. doi: 10.1145/3701716.3715290. URL https://doi.org/10.1145/3701716.3715290.

Evelien Riddell, James Riddell, Gengyi Sun, Michał Antkiewicz, and Krzysztof Czarnecki. Stalled, biased, and confused: Uncovering reasoning failures in LLMs for cloud-based root cause analysis. In *Proceedings of the 2026 IEEE/ACM Third International Conference on AI Foundation Models and Software Engineering (FORGE)*, pp. 172–183, 2026. doi: 10.1145/3793655.3793732.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. Evidence contextualization and counterfactual attribution for conversational QA over heterogeneous data with RAG systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 1040–1043, 2025. doi: 10.1145/3701551.3704126. arXiv:2412.10571, first posted 2024-12-13.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. Faithcot-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations (ICLR)*, 2026.

Riccardo Terrenzi, Maximilian von Zastrow, and Serkan Ayvaz. Why neighborhoods matter: Traversal context and provenance in agentic GraphRAG. *arXiv preprint arXiv:2605.15109*, 2026.

Yifang Tian, Yaming Liu, Zichun Chong, Zihang Huang, Yiran Li, and Hans-Arno Jacobsen. GALA: Graph-augmented LLM agents for root cause analysis and incident response in microservices. *arXiv preprint arXiv:2608.08968*, 2026.

Benedikt J. Wagner. Two axes of LLM abstention: Answer correctness and question answerability. *arXiv preprint arXiv:2607.08456*, 2026.

Yilun Wang, Guangba Yu, Haiyu Huang, Yujie Huang, Zirui Wang, Pengfei Chen, and Michael R. Lyu. Cloud-opsbench: A reproducible benchmark for agentic root cause analysis in cloud systems. *arXiv preprint arXiv:2603.00468*, 2026.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. Openrca: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations (ICLR)*, 2025.

Dylan Zhang, Xuchao Zhang, Chetan Bansal, Pedro Las-Casas, Rodrigo Fonseca, and Saravan Rajmohan. LM-PACE: confidence estimation by large language models for effective root causing of cloud incidents. In Marcelo d'Amorim (ed.), *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, FSE 2024, Porto de Galinhas, Brazil, July 15-19, 2024*, pp. 388–398. ACM, 2024. doi: 10.1145/3663529.3663858. URL https://doi.org/10.1145/3663529.3663858.

A ADDITIONAL RESULTS AND COMPARISONS

This appendix carries the two deployments Section 7 summarises without tabulating, the three figures its paragraphs point at, and one positioning table. Table 5 places the probe against the concurrent comparator; Figures 2–4 give within-family scaling, citation breadth and the channel split; Tables 6–8 give Bank, Telecom, and the coverage that leaves one model with no specificity control to measure.

Table 5: **Row-by-row delta against the concurrent comparator.** The conjunction of the middle four rows is our contribution; the first row is theirs, and the removal primitive itself is Saha Roy et al. (2025)'s. This is qualitative positioning read off the cited paper, not an empirical head-to-head.

	Terrenzi et al. (2026)	**this work**
intervention	Cited / Random Removal, Full Isolation	identical (their prior claim)
scoring	accuracy vs. reference answers	**label-free**; no gold cause exists
domain	30 multi-hop text questions	335 incidents of telemetry
what removal means	one aggregate accuracy drop	**decomposed** into flip / abstain
control coverage	not reported	reported (`no_control_rate`)
artifact	internal analysis of a system	released evaluation instrument

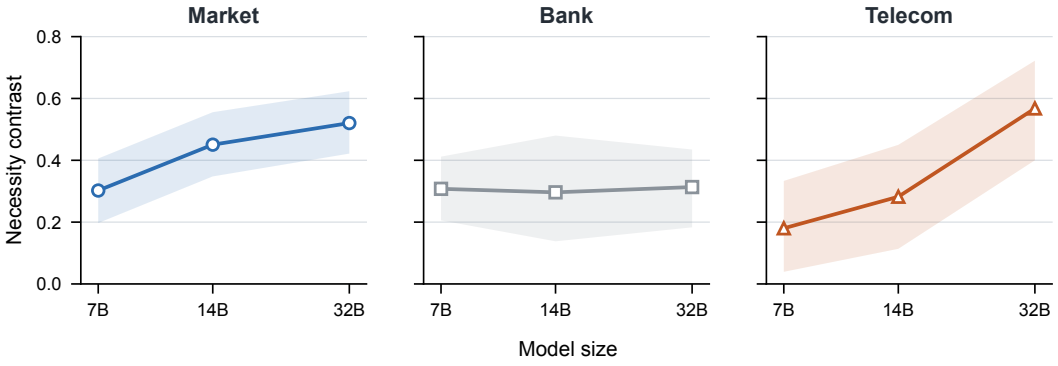

Figure 2: Within-family scaling, all three deployments. Shaded regions denote 95% bootstrap confidence intervals. The Qwen sizes separate on Market (+0.218) and on Telecom (+0.388) but are not separated on Bank (+0.006). The figure shows *where* the separation appears; it is not evidence that scaling buys evidence dependence, which this paper files as EXPLORATORY (Table 12) because the Market separation does not survive a worst-case selection bound and disappears once Market's log modality is dropped. A line missing here would contradict the text, which is why the generator refuses to emit this figure with fewer than three. Each point rests on that model's paired incidents, between 27 and 139 depending on cell (the n_{pair} column of Tables 2, 6 and 7); the flat Bank line is the thinnest, 27 incidents at 14B, which is why its band is the widest of the nine.

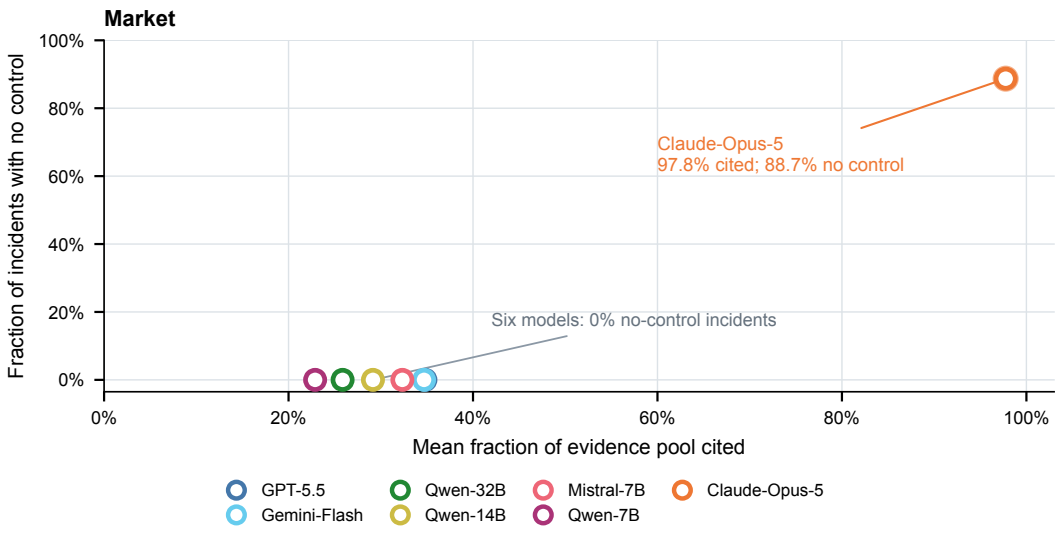

Figure 3: Citation breadth against control coverage, Market. One model sits alone in the top-right: citing nearly the whole pool leaves almost no incident on which specificity can be measured at all. The other six sit on top of one another at zero, so the ordering that separates them is the horizontal one, and Table 8 gives it to three decimals.

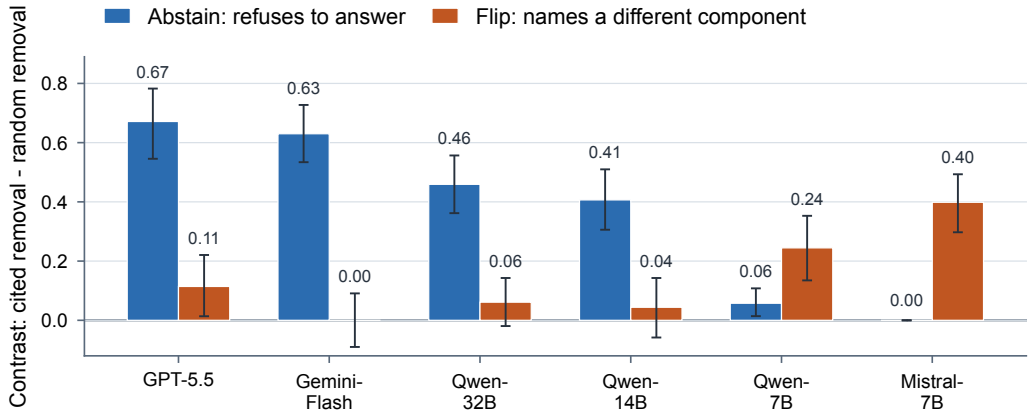

Figure 4: Dependence expressed through abstention (blue) vs. through switching to a different root cause (orange), Market, with 95% bootstrap CIs. The small models carry essentially all of their signal in the orange channel. Mistral-7B's abstain contrast is a true zero rather than a missing bar: it never abstains under either arm, so the interval is $[0.00, 0.00]$. Round 2 could not recover that channel above rater habit; round 4's larger set does (0.269 against 0.144), against the direction we pre-registered. We still do *not* label it confabulation, because whether a re-diagnosis is invention or a legitimate reading of what remains is unresolved (Section 6).

Table 6: **Second deployment (Bank, 136 incidents, no diagnosable log signal).** The models spread rather than blur together here: the frontier systems rise while the open ones fall toward chance under random removal. We do *not* attribute that to sparse evidence: dropping the log modality from Market's own pools leaves the within-family separation at zero (Section 7), so the spread is a deployment effect. CIs are 95% percentile bootstrap over probed rows. †Too few incidents admit a specificity control for a contrast to be estimable (structurally unmeasurable, not zero). ‡Estimated on a subset the model selected for itself by abstaining at the base step (n_{prb}/n below one half), so it is not measured on the same incidents as the unmarked rows. Rows are per-model descriptives on each model's own eligible incidents; cross-model claims use the matched-population comparisons instead (Table 10).

	Co-primary channels		Other readouts		Population		
Model	**abstain-c**	**flip-c**	**contrast (= sum)**	**suff.**	n	n_{prb}	n_{pair}
Gemini-Flash	0.571 [0.48,0.67]	0.286 [0.19,0.38]	0.857 [0.78,0.92]	0.86	136	110	98
Qwen-14B‡	0.519 [0.26,0.76]	-0.222 [-0.50,0.07]	0.296 [0.14,0.48]	0.32	136	31	27
GPT-5.5	0.370 [0.26,0.49]	0.479 [0.36,0.61]	0.849 [0.76,0.93]	0.76	136	75	73
Qwen-32B‡	0.119 [-0.02,0.25]	0.194 [0.04,0.34]	0.313 [0.18,0.43]	0.24	136	67	67
Qwen-7B	0.029 [-0.08,0.13]	0.279 [0.17,0.39]	0.308 [0.21,0.41]	0.26	136	104	104
Mistral-7B	0.007 [0.00,0.02]	0.235 [0.16,0.32]	0.243 [0.16,0.32]	0.38	136	136	136
Claude-Opus-5	n/a†	n/a†	n/a†	0.70	136	113	5

Table 7: **Third deployment (Telecom, 51 incidents, a different vendor's telemetry).** The two frontier models decline to diagnose 63% of incidents here, so their contrasts are measured on a self-selected subset ($n_{\text{prb}} = 19$) and are not directly comparable to the open models'. Claude-Opus-5 cites the whole pool in 97.5% of incidents, leaving $n_{\text{pair}} = 1$: structurally unmeasurable rather than zero. CIs are 95% percentile bootstrap over probed rows. †Too few incidents admit a specificity control for a contrast to be estimable (structurally unmeasurable, not zero). ‡Estimated on a subset the model selected for itself by abstaining at the base step (n_{prb}/n below one half), so it is not measured on the same incidents as the unmarked rows. Rows are per-model descriptives on each model's own eligible incidents; cross-model claims use the matched-population comparisons instead (Table 10).

	Co-primary channels		Other readouts		Population		
Model	**abstain-c**	**flip-c**	**contrast (= sum)**	**suff.**	n	n_{prb}	n_{pair}
Gemini-Flash‡	0.833 [0.64,1.00]	0.000 [-0.17,0.15]	0.833 [0.64,1.00]	0.79	51	19	18
GPT-5.5‡	0.421 [0.13,0.69]	0.158 [0.00,0.33]	0.579 [0.35,0.80]	0.47	51	19	19
Qwen-14B	0.410 [0.26,0.56]	-0.128 [-0.33,0.10]	0.282 [0.11,0.45]	0.38	51	39	39
Qwen-32B	0.297 [0.12,0.49]	0.270 [0.10,0.44]	0.568 [0.40,0.72]	0.38	51	37	37
Qwen-7B	0.040 [-0.04,0.12]	0.140 [-0.02,0.30]	0.180 [0.04,0.33]	0.22	51	50	50
Mistral-7B	0.000 [0.00,0.00]	0.102 [-0.02,0.23]	0.102 [-0.02,0.23]	0.20	51	51	49
Claude-Opus-5	n/a†	n/a†	n/a†	0.68	51	40	1

Table 8: **Citation breadth and control coverage (Market).** Bank and Telecom are in Table 22, where several models have non-zero whole-pool rates. Citing the entire pool leaves no non-cited item to remove, so specificity cannot be measured at all: the contrast is undefined for that model rather than low. Lavender highlights the whole-pool citation pattern that prevents a specificity control.

model	**cited fraction of pool**	**cites whole pool**	**no specificity control**
Claude-Opus-5	**0.978**	**0.887**	**0.887**
GPT-5.5	0.348	0.000	0.000
Gemini-Flash	0.347	0.000	0.000
Mistral-7B	0.323	0.000	0.000
Qwen-14B	0.291	0.000	0.000
Qwen-32B	0.259	0.000	0.000
Qwen-7B	0.229	0.000	0.000

B PITFALLS IN NAIVE EVIDENCE-DEPENDENCE METRICS

Each was found on real data: (i) necessity without a random control inflates the signal, as Mistral-7B on Telecom removes its cited evidence and changes its answer 89.8% of the time on the paired subset, which alone reads as near-perfect dependence, but it does the same on 79.6% of *random* removals, leaving a contrast whose interval includes zero (0.102 [-0.020, 0.234]). **Both rates here are on the paired subset**: the unpaired necessity rate is 90.2% over all 51 probed incidents, and differencing that against the paired control is pitfall (iii) below; (ii) strict string matching of conclusions deflates it (`adservice` vs `adservice.ts:8088`); (iii) differencing rates computed on different denominators produces a deceptive $0.000\,[0.000, 0.000]$; and (iv) API/thinking models require a null (identical re-ask) control, and on Market Claude-Opus-5 changes its diagnosis or declines on 90.4% of identical re-asks (Telecom 0.350, Table 11).

C CONCLUSION MATCHING AND MATCHED POPULATIONS

Table 9 gives the salience-matched control on Market for the three models probed on the full pool, each on its own probed set. The nine-cell version over Market and Telecom, which restricts every cell to size-matched rows and is what Section 7 quotes, is Table 23.

Table 9: **Salience-matched control (Market).** Rows are the three models probed on the full pool; GPT-5.5's Market run is on its self-selected 67 of 148 and appears in Table 23, which restricts to the 65 of those 67 whose salience-matched removal was the same size as the cited one and therefore differs slightly. * excludes zero.

model (Market)	vs. random	vs. salience-matched	difference
Gemini-Flash	+0.630 [+0.54,+0.72]*	+0.510 [+0.41,+0.61]*	+0.120 [+0.01,+0.23]*
Qwen-32B	+0.520 [+0.42,+0.62]*	+0.490 [+0.38,+0.59]*	+0.031 [-0.04,+0.10]
Mistral-7B	+0.399 [+0.29,+0.50]*	+0.284 [+0.18,+0.39]*	+0.115 [+0.03,+0.20]*

How much does one random draw decide? Every contrast above uses a single realisation of R, and the incident-level bootstrap resamples incidents, not draws. Three draws per incident on Market give per-draw spreads of 0.031 (Qwen-32B) and 0.098 (Mistral-7B); averaging the control over draws and bootstrapping over *both* leaves both contrasts significant (+0.531 [+0.429, +0.633] and +0.343 [+0.235, +0.455]) and changes no conclusion.

Is the contrast already noise-corrected? API models change their diagnosis on a large fraction of *identical* re-asks, so a null-corrected estimator is a fair request. Rather than assume a noise model, we measure where the control sits: on Telecom the random arm changes the conclusion at *exactly* the null rate for both models with enough null-controlled incidents to test (difference 0.000, 95% CI ± 0.17 and ± 0.21; a third confirmation under the capped contract above) while the cited arm sits 0.833 and 0.579 above it. **The random control is empirically the nondeterminism floor**, so the contrast already subtracts the noise. That is consistency at these n, not proof.

The stronger control also changes a model comparison. Matching incidents by position and bootstrapping the paired difference, Gemini-Flash beats Qwen-32B by +0.226 [+0.07, +0.38] against the random control but only +0.107 [-0.04, +0.26], not distinguishable, against the salience-matched one. Half of the frontier model's apparent advantage on Market was that it cites more salient evidence, not that its conclusions depend on the citation more. Gemini-Flash still separates from Mistral-7B under both controls (+0.210 [+0.06, +0.35]). This is the capability-does-not-predict-dependence result again, and it is the strongest form of it in this paper: primary deployment, matched populations, stronger control. This check covers the four models it was run on.

Does the metric depend on ignoring fault type? Conclusions are compared at the root-cause component, which would record a switch of mechanism on the same component as evidence-invariance. Comparing raw fault strings instead is not the fix: the null control shows the fault wording changes on 68–100% of *identical* re-asks (versus 5–33% for the component), so it inflates

both arms and collapses every contrast. GPT-5.5 on Market falls from 0.786 to 0.029. Mapping fault descriptions to coarse mechanism classes (cpu / memory / disk / network / process / config), so that paraphrase collapses but a genuine mechanism switch counts, leaves the headline numbers essentially unchanged (Market: Mistral-7B 0.399→0.385, Qwen-32B 0.520→0.531, GPT-5.5 unchanged; largest move anywhere 0.080) and preserves every ordering. The component-level rule is therefore a reporting convenience, not a load-bearing assumption.

Table 10: **Matched populations on Market, the evidence for every cross-model claim on that deployment.** Each pair is recomputed on the incidents where *both* models were eligible and bootstrapped as a paired difference; the per-model tables in Section 7 are descriptive only, because each of their rows uses a different, self-selected set of incidents. Only 6 of 148 incidents are eligible for all seven models at once, so a single seven-way ranking does not exist. Claude-Opus-5 is excluded: its contrast is structurally unmeasurable (cite-everything), so differences against it measure nothing. Both difference columns are computed from unrounded estimates, so they can differ by 0.001 from the difference of the rounded per-model figures in Table 2. * excludes zero.

		Contrast difference	
Pair	n	**Own subsets**	**Matched [95% CI]**
Qwen-32B − Qwen-7B	98	+0.218	+0.235 [+0.071, +0.398]*
Qwen-14B − Qwen-7B	89	+0.148	+0.169 [−0.011, +0.360]
Qwen-14B − Qwen-32B	85	−0.070	−0.035 [−0.188, +0.118]
Mistral-7B − Qwen-7B	139	+0.096	+0.094 [−0.050, +0.237]
Mistral-7B − Qwen-14B	91	−0.052	−0.066 [−0.242, +0.110]
Mistral-7B − Qwen-32B	98	−0.122	−0.112 [−0.255, +0.041]
Gemini-Flash − Qwen-32B	84	+0.110	+0.214 [+0.060, +0.357]*
Gemini-Flash − Qwen-14B	77	+0.179	+0.325 [+0.182, +0.468]*
Gemini-Flash − Qwen-7B	99	+0.328	+0.374 [+0.212, +0.535]*
Gemini-Flash − Mistral-7B	100	+0.231	+0.200 [+0.030, +0.360]*
GPT-5.5 − Qwen-32B	58	+0.265	+0.345 [+0.172, +0.517]*
GPT-5.5 − Qwen-14B	57	+0.335	+0.509 [+0.351, +0.649]*
GPT-5.5 − Qwen-7B	68	+0.484	+0.485 [+0.309, +0.662]*
GPT-5.5 − Mistral-7B	70	+0.387	+0.429 [+0.243, +0.629]*
GPT-5.5 − Gemini-Flash	63	+0.156	+0.095 [−0.063, +0.254]

Class versus instance. Diagnoses sometimes name a service class where the evidence names an instance (`redis` vs `redis_001`, `Tomcat` vs `Tomcat01`), and sometimes answer with a phrase ("adservice on node-6"). The headline rule keeps class and instance apart, because OpenRCA's pod indices are genuine identities. A lenient rule merges them only across a letter-to-digit boundary with a separated or zero-padded digit run, so `adservice` and `adservice2` stay distinct services. That rule, which also matches a component named inside a phrase, moves two models on Telecom (Qwen-7B 0.180→0.260, Mistral-7B 0.102→0.082) and leaves the other four bit-identical, preserving every ordering. We report the sensitivity rather than redefine the outcome after the fact.

All four checks are recomputed offline from the persisted traces by `analysis/robustness.py`; no model was re-queried.

D TOP-k CITATION CAP

The cap replaces the contract clause that licenses citing the pool with "cite at most k evidence items: the ones your conclusion most depends on, ordered most load-bearing first. If more than k items support the conclusion, cite only the k whose removal would most change your answer." Asking for a ranking makes the cap a claim about relative dependence rather than an arbitrary truncation. The default $k = 0$ is an internal sentinel for *no budget* rather than a cap of zero, and leaves the contract byte-identical, so all other results are unaffected. Run on Telecom with $k = 3$ and the null control; Table 11. Intervals are incident-level bootstrap, $B = 20000$.

Three cautions. (i) Claude-Opus-5's uncapped channel split is computed on $n_{\text{paired}} = 1$ and is marked not estimable; the capped run must not be described as having moved that model from one

Table 11: **Top-k citation cap on Telecom ($k = 3$, 51 incidents).** "removed" is the mean number of evidence items deleted by the intervention; the random control removes the same number. * excludes zero. †Not estimable: the uncapped Claude-Opus-5 column has $n_{\text{paired}} = 1$. "—" marks a cell the row does not apply to. Blue marks the zero no-control rates under the citation cap.

	Claude-Opus-5		Gemini-Flash	
	no cap	**cap k=3**	**no cap**	**cap k=3**
items removed (pool = 16)	15.975	3.000	8.74	3.000
cites whole pool	0.975	0.000	0.053	0.000
no-control rate	0.975	**0.000**	0.053	**0.000**
n_{probed} / n_{paired}	40 / 1	42 / 42	19 / 18	22 / 22
base abstentions (of 51)	10	9	32	29
null instability	0.350	0.262	0.211	0.091
contrast	n/a†	$+0.429^{*}$	$+0.833^{*}$	$+0.455^{*}$
[95% CI]		[0.262, 0.595]	[0.667, 1.000]	[0.182, 0.727]
flip channel	n/a†	$+0.429^{*}$	$+0.000$	$+0.136$
abstain channel	n/a†	$+0.000$	$+0.833^{*}$	$+0.318^{*}$

channel to the other. (ii) The Gemini-Flash contrast drop is confounded with intervention size by construction, and on the 15 incidents eligible in both runs it is $+0.333$ [0.000, 0.733], too imprecise to establish either that the cap lowers a well-behaved model's contrast or that it does not. (iii) Gemini-Flash's apparent shift toward the flip channel under the cap is consistent with three of sixteen items leaving enough evidence to answer differently rather than not at all, but its interval touches zero.

E THE CLAIM LEDGER

Table 12 is the ledger that Table 1 in the main text summarises: every claim this paper makes, the population it is estimated on, and its status. The three tables here belong together: a reader deciding how far to trust a number should be able to see what we claim, what we stopped claiming, and what we would have to run next, without assembling it from three appendices.

Table 12: **Every claim this paper makes, the population it is estimated on, and its status.** SUPPORTED means an interval excluding zero on the stated cells after the corrections in Appendix F; EXPLORATORY means a direction we observe but cannot defend at the available n or without an untested control; REFUTED means measured and absent, not merely unproven; NOT ESTABLISHED means measured but inconclusive under the pre-registered validity gate; NO EVIDENCE means the required comparison population does not exist in these data; WITHDRAWN means we asserted it in an earlier version and the measurement did not survive; those are in Table 13, listed there rather than only in the text that used to contain them.

Claim	**Estimated on**	**Status**	**Where**
Removing an agent's cited evidence changes its conclusion more than removing a size-matched non-cited set	3 deployments, 6 measurable models, incidents with a concrete cited diagnosis; Mistral-7B on Telecom is the one such cell whose interval includes zero	SUPPORTED	§7, T2, 6, 7

Table 12, continued from the previous page

Claim	Estimated on	Status	Where
Frontier models show more dependence than open ones	Market only; under clustering *and* Holm applied jointly, only the two against Qwen-7B survive, at Holm 0.040 and 0.046; of those only Gemini-Flash also survives selection bounds; gone under the salience-matched control	EXPLORATORY	§7, App. F
The contrast survives matching on the measured salience proxies	7 of 9 model×deployment cells; the 2 exceptions are the 2 highest random-removal rates. Matches modality, trace status and rank; not entity, temporal position or semantic relevance	SUPPORTED *where tested*	§7, T23
It is not merely the diagnosed component ceasing to be mentioned	Market and Bank, both strata; Gemini-Flash runs the other way	SUPPORTED	App. F
Dependence runs through two channels (flip, abstain) whose relative size reverses the naive ranking	Market and Telecom; *not* Bank, where GPT-5.5 and Gemini-Flash have the two highest flip contrasts. The reversal is a property of the unconditional marginals: the outcomes are mutually exclusive, so a high necessity abstention rate censors the flip channel, and conditioning on non-abstention removes the reversal everywhere (App. F.10)	SUPPORTED *where tested*	§5, F4, App. F.10
Results are conditional on the pool we built, but the *comparisons* are not	Market, 3 open models, per-modality cap halved: levels rise 0.08–0.27, ordering and within-family separation unchanged	SUPPORTED	App. F
Set-level removal is better conditioned than per-item attribution on this task	50 Market incidents, 2 models: 27% of one model's dependence is joint, 12% of the other's per-item outcomes are incoherent	EXPLORATORY	App. F
Citing the whole pool makes specificity structurally unmeasurable	Claude-Opus-5, all 3 deployments; a k=3 budget removes it on Telecom	SUPPORTED	§7, T8
Within-family scaling improves dependence	Market only (Qwen-32B − 7B, Holm-corrected); zero on Bank, absent once Market's logs are dropped, and not robust to selection bounds on Market though it is on Telecom	EXPLORATORY	§7, T10
Dependence is ordered by model tier	—	NO EVIDENCE	§7

Table 12, continued from the previous page

Claim	Estimated on	Status	Where
The metric is usable as a screen for correctness	21 model×deployment cells; recall is near 1.000 and the screen retains 92% of rows, so thresholding on it separates almost nothing	REFUTED	App. F
The metric is *associated* with correctness	the same cells; no lift excludes zero, but 19 of 21 carry fewer than 8 correct incidents, so this is a null under low power and not a demonstrated absence	NOT ESTABLISHED	App. F
Item-level human validity of the metric	4 rounds, 9 annotators; the 72-item main set fails its reliability gate ($\kappa = 0.369$), so per-rater agreement with the metric, though above chance, rests on a judgment the raters do not share	NOT ESTABLISHED	§6

Table 13: **Claims from earlier versions of this paper that the measurements did not support.** WEAKENED means a direction we committed to in a pre-registration *before* collecting the data, which the data then contradicted, so the reading survives in attenuated form rather than intact; WITHDRAWN means the claim is gone. Each was overturned by a check that could have gone the other way.

Claim	What the measurement showed	Status	Where
Flips are confabulation, not a visible property of the evidence	72 items, 3 raters; the direction was fixed in advance and *flip* cleared its permutation baseline (0.269 vs 0.144)	WEAKENED	§6, App. K
Eligibility selection does not run in the direction that would inflate cross-model differences	held on Market and Bank; on Telecom GPT-5.5 answers where the contrast is higher ($+0.190$) and Qwen-14B where it is lower (-0.306), both widening the gap	WEAKENED	App. F
Cite-everything worsens as evidence thins	was 3 deployments; a pool-size artifact	WITHDRAWN	§7
Gold settles whether abstention is evidence-sensitivity or capability	was 3 deployments; rests on 2 correct incidents for GPT-5.5 on Market	WITHDRAWN	§7
A sparser evidence regime widens the gap	was Market vs Bank; within-system manipulation leaves the separation at zero	WITHDRAWN	§7

What we stopped claiming. Table 13 lists the claims we asserted in earlier versions and then withdrew or weakened, each with the check that overturned it. Two were weakened rather than dropped, meaning the direction we pre-registered survives in attenuated form; three are gone. We keep them visible rather than deleting them because a reader who compares versions should not have to reconstruct what changed, and because two of the five were overturned by checks that could have gone the other way.

What would move each claim. Table 14 pairs every surviving claim with the measurement that would raise its status, in the ledger's order. Two requirements recur: telemetry from outside the OpenRCA family, since every deployment here is OpenRCA-derived, and a salience-matched control on the deployments that lack one. Neither is expensive relative to the suite already run, and we name them rather than leave a reader to infer what we would do next.

Table 14: **What would raise each claim's status, in the ledger's order.** Two requirements recur: telemetry from outside the OpenRCA family, since every deployment here is OpenRCA-derived, and a budgeted rerun. Neither is expensive relative to the suite already run (Appendix F.9 prices the arms), and we name them rather than leave a reader to infer what we would do next.

Claim	What would strengthen it
Removing cited evidence changes the conclusion more than a size-matched non-cited removal	Telemetry from outside the OpenRCA family; every deployment here is OpenRCA-derived
Frontier models depend on their citations more than open ones	Salience-matched controls on Bank and Telecom, and a budgeted rerun that makes the seventh model measurable
The contrast survives matching on the measured salience proxies	Telecom is done on all five eligible models and yields a boundary rather than a confirmation. Bank remains and needs its incident index rebuilt from the upstream telemetry; more cells in the high-random-removal regime would test whether that rate really is what predicts separability
It is not merely the diagnosed component ceasing to be mentioned	Pools whose modality composition is held fixed across deployments, which Appendix F.7 shows OpenRCA does not offer
Dependence runs through two channels whose relative size reverses the naive ranking	The survival mediation of Appendix F.8 repeated where pools are released for all three deployments, not only Telecom
Results are conditional on the pool, but the comparisons are not	A cap raised rather than lowered, which needs the raw telemetry and a rebuilt pool
Set-level removal is better conditioned than per-item attribution	Per-item leave-one-out on more than two models
Citing the whole pool makes specificity structurally unmeasurable	A budgeted rerun as the primary condition, which Appendix F.4 recommends to adopters and this paper deliberately does not do
Within-family scaling improves dependence	A deployment where pool composition and system identity are not confounded, and a selection bound that keeps its sign on Market rather than only on Telecom
Dependence is ordered by model tier	A population on which all seven models are eligible. None exists in this data, which is why the ranking is abandoned rather than reported as unsupported
The metric predicts whether a diagnosis is correct	A setting whose base accuracy is above 2–11%, where a screen has room to separate anything
Item-level human validity of the metric	Raters who share a sufficiency threshold. More raters would estimate the disagreement more precisely; round 4 shows they would not remove it

F THREE RECOMPUTATIONS ON THE PERSISTED TRACES

Each of these was run on already-persisted traces, with no new model calls. All three are reported whichever way they came out.

The ablation removes the items an agent cited, not every trace of the component it named, so the contrast might detect the component ceasing to be *mentioned* rather than the conclusion depending on the citation. Splitting by whether the diagnosed component survives answers it: on Market and Bank the contrast excludes zero in *both* strata, including the majority one where it is still named (Mistral-7B $+0.387$ [$+0.261$, $+0.505$], $n = 111$), and Gemini-Flash runs the other way, higher where the component survives ($+0.718$) than where it was erased ($+0.318$), which a mentions-only account cannot produce. Of the 21 pairwise between-model tests per deployment, 16/8/15 are sig-

nificant uncorrected and 13/8/11 survive Holm; the step the scaling claim rests on does (Qwen-32B vs Qwen-7B, Holm 0.040), three Market and four Telecom separations do not. Market's 148 rows share 14 incident ids, so resampling *clusters* rather than rows is the honest interval for a between-model comparison: no conclusion changes, but the closed models' intervals widen (GPT-5.5 $\times 1.94$, Gemini-Flash $\times 1.63$), and the clustered version is what every cross-model claim in this paper is tested on. **The per-model intervals of Tables 2–8 are row-level and therefore narrower than their clustered equivalents by roughly those factors**; they are descriptives of one model on its own eligible incidents, not the basis of any comparison, and should not be read as one.

Do the channels differ on correct diagnoses? They may, and the data cannot say. Base accuracy is low enough that only one of the 21 model-deployment cells has at least eight correct diagnoses to stratify on: Qwen-32B on Telecom, 12 correct against 25 wrong. There, the whole difference sits in one channel. Abstain-c is $+0.500$ where the model was right against $+0.200$ where it was wrong, while flip-c is unchanged ($+0.250$ vs $+0.280$). That would read as a model that gives up when its evidence is pulled out from under a correct answer but re-diagnoses when it was heading somewhere wrong anyway. The interval on that difference is $[-0.067, +0.667]$ and contains zero. We record the pattern because it is the shape a better-powered study should look for, not because 12 incidents establish it.

Does eligibility select favourable incidents? Cross-model claims are recomputed where both models answered, which fixes the denominator but not the selection: eligibility is an outcome, since a model is eligible where it did not abstain at the base step. The model that probes nearly everything is the instrument. Comparing Mistral-7B's contrast on the incidents a pickier model was eligible for against those it was not gives -0.036, $+0.028$, -0.049, $+0.006$, $+0.097$, -0.079 on Market, where the largest magnitude is under 0.10 and the lean is negative. Bank behaves the same way (-0.064, -0.066, -0.010, $+0.030$, $+0.110$, -0.006; every interval contains zero). **Telecom does not, and an earlier version of this paper said it did.** There the differences are three times larger and the two that bear on the frontier/open comparison both widen it: GPT-5.5 answers where the reference model's contrast is already higher ($+0.190$ [-0.052, $+0.450$]) while Qwen-14B answers where it is lower (-0.306 [-0.612, -0.013], the one value here excluding zero), so eligibility flatters the frontier side and penalises that open model at the same time. Neither interval on GPT-5.5 establishes the effect, and no Telecom claim in this paper rests on it. We already report the frontier/open pair as not distinguishable there under either control. But **on Telecom we cannot say what we previously said**, that selection does not run in the direction that would inflate our cross-model differences. On Market and Bank, where the cross-model claims actually live, it does not.

How much of this is the pool we built? Four reviews asked, and the honest answer needs the pool changed rather than defended. We have one check at the permissive end already, rebuilding without the salience ranking, and this is the other: halve the per-modality cap from eight to four, which is derivable from the released pools because ids encode within-modality rank, and re-probe the three open models on Market. Mean pool size falls from 21.6 to 8.0.

Absolute contrasts move a lot and the comparisons do not. Every model rises: Qwen-32B $0.520 \rightarrow 0.789$ [0.696, 0.872], Qwen-7B $0.302 \rightarrow 0.431$, Mistral-7B $0.399 \rightarrow 0.480$. The rise comes from the cited arm rather than the control. Mistral's random-removal rate is unchanged ($0.399 \rightarrow 0.405$) while its comprehensiveness goes $0.797 \rightarrow 0.885$, which is what a smaller pool should do, since the cited set is then a larger share of the evidence. The ordering is identical under both caps, and the within-family separation survives and grows: Qwen-32B $-$ Qwen-7B is $+0.303$ [$+0.157$, $+0.449$] on the 89 incidents eligible for all three, against $+0.218$ [$+0.060$, $+0.357$] at the original cap. So the paper's decision to report comparisons and not levels is not a hedge. It is the only part of this that transfers across a pool we could have built differently.

The within-system sparsity manipulation. Bank's spread is striking on its own terms: the frontier models rise (Gemini-Flash 0.630→0.857, GPT-5.5 0.786→0.849) while the open models fall (Qwen-32B 0.520→0.313, Mistral-7B 0.399→0.243), and under random removal the open models are almost as unsettled as under removal of their own citations (rand 0.54–0.70) where the frontier models barely move (0.14–0.15). Reading that as an effect of evidence sparsity requires Market and Bank to differ only in sparsity, and they differ in upstream system, fault taxonomy, entity naming and per-model eligibility as well.

The manipulation that isolates it drops the log modality from Market, leaving 8 metric and 8 trace items in every pool, uniformly 16, which also removes the pool-size variation that turned out to explain the cite-everything trend. Ids are not renumbered, since dropping a whole modality leaves the surviving sequences contiguous. On the 55 incidents where all three open models remained eligible, every separation goes to zero: Qwen-32B − Qwen-7B 0.000 $[-0.218, +0.218]$, Mistral-7B − Qwen-32B 0.000 $[-0.196, +0.196]$, Mistral-7B − Qwen-7B $+0.021$ $[-0.110, +0.152]$. Qwen-32B also stops answering: base-step abstention takes it from 98 eligible incidents to 56, all 92 skips recorded as `base_abstain`, so it needs the log signal to commit to a diagnosis at all, and its flip contrast turns *negative* (-0.179 $[-0.314, -0.036]$) leaving abstention as its entire signal. That last point is a caveat on the comparison as well as a finding: the surviving 56 are a self-selected subset.

Why set-level removal rather than per-item attribution. The probe removes the agent's cited *set* in one intervention. The obvious alternative is per-item leave-one-out, and beyond that a Shapley-style attribution over the pool; both answer a finer question, which item carries the weight, and both cost far more. Measured on our own runs, the three arms plus the base step average 3.03 model calls per incident per model, or 7109 for the whole benchmark (Table 27). Leave-one-out needs one call per pool item, 46999 calls, $\times 6.6$; a Monte-Carlo Shapley estimate at a very optimistic 20 permutations needs 893088, $\times 126$. **Cost is not the only argument, and we ran the finer probe rather than assert it.** On 50 Market incidents we removed each cited item on its own as well as the set, at $\times 2.7$–$\times 3.6$ the calls (leave-one-out over the *cited set* is cheaper than over the whole pool, which is the $\times 6.2$ figure above). Two things fall out. **For Qwen-32B, 27% [0.11, 0.52] of dependent incidents are *joint***: no single citation is load-bearing on its own, but removing the set changes the conclusion. A per-item attribution scores every one of those zero and misses them, which set-level removal does not. **For Mistral-7B, 12% are *incoherent***. A single removal changes the conclusion while removing all of them does not, which no monotone account of evidential support permits, so at least an eighth of its per-item outcomes measure instability rather than structure. The finer probe is not strictly more informative here; on one model it misses a quarter of the effect and on the other an eighth of what it reports is noise (Appendix F).

Would the metric help someone decide? The paper motivates the probe with an engineer deciding whether to trust a diagnosis, so it should be asked whether the measure would help them. Treating "the conclusion changed when its cited evidence was removed" as a screen and scoring the kept diagnoses against gold, the lift over the base rate is $+0.005$, $+0.005$, $+0.037$, -0.006, 0.000, 0.000, 0.000 on Market and comparably small on Bank and Telecom; **no interval in any of the 21 cells excludes zero**. Recall is 1.000 almost everywhere because the screen keeps nearly every diagnosis (118 of 148 for Mistral-7B, 115 of 115 for Claude-Opus-5): being load-bearing is the common case, so thresholding on it discards almost nothing and separates almost nothing. This is a negative result and we report it as one. It says what the metric is for. Evidence dependence is a property of the reasoning, not a proxy for its correctness, and a reader who wanted the second should not take the first for it.

Hole position, and why the id gaps do not explain the abstain channel. Ablation filters the pool without renumbering it, so a model re-answers a catalogue reading `M#2, M#3, M#5` with `M#1` and `M#4` visibly absent. It could infer that evidence was withheld and abstain for that reason alone, which would manufacture abstain-contrast with no dependence on its own citation, and abstain is the whole signal for the two closed models, so this is not a footnote. Three things answer it on data already collected. The size-matched random arm leaves holes too and the contrast subtracts it, so only a difference in *where* the holes fall can survive; on Telecom the mean rank of removed items differs between the cited and random arms by 0.05 for Gemini-Flash and 0.09 for GPT-5.5, that is, not at all for exactly the two models the objection targets. Within the random arm, where nothing cited was removed and authorship is held fixed, no model's abstention separates by hole position. And the salience-matched control, which picks the non-cited item at the nearest index in the same modality, places its holes within one rank of the cited arm's for 77–92% of items, so wherever that control ran, hole position is already matched. GPT-5.5 has since been run against that control, which is what closes this objection where it mattered most: the salience arm's holes sit within one rank of the cited arm's, so if abstention were a response to seeing high-rank items missing rather than to losing its own evidence, GPT-5.5's contrast would collapse against it. It falls from 0.851 to 0.836.

Both figures are on all 67 rows of its salience run; on the 65 of those that are size-matched, which is what Table 23 reports, the pair is 0.877 and 0.862. The comparison is the same either way and only the sample differs. Since 88% of GPT-5.5's signal is in the abstain channel, the channel the objection targets, that is the strongest test of it we can run, and it does not hold.

Per-stratum detail. On Market the share of probed rows in which the diagnosed component survives is 75–88% for the open models and 76% for GPT-5.5, and the contrast in that stratum is $+0.387$, $+0.450$, $+0.500$, $+0.267$, $+0.718$ and $+0.811$ for Mistral-7B, Qwen-14B, Qwen-32B, Qwen-7B, Gemini-Flash and GPT-5.5. Every interval excludes zero. The pooled 68/29/35% figures in Section 3 hide wide per-model variation (Market ranges from 0.04 for Claude-Opus-5, whose cite-everything behaviour erases the pool, to 0.88 for the two larger Qwen models), which is why we give the split per model rather than per deployment.

The two corrections, applied jointly. The multiplicity p-values above resample rows, and clustering widens GPT-5.5's own interval by $\times 1.94$ on Market, so a between-model p involving it is too small and Holm applied to those p-values is optimistic for exactly the pairs the paper leans on. An earlier version flagged that interaction and left it there. Table 15 runs it: every pairwise p recomputed by resampling the 14 incident *clusters*, then Holm across that family.

Table 15: **Market, 21 pairwise comparisons,** $B = 20000$**.** *row* is the correction the paper reported before; *cluster* recomputes each p by resampling incident clusters and then applies Holm to those. Comparisons against Claude-Opus-5 are included in the counts for completeness and carry no meaning, since its contrast is structurally unmeasurable.

	raw < 0.05	**Holm**	**Benjamini–Hochberg**
row-resampled p, then Holm	16	13	15
cluster-resampled p, then Holm	12	9	10

The joint correction does not overturn what the paper claims; it confirms the narrowing already made and thins the margin. The three comparisons Section 7 rests on all survive: Gemini-Flash − Qwen-7B (Holm 0.040), GPT-5.5 − Qwen-7B (0.046) and the within-family step Qwen-32B − Qwen-7B (0.041). Under row resampling the first two sat at $p < 0.001$; under clustering they sit just inside 0.05, which is the optimism the caveat warned about, made numeric.

What drops is exactly the set the paper had already declined to claim. Four comparisons lose significance: Mistral-7B − GPT-5.5 (Holm 0.160), Qwen-14B − GPT-5.5 (0.300), Mistral-7B − Gemini-Flash (0.301) and Qwen-32B − GPT-5.5 (0.492). Those are precisely the frontier/open pairs other than the two against Qwen-7B. Six of the nine survivors are comparisons against Claude-Opus-5, which are significant without being meaningful and which enlarge the family; excluding them leaves the three above.

One pair survives every correction in this paper. Gemini-Flash − Qwen-7B keeps its sign under clustering with Holm, under the salience-matched control, and under the worst-case selection bounds of Appendix F.1. The other two survive the first but not the third: GPT-5.5 − Qwen-7B has selection bounds of $[-0.029, +1.000]$ and Qwen-32B − Qwen-7B $[-0.041, +0.510]$. Three corrections that share no assumptions therefore leave overlapping but not identical survivors, and the intersection is a single pair.

F.1 Selection bounds: how far could eligibility alone move a cross-model contrast?

Four review rounds made the same objection, and the collider check above answers only a weaker form of it: eligibility is an *outcome*, since a model is eligible where it did not abstain at the base step and did emit a resolvable citation, so recomputing on the common subset fixes the denominator and leaves the selection. This is the stronger question. For a pair (A, B) with eligible sets E_A, E_B and intersection I, the reported estimate is $\mathrm{mean}_A(I) - \mathrm{mean}_B(I)$, where $d_i^A \in \{-1, 0, 1\}$ is incident i's contribution to A's contrast. We replace I with the extremal equally-sized subsets of those contributions,

$$\text{lower} = \frac{1}{|I|}\sum_{i\in\underline{S}_A} d_i^A \;-\; \frac{1}{|I|}\sum_{i\in\overline{S}_B} d_i^B, \qquad \text{upper} = \frac{1}{|I|}\sum_{i\in\overline{S}_A} d_i^A \;-\; \frac{1}{|I|}\sum_{i\in\underline{S}_B} d_i^B,$$

where $\underline{S}_X$ and $\overline{S}_X$ are the $|I|$ smallest and $|I|$ largest values of d^X over E_X. This bounds the published difference over every possible subset each model could have contributed. This is a trimming bound in the spirit of the sample-selection literature rather than a known estimator from it: the standard construction bounds an average effect for the always-selected under monotone selection, whereas this assumes nothing about *why* a model was eligible and is correspondingly wider. We describe it in full rather than name it so that a reader checks the construction and not a label. A sign that survives is robust to any selection story acting *among the incidents that were measured*. **It does not bound selection into that set**: incidents where a model abstained at the base step never enter E_X, and the bound assigns them no value. **The same coarseness limits the intervals in Table 3**: GPT-5.5 − Mistral-7B has 18 paired incidents, so resampled means fall on multiples of 0.056 and its 2.5th percentile sits on the boundary between 0.000 and 0.056, landing on either side depending on the resampling seed. We leave that pair unstarred, the conservative reading of a bound the seed decides; it is also the pair the selection bounds above find sign-robust, so the two diagnostics disagree about it and we report the weaker. The per-incident quantity is the one `necessity_contrast` aggregates, $\mathbf{1}[\text{flip or abstain on the necessity arm}] - \mathbf{1}[\text{flip or abstain on the control arm}] \in \{-1, 0, 1\}$, and a test asserts that averaging it reproduces the published metric exactly. Table 16 gives the resulting bounds.

Table 16: **Worst-case bounds over which incidents each model contributed, Market and Telecom.** *matched* is the number the paper reports. Bootstrap intervals elsewhere quantify sampling; these quantify *who got measured*, which is a different and previously unquantified source of uncertainty. ✓ (yes) and × (no) indicate sign robustness to these bounds. Lavender marks the sole sign-robust Market comparison.

pair	$\|I\|$	matched	selection bounds	sign robust
Market (the headline deployment)				
Gemini-Flash − Qwen-7B	99	+0.374	**[+0.071, +0.616]**	✓ yes
GPT-5.5 − Qwen-7B	68	+0.485	[-0.029, +1.000]	× no
Gemini-Flash − Qwen-32B	84	+0.214	[-0.060, +0.310]	× no
GPT-5.5 − Qwen-32B	58	+0.345	[-0.155, +0.759]	× no
Gemini-Flash − Mistral-7B	100	+0.200	[-0.080, +0.520]	× no
GPT-5.5 − Mistral-7B	70	+0.429	[-0.214, +0.957]	× no
Qwen-32B − Qwen-7B	98	+0.235	[-0.041, +0.510]	× no
Telecom				
Qwen-32B − Qwen-7B	36	+0.306	[+0.194, +0.694]	✓ yes
Gemini-Flash − Mistral-7B	18	+0.778	[+0.389, +1.000]	✓ yes
GPT-5.5 − Mistral-7B	18	+0.389	[+0.111, +0.778]	✓ yes
Gemini-Flash − Qwen-7B	17	+0.529	[+0.059, +1.118]	✓ yes
Gemini-Flash − Qwen-32B	16	+0.125	[-0.188, +0.938]	× no
GPT-5.5 − Qwen-32B	15	+0.067	[-0.533, +0.733]	× no

On Market exactly one separation survives, and it is one the paper had already narrowed to. Gemini-Flash − Qwen-7B keeps its sign; every other Market pair, including both against Qwen-32B, does not. Section 7 reached almost the same place by a different route, resampling the 14 incident clusters and Holm-correcting left "the only frontier/open separations left on Market are the two against Qwen-7B". So a multiplicity-and-clustering correction and a worst-case selection bound, which share no assumptions, leave overlapping survivors. We read that convergence as the strongest available evidence that the narrowed claim is not an artefact of the correction we happened to choose.

One claim moves the other way and we restate it. Within-family scaling on Market (Qwen-32B − Qwen-7B) has a bootstrap interval excluding zero, $+0.235$ [+0.071, +0.398], and a selection bound that does not, [-0.041, +0.510]. Sampling uncertainty on that pair is small and selection uncertainty is not. On Telecom the same pair is selection-robust ([+0.194, +0.694]), the reverse of

the deployment pattern Section 7 reports from bootstrap alone, and a reason to keep within-family scaling EXPLORATORY rather than to promote it.

Bank is the deployment where these bounds are widest in the informative direction: six of seven pairs keep their sign, including both frontier models against Qwen-32B. We do not lean on that. Bank is also where 64% of incidents fall back to an undersized control, so its arms are the least comparable in the paper, and a bound that ignores that defect is not evidence for a claim the defect undermines.

F.2 ONE INCIDENT, END TO END

Every quantity in this paper is an average over incidents like this one. It is `openrca-Telecom-021-task_7` answered by Qwen-7B, chosen because the necessity arm flips and the control arm does not, so both halves of the contrast are visible in a single case. Nothing is paraphrased: the pool is the released pool and the two outputs are the released traces. **This appendix shows how the metric is computed and nothing else**; what the human raters were asked, and how four of them answered, is the separate job of Appendix M. No human judgment appears below and none enters the number.

The panel below is generated, not typed. Every case-specific value in it — the incident id, the model, all 16 pool lines, the cited set C, the sampled control set R, both components, both fault strings, each arm's score and the contrast — is read off `dataset/inc_telecom.jsonl` and the released outcomes trace by `analysis/make_worked_example.py`. Only the static labels are written by hand, and a test regenerates the file and requires it byte-identical.

How to read the colour. A fill marks the role a line plays and every band names that role in words, so the panel reads the same in greyscale. **Orange marks what the arm in front of you removes, not "cited evidence"**: the specificity arm removes items the agent never cited, so a fill that meant *cited* could not express that arm at all. C is therefore pale blue where it is introduced, as an input, and turns orange only inside the necessity arm. Grey is what an arm retains, including the pool $E \setminus C$ the control is drawn from; purple is outcome coding. Appendix M uses the same five fills for the same five things. *med* abbreviates the series median, which the released pool spells in full.

Evidence pool E — 16 items as released, 8 metric and 8 trace

C, the agent's own cited set — an input to the probe, not yet an intervention

```
M#1 [m|redis_002]  used_memory=2.59854e+07 (med 2.17588e+07)
M#2 [m|redis_002]  used_memory_rss=2.6964e+07 (med 2.49979e+07)
```

$E \setminus C$ — the 14 non-cited items; this incident's control draw R took `T#3` and `T#8`

```
M#3 [m|redis_001]  keyspace_hits=7.46112e+09 (med 7.46286e+09)
M#4 [m|redis_002]  keyspace_hits=7.39558e+09 (med 7.3973e+09)
M#5 [m|redis_009]  keyspace_hits=7.32168e+09 (med 7.32339e+09)
M#6 [m|redis_001]  used_memory=1.98206e+07 (med 1.82934e+07)
M#7 [m|redis_010]  used_memory=1.54974e+07 (med 1.40763e+07)
M#8 [m|redis_009]  used_memory=1.95213e+07 (med 1.8103e+07)
T#1 [t|docker_006]  db_003 dur=5 status=fail
T#2 [t|docker_005]  db_003 dur=2 status=fail
T#3 [t|docker_005]  db_003 dur=1 status=fail
T#4 [t|docker_006]  db_003 dur=1 status=fail
T#5 [t|docker_003]  csf_002 dur=56 status=fail
T#6 [t|docker_003]  csf_002 dur=55 status=fail
T#7 [t|docker_006]  local_method_011 dur=9 status=fail
T#8 [t|docker_005]  local_method_011 dur=7 status=fail
```

openrca-Telecom-021-task_7 · Qwen2.5-7B-Instruct

16 evidence items: 8 metric items and 8 trace items. Telecom provides no log modality.

Base output

Root-cause component: `redis_002`

Fault description: `memory_pressure`

Cited evidence IDs: $C = \{\texttt{M\#1}, \texttt{M\#2}\}$

Necessity

C is the agent's own citation, not sampled.

Remove $C = \{\texttt{M\#1}, \texttt{M\#2}\}$

Re-run output: `db_003 / failure`

Component changed: `redis_002` → `db_003`

Outcome: FLIP

Necessity $= 1$

Specificity

Sample R from $E \setminus C$, with $|R| = |C|$.

Remove $R = \{\texttt{T\#3}, \texttt{T\#8}\}$

Re-run output: `redis_002 / resource_exhaustion`

Component unchanged: `redis_002` → `redis_002`

Outcome: NO CHANGE

Specificity $= 0$

Outcome coding

Reported contrast = necessity − specificity $= 1 - 0 = +1$

The cited-evidence removal changes the predicted component; the size-matched non-cited removal does not.

The three released outputs, verbatim

Base

```
[REASONING] The traces indicate failures in the 'csf_002' and 'local_method_011' components, which are more frequent and have longer durations compared to other components. Additionally, the metrics show that 'redis_002' has a higher 'used_memory' and 'used_memory_rss' compared to the median values, suggesting potential memory pressure. Given the timing of the failures, it is likely that the high memory usage in 'redis_002' led to performance degradation, causing the downstream services ('csf_002' and 'local_method_011') to fail. [ROOT_CAUSE] entity: redis_002; fault_type: memory_pressure [CITED_EVIDENCE] M#1, M#2 [CONFIDENCE] 0.8
```

Necessity re-run

```
[REASONING] The traces show multiple failures related to 'db_003' and 'csf_002', indicating potential issues with these components. The metric data does not show any significant anomalies that would suggest an immediate cause, but the trace failures provide more direct evidence of the problem's occurrence and component involved. [ROOT_CAUSE] entity: db_003; fault_type: failure [CITED_EVIDENCE] T#5, T#6 [CONFIDENCE] 0.8
```

Specificity re-run

```
[REASONING] The traces indicate failures in the 'csf_002' and 'db_003' components, which are the most frequent failures during the specified time period. The metric data shows that 'redis_002' has higher memory usage compared to other Redis instances, suggesting potential resource exhaustion. Given the failures and the high memory usage, it is likely that the 'redis_002' instance is under significant load, leading to performance degradation and subsequent failures in dependent services. [ROOT_CAUSE] entity: redis_002; fault_type: resource_exhaustion [CITED_EVIDENCE] M#1, M#2, T#5, T#6 [CONFIDENCE] 0.8
```

No gold root cause appears anywhere above, because none enters the probe. The purple rows are the metric's own arithmetic on the two arms; $+1$ records that the citation was load-bearing on this incident, **not** that the diagnosis was correct. R is one draw from $E \setminus C$ and not the whole of it, which is why the two are written separately.

Three things this case teaches that a table cannot. First, **why the control is not optional**: without it, "the answer changed when we deleted two items" would have been the whole finding, and it is not.

Second, **why conclusions are compared at component level**. Scoring the fault string would have recorded the control arm's `memory_pressure` → `resource_exhaustion` as a change and inflated the control rate, which deflates every contrast, one of the pitfalls Appendix B quantifies.

Third, **why the salience-matched control exists**. The uniform draw removed two *trace* items against a citation of two *metric* items. It is size-matched and nothing more, so on this incident the two arms are not comparable in modality. The salience-matched control fixes exactly that, and Section 7 reports what happens to the headline when it is used: on GPT-5.5 the contrast loses 1.8%, and half of one frontier model's margin over the best open model disappears.

F.3 Normalisation, channel independence, pool coverage and clustering

Each of these re-reads the data already collected under a different, defensible convention, to see whether a published reading survives it. None calls a model; `analysis/make_review_recomputations.py` emits every number below from the released traces. Two of the four cut against us and are reported as such.

Normalisation does change the cross-model ordering. `contrast` is a raw difference, so its ceiling is $1 -$ control and cells with different control rates are on different scales. Dividing by that ceiling, the normalisation Carton et al. (2020) apply to the same baseline we borrowed the control from, **moves the model ordering on all three deployments** (Table 17). On Market, Mistral-7B rises above both Qwen-32B and Qwen-14B. **We therefore withdraw any reading of the raw column as a cross-model ranking**, which Section 7 already declines to give but Table 2 makes easy to take. The normalised form has its own degeneracy — it reads 1.000 whenever comprehensiveness is 1.000, regardless of the control — so neither column is a ranking and we print both.

The channel split is not shown to be a restatement of abstention, and not cleared of it either. A model that never declines at the base step cannot express dependence through the abstain channel whatever its citations do, so the split could be a relabelling of that disposition. Across the seven models the association is positive on all three deployments and significant on none (Table 18; Market $\rho = 0.679$, $p = 0.094$). **Seven models cannot settle this**, and we do not claim the split is independent of abstention disposition. What does not depend on it is the mechanism evidence of Appendix F.8, which conditions on component survival within a model rather than comparing models.

Most pools do not name the gold component, and the contrast barely notices. A reviewer asked whether the probe runs on evidence that could have supported a correct answer. It mostly does not: **the gold component is named somewhere in the pool for 43% of Telecom incidents and under a third of the Market variants we can inspect** (Table 19). Splitting the contrast by that stratum moves it very little, and on one variant moves it the wrong way. Two readings survive this and we cannot separate them: the contrast is orthogonal to whether the pool could support a correct answer, which is what a label-free behavioural measure should be; or the probe is largely measuring self-consistency over evidence unrelated to the fault. **We report the coverage rather than choose**, and note that it bears directly on Section 8's statement that the typical measured diagnosis is wrong.

Clustering the per-model intervals does not overturn the headline. Market's 148 rows carry 14 distinct incident ids, and the per-model intervals of Table 2 resample rows while the cross-model tests already resample clusters. Recomputed on clusters, **every measurable model's interval still excludes zero** (Table 20). The two closed models widen most — GPT-5.5 from [0.686, 0.871] to [0.535, 0.904] — so the row-level intervals are optimistic, but not by enough to change the sign anywhere. Bank and Telecom need no such recomputation: their rows carry 136 and 51 distinct ids respectively, one per row.

Table 17: **Raw contrast against the same contrast normalised by its own ceiling.** `contrast` = comp − control has upper bound 1−control, so a cell whose control rate is high cannot reach values a cell with a low one can. The last column divides by that ceiling, the normalisation Carton et al. (2020) apply to the same baseline. **The model ordering changes on all three deployments** (positions moved: Market 3, Bank 3, Telecom 2). *n/a*: no size-matched control exists, so neither form is estimable. A normalised 1.000 only means comp = 1.000.

deployment	model	comp	control	raw	normalised
Market	GPT-5.5	1.000	0.214	0.786	1.000
	Gemini-Flash	1.000	0.370	0.630	1.000
	Qwen-32B	0.622	0.102	0.520	0.580
	Qwen-14B	0.670	0.220	0.451	0.577
	Mistral-7B	0.797	0.399	0.399	0.663
	Qwen-7B	0.626	0.324	0.302	0.447
	Claude-Opus-5	1.000	1.000	n/a	n/a
Bank	Gemini-Flash	1.000	0.143	0.857	1.000
	GPT-5.5	1.000	0.151	0.849	1.000
	Qwen-32B	0.955	0.642	0.313	0.875
	Qwen-7B	0.846	0.538	0.308	0.667
	Qwen-14B	1.000	0.704	0.296	1.000
	Mistral-7B	0.846	0.603	0.243	0.611
	Claude-Opus-5	1.000	0.600	n/a	n/a
Telecom	Gemini-Flash	1.000	0.167	0.833	1.000
	GPT-5.5	1.000	0.421	0.579	1.000
	Qwen-32B	1.000	0.432	0.568	1.000
	Qwen-14B	0.923	0.641	0.282	0.786
	Qwen-7B	0.800	0.620	0.180	0.474
	Mistral-7B	0.898	0.796	0.102	0.500
	Claude-Opus-5	1.000	1.000	n/a	n/a

Table 18: **Does the channel split restate each model's willingness to abstain?** Spearman ρ across the 7 models between the share of incidents a model declines at the *base* step and its two channel contrasts. The association is positive on all three deployments and **significant on none of them at** $n = 7$, so this neither confirms nor clears the concern.

deployment	models	ρ(base-abstain, abstain-c)	p	ρ(base-abstain, flip-c)	p
Market	7	+0.679	0.094	-0.321	0.482
Bank	7	+0.286	0.535	-0.214	0.645
Telecom	7	+0.582	0.170	+0.218	0.638

Table 19: **How often does the evidence pool contain anything naming the gold component?** Telecom as released, and the two Market pool variants whose composition we can reconstruct. Computed per row, since twelve Market rows can share one incident id while carrying different questions and different pools. The last two columns split the contrast by that stratum. Gold labels are used here only to ask the question; they enter no metric.

pool	incidents	name gold	contrast if named	if not
Telecom (as released)	51	22 (43%)	0.400 ($n = 95$)	0.291 ($n = 117$)
Market, cap-4 variant	148	46 (31%)	0.598 ($n = 117$)	0.510 ($n = 251$)
Market, log-free variant	148	43 (29%)	0.495 ($n = 101$)	0.524 ($n = 248$)

Table 20: **Market per-model intervals, resampling rows against resampling incident clusters.** Market's rows are not independent: they carry 14 distinct incident ids. The cross-model tests already resample those clusters; this recomputes the per-model intervals the same way. **Every interval still excludes zero**, and the two closed models widen most. $B = 20000$.

model	rows	ids	contrast	row-level [95% CI]	cluster-level [95% CI]
GPT-5.5	70	11	0.786	[0.686, 0.871]	[0.535, 0.904]
Gemini-Flash	100	13	0.630	[0.530, 0.720]	[0.455, 0.760]
Qwen-32B	98	12	0.520	[0.418, 0.622]	[0.444, 0.618]
Qwen-14B	91	14	0.451	[0.352, 0.549]	[0.341, 0.583]
Mistral-7B	148	14	0.399	[0.297, 0.500]	[0.298, 0.510]
Qwen-7B	139	14	0.302	[0.194, 0.403]	[0.208, 0.418]

F.4 Recommended protocol, and the reporting schema it emits

This paper runs LOCA-RCA in its *diagnostic* form: an unbudgeted citation contract, because the question was what agents do when nothing constrains them, and the answer includes a failure mode we could not have seen under a budget. That is the wrong configuration to adopt. We therefore separate two versions and recommend the second.

Diagnostic LOCA-RCA (this paper). Unbudgeted contract, base plus necessity plus specificity plus sufficiency. Use it to characterise elicitation behaviour, including whether a model's citation contract leaves any control to measure against.

Budgeted LOCA-RCA (recommended for deployment, and for a leaderboard only as the multi-column report of Table 21): **there is no direction to rank on**. Section 8 states that dependence is not a good in itself and that absolute levels do not survive a change of elicitation format, so a single sorted column would be a misuse:

1. **Base, necessity, specificity, sufficiency.** Four calls per probed incident. The size-matched random control is not optional. Without it the number is sensitivity to a smaller pool.
2. **A top-k citation budget by default** ($k = 3$ in our test). This is what stops a model making its own specificity unmeasurable. Capped and uncapped runs measure different interventions and must not be pooled.
3. **A salience-matched control whenever models are compared, and treat a high random-removal rate as a reason not to skip it.** It cost 1.8% of the contrast on the model our headline belongs to and removed half of one frontier margin. Both cells where it overturned the authorship reading carried the paper's highest random-removal rates, and that rate comes off the same run as the contrast, so it is available before the decision (Appendix F.5).
4. **A null re-ask control for nondeterministic or API models**, establishing where the nondeterminism floor sits before any contrast is read as signal.
5. **Run the base step twice on API models before trusting the eligible set.** On Telecom, GPT-5.5's eligible set moved by 8 of 51 incidents between two runs of an identical protocol, while both local models reproduced exactly (Appendix F.6). Nothing else in the protocol detects this.
6. **Report cite-everything as a failure category, not a missing row.** A model that cites the whole pool has not been shown to be less dependent on its citations; it has made the question unanswerable. Those are different findings and one number merges them.

Reporting schema. The released `metrics.stats.summarize` emits every field of Table 21; none is computed by hand. The last two are mandatory, and the reason is the row they expose.

Table 21: **The fields a LOCA-RCA report must carry.** Reporting a contrast without the last two prints a number while hiding how much of the population it could not be computed on; `n_paired` is the denominator itself.

field	what it is, and why it is in the schema
`necessity_contrast`	the headline: disruption from removing the cited set minus disruption from a size-matched non-cited set
`flip_contrast`	co-primary; the conclusion moved to a different root cause
`abstain_contrast`	co-primary; the model declined. Summing the two hides which mechanism produced the number and they call for different responses
`sufficiency_rate`	the conclusion survives when *only* the cited set is kept
`n_paired`	incidents on which both arms actually ran; every contrast's denominator
`no_control_rate`	**mandatory.** Share of probed incidents where the model cited the whole pool, so no control existed at all
`unmeasurable_rate`	**mandatory.** Share where the contrast is not measurable as documented, either no control or an undersized one

Table 22: **The two mandatory fields on our own runs. whole-pool** is `no_control_rate` and **unmeas.** is `unmeasurable_rate` of Table 21. They are the reason the schema needs them: Claude-Opus-5 is unmeasurable everywhere; on Bank more than half of Gemini-Flash's and GPT-5.5's incidents are too (0.68 and 0.51), and on Telecom Gemini-Flash's still are (0.53) while GPT-5.5's are not (0.16). A leaderboard printing only a contrast would show none of this.

	Market		Bank		Telecom	
model	**whole-pool**	**unmeas.**	**whole-pool**	**unmeas.**	**whole-pool**	**unmeas.**
Mistral-7B	0.00	0.05	0.00	0.08	0.04	0.41
Qwen-7B	0.00	0.01	0.00	0.01	0.00	0.08
Qwen-14B	0.00	0.08	0.13	0.65	0.00	0.26
Qwen-32B	0.00	0.00	0.00	0.25	0.00	0.03
Gemini-Flash	0.00	0.05	0.11	0.68	0.05	0.53
GPT-5.5	0.00	0.06	0.03	0.51	0.00	0.16
Claude-Opus-5	0.89	1.00	0.96	1.00	0.97	1.00

Table 22 reports those two fields on our own runs, which is where they earn their place in the schema. **Cite-everything stops being an embarrassing exception once it is a field.** Under the schema, Claude-Opus-5 does not vanish from the comparison with a footnote; it reports `unmeasurable_rate` 1.00 on all three deployments, which is a result about elicitation behaviour and is legible next to everyone else's. The same field also disciplines numbers that look clean: Gemini-Flash's Bank contrast of +0.857 carries an unmeasurable rate of 0.68, and reporting the first without the second is the failure the schema exists to prevent.

F.5 WHERE THE AUTHORSHIP CLAIM HOLDS, AND WHAT PREDICTS IT

Four review rounds named extending the salience-matched control as the highest-return next step. We ran it on Telecom, the one deployment here that is a genuinely different system rather than a second schema, on all five models whose pools admit it. It did not confirm the claim across the board. What it produced instead is a *boundary* with a candidate predictor, which is more useful than another confirmation would have been. Bank is untested: its incident index is not part of the release and rebuilding it needs the upstream telemetry.

Every figure is restricted to incidents where the salience-matched removal was the *same size* as the cited removal, the discipline the paper already applies to the random control. The restriction bites on Telecom and barely on Market: a Telecom pool is exactly 16 items, so a citation often exhausts a modality and the matcher returns fewer stand-ins than there were cited items.

Table 23: **Salience-matched control, size-matched rows only, ordered by each cell's random-removal rate.** n and the salience contrast are on the size-matched rows; **the random-removal rate is the cell's rate over all rows that run probed**, since it is used to *order* the cells and not differenced against them. Computing it on the same size-matched rows instead gives 0.397/0.102/0.381/0.123 on Market and 0.767/0.690/0.444/0.587/0.462 on Telecom, the same ordering and the same ρ. Percentile bootstrap, $B = 20000$, one generator per model so a point does not depend on processing order.

deployment	model	rand. rate	n	salience contrast
Market	Qwen-32B	0.102	98	+0.490 [+0.388, +0.592]
Market	GPT-5.5	0.149	65	+0.862 [+0.769, +0.938]
Market	Gemini-Flash	0.370	97	+0.505 [+0.402, +0.608]
Market	Mistral-7B	0.399	141	+0.277 [+0.177, +0.383]
Telecom	Qwen-32B	0.432	36	+0.472 [+0.306, +0.639]
Telecom	GPT-5.5	0.467	13	+0.692 [+0.462, +0.923]
Telecom	Qwen-7B	0.620	46	+0.239 [+0.087, +0.391]
Telecom	Qwen-14B	0.641	29	+0.172 [-0.034, +0.379]
Telecom	Mistral-7B	0.796	30	+0.100 [-0.033, +0.233]

Seven of nine cells hold; two do not, and they are not the two smallest models. Qwen-7B holds while Qwen-14B, which is larger, does not, and Mistral-7B fails where the same-sized Qwen-7B succeeds. Model tier does not organise this. **The random-removal rate does**: ordering the nine cells by it gives Spearman $\rho = -0.750$ (exact permutation $p = 0.0127$), and the two cells whose interval includes zero carry the two highest rates, 0.641 and 0.796, against 0.620 for the highest elsewhere.

The mechanism is legible. When random removal already changes the conclusion on most incidents, the model is broadly sensitive to losing any evidence, and removing comparably salient non-cited items removes nearly as much as removing its own citations. Mistral-7B's salience arm disrupts 0.867 of incidents against 0.967 for the necessity arm, where Qwen-32B's disrupts 0.528 against 1.000. There is no authorship margin left to detect.

Scope of the ordering. It is an ordering over nine points, not a threshold: the two cells either side of the apparent cut differ by 0.021 in random-removal rate and their intervals overlap, and within Market alone the association is absent ($\rho = -0.400$, $n = 4$) because Market's whole range of rates lies below Telecom's lowest, so it contains no cell in the regime where failures occur. Restricting to size-matched rows also moved Mistral-7B's figure *up*, from $+0.041$ across all 49 rows to $+0.100$ on the 30 size-matched ones, so the looser reading is worse for the claim rather than better.

What it licenses. A practitioner reusing this instrument can read the random-removal rate off the same run that produces the contrast, before deciding whether the contrast needs a salience-matched control to be interpretable. Appendix F.4 states it as a protocol step. We would rather hand over a rule with a stated failure mode than a claim that held everywhere we happened to look.

F.6 An API model's eligible set is not stable across runs

The salience arm is a fresh run of every arm, so it re-executes the base step, and comparing the two runs of the same model on the same incidents is free. It should be a formality: the paper runs every agent greedily and states that comparisons are between deterministic outputs. For the local weights it is a formality. For the API model it is not (Table 24).

Table 24: **The same model, the same 51 Telecom incidents, run twice.** Local weights reproduce exactly; the API model does not, and what moves most is not the rate (0.579 to 0.533) but *which incidents were measurable at all* (19 to 15).

	Probed incidents		
Model	**Run 1**	**Run 2**	`necessity_contrast`
Mistral-7B (local)	51	51	0.102 both runs
Qwen-32B (local)	37	37	0.568 both runs
GPT-5.5 (API)	19	15	0.579 then 0.533

The local runs reproduce to the last decimal. GPT-5.5 does not, and the instability sits in the base step: 8 of 51 incidents changed eligibility between runs, 6 going from answered to abstained and 2 the other way, leaving only 13 probed in both. Of those 13, 2 returned a different root-cause component the second time.

This is a different problem from the one the null control measures. That control re-asks an identical prompt inside one run and shows the random arm changing conclusions at the same rate, which bounds how much of a *contrast* is nondeterminism. What moves here is the *population*: the set of incidents on which an API model is measurable is itself unstable, so two runs of the same protocol estimate the same quantity on different samples. Every review round objected that eligibility is self-selected; this says the selection is not even fixed for a given model.

We do not have enough runs to quantify it. Two runs of one model is an observation, not a rate, and we did not spend more on establishing it, because the finding is qualitative and already consequential: **a cross-model comparison involving an API model carries a source of variance that repeating the run would expose and that no interval in this paper includes.** The local comparisons do not carry it. Anyone reusing this instrument on API models should run the base step more than once before trusting an eligible set, and Appendix F.4 now says so.

F.7 Pool composition moderates the contrast, and Market can test it

A review asked for the contrast stratified by pool size and modality counts, since those differ across deployments and could drive the cross-model comparison. Telecom cannot answer it: every one of its 51 pools is exactly 16 items, 8 metric and 8 trace, so there is no variation to stratify on. Market can, and without building a new variant. The released cap-4 pools recover each Market incident's native log count, because capping at four per modality gives $\min(4, n_{\log}) + 4 + 4$ once the log-free variant establishes that metric and trace are 8 apiece on every Market pool. Market therefore contains 35 incidents that *never had log evidence*, in pools of 16 matching Telecom's exact composition, against 105 whose native log count is at least four. The remaining 8 have one to three logs, where the cap-4 pool recovers the exact count rather than censoring it, and fall in neither stratum. Table 25 splits the contrast by that stratum.

Table 25: **Market split by native log presence.** Three models have a stratum below 8 and are omitted. Claude-Opus-5 is also not shown: its contrast is unmeasurable in both strata, so the corresponding 0.000 values are not evidence that composition does not matter to it.

	Log-free		Log-present		
Model	n	**Contrast**	n	**Contrast**	**Difference**
Mistral-7B	35	+0.543	105	+0.381	+0.162
Qwen-7B	34	+0.324	99	+0.273	+0.051
Gemini-Flash	18	+0.167	78	+0.744	-0.577

The composition is a strong moderator for the frontier model and not for the open ones. Gemini-Flash's Market contrast is +0.744 where the pool contains logs and +0.167 where it never did, with non-overlapping intervals ([+0.641, +0.833] against [+0.000, +0.333]); the two open models move by +0.162 and +0.051. So Gemini-Flash's Market advantage over the open models lives in

the log-present stratum, and a cross-model comparison pooled over both strata mixes that in. This is a further reason the frontier/open claim stays EXPLORATORY.

It also corroborates a result the paper reaches by manipulation. Section 7 drops the log modality from Market's pools by construction and finds the within-family separation disappearing rather than widening, against Telecom's $+0.388$ for the same pair at the same composition. Here the same conclusion follows from incidents that were *natively* log-free, so it is not an artefact of removing a modality after the fact: Gemini-Flash is +0.167 on Market's log-free pools and +0.833 on Telecom's identically composed ones. Same size, same modalities, opposite result. The deployment, not the pool, carries it.

F.8 COMPONENT SURVIVAL SEPARATES FLIP FROM ABSTAIN

A review asked whether the two channels differ by whether the diagnosed component still appears in what the ablation left behind, and whether that mediates them. The question needs the evidence pools, so it runs on Telecom, the one deployment whose pools this repository ships. Survival is decided by the definition the paper already uses rather than a fresh one, and Table 26 cross-tabulates the outcome against it.

Table 26: **Telecom necessity-arm outcomes by component survival, pooled over 7 models on probed rows ($n = 255$; 87/255=0.341 have a surviving component).** Rows are not independent across models and this is one deployment, so we read the pattern and not the precise shares. Blue marks the most frequent outcome in each column.

	Component after cited-evidence removal	
Necessity outcome	**Survives ($n = 87$)**	**Erased ($n = 168$)**
FLIP	**58 (66.7%)**	61 (36.3%)
ABSTAIN	12 (13.8%)	**106 (63.1%)**
POST-HOC	17 (19.5%)	1 (0.6%)

Survival separates the channels sharply. Where the component survives, FLIP is the modal outcome and ABSTAIN is rare; where it is erased, the order reverses. That is a *mechanistic* reason to report the channels separately, which the paper had previously justified only by their operational meaning, and it is partial evidence against the competing reading of abstention we say we cannot exclude. Abstention concentrates where nothing names the component any more, which is what evidence-sensitivity predicts and not what a model giving up on a wrong answer predicts.

It does not settle whether a flip is confabulation. Flips occur in both strata. A flip where the component survives is a change away from a component that something remaining still names, which is consistent with the *cited* support having been load-bearing and equally consistent with instability; the stratification cannot choose between them, and neither could round 4.

The POST-HOC row of Table 26 replaces a check we withdrew. Section 7 withdraws an argument that POST-HOC is genuine invariance. We had observed that post-hoc outputs never name a deleted id, which they cannot, since deleted ids are absent from the prompt. This one is not circular: a model *could* keep its conclusion with nothing left naming the component, and across 168 erased-stratum cases it does so once. POST-HOC therefore requires surviving mention of the component; it is not the model ignoring its evidence.

F.9 WHAT EACH ARM AND CONTROL COSTS

An instrument paper should say what adopting it costs, and this one did not. Every arm is one re-run per probed incident, so cost is countable in calls; the released traces carry token usage for the three API models. Table 27 totals both.

Table 27: **Cost of the full suite over 335 incidents per model.** Calls per incident sit below four because an incident that abstains at the base step stops after one call, and that share differs by model, so the cheapest models to probe are the ones that answer least.

			Tokens		
Model	**Kind**	**Calls**	**Input**	**Output**	**Calls/incident**
Mistral-7B	local	1338	1149888	304883	3.99
Qwen-7B	local	1214	970426	168424	3.62
Qwen-32B	local	941	774638	142822	2.81
Qwen-14B	local	814	695887	148148	2.43
Gemini-Flash	API	1028	825226	226846	3.07
Claude-Opus-5	API	930	957894	1077128	2.78
GPT-5.5	API	844	604105	474326	2.52
API subtotal		2802	2.39M	1.78M	
local subtotal		4307	—	—	

A probed incident costs four calls: base, necessity, specificity, sufficiency. Every further control is $+1$ call per probed incident, except multi-draw at $+(k-1)$. Ranked by what they changed in this paper per call spent:

- **The size-matched random control ($+1$) is not optional.** Without it the metric is sensitivity to a smaller pool; Appendix B shows what that measures instead.
- **The salience-matched control ($+1$) bought the most per call.** It removed half of one frontier model's margin over the best open model and cost 1.8% of the contrast on the model the headline belongs to. Anyone reusing this instrument should run it first.
- **The null re-ask control ($+1$) is cheap insurance.** It established that the random arm changes conclusions at the rate an identical re-ask does, i.e. the contrast sits at the nondeterminism floor, one number that forecloses a whole class of objection.
- **Multi-draw random control ($+2$ for three draws) changed nothing** on the two models where we ran it, per-draw spreads 0.031 and 0.098. Lowest return of the four; we would spend those calls on salience-matching a further deployment instead.
- **The sufficiency arm ($+1$) is the one we cannot currently price**, because it has no paired nondeterminism control of its own: pricing it would need a second cited-only re-ask per incident, whose disagreement rate is the floor its retention rate should be read against. We did not run it, so the sufficiency column is an uncalibrated descriptive rate.

F.10 The correctness contingency on the net partition, and the channels uncensored

Two objections to the way Section 8 and Section 5 read their own numbers. Both are answerable from the persisted traces and neither needs a model call; `analysis/make_operator_appendix.py` produces everything here.

The partition in Table 4 was the single arm, and the metric is a difference. Splitting incidents on whether removing the *cited* evidence disrupts the answer is the necessity arm on its own, and Appendix B's first pitfall is that the necessity arm alone conflates dependence on the citation with broad sensitivity to losing any evidence. The partition that matches the metric is *disrupted by the cited removal and not by the random one*. Table 4 now reports both. The single arm admits 91.5% of rows, so as a screen it separates almost nothing; the net partition admits 38.0%.

The net screen does carry a correctness signal, and it is not the citation's. Table 28 scores each partition as a screen against the gold component. The single arm lifts precision by +0.009 [-0.003, +0.021], a null. The net partition lifts it by +0.129 [+0.040, +0.218], which excludes zero and would, taken alone, overturn this paper's REFUTED verdict on the metric as a correctness screen. It does not, because the lift decomposes. *Surviving the random removal* on its own lifts precision by +0.112 [+0.032, +0.192]; conditional on that, whether the cited removal disrupts adds +0.017

[-0.009, +0.043]. **What predicts correctness here is robustness to losing arbitrary evidence, not dependence on the evidence the model chose to point at.** That is the same verdict the paper already reports, reached on the partition that matches the metric rather than on the single arm, and it is a sharper statement of it. We file it as EXPLORATORY: one deployment, one gold matcher, and the two indicators are not independent.

Stratifying changes the size of the effect and not its source. The lift is carried by the open-weight rows (+0.129 [+0.048, +0.217]) and is not separated from zero on the 38 frontier rows. It is largest where the pool names the gold component at all (+0.214 [+0.081, +0.343] on those 95 rows) and is identically zero on the other 118, where no diagnosis can be scored correct because the component the gold names is not in the prompt. **More than half of the rows behind Table 4 are incidents in which correctness was unavailable by construction**, which is a property of the bounded pool (Appendix F.3) and not of the models.

FLIP and ABSTAIN are competing risks, so the channel split is censored. Equation (4) makes the three outcomes mutually exclusive, so an arm that abstains on a share a of incidents can flip on at most $1 - a$ of them. A model that answers cited-evidence removal mostly by declining therefore has its flip channel bounded from above by its own abstention, and if the control arm meanwhile shifts components at any rate at all, flip-c goes *negative* with nothing behavioural behind it. That is what Qwen-14B's -0.222 on Bank and -0.128 on Telecom are.

Table 29 reports each channel again conditional on the arm not abstaining. **The conditional flip contrast is positive in every cell**, including both negative ones, and **it reverses the ordering the unconditional split produces**: the frontier models lead the flip channel on all three deployments, where unconditionally they are the ones said to express dependence by abstaining. Neither form is the disposition it looks like — where a model's necessity arm never returns POST-HOC its conditional flip rate is 1.000 by construction and the column is reporting the control arm. **We therefore narrow the channel claim** (Table 12): the channels carry different operational meanings and the decomposition is worth reporting, but the *ranking* it induces is a property of the unconditional marginals under censoring and not of the models. The same censoring is why the budgeted contract of Appendix D moves Claude-Opus-5's whole contrast into the flip channel: at $k = 3$ the intervention leaves thirteen of sixteen items standing, which is enough context to name something else instead of declining.

Table 28: **Read as a screen for correctness, against the base rate of the stratum it runs on.** `keeps` is how many of the stratum's rows the screen admits, `prec` the share of those that are right, `lift` that minus the base rate. Intervals resample *incidents*, carrying all of an incident's model rows together. * excludes zero. The two gold strata are exhaustive: where the pool names no item bearing the gold component, no diagnosis in it is scored correct, so both screens are trivially flat there.

stratum	indicator	n	keeps	base	prec	lift [95% CI]
all rows	single arm	213	195	0.155	0.164	+0.009 [-0.003, +0.021]
	net contrast	213	81	0.155	0.284	+0.129 [+0.040, +0.218]*
frontier	single arm	38	38	0.289	0.289	+0.000 [+0.000, +0.000]
	net contrast	38	26	0.289	0.346	+0.057 [-0.044, +0.156]
open-weight	single arm	175	157	0.126	0.134	+0.008 [-0.005, +0.020]
	net contrast	175	55	0.126	0.255	+0.129 [+0.048, +0.217]*
pool names gold	single arm	95	89	0.347	0.360	+0.012 [-0.014, +0.045]
	net contrast	95	41	0.347	0.561	+0.214 [+0.081, +0.343]*
pool does not	single arm	118	106	0.000	0.000	+0.000 [+0.000, +0.000]
	net contrast	118	40	0.000	0.000	+0.000 [+0.000, +0.000]

Table 29: **Each channel again, conditional on the arm not abstaining.** FLIP and ABSTAIN partition the same outcome space, so a necessity arm that abstains often has little room left to flip and `flip-c` can go negative with nothing behavioural behind it. The last column removes that censoring: it differences the flip rate among non-abstaining runs of each arm. **It is positive in every cell**, including both where `flip-c` is negative, and it reverses the published channel ordering — the frontier models lead it on all three deployments. Where a model's necessity arm never produces POST-HOC, its conditional flip rate is 1.000 by construction and the column is reporting the control arm; that is a further reason not to read either form as a disposition. Cells with fewer than 8 paired incidents, and models whose contrast is structurally unmeasurable, are omitted. Intervals resample incident *ids* (Market's 148 rows carry 14), the more conservative unit; $B = 20000$.

deployment	model	n_{pair}	abstain-c	flip-c	flip | ¬abstain, both arms	[95% CI]
Market	GPT-5.5	70	+0.671	+0.114	+0.948	[+0.857, +1.000]
	Gemini-Flash	100	+0.630	+0.000	+0.797	[+0.718, +0.878]
	Qwen-32B	98	+0.459	+0.061	+0.191	[+0.119, +0.345]
	Qwen-14B	91	+0.407	+0.044	+0.216	[+0.085, +0.430]
	Qwen-7B	139	+0.058	+0.245	+0.287	[+0.192, +0.408]
	Mistral-7B	148	+0.000	+0.399	+0.399	[+0.299, +0.511]
Bank	GPT-5.5	73	+0.370	+0.479	+0.969	[+0.921, +1.000]
	Gemini-Flash	98	+0.571	+0.286	+0.977	[+0.941, +1.000]
	Qwen-32B	67	+0.119	+0.194	+0.532	[+0.334, +0.707]
	Qwen-14B	27	+0.519	-0.222	+0.381	[+0.176, +0.591]
	Qwen-7B	104	+0.029	+0.279	+0.378	[+0.259, +0.499]
	Mistral-7B	136	+0.007	+0.235	+0.242	[+0.162, +0.321]
Telecom	GPT-5.5	19	+0.421	+0.158	+0.917	[+0.727, +1.000]
	Gemini-Flash	18	+0.833	+0.000	+0.938	[+0.800, +1.000]
	Qwen-32B	37	+0.297	+0.270	+0.808	[+0.643, +0.957]
	Qwen-14B	39	+0.410	-0.128	+0.236	[+0.007, +0.448]
	Qwen-7B	50	+0.040	+0.140	+0.182	[+0.019, +0.343]
	Mistral-7B	49	+0.000	+0.102	+0.102	[-0.020, +0.224]

G LIMITATIONS IN DETAIL

Every intervention here is *subtractive*, so we test whether a conclusion moves when evidence is removed, never whether it moves *toward* what remains. A model re-answering near-randomly under any information loss is indistinguishable from one whose citations are load-bearing.

Absolute levels are conditioned on the salience heuristic and the citation scaffold (Qwen-7B 0.302→0.175 unfiltered and →0.430 scaffolded), and halving the per-modality cap raises every model's contrast by 0.08–0.27 while leaving the ordering and the within-family separation intact, so we report comparisons and not levels.

Claude-Opus-5's cite-everything behaviour leaves n_{paired} of 13, 5 and 1; those cells are reported as unmeasurable (†) rather than as contrasts of zero and support no comparative claim.

Market's incident ids are not unique (148 incidents, 14 distinct ids) and the random control is seeded on that id, so all analyses pair by position; the salience-matched control uses no RNG and agrees. Nothing in this paragraph is absent from the submission build; it is disclosure, relocated. Base-step abstention makes several cells self-selected, and not only on Telecom, since Qwen-14B is probed on 31 of 136 Bank incidents, a narrower base than any Telecom cell. Every such cell is marked ‡ (Table 30).

Open models run 4-bit against closed models at the providers' precision, which confounds capability with quantization; we measured it rather than only disclosing it and it does not carry the cross-tier differences (paired contrast +0.036 [-0.094, +0.167], every arm's interval including zero; Table 31), though it remains untested for the larger open models.

Moved from the main-text Limitations for space; nothing here is omitted from the submission build. Each item is stated in the Limitations section; this appendix carries the numbers behind two of them, which do not fit the main text's page limit and are not optional reading.

Base accuracy, and what the contrast is therefore measured on. Exact matching against the gold component gives 0.000–0.070 on Market, 0.027–0.104 on Bank and 0.020–0.324 on Telecom. Checked against a lenient matcher before being quoted: the gap is at most 0.113 and the residual misses are level mismatches (`adservice-1` vs `adservice`) or class/instance pairs (`Tomcat01` vs `Tomcat`) that are arguable in either direction, so the low accuracy is a property of the answers rather than of the comparison. The 25% quoted for published on-call agents is a tool-using agent on a live system, not a single-shot answer over a fixed 16-item pool. Consequently **no model on Market or Bank reaches $n = 8$ correct-at-base paired incidents**, so the contrast cannot be shown to survive on correct answers there; the one cell that exists is Telecom Qwen-32B, $n = 12$, contrast +0.750 [+0.500, +1.000]. A wrong diagnosis that does not rest on its evidence is arguably the more dangerous case for an operator, but no claim here should be read as being about correct reasoning.

Quantization. Precision changes base-step refusal. 145 of 148 incidents reach the probe at full precision against 139 at 4-bit, which is the direction the confound was predicted to run, and the reason the comparison must be paired: on own subsets the gap reads +0.029, mixing a behavioural change with a change in which incidents were measurable at all. Even at the top of the paired interval, quantization covers at most 0.167 of the 0.484 GPT-5.5 − Qwen-7B gap.

H BASE-STEP FUNNEL, MODALITY STRATA, AND MECHANISM CHECKS

Every number here is recomputed offline from the persisted traces by `analysis/robustness.py` and `analysis/mechanism.py`; no model was re-queried.

Table 30: **Why incidents never reached the probe.** *probed* excludes base-step skips; *paired* additionally requires that a size-matched control existed, which is what cite-everything destroys. Parenthesised counts are base-step abstentions; the only other skip reason anywhere is an unparseable output (Market 2, Bank 2, Telecom 2).

	Market (n=148)			**Bank** (n=136)			**Telecom** (n=51)		
model	**probed**	**paired**	**(abst)**	**probed**	**paired**	**(abst)**	**probed**	**paired**	**(abst)**
Mistral-7B	148	148	(0)	136	136	(0)	51	49	(0)
Qwen2.5-7B	139	139	(8)	104	104	(30)	50	50	(0)
Qwen2.5-14B	91	91	(56)	31	27	(105)	39	39	(12)
Qwen2.5-32B	98	98	(50)	67	67	(69)	37	37	(14)
Gemini-Flash	100	100	(48)	110	98	(26)	19	18	(32)
GPT-5.5	70	70	(78)	75	73	(61)	19	19	(32)
Claude-Opus-5	115	**13**	(33)	113	**5**	(23)	40	**1**	(10)

The contrast is not confined to one telemetry modality. Evidence ids carry their modality, so the contrast can be stratified by what the agent actually cited. On Market the metric-only, log-only and mixed strata are all positive and exclude zero for every measurable model (GPT-5.5 log-only +0.86, mixed +0.82; Qwen-32B +0.50 and +0.55), and on Bank the metric-only and mixed strata behave the same way (Gemini-Flash +0.86 in both). One exception is worth reporting rather than smoothing: on Telecom the two flip-dominant small models go *negative* on mixed-modality citations (Mistral-7B -0.04, Qwen-7B -0.13) while staying positive on single-modality ones. Citation profiles differ sharply by deployment: Market is mixed 480 / log-only 181 / metric-only 75, Bank mixed 340 / metric-only 239, Telecom mixed 123 / trace-only 82. The strata are therefore not comparable across deployments in absolute terms.

Gold matching, exact and lenient. Both gold-using checks in Section 7 use the gold component, which enters no metric. Full per-model numbers, including the contrast recomputed on correct-at-base incidents under exact and lenient gold matching, are produced by `analysis/mechanism.py`; the lenient matcher changes base accuracy by at most 0.113 and its residual misses are level mismatches (`adservice-1` vs `adservice`) that are arguable in either direction, which is why the low accuracy is reported as real rather than as an artifact.

Table 31: **Quantization ablation: Qwen-7B on Market, 4-bit vs full precision, paired by incident on the 138 of 148 incidents both runs probed.** "own" columns are each run's own self-selected subset, the quantity a two-level report would show; the paired column is the one to quote. No interval excludes zero.

	Own subsets		Paired full precision − 4-bit	
Metric	**4-bit**	**Full precision**	**Difference**	**95% CI**
contrast	+0.302	+0.331	+0.036	[−0.094, +0.167]
flip-c	+0.245	+0.297	+0.065	[−0.065, +0.203]
abstain-c	+0.058	+0.034	−0.029	[−0.080, +0.022]
comprehensiveness	+0.626	+0.669	+0.036	[−0.051, +0.123]
random arm	+0.324	+0.338	+0.000	[−0.094, +0.094]
sufficiency	+0.568	+0.600	+0.051	[−0.058, +0.159]

I METHOD ABLATIONS

Robustness to evidence selection. Rebuilding the pools without salience filtering (the earliest k items per modality in the window) preserves both the ordering (Qwen-32B − Qwen-7B = +0.325 [0.038, 0.615], vs +0.218 with salience) and the channel profile (Qwen-32B stays abstention-dominant, 0.812 abstain vs 0.062 flip). With unfiltered pools models abstain at the base step far more often (skip 61% and 89% vs 6% and 34%), showing the probe does not measure indiscriminate confabulation. Models do respond to evidence quality. Absolute levels are not comparable across modes because the usable sample differs (Qwen-32B window: $n = 16$).

Design sensitivity (two open models only). A citation scaffold raises both models' contrast without changing their order or channel profile, and re-asking Qwen-7B's post-hoc cases after stating the cited evidence is gone changes its answer in only 25% of 52 cases. Both checks cover 2 of 7 models on one deployment, so they show the ordering *can* be insensitive to these choices, not that it always is. The frontier models, and the cite-everything behaviour in particular, are untested here.

Stated confidence does not warn about decorative citation. Every model reports a confidence next to its diagnosis, and it is the only self-assessment an operator gets. It is close to constant, with standard deviations of 0.013–0.105 and Gemini-Flash answering 0.946±0.013 almost every time. Where it does vary it points the wrong way. Across all four open models the point-biserial correlation between stated confidence and post-hoc citation is *positive* (0.072–0.227) and every binned curve rises (Figure 5; Qwen-32B goes from 0.12 post-hoc below confidence 0.8 to 0.47 above it). Only Qwen-32B's correlation reaches $p < 0.05$ individually, but the sign is the same in four of four models. For the frontier models the correlation is undefined because they produce no post-hoc cases on Market at all. The practical reading is that a high stated confidence is not evidence of grounding, and on the open models it is mildly evidence against it.

Citation scaffold (prompt sensitivity). Appending the list of valid evidence ids to the prompt raises contrast for both models (Qwen-32B 0.520→0.727; Qwen-7B 0.302→0.430). We report the two levels rather than a difference, because the scaffold and no-scaffold runs are not paired on the same eligible incidents. Yet the ordering (+0.298 [0.167, 0.427] vs +0.218 without) and the channel profile (7B flip-dominant, 32B abstention-dominant) are unchanged. **Comparative claims are robust to how citations are elicited; absolute levels are not**, and we therefore report model comparisons rather than absolute dependence scores.

Self-correction. Re-asking the post-hoc cases once, stating explicitly that the cited evidence is no longer available, Qwen-7B revises or abstains in only 25% of 52 cases. **Told outright that its evidence is gone, the model keeps its original root cause three times out of four**. Post-hoc citation is not merely a passive invariance but survives direct contradiction.

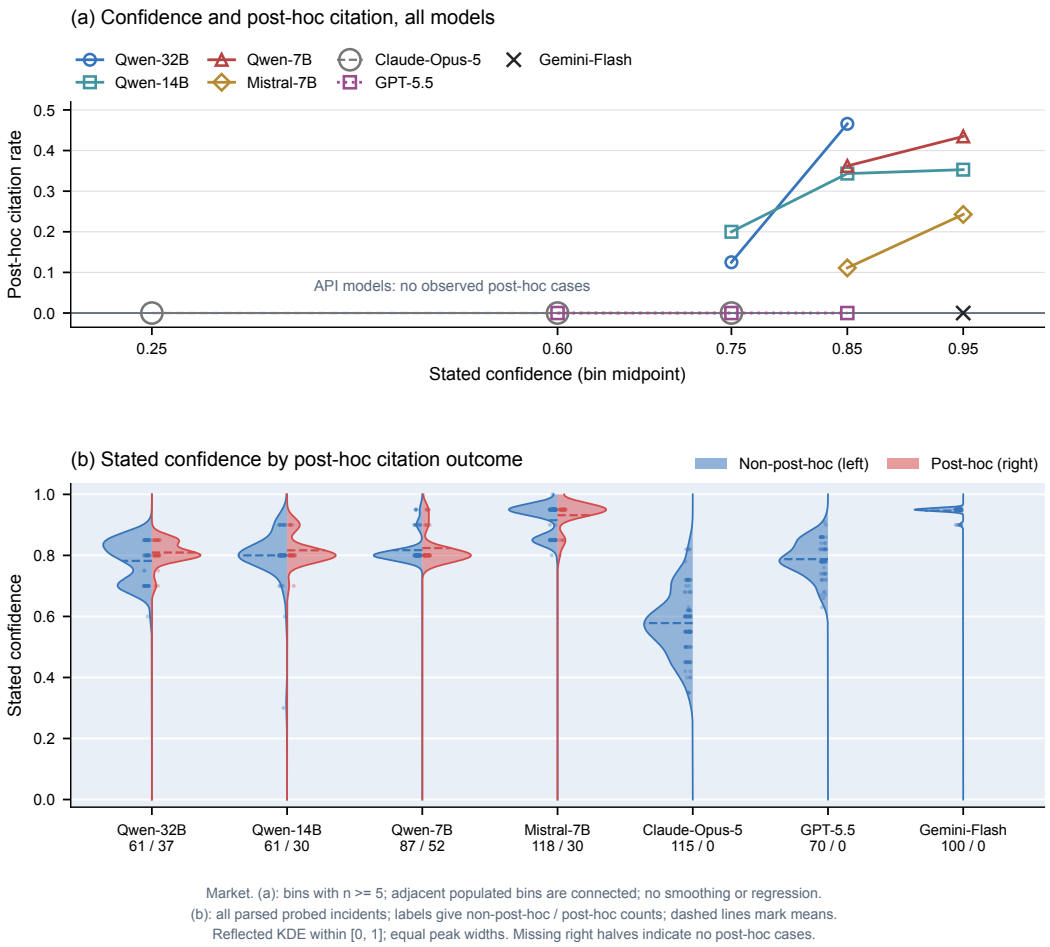

Figure 5: Stated confidence against decorative citation, Market. **(a)** Binned post-hoc rate. All four open-weight curves rise, so a higher stated confidence goes with a citation that is *more* often decorative, and the three API models sit on zero throughout. **(b)** The same confidences split by outcome, over every parsed probed incident rather than three bins; the counts under each model are non-post-hoc / post-hoc. The two distributions overlap almost completely, which is why the stated number cannot separate a grounded diagnosis from a decorative one. **The API models produce 0 post-hoc cases in 285 probed incidents**, so no correlation is defined for them and their zero is a measured rate rather than a missing measurement.

Table 32: **Every fault-type group with at least eight probed incidents, for every model and deployment – including the groups that cut against the trend described in the text.** Gold fault labels are used here for grouping only and enter no metric. Per-cell n is small throughout, so this table is descriptive. †No estimable contrast: the model left no non-cited item to form a specificity control (Section 7).

deployment	fault type	model	n_{prb}	contrast
Bank	high CPU usage	Gemini-Flash	27	0.958
		GPT-5.5	18	0.889
		Mistral-7B	33	0.333
		Qwen-32B	15	0.333
		Qwen-7B	28	0.286
		Claude-Opus-5	27	n/a†
	high disk I/O read usage	GPT-5.5	9	0.889
		Gemini-Flash	16	0.846
		Qwen-7B	15	0.400

Table 32, continued from the previous page

deployment	fault type	model	n_{prb}	contrast
		Mistral-7B	19	0.105
		Claude-Opus-5	17	n/a†
	high memory usage	Mistral-7B	10	0.400
	network latency	Gemini-Flash	18	0.938
		GPT-5.5	16	0.938
		Claude-Opus-5	22	n/a†
		Qwen-7B	24	0.375
		Mistral-7B	27	0.296
		Qwen-32B	15	0.000
	network packet loss	Gemini-Flash	29	0.760
		Qwen-32B	23	0.435
		Qwen-14B	14	0.333
		Qwen-7B	21	0.333
		Mistral-7B	32	0.219
		Claude-Opus-5	30	n/a†
		GPT-5.5	17	0.733
Market	container CPU load	Qwen-32B	8	0.500
		Mistral-7B	13	0.385
		Claude-Opus-5	11	n/a†
		Qwen-7B	12	-0.083
	container memory load	GPT-5.5	8	0.750
		Mistral-7B	13	0.538
		Qwen-7B	11	0.455
		Gemini-Flash	9	0.444
		Qwen-32B	9	0.222
		Qwen-14B	8	0.125
		Claude-Opus-5	10	n/a†
	container network latency	Mistral-7B	8	0.625
		Qwen-7B	8	0.500
	container network packet corruption	Qwen-14B	10	0.500
		Mistral-7B	13	0.462
		Qwen-32B	11	0.455
		Qwen-7B	13	0.231
		Claude-Opus-5	10	n/a†
	container network packet retransmission	Qwen-14B	9	0.556
		Qwen-32B	8	0.250
		Qwen-7B	13	0.154
		Mistral-7B	13	0.077
		Claude-Opus-5	10	n/a†
		Gemini-Flash	8	1.000
	container packet loss	Gemini-Flash	8	0.625
		Qwen-32B	8	0.375
		Qwen-7B	8	0.375
		Mistral-7B	8	0.125
	container read I/O load	Mistral-7B	17	0.647
		Qwen-32B	11	0.545
		Qwen-14B	9	0.444
		Qwen-7B	16	0.188
		Claude-Opus-5	13	n/a†
		Gemini-Flash	12	0.667
		GPT-5.5	10	0.600
	node CPU load	Qwen-7B	8	0.375
		Mistral-7B	9	0.333
	node disk read I/O consumption	Mistral-7B	9	0.444
		Qwen-7B	9	0.222
	node disk space consumption	Mistral-7B	10	0.500
		Qwen-7B	10	0.500
		Claude-Opus-5	8	n/a†
	node disk write I/O consumption	Gemini-Flash	8	0.750
		Mistral-7B	10	0.300
		Qwen-7B	10	0.300
		Claude-Opus-5	8	n/a†
	node memory consumption	Qwen-7B	8	0.250
		Mistral-7B	9	0.000
Telecom	CPU fault	Qwen-32B	11	0.364
		Mistral-7B	19	0.222
		Qwen-7B	19	0.158
		Qwen-14B	13	0.077
		Claude-Opus-5	12	n/a†
	network delay	Qwen-32B	8	0.500
		Qwen-14B	8	0.250
		Mistral-7B	13	0.000

Table 32, continued from the previous page

deployment	fault type	model	n_{prb}	contrast
		Qwen-7B	13	0.000
		Claude-Opus-5	12	n/a†

J ROUND 1 ANNOTATION, AND THE UNCORRECTED ROUND-2 CONFUSION

Round 2 without the author-rater. Holm-corrected across the three classes, the three-rater lifts clear their baselines at the Holm levels 0.0167 and 0.025 and FLIP does not. Recomputed on the two raters who are not authors, ABSTAIN falls to $+0.154$ ($p = 0.037$, which does *not* survive Holm) and POST-HOC to $+0.090$ ($p = 0.116$, not significant even uncorrected); those two raters agree at $\kappa = -0.042$. Between-model contrasts elsewhere in the paper are **uncorrected and exploratory**. We state that here once rather than per table. Table 33 gives the per-class rates and Figure 6 the confusion matrix they are the diagonal of.

Table 33: **Round-2 per-class recovery against a null that preserves each rater's own marginals.** * clears the permutation baseline. Lavender marks the existing starred results and their lifts, not item-level validation.

	Per-class recovery		Against rater baseline	
Metric class	**Rate**	**Rater's own baseline**	**Lift**	**Permutation p**
ABSTAIN	0.821	0.641	$\mathbf{+0.179}$	**0.004** *
POST-HOC	0.359	0.214	$\mathbf{+0.145}$	**0.006** *
FLIP	0.205	0.145	$+0.060$	0.15

Annotators judged whether the reasoning *textually* relies on the cited evidence. Agreement with the per-item flag was $\kappa = 0.311$ (strict); across four metric definitions the best alignment was with *abstention* ($\kappa = 0.467$) and the worst with *flip* ($\kappa = -0.149$). Eight combinations were compared post hoc; we report this round as exploratory only.

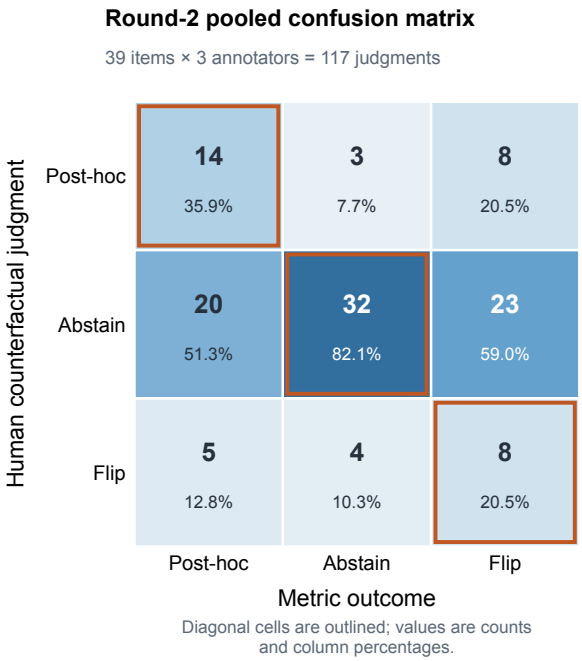

Figure 6: Round-2 confusion. The 39 items are balanced 13/13/13 by metric outcome, so each column pools 39 judgments and **the outlined diagonal is the *rate* column of Table 33**. Abstentions are recovered and flips are not, and most judgments on metric-flip items say the remaining evidence warrants abstention instead. Pooling is deliberate: with agreement at chance, any single annotator's confusion would present one noisy view as the human verdict.

K ROUND 4: THE THIRD PRE-REGISTERED INSTRUMENT

Round 3's pre-registration said *no third instrument*. We built one anyway, and disclose that rather than let the earlier document stand as though it had been honoured. Round 4 fixed those three defects in a pre-registration written before any data was collected, drew items disjoint from everything round 3's raters saw, and ran three raters through a hand-confirmed calibration set before measuring anything.

It **reached its pilot gate** and then **failed the main set it gated**. The pilot, 12 items, gave mean pairwise $\kappa = 0.643$ on the sufficiency question (Fleiss 0.639; 9 of 12 unanimous) against the same 0.40 threshold round 3 missed. The pre-registered 72-item main set then gave $\kappa = 0.369$ [0.200, 0.521], Fleiss 0.323, below the gate. Both numbers come from the same three raters on the same instrument, and the pilot's own bootstrap interval, $[0.185, 1.000]$, had contained the main-set value all along. We record the pilot result as what it was, a rule not violated at $n = 12$, rather than as a reliability finding that the main set later contradicted.

Where the disagreement lives. The three pairs are 0.503, 0.379 and 0.226, while raw agreement stays high throughout (0.625–0.778). That gap between raw agreement and κ is *consistent with* the raters differing in *marginals* rather than in which items they find hard, and we cannot rule out wording: the boundary Appendix M reports as unspecified would move exactly these marginals. One judged the surviving evidence sufficient on 33 of 72 items where the other two did so on 13 each, so two raters answering "no" most of the time accrue high chance agreement and a low κ, and the rater with the looser threshold disagrees with both. Where a rater puts the sufficiency line is a property of that person, not of our wording, and we do not think a fourth rewording reaches it. This is the reason the paper stops here rather than iterating: PREREG_ROUND4 §8 fixed that in advance, and the diagnosis we can offer for the failure is not one an instrument fixes.

Two pre-registered readings the larger sample reversed. Both are reported because the pre-registration required it, and one cuts against us.

Agreement with the metric no longer contains zero. Per-rater κ against `metric_outcome` is 0.429 [0.263, 0.591], 0.292 [0.133, 0.449] and 0.256 [0.120, 0.400] on $n = 71$, 72 and 59 mapped responses. At $n = 12$ every one of these intervals had contained zero, and the paper said so. They no longer do. We nevertheless do not claim item-level validity, for a reason that is not modesty: a validity coefficient is bounded by the reliability of the measure it is computed against, and at $\kappa = 0.369$ the human judgment is not reliable enough for the bound to mean much. What these three numbers establish is that the raters are not answering at random with respect to the metric, which is worth recording, and is less than validation.

Flip cleared the baseline we pre-registered it to miss. PREREG_ROUND4 §7 states the direction explicitly: we expected *flip* **not** to clear its own permutation baseline, because the paper reads flips as confabulation, and committed to reporting a clearance as weakening that reading. Against a null that shuffles each rater's own labels across items, all three classes clear: post-hoc 0.574 against 0.292, abstain 0.821 against 0.564, flip 0.269 against 0.144 ($p = 0.0006$). Flip is the weakest of the three by a wide margin, and the confusion structure is consistent with the original reading. 37 of 72 pooled flip judgments were called *abstain* by a rater, i.e. the evidence looked gone rather than redirected. But weakest is not absent, and absent is what we pre-registered. A reader entitled to hold us to the stated direction should read this as evidence that a flip is at least partly visible in the evidence itself.

"Cannot tell", and what one rater's willingness is worth. The option was added because we believed round 3's forced binary was part of its failure. It was used **zero** times in the pilot's 36 responses, which the earlier version of this paper reported as the option contributing nothing. On the main set it was used, but by **one rater of three**, who chose it on 20.7% of the items where support had gone, while the other two never chose it at all. Two consequences. The option is not inert: the pre-registered primary ($\kappa = 0.369$) and the yes/no secondary (0.390) now differ, where in the pilot they were numerically identical. And its effect is smaller than the between-person difference in willingness to use it, so we cannot attribute anything about the instrument to it. By metric class the option concentrates entirely on *abstain* (rate 0.028; 0.000 for post-hoc and flip), so where it absorbs uncertainty is where the evidence is gone and the question of what replaced it does not arise.

Intra-rater agreement: still unavailable, for a new reason. The pilot had no duplicates, so the pre-registered check that separates item ambiguity from rater inconsistency could not run. The main set carried 6 duplicated items at re-randomised positions, at least 20 presentation slots after their originals, and it still cannot run: every rater gave a single label to all six, so κ is 0.000, undefined and 0.000 against raw agreement of 0.833, 1.000 and 0.833. With six duplicates and marginals this skewed, no value of κ was reachable that would have distinguished the two explanations. The fourth pre-registered hypothesis, that raters would agree on the abstention class, is therefore unevaluable, and that is a limit of our design, six duplicates being too few, and not a result about the raters.

What the instrument gate caught before any rater saw an item. The generator refuses to emit annotator workspaces until every calibration answer carries an explicit author confirmation, and that refusal caught five defects in succession that would each have reached the raters: answers derived automatically from the metric's own outcome (the mechanism that produced round 3's failure); a candidate set in which five of eight items were the same diagnosis on the same deployment; three items whose evidence pool was emptied by the removal, making "would what remains suffice" trivially answerable; a set in which sufficiency and mere mention agreed on every item, which would have taught round 3's recognition rule again; and a version of the gate that accepted its own drafts as confirmations and reported them as confirmed. We record these because the fifth was caused by the gate's own design and caught by its own second check, and because the fourth was found only by working the eight intended answers through by hand. No automated check we had would have surfaced it.

L FAILURE TAXONOMY AND FAULT-TYPE STRUCTURE

We propose no taxonomy. FORGE's RF-01–16 covers the observational space, and the four labels below are simply names for outcomes already defined and measured in Sections 5 and 7, offered for cross-reference with that scheme rather than as a second contribution. Unlike FORGE's cate-

gories they carry no independent reliability check. Each is defined by an *intervention* rather than by inspecting a trace: **FF-1** evidence-invariance under ablation (the conclusion is unmoved by removing its own support), **FF-2** unfaithful self-correction (the model does not revise even when told the evidence is gone; 75% of post-hoc cases, Section 7), **FF-3** post-hoc citation, and **FF-4** citation of the entire pool. FF-1 and FF-2 have no FORGE equivalent; FF-3 and FF-4 partially intersect RF-01/07/08 (fabricated, arbitrary, insufficient evidence), which flag *fake or weak* citations rather than real-but-non-load-bearing ones.

Fault type does not induce a model-independent difficulty ordering. Grouping incidents by their gold fault type, used for *grouping only* and never in the metric, produces large per-cell variation, and it is tempting to read a story into it: faults with a sharp, localized signature (*container read I/O load*: 0.647 Mistral-7B, 0.545 Qwen-32B) appear to yield higher contrasts than diffuse ones (*network packet retransmission*: 0.077 Mistral-7B). **That reading does not survive the full table.** Across the 10 model pairs on Market with at least four shared fault groups, the mean pairwise Spearman correlation between per-fault-type contrast rankings is $\rho = +0.021$ (range -0.90 to $+0.46$), no better than chance. Fault types that are hard for one model are not hard for another, so the apparent gradient above is an artifact of which cells one quotes. Table 32 lists every group for every model so this can be checked rather than taken on trust. What remains is weaker and does not support the deployment-level analogy we previously drew: the model–fault interaction is large, but it does not decompose into a fault-difficulty axis that transfers across models. Per-cell n is 8–19 throughout, which is part of why.

M THE ROUND-4 ANNOTATION INSTRUMENT, AND FOUR CASES AS THE RATERS SAW THEM

This appendix is round 4's audit trail and carries no part of the metric. Everything in it is a human judgment about round-4 items; how the metric itself is computed is Appendix F.2, which shares this appendix's five fills and their meanings.

Section 6 reports that item-level reliability is established under none of the four instruments, and Appendix K argues the round-4 failure is a difference in where each rater puts the sufficiency line rather than a defect in our wording. Log items below read `EVERE:` rather than `SEVERE:` because the severity field arrives that way in the upstream telemetry; the items are reproduced as the raters and the models saw them. That argument is about the wording, so the wording belongs in the paper. Everything below is read off the released material by `analysis/make_instrument_appendix.py`: the questions and options come out of the rater workspace `annotation_round4/C1/main.html`, the answers out of the three `round4_main_C*.csv` files, the key out of `annotation_round4/master/main_key.jsonl`. Nothing is retyped and nothing is rephrased.

What the rater is asked. Each item shows the incident report, the agent's baseline diagnosis and reasoning, and the evidence pool with the cited items struck through. Then:

> **Q1.** If the struck-through evidence were removed, would what remains still be enough to support this diagnosis's root cause (*component*)?
> *Judge sufficiency, not whether the component is mentioned anywhere. If the remaining evidence does not let you decide, say so, that is a real answer here.*
>
> **Q2** (only when Q1 = `no`). Would what remains be enough to support any specific root cause?
> *You are not asked which one. Only whether the remaining evidence would be enough to support some specific cause at all.*

The first hint is the whole difference between round 3 and round 4: round 3 asked whether anything remaining still *pointed to* the component, which a mention satisfies, and round 4 asks whether what remains *suffices*. Answers map to outcomes by `yes` → POST-HOC, `no`/`no` → ABSTAIN, `no`/`yes` → FLIP; a `cannot` on either question maps to nothing and the item is dropped from that rater's agreement.

Where the instrument is silent, and why that matters. Three rules are stated to the rater: judge sufficiency rather than mention, answer about this diagnosis's component rather than any component, and use `cannot tell` rather than guess. **One boundary is not specified anywhere in the instrument**: what to answer when the surviving evidence supports the same component under a *different* fault mechanism. Q1 names only the component, so a rater meeting that case decides it privately. We report this as an omission in our design rather than supply a rule after the fact, because it is one candidate explanation for the marginal differences of Appendix K and we cannot distinguish it from the others.

The calibration key is not a gold standard. Before any item, raters work eight calibration items whose intended answers carry an explicit provenance flag. Those answers are the definitions being taught, so one is reproduced below in full. It was drafted by the assistant and then confirmed by an author; it is a teaching answer, not a human ground truth, and the distinction is the reason the flag exists.

Four cases. Selected by a rule fixed in the generator, not by inspection: one unanimous POST-HOC, one unanimous ABSTAIN, one that drew a `cannot tell` together with a three-way spread, and one where the human majority differs from the metric outcome; within each category, the first item in item-id order. **No majority vote is a human label** and none of these is evidence for the metric. Each case is one block: a band names a role in words and the lines beneath it carry that role's fill, so the block reads the same printed in greyscale. Only the lines that bear on the judgment are shown, the removed set and whatever survives that still names the component; the full pools are in the repository.

Round 4: the two answers, with the options exactly as the workspace shows them

Q1 — asked on every item

`yes`	yes — what remains would still be enough to support it
`no`	no — what remains would not be enough
`cannot`	I cannot tell from this evidence

Q2 — asked only when Q1 = `no`

`yes`	yes — enough to support some specific root cause
`no`	no — not enough to support any specific root cause
`cannot`	I cannot tell

Round 4 calibration item calibration-000 — a taught answer, not a human ground truth

Confirmed answer `no`, drafted by the assistant and confirmed by an author

Nothing left is about adservice on node-6: what remains is on adservice.ts:8088, cartservice-0, cartservice-1, cartservice-2. adservice.ts:8088 is the same kind of component but not adservice on node-6, so it does not count. With no surviving anomaly on adservice on node-6 itself, the remaining evidence cannot support it as the root cause.

Round 4, main-028 — unanimous post-hoc

Qwen2.5-7B-Instruct, diagnosis **adservice**, metric outcome POST-HOC; pool of 24 items, 4 cited

Removed — the 4 cited items the intervention deletes

`L#1 [frontend-2]`	severity: warning, message: failed to retrieve ads
`L#4 [frontend-0]`	severity: warning, message: failed to retrieve ads
`L#7 [frontend-1]`	severity: warning, message: failed to retrieve ads
`T#8 [adservice-2]`	hipstershop.adservice/getads dur=13 status=OK

Remaining — 10 of the 20 survivors still name the component

`L#2 [adservice-0]`	EVERE: Exception while executing runnable…
`L#3 [adservice-0]`	EVERE: Exception while executing runnable…
`L#5 [adservice-0]`	at hipstershop.AdService.access$500(AdService.java:64)

and 7 further, all naming it.

C1 `yes` → POST-HOC **C2** `yes` → POST-HOC **C3** `yes` → POST-HOC

Round 4, main-004 — unanimous abstain

gpt-5.5, diagnosis **emailservice**, metric outcome ABSTAIN; pool of 16 items, 2 cited

Removed — the 2 cited items the intervention deletes

`M#7 [node-6.emailservice-1]` container_memory_cache=7.786e+08 (median 4.58916e+07)

`M#8 [node-6.emailservice-2]` container_memory_cache=7.77103e+08 (median 4.4929e+07)

Remaining — nothing that survives names the component at all

C1 `no / no` → ABSTAIN **C2** `no / no` → ABSTAIN **C3** `no / no` → ABSTAIN

Round 4, main-015 — cannot tell with a three-way spread

Qwen2.5-14B-Instruct, diagnosis **adservice**, metric outcome ABSTAIN; pool of 24 items, 9 cited

Removed — the 9 cited items the intervention deletes

`L#1 [frontend-0]` severity: warning, message: failed to retrieve ads

`L#2 [adservice-0]` at hipstershop.AdService.access$500(AdService.java:64)

`L#3 [adservice-0]` EVERE: Exception while executing runnable…

`L#5 [frontend-2]` severity: warning, message: failed to retrieve ads

and 5 further removed.

Remaining — 3 of the 15 survivors still name the component

`L#4 [adservice-0]` EVERE: Exception while executing runnable…

`L#8 [adservice-0]` EVERE: Exception while executing runnable…

`T#1 [adservice2-0]` hipstershop.adservice/getads dur=16 status=OK

C1 `yes` → POST-HOC **C2** `no / cannot` → CANNOT TELL **C3** `no / yes` → FLIP

Round 4, main-000 — human majority differs from the metric

Mistral-7B-Instruct-v0.3, diagnosis **adservice-0**, metric outcome FLIP; pool of 24 items, 8 cited

Removed — the 8 cited items the intervention deletes

`L#1 [adservice-0]` EVERE: Exception while executing runnable…

`L#2 [adservice-0]` EVERE: Exception while executing runnable…

`L#6 [adservice-0]` EVERE: Exception while executing runnable…

`L#7 [adservice-0]` at hipstershop.AdService.access$500(AdService.java:64)

and 4 further removed.

Remaining — 3 of the 16 survivors still name the component

`M#2 [adservice2.ts:8088]` java_lang_OperatingSystem_ProcessCpuTime=7.88518e+11 (median 7.91407e+11)

`T#3 [adservice-1]` hipstershop.adservice/getads dur=13 status=OK

`T#4 [adservice-0]` hipstershop.adservice/getads dur=11 status=OK

C1 `no / yes` → FLIP **C2** `no / no` → ABSTAIN **C3** `no / no` → ABSTAIN

The third case is the boundary the instrument does not settle. Nine of twenty-four items were cited and removed, and three survivors still name `adservice`, two of them `SEVERE` exception traces on `adservice-0`. One rater read that as sufficient, one could not tell, and one judged it enough for some other cause. The fourth case shows the same thing from the other side: what survives names `adservice2.ts:8088`, a different instance of the same kind of component, which the calibration answer above says explicitly does not count, and two of three raters treated the diagnosis as unsupported where the metric recorded a FLIP.

N AN OPEN REASONING MODEL: QWQ-32B

The three frontier systems in the main results are reasoning models, but none is open-weight, so none can be re-run by others. As an open counterpart we probed QwQ-32B, same parameter count and family as Qwen2.5-32B but reasoning-trained, which lets us ask whether explicit reasoning changes *how* evidence dependence is expressed rather than merely how much of it there is.

Inference cost limited this to a 40-incident subset: with a 2048-token budget the model spends most of it on its thinking trace, and a single incident (four calls) takes about 16 minutes, i.e. roughly 37 hours for the full set. Of the 40, 20 yielded a concrete cited diagnosis and could be probed (Table 34).

Table 34: **QwQ-32B and its instruction-tuned counterpart of the same size and family.** QwQ-32B is evaluated on the 20 of 40 subset incidents that yielded a concrete cited diagnosis; the instruction-tuned row reports its Market estimate on 98 probed incidents. The flip and abstain columns are *raw necessity-arm rates*, not channel contrasts: they sum to comprehensiveness, not to the contrast beside them. The rows are descriptive rather than a matched-population comparison, and at $n = 20$ the QwQ-32B channel contrast is not significant.

model	**contrast [95% CI]**	**necessity-arm flip / abstain**	$n_{\textbf{prb}}$
Qwen2.5-32B (instruct)	0.520 [0.422, 0.624]	0.143 / 0.480	98
QwQ-32B (reasoning)	0.450 [0.227, 0.684]	0.600 / 0.350	20

The overall level of dependence is not separated between the two (the intervals overlap almost entirely, which is a null and not an equivalence result at $n = 20$), but the *raw necessity rates* look inverted: the instruction-tuned model disrupts mainly by abstaining (0.480 vs 0.143 flips), the reasoning-trained one mainly by naming a different component (0.600 vs 0.350). **That inversion does not survive the control.** Those are unadjusted rates, and subtracting the random arm leaves QwQ-32B with flip-c $= 0.150$ against a total contrast of 0.450, so *twice* as much of its measured dependence runs through abstention as through flips — the same orientation as its instruction-tuned counterpart, not the opposite one. If it held, that would place reasoning training on the channel our raters recover least reliably (Section 6), a natural target for the CoT-faithfulness literature (Shen et al., 2026; Han et al., 2026).

We report this as an observation, not a claim. With $n = 20$ the channel contrast is not significant (flip-c $= 0.150$, CI [-0.167, 0.444]) and the contrast intervals of the two models overlap; establishing the effect would need roughly 100 incidents, about 27 hours of inference, which we leave to future work.