NOT UNTIL THE EVIDENCE SAYS SO: TEACHING LLM INVESTIGATORS WHEN TO CLOSE A CASE

Anonymous authors
Paper under double-blind review

ABSTRACT

Accident, defect and outage investigations end with a decision that ordinary question answering never faces: whether the evidence gathered so far is enough to close the case. We study this decision for LLM investigators, which request evidence from a case file, revise their hypotheses, and either close the case with a conclusion grounded in what they read or leave it open and name what is missing. This judgment does not come with capability: an untrained 9B model overstates its evidence in 97% of its answers, and a frontier model that identifies the right cause in 84% of cases still overstates in 91% and closes 17 of the 41 cases whose official finding is "cause undetermined". Measuring it is also non-trivial: the source of a case largely predicts its label, and a rule that reads only the source reaches 83.0 balanced accuracy on our test cases. We therefore evaluate closure with three tests: *closure accuracy*, reported against this rule and within each source; *evidence dependence*, which removes the grounds of a conclusion and checks whether the model stops closing; and *conclusion and gap quality*, a judged checklist of what the model asserts and what it says is missing. We build Nautil, 731 audited cases from aviation, rail, maritime, chemical-safety and vehicle-defect reports and production server incidents, with teacher trajectories, an out-of-distribution test set and counterfactual evidence versions. Fine-tuning a 9B model on these trajectories makes its closures follow the evidence: removing the grounds lowers its closure rate by 26 points relative to a matched control (base model: 6; frontier model: 7), overstatement falls from 97% to 35%, and conclusions that are both correct and not overstated rise from 3% to 43% (frontier model: 9%). Reinforcement learning that rewards only the closure decision then raises balanced accuracy from 69.2 to 83.3, on par with the teacher, and within-source accuracy from 60.4 to 74.1, at some cost in evidence dependence. We release the models, the dataset and the evaluation suite.

1 INTRODUCTION

Investigating an aircraft accident, a vehicle defect or a server outage means collecting evidence piece by piece and then deciding whether it is enough to state a cause. Official investigations often decide that it is not: many reports end with "cause undetermined". An LLM investigator therefore has to make two decisions, not one: what the cause is, and whether the evidence at hand establishes it.

Current models are much better at the first decision than at the second (Figure 1). On our test cases, a frontier model (Gemini 3.8 Flash) identifies the reference cause in 84% of cases, yet 91% of its answers present as established what the evidence only suggests, and it closes 17 of the 41 cases that the official reports leave undetermined. An untrained 9B model overstates in 97% of its answers. Knowing when to close does not come with the ability to find causes; it has to be taught and measured separately.

Abstention research asks whether to answer a single question whose information is given up front (Kirichenko et al., 2025; Zhai et al., 2026), and diagnosis agents are scored on finding the cause (Xu et al., 2025; Pham et al., 2025), not on whether they should have concluded. Our cases hold tens to hundreds of evidence items that must be requested over several turns, and "the cause cannot be determined from this record" is a correct, scored outcome.

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

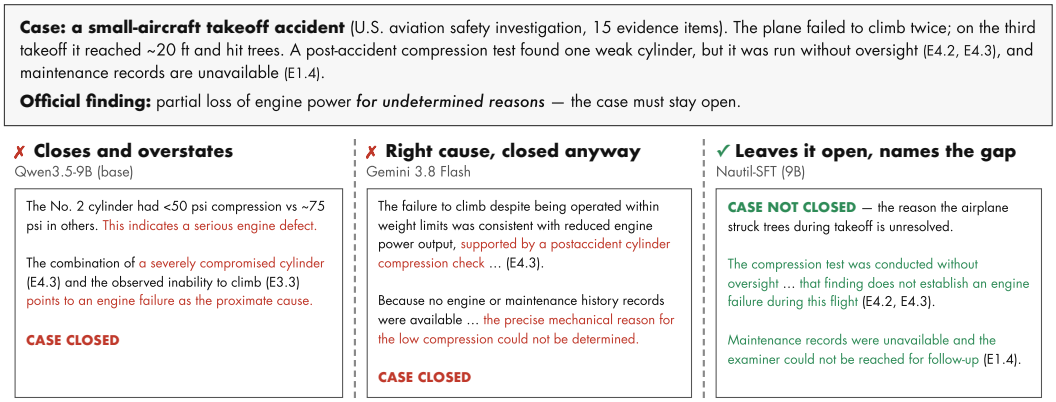

Figure 1: Three investigators on a test case whose official finding is a partial loss of engine power *for undetermined reasons*. Left: the untrained base model turns an unverified reading into an established engine failure and closes the case. Middle: a frontier model identifies the right cause and even writes that it could not be determined, yet still closes the case. Right: Nautil-SFT keeps the case open and names the missing evidence. Excerpts are the models' own words, shortened (. . .); one sample per model.

Measuring this decision is harder than it looks. In real investigation data the source of a case largely predicts its label: most NTSB and NHTSA cases in our data end undetermined, most board and server cases end with a cause. A rule that ignores the evidence and answers by source reaches 83.0 balanced accuracy on our non-host test cases. Any closure accuracy on such data must therefore be read against this rule, within each source, and together with tests that change the evidence while keeping the source fixed.

We make four contributions.

- **Task and evaluation.** We define investigative closure and evaluate it with three complementary tests: closure accuracy (reported against the source-only rule and within each source), evidence dependence (counterfactual removal of a conclusion's grounds against a matched control), and conclusion and gap quality (a judged checklist).
- **Data.** Nautil[1] contains 731 audited cases from four sources, including real production server incidents, with teacher investigation trajectories, a 141-case out-of-distribution test set and counterfactual evidence versions.
- **Models.** Nautil-SFT and Nautil-RLVR, two 9B investigators. Supervised fine-tuning on audited trajectories makes closure follow the evidence and cuts overstatement; reinforcement learning with a reward on the closure decision alone raises closure accuracy to the teacher's level.
- **Findings.** Training teaches restraint and grounding rather than cause-finding; rewarding judged conclusion quality during RL collapses the policy into never closing; and an outcome-only reward trades evidence dependence for accuracy.

2 Related Work

Abstention and unanswerable questions. Models answer questions they cannot answer (Rajpurkar et al., 2018; Yin et al., 2023; Kadavath et al., 2022), even when the context is insufficient (Joren et al., 2025; Kirichenko et al., 2025), partly because binary grading rewards guessing (Kalai et al., 2025). Abstention has been trained with SFT (Zhang et al., 2024; Yang et al., 2024) and RL (Wei et al., 2026; Damani et al., 2026; Zhai et al., 2026); RL fine-tuning can also erode it (Song et al., 2025b), and abstention rewards can collapse into refusal (Che et al., 2026). Recent work asks models to state what is missing (Zhai et al., 2026; Li et al., 2025) or to stop acting when they should

[1]Read as *not until*.

(Liu et al., 2026; Luo et al., 2026; Liu et al., 2025). These settings are single-turn or give the information up front and grade whether to answer. In ours, the decision follows multi-turn evidence gathering, the "cannot determine" outcome is set by real investigation reports, and an open case must name the missing evidence.

LLMs for incident investigation and diagnosis. Root-cause analysis benchmarks and agents for cloud incidents (Xu et al., 2025; Pham et al., 2025; Chen et al., 2025; Wang et al., 2024), auditable multi-agent diagnosis such as JustDiag (Bi et al., 2026), evidence-first investigators (Marozzo & Liò, 2026; Chen et al., 2026), and sequential medical diagnosis (Nori et al., 2025; Li et al., 2024) all score whether the cause is found. Premature closure has been measured in frontier models (Handler et al., 2026). We add closure itself as a scored, trainable decision across domains.

RLVR and agentic RL. Group-relative policy optimisation (Shao et al., 2024; Guo et al., 2025) and tool-using RL agents (Jin et al., 2025; Song et al., 2025a; Wang et al., 2025) assume a solvable task with a checkable answer. We make "not closed" one of the verifiable outcomes and show that judge-scored rubric rewards (Gunjal et al., 2026), which can be gamed (Skalse et al., 2022; Gao et al., 2023), collapse closure in this setting.

Shortcuts and faithful evaluation. Annotation artefacts and shortcuts let models score without using the input (Gururangan et al., 2018; Geirhos et al., 2020); contrast sets and counterfactual edits test whether predictions follow it (Gardner et al., 2020; Kaushik et al., 2020), though constructed insufficiency has its own artefacts (Qiu et al., 2026). LLM judges are useful but biased, including towards their own outputs (Zheng et al., 2023; Panickssery et al., 2024). These motivate our source-only baseline, the matched-control counterfactual, and cross-judge checks.

3 INVESTIGATIVE CLOSURE

Protocol. An investigator receives a case: a fixed question (what caused the incident, citing evidence for each key claim), a brief, an opening record E_1 shown in full (median 4 items), and an index that lists only the identifier and a neutral title of every other evidence item (over all 731 cases: median 115 items per case, up to 682). Titles are written so that they do not reveal conclusions. In each turn the investigator writes a working note (current hypotheses, what the new evidence supports or weakens, a hypothesis ledger, and the next discriminating check) and may call `request_evidence` for 1–12 unread items; the items arrive in the next message. An episode has at most 15 turns and ends with an answer that starts with CASE CLOSED, followed by a causal conclusion that cites the evidence, or CASE NOT CLOSED, followed by what is established, what is missing, and a provisional assessment. The label of a report case is whether the official report determined the cause.

Three tests. A single accuracy can be inflated either by the source of a case or by a conservative policy, so we evaluate closure with three tests, each closing one loophole: *closure accuracy* guards against recognising the source instead of reading the evidence; *evidence dependence* guards against closing out of habit, or stopping merely because evidence is scarce; and *conclusion and gap quality* guards against a right decision with a wrong or overstated conclusion.

Closure accuracy. We report the rate of closing among cases that should close, the rate of not closing among cases that should not, and their mean (balanced accuracy), so that never closing is not rewarded. An answer without a closure marker counts as wrong on both sides. Every result is shown next to the *source-only rule*, which gives each source its majority training label. *Within-source balanced accuracy* computes balanced accuracy separately inside each report source and averages them; the source-only rule scores exactly 50 on it, so only reading the evidence can raise it.

Evidence dependence. For cases labelled closed, we build a version without the items the conclusion rests on and a control without the same number of other items; the brief and question are unchanged, and removed items simply drop out of the index. The two versions differ only in whether the grounds are present, so the source is held fixed. We report the difference in closure rate between them (removed minus control; more negative means closure follows the evidence) and a two-class balanced accuracy per case, $\frac{1}{2}[\mathbb{1}(\text{control closed}) + \mathbb{1}(\text{removed explicitly not closed})]$.

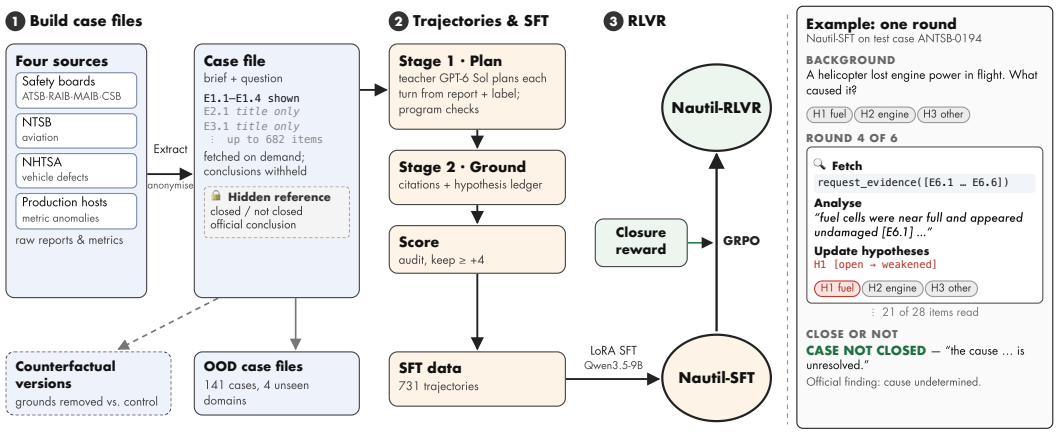

Figure 2: Overview of Nautil. (1) Case files are built from four sources: evidence items with neutral titles are shown on demand, conclusions are withheld, and a hidden reference records whether the case should close; counterfactual versions and the OOD set follow the same format. (2) The teacher plans each investigation from the report and its label, a program assembles and checks it, Stage 2 adds citations and the hypothesis ledger, and audited trajectories train Nautil-SFT. (3) RLVR rewards only the closure decision, checked by a program. Right: one round of an actual Nautil-SFT investigation on a test case whose official finding is undetermined; after its checks weaken the fuel and engine explanations, it leaves the case open.

Conclusion and gap quality. A judge answers a fixed checklist about the final answer against the reference: whether the initiating cause is right, whether the causal chain matches, whether the answer overstates (presents as established what the evidence only suggests), what share of its causal claims cite evidence that was read and supports them, and whether it addresses alternative explanations. Our summary metric is the share of answers whose cause is right *and* not overstated. For unclosed answers the judge also checks whether the gap names specific evidence, whether that evidence is truly absent from the case, and whether it matches the reference's reason for indeterminacy. The main judge is GPT-6 Sol; because it is also the teacher (§4), we repeat the checklist with Claude Sonnet 5 and Grok 4.7 (§6.3, Appendix C).

4 The Nautil Dataset

Sources. Nautil draws on four kinds of investigation (Table 1). *Boards*: reports of four transport and chemical-safety boards (ATSB, RAIB, MAIB, CSB). *NTSB*: aviation final reports 2014–2024. *NHTSA*: vehicle defect investigations, with opening and closing summaries. *Host*: anomaly windows from a year of production server monitoring at a large internet company. Report cases take their label from whether the official report determined the cause. Host incidents have no report; their label is the teacher's closure decision (below), which we use for training and report only in the appendix.

Case files. For reports, an extraction pipeline (GPT-5.6 Sol) splits the text into factual, analytical and conclusion passages, turns factual passages into atomic evidence items with neutral titles, withholds any sentence that hints at the conclusion, writes the brief, anonymises identifiers, and records a hidden reference: whether the cause was determined, the official conclusion, rejected explanations and alternatives left open. E_1 is fixed to the first four observation or measurement items of the primary source. For hosts, selected metric series are rendered as charts and described in text by a vision-language model that sees each chart with its numbers table and may quote only those numbers; a program flags any number not in the table. Service names are replaced by neutral roles, and addresses, alert IDs and dates are removed.

Teacher trajectories. The teacher (GPT-6 Sol) sees the case file, the official analysis and the label, and plans an investigation: per turn, a note, the items to request and why, then a final answer (Stage

Table 1: Nautil by source. Items: median evidence items per case, over all cases of the source.

Source	Candidates	Cases	Closed / not closed	Items	Train / Val / Test
Boards (4 agencies)	508	140	110 / 30	173	102 / 15 / 23
NTSB (aviation)	363	142	19 / 123	23.5	107 / 14 / 21
NHTSA (vehicle defects)	373	130	7 / 123	26	97 / 13 / 20
Host (servers)	1,713	319	259 / 60	145	239 / 32 / 48
Total		731	395 / 336	115	545 / 74 / 112

1). A program assembles the dialogue with tool returns copied verbatim, rejecting plans that cite unrequested items, contradict the label or leak withheld identifiers. Stage 2 adds exact citations and a ledger marking each hypothesis open, favoured or weakened, with every change citing visible evidence. For hosts the teacher also sees the closing report of a run of the JustDiag diagnostic harness (Bi et al., 2026) on the incident. Trajectories are retraced from the report's argument; their fetch order is not free exploration.

Audit. The teacher scores each trajectory from -5 to $+5$ on alignment with the source, evidence grounding, active fetching, hypothesis updates and honest closure; we keep scores ≥ 4 without hard failures and allow one revision of the final answer for $+3$. Of 1,345 generated report and host cases, 996 pass Stage 1, 818 pass Stage 2 and 518 (38.5%) pass the audit; with supplements for host not-closed cases, a salvage round and the removal of 26 cases whose brief and evidence came from different incidents, the final set has 731 cases. It is split 545 / 74 / 112 within each source, balancing labels and keeping each incident in one split.

Source and label. Table 1 shows the structure that makes closure hard to measure: boards and hosts mostly close, NTSB and NHTSA mostly do not. Answering by source alone gives 83.0 balanced accuracy on the 64 non-host test cases (23 should close, 41 should not). This is why every closure result below is paired with the source-only rule and the within-source metric, and why the headline test set is non-host: host labels are teacher judgments, and the teacher itself, rerun under the same protocol, closes only 5 of the 39 host test cases labelled closed.

Out-of-distribution test set. 141 cases from unseen domains: 83 reports (NTSB highway and pipeline, TSB Canada pipeline), 20 medical cases (Wu et al., 2025) relabelled by whether the record contains the confirming finding, 18 microservice faults (Pham et al., 2025), and 20 Tennessee Eastman process runs (Downs & Vogel, 1993; Rieth et al., 2017). Domain and label are again tied, so we report the OOD set per domain and as a secondary column.

Counterfactual versions. For the 62 test cases the teacher closes, the grounds are the items its final answer cites; versions remove 40%, 75% or 100% of them, and the control removes as many other items. A leak check flags removed items that may be restated elsewhere; 20 cases have none.

5 Training Nautil Investigators

Supervised fine-tuning. We fine-tune Qwen3.5-9B (Qwen Team, 2026) with LoRA (Hu et al., 2022) (rank 16, $\alpha = 32$, all linear layers) on the 545 training trajectories, with the native tool-call format, a 32k-token context, and loss on assistant turns only. We train two epochs (546 steps) and select the epoch-2 checkpoint before touching the test set: validation loss flattens after it, validation closure accuracy is highest, and its samples disagree most on the closure decision, which leaves room for RL. We call it Nautil-SFT.

RLVR with a closure-only reward. Starting from Nautil-SFT, we run GRPO (Shao et al., 2024) in the same case environment. The reward is deliberately simple and fully checked by a program:

$$r = D + s, \qquad D = \begin{cases} +1 & \text{closure decision matches the label,} \\ -1 & \text{it does not, or no closure marker,} \end{cases} \tag{1}$$

Table 2: Closure accuracy. Test: 64 non-host cases (23 should close, 41 should not). OOD: 141 cases in four unseen domains. Within-src.: balanced accuracy inside each report source, averaged; Boards: balanced accuracy inside boards, the only source with enough cases on both sides (18 / 5). The source-only rule scores 50 on both by construction.

	Test (non-host)					OOD
Model	Close→cl.	Open→op.	Balanced	Within-src.	Boards	Balanced
Qwen3.5-9B (base)	81.2	18.7	49.9	48.2	47.4	50.6
+ ReAct prompt	63.8	31.7	47.7	47.9	37.2	44.7
+ ledger-format prompt	92.8	17.9	55.3	56.9	52.0	56.9
Qwen2.5-3B-Instruct	10.1	68.3	39.2	40.9	33.7	31.9
Abstain-R1 (3B)	8.7	47.2	27.9	25.3	21.3	39.1
Gemini 3.8 Flash[1]	100.0	58.5	79.3	77.0	70.0	68.9
Nautil-SFT	44.9	**93.5**	69.2	60.4	70.2	61.7
Nautil-RLVR[3]	81.2	85.4	**83.3**	74.1	**78.7**	67.0
GPT-6 Sol (teacher)[1]	78.3	85.4	81.8	73.0	71.7	–
Source-only rule	78.3	87.8	83.0	50.0	50.0	–[2]

[1]One sample per case. [2]Not applicable: OOD domains are unseen in training. [3]Its checkpoint was chosen on a screening set that included 30 of these 64 test cases and 20 OOD cases (§5); Nautil-SFT was selected before any test-set use.

where $s \leq 0$ is a format penalty (no citation in the final answer, citing unread or non-existent items, turns beyond six, repeated requests, malformed calls, a missing marker). Within each group, s is centred separately among closing and non-closing rollouts, so that only D separates the two decisions. No judge is called during training. Cases are drawn uniformly from the eight source $\times$ label strata, so that no source dominates the updates; host cases, whose labels are teacher judgments, make up a quarter of the draws. Each step samples 22 cases with 8 rollouts each (temperature 1.0, up to 15 turns); we use learning rate 5×10^{-6}, asymmetric clipping (0.2 / 0.28) (Yu et al., 2025), no KL term, and 50 steps on two H100 GPUs. We select step 30 as Nautil-RLVR with a screening evaluation that included 30 of the test cases; Appendix B reports results with those cases removed. We arrived at this reward after two judge-based variants performed worse (§6.4); full settings are in Appendix B.

6 Experiments

Setup. All investigators run the same protocol, prompt and tool. We compare the untrained base model (Qwen3.5-9B), the base model prompted with ReAct (Yao et al., 2023) or with our note-and-ledger format, Abstain-R1 (Zhai et al., 2026) and its 3B base model, Gemini 3.8 Flash as a frontier model, Nautil-SFT, Nautil-RLVR, and GPT-6 Sol as a teacher reference. 9B and 3B models are sampled three times per case; API models once, except Gemini on the non-host counterfactual cases (three times). Unless noted, numbers are on the 64 non-host test cases, averaged over samples per case; intervals and host results are in Appendix D.

6.1 Closure accuracy: are closure decisions right, and not by source?

Yes: training turns a model that closes almost everything into one that separates the two kinds of case, and not only by source (Table 2). The base model closes 81% of cases that should close but leaves only 18.7% of undetermined cases open, for 49.9 balanced accuracy. Nautil-SFT leaves 93.5% of undetermined cases open, the highest of all models, and reaches 69.2; Nautil-RLVR balances both sides (81.2 / 85.4) and reaches 83.3, on par with the teacher (81.8) and Gemini (79.3). Gemini reaches its score by closing every case that should close and 41% of those that should not. Within each source, where the source-only rule is at 50, Nautil-RLVR reaches 74.1 (Nautil-SFT 60.4, base 48.2), so its gains come from separating cases of the same source. Gemini's 77.0 is helped by NHTSA and NTSB having only 2 and 3 should-close cases; on boards, the only source with enough cases on both sides, Nautil-RLVR reaches 78.7, Gemini 70.0 and Nautil-SFT 70.2 (differences within noise). Over the whole test set, Nautil-RLVR ties the source-only rule (83.3 vs.

Table 3: Evidence dependence. Closure rate (%) on the 23 non-host counterfactual cases: full case, control (non-grounding items removed), and all grounds removed. Δ = removed − control (more negative = closure follows the evidence). Balanced: two-class balanced accuracy (control closed, removed explicitly not closed). Right: Δ on all 62 cases (incl. host) and on 20 cases without flagged leaks.

	Non-host (23)					All (62)	No leak (20)
Model	Full	Control	Removed	Δ	Balanced	Δ	Δ
Qwen3.5-9B (base)	87.0	92.8	87.0	−5.8	47.8	−1.6	−8.3
Gemini 3.8 Flash	97.1	100.0	92.8	−7.2	53.6	−4.3	+6.7
Nautil-SFT	50.7	52.2	26.1	**−26.1**	**62.3**	**−16.7**	**−23.3**
Nautil-RLVR	75.4	76.8	60.9	−15.9	58.0	−9.1	−5.0
GPT-6 Sol (teacher)[1]	87.0	82.6	34.8	−47.8	71.7	−25.8	−35.0

Gemini: 3 samples per case on non-host cases and 1 on host cases, so its All and No-leak columns mix the two. [1]One sample per case; the grounds are the items it cited as teacher, which favours it.

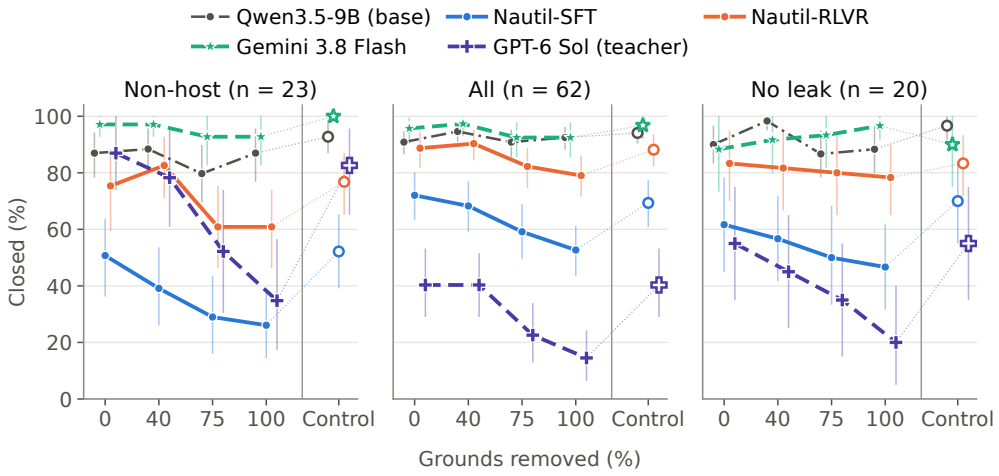

Figure 3: Closure rate as more of a conclusion's grounds are removed (0–100%), and on the control that removes as many non-grounding items (hollow markers). Nautil-SFT closes less as its grounds disappear; the base model and Gemini keep closing. Dashed: API models (Gemini: 3 samples per case on non-host cases, 1 on host; GPT-6 Sol: 1). Bars: 95% case-bootstrap intervals.

83.0), which is why the next two tests matter. The Nautil-SFT numbers are unaffected by checkpoint selection; the Nautil-RLVR checkpoint was screened on a set that included 30 of the 64 test cases (Appendix B).

6.2 Evidence dependence: do closures follow the evidence?

Yes: after training, removing the grounds of a conclusion stops the model from closing (Table 3, Figure 3). With all grounds removed, Nautil-SFT closes 26 points less than on the control, whose size is identical; the base model drops by 6 and Gemini by 7, still closing 87% and 93% of cases whose grounds are gone. The effect grows with the share removed, holds on all 62 cases including hosts (−16.7 vs. −1.6 and −4.3), and on the 20 cases without any flagged leak (−23.3 vs. −8.3 and +6.7). Because the two versions share source, brief and number of items, this test rules out both explanations that accuracy alone leaves open: recognising the source, and stopping just because evidence is scarce. In two-class terms, Nautil-SFT scores 62.3, against 47.8 for the base model and 53.6 for Gemini. As the grounds are those the teacher cited, a model might close a reduced case on other sufficient evidence; the control and the leak-free subset limit but do not exclude this. Nautil-RLVR keeps a clear but smaller effect (−15.9), a trade-off we examine in §6.4. Appendix Figure 6

Table 4: Conclusion and gap quality on the 64 non-host test cases (first sample). Left: GPT-6 Sol judge; right: Claude Sonnet 5 judge on the same answers. Overst.: presents as established what the evidence only suggests (lower is better). Right & not ov.: cause right and not overstated. Supp.: share of causal claims citing read, supporting evidence. Alt.: addresses alternative explanations. Gap: gap score of unclosed answers (0–1).

	GPT-6 Sol judge							Sonnet 5 judge	
Model	Score	Overst.↓	Cause	Right & not ov.	Supp.	Alt.	Gap	Overst.↓	Right & not ov.
Qwen3.5-9B (base)	0.30	97	54	3	26	19	0.80	81	10
Gemini 3.8 Flash	0.55	91	84	9	56	53	0.96	55	42
Nautil-SFT	**0.58**	**35**	56	**43**	**78**	38	0.38	**11**	**63**
Nautil-RLVR	0.53	38	62	36	76	17	0.39	20	50
GPT-6 Sol (teacher)	–	–	–	–	–	–	–	5	86

Table 5: Components. SFT ablations are trained from the base model with the same two-epoch schedule; RL variants start from Nautil-SFT and differ only in the reward. Δ_{CF}: removed − control on the 23 non-host counterfactual cases. Overst. and Score: GPT-6 Sol judge.

Model	Close→cl.	Open→op.	Balanced	Within-src.	Δ_{CF}	Overst.↓	Score
Nautil-SFT	44.9	93.5	69.2	60.4	−26.1	35	0.58
w/o hypothesis ledger	62.3	90.2	76.3	61.2	−8.7	39	0.58
w/o not-closed trajectories	94.2	0.0	47.1	48.8	−8.7	84	0.41
50% of the data	36.2	93.5	64.9	56.9	–	36	0.56
10% of the data	7.2	84.6	45.9	39.3	–	31	0.58
RL: checklist always counted	17.4	95.9	56.7	53.8	−15.9	19	0.61
RL: checklist if decision right	56.5	89.4	73.0	63.0	−20.3	33	0.60
RL: decision only (Nautil-RLVR)	81.2	85.4	**83.3**	**74.1**	−15.9	38	0.53

shows a case where Nautil-SFT closes on the control and stops when the recall record and crash test behind its conclusion are removed.

6.3 Conclusion and gap quality: what does training change in the answers?

Training changes restraint and grounding, not cause-finding (Table 4). Overstatement falls from 97% (base) to 35% (Nautil-SFT); the share of causal claims backed by evidence the model actually read rises from 26% to 78%; and answers that are both right about the cause and not overstated rise from 3% to 43%, against 9% for Gemini. The rate of naming the right cause barely moves (54% to 56%), and Gemini's is higher (84%): the frontier model finds causes but states them beyond what the file supports. The contrast is sharpest on officially undetermined cases: Gemini closes 17 of these 41 cases and the judge rates all 17 conclusions as overstated, while Nautil-SFT leaves 15 of the same 17 open. Nautil-SFT answers that make no causal claim at all are, without exception, CASE NOT CLOSED answers. Overstatement rates depend on the judge (Sonnet 5: 11% for Nautil-SFT, 55% for Gemini), but the ranking holds, and all three judges find trained models overstate far less (Appendix C). The limitation is on the other side of the decision: when Nautil-SFT leaves a case open, its gap is often generic (gap score 0.38 against Gemini's 0.96), naming the specific missing evidence in 32% of unclosed answers.

6.4 Which components matter, and what does RL add?

Not-closed demonstrations are the source of restraint. Without them, the model never leaves a case open (0.0%) and overstates in 84% of its answers (Table 5). *The hypothesis ledger makes closure follow the evidence.* Removing it leaves accuracy slightly higher but within noise of Nautil-SFT (76.3 vs. 69.2), while it shrinks the counterfactual effect from −26.1 to −8.7. *Data scale matters:* 10%, 50% and 100% of the trajectories give 45.9, 64.9 and 69.2 (the number of steps grows with the data).

A closure-only reward gives the largest and cleanest RL gain. It raises balanced accuracy by 14.1 points and within-source accuracy from 60.4 to 74.1, removes missing closure markers (5.2% to 0), and shortens investigations (5.5 to 4.6 turns); the OOD gain (5.3 points) is within noise. Adding the judged checklist to the reward is harmful: counted on every rollout, it drives the policy towards never closing (17.4% of should-close cases closed at step 10); counted only when the decision is right, it avoids the collapse but gains only 3.8 points. A likely reason: when every rollout in a group is wrong, the checklist still ranks the wrong answers, and this spills over into reluctance to close. The closure-only reward has a measurable cost: the counterfactual effect shrinks (-26.1 to -15.9; -23.3 to -5.0 without leaks), addressing alternatives drops (38% to 17%), and the conclusion score falls by 0.05. We read this as a trade-off between decision accuracy and evidence dependence; Nautil-SFT and Nautil-RLVR sit at its two ends.

6.5 Further results

Prompting and single-turn abstention training do not substitute. ReAct prompting (47.7) and our ledger format as a prompt (55.3, with 95% overstatement) leave the base model's closing habit intact (Table 2). Abstain-R1 (27.9) and its base model (39.2) mostly fail to finish a multi-turn investigation (44% and 32% of runs end without a closure marker): abstention learned on single questions does not carry over to evidence gathering.

Out of distribution. OOD balanced accuracy rises from 50.6 (base) to 61.7 (Nautil-SFT) and 67.0 (Nautil-RLVR), close to Gemini (68.9). Because domain and label are tied in this set, this mostly reflects the direction of the decision; within individual domains no model is clearly above chance (Appendix D).

Own fetching versus reading everything. On the 20 largest case files (18 should close), Nautil-RLVR closes 93% of the should-close cases when it fetches its own evidence and 61% when the whole file is given at once, while it matches the main cause more often when given everything, since it reads only a small share of these files (Appendix E). Appendix Figure 7 shows a failure: Nautil-SFT misreads a valve test and leaves open a case the report closes.

7 Limitations

Labels and sources. The source of a case still predicts its label, so accuracy is paired with within-source and counterfactual tests; the within-source metric rests mostly on board cases. Host labels are teacher judgments the teacher does not reproduce; host results are in the appendix. *Judge, teacher and samples.* GPT-6 Sol is both teacher and main judge, so its investigator row is only a reference; two further judges agree on direction but not on absolute rates. API models are mostly sampled once; the non-host counterfactual set has 23 cases, and its grounds are defined by the teacher's citations. The Nautil-RLVR checkpoint was screened on 30 test cases, and a 33-case pilot set consulted while choosing the SFT epoch overlaps the OOD test set. *Data.* About 4% of references have extraction errors, some anonymised words are damaged, and trajectories retrace the report rather than explore. *Open side of the decision.* When our models leave a case open they often do not name the specific missing evidence, where frontier models are far ahead; Nautil-RLVR depends less on the evidence than Nautil-SFT; and within individual OOD domains no model is clearly above chance.

8 Conclusion

Closing a case only when the evidence supports it is a judgment that does not come with capability: a frontier model finds causes more often than ours and an untrained model about as often, yet both close cases their reports leave undetermined and overstate what the evidence shows. Trained on audited investigation trajectories, a 9B model gains restraint and grounding: its closures move with the evidence, its claims cite what it read, and its conclusions are far less often overstated. A closure-only RL reward lifts accuracy to the teacher's level at some cost in evidence dependence, whereas rewarding judged conclusion quality collapses the policy into never closing. Because the source of a case can stand in for its evidence, closure accuracy should be reported with a source-only baseline, within each source, and alongside evidence dependence.

AI USE STATEMENT

In this work, we used generative AI tools for literature search, discussion of research ideas, manuscript editing, and generation of experiment code. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code, prompts and training configurations of the whole pipeline, and no data: case extraction for reports and for hosts, trajectory construction and auditing, splitting and counterfactual construction, the investigation environment, the investigator prompt and tool definition, and the SFT and RLVR code with its reward, judge prompts, configurations and unit tests. We will release the Nautil dataset and the weights of Nautil-SFT and Nautil-RLVR.

REFERENCES

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, 2026. URL https://arxiv.org/abs/2606.19407.

Xujun Che, Yuchen Yuan, Weida Zhao, and Chenyang Yu. Abstention as an Action Can Kill Both the Reward Gradient and the KL Anchor: Collapse Law and Repair for Error-Penalized Reinforcement Learning, 2026. URL https://arxiv.org/abs/2608.00301.

Haodong Chen, Ziheng Zhang, Jinghui Jiang, Qiang Su, and Qiao Xiang. Holmes: An Evidence-Grounded LLM Agent for Auditable DDoS Investigation in Cloud Networks, 2026. URL https://arxiv.org/abs/2601.14601.

Yinfang Chen, Manish Shetty, Gagan Somashekar, Minghua Ma, Yogesh Simmhan, Jonathan Mace, Chetan Bansal, Rujia Wang, and Saravan Rajmohan. AIOpsLab: A Holistic Framework to Evaluate AI Agents for Enabling Autonomous Clouds. In M. Zaharia, G. Joshi, and Y. Lin (eds.), *Proceedings of Machine Learning and Systems*, volume 7. MLSys, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/file/d1f9e4a9f109b6e8b75ed362736f22ec-Paper-Conference.pdf.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 59456–59493, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/615675cc6e94ddb1a783904fb178b5f6-Paper-Conference.pdf.

J.J. Downs and E.F. Vogel. A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3):245–255, March 1993. ISSN 0098-1354. doi: 10.1016/0098-1354(93)80018-i. URL http://dx.doi.org/10.1016/0098-1354(93)80018-I.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association

for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cfd7bee7a651ee9af525098ef67a9e45-Paper-Conference.pdf.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Rebecca Handler, Suhana Bedi, and Nigam Shah. Quantifying and Mitigating Premature Closure in Frontier LLMs, 2026. URL https://arxiv.org/abs/2605.15000.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=OkHC30LLpO.

Belinda Z. Li, Been Kim, and Zi Wang. QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=gpwA9aZLTZ.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888. Curran Associates, Inc., 2024. doi: 10.52202/079017-0908. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/32b80425554e081204e5988ab1c97e9a-Paper-Conference.pdf.

Grace Liu, Yuxiao Qu, Jeff Schneider, Aarti Singh, and Aviral Kumar. CaRT: Teaching LLM Agents to Know When They Know Enough, 2025. URL https://arxiv.org/abs/2510.08517.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM Agents Know When Not to Act?, 2026. URL https://arxiv.org/abs/2607.10059.

Han Luo, Bingbing Wen, and Lucy Lu Wang. Agentic abstention: Do agents know when to stop instead of act?, 2026. URL https://arxiv.org/abs/2606.28733.

Fabrizio Marozzo and Pietro Liò. LLM-as-an-Investigator: Evidence-First Reasoning for Robust Interactive Problem Diagnosis, 2026. URL https://arxiv.org/abs/2606.13220.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

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 *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 777–780. ACM, May 2025. doi: 10.1145/3701716.3715290. URL http://dx.doi.org/10.1145/3701716.3715290.

Jingxi Qiu, Zeyu Han, and Cheng Huang. Evidence Absence Is Not Evidence Insufficiency: Diagnosing NEI Construction Artifacts in Fact Verification, 2026. URL https://arxiv.org/abs/2605.26663.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Cory A. Rieth, Ben D. Amsel, Randy Tran, and Maia B. Cook. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation, 2017. URL https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/6C3JR1.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL https://arxiv.org/abs/2402.03300.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471. Curran Associates, Inc., 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d719fee332caa23d5038b8a90e81796-Paper-Conference.pdf.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning, 2025a. URL https://arxiv.org/abs/2503.05592.

Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2105–2120, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.112. URL https://aclanthology.org/2025.findings-emnlp.112/.

Zefan Wang, Zichuan Liu, Yingying Zhang, Aoxiao Zhong, Jihong Wang, Fengbin Yin, Lunting Fan, Lingfei Wu, and Qingsong Wen. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 4966–4974. ACM, October 2024. doi: 10.1145/3627673.3680016. URL http://dx.doi.org/10.1145/3627673.3680016.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding Self-Evolution

in LLM Agents via Multi-Turn Reinforcement Learning, 2025. URL https://arxiv.org/abs/2504.20073.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TnJMJm3Fbi.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports, 2025. URL https://arxiv.org/abs/2505.11733.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 84562–84590, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d29b8d53678015079e1d245c023e49d2-Paper-Conference.pdf.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=67K3Xlvw8L.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Haotian Zhai, Jingcheng Liang, and Dongyeop Kang. Abstain-R1: Calibrated Abstention and Post-Refusal Clarification via Verifiable RL. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19674–19695, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.985. URL https://aclanthology.org/2026.findings-acl.985/.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL https://aclanthology.org/2024.naacl-long.394/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

A Dataset construction

Sources and licences. Candidate pools: boards 508 reports (ATSB 86, RAIB 158, MAIB 170, CSB 94), NTSB 363 aviation final reports (2014–2024), NHTSA 373 defect investigations, and 1,713 host anomaly windows. ATSB reports are CC BY 4.0, RAIB and MAIB reports are under the Open Government Licence v3, and CSB, NTSB and NHTSA documents are US government works. Host windows were flagged by a rolling median/MAD detector, so they are anomalies, not verified root-cause labels. Medical case reports were not used for training because their licences are mixed; they appear only in the OOD test set.

Report extraction. The extractor (GPT-5.6 Sol) runs seven steps: (1) label each passage as factual, analytical or conclusion; (2) split factual passages into atomic evidence items with neutral titles, moving any sentence that hints at a conclusion to a withheld list; (3) build the source table; (4) write the brief; (5) anonymise; (6) write the hidden reference (determined or not, closure object, official conclusion verbatim, rejected explanations, alternatives left open); (7) review. E_1 is regrouped to the first four observation or measurement items of the primary source, so that the opening record does not depend on document layout.

Host textualisation. For each window we select metrics, render one chart per six metrics (baseline median, spread band, largest departure), describe each chart with a vision-language model that also receives the numbers table, and assemble the case with the alert record as E_1. A program flags any number in a description that is not in the table.

Anonymisation. For reports, a model lists identifiers (dates, locations, operators, registrations, report numbers) and a program replaces them in two passes, followed by a forced sweep. The replacement matched substrings rather than whole words, which damaged some ordinary words (e.g. “thunder” became “thundethe”) in at least 63 cases. We did not alter generated trajectories; items never fetched by the teacher were re-anonymised with whole-word matching (942 items). For hosts, service names become neutral roles; IP addresses, alert IDs, links and calendar dates are removed.

Host labels. The JustDiag diagnostic harness (Bi et al., 2026) (with GLM-5.3-Flash) was run on every host window; runs were scored by GPT-5.6 Sol and those scoring $\geq +1$ entered the pool. When writing a host trajectory, the teacher sees this run’s closing report as a reference, so the closure marker of a host trajectory is the teacher’s judgment, not a ground truth. Not-closed host examples come from 79 unfinished diagnosis runs, after removing hosts that also appear in the test set. Rerun under our protocol, the teacher closes only 5 of the 39 host test cases labelled closed, so host labels are not reproducible and host results appear only in Table 13.

Trajectories and audit. Stage 1 asks the teacher (GPT-6 Sol) for a plan of 3–12 rounds (4–16 for hosts), each with a note, 1–12 new evidence IDs and a reason, and a final answer. A program assembles the dialogue, copying tool returns verbatim, and rejects plans that cite items before they are fetched, use undeclared hypotheses, cite unread items in the final answer, contradict the determined/undetermined reference, or contain withheld identifiers. Stage 2 keeps the prose and fetches, inserts exact-anchor citations, and adds a ledger with states open, favoured and weakened. The audit returns a score from -5 to $+5$, an action (keep, revise, skip), a hard-failure flag and five checks (source alignment, evidence grounding, active fetching, hypothesis trajectory, honest closure). Funnel over 1,345 non-medical cases: 996 pass Stage 1, 818 pass Stage 2, 458 are kept directly and 60

after one revision of the final answer (518). Adding 21 pilot cases gives 539; we then added 113 supplements (50 host not-closed, 7 oversized board cases, 56 cases wrongly blocked by the identifier check) and 105 failed report cases whose plans were rewritten by a second LLM and then passed the same program checks and audit, and removed 26 board cases whose brief and evidence came from different incidents, for 731.

Splits. The split (seed 20260925) is balanced by source subtype and label and keeps each report incident in one split. Host cases are split by case: 21 of 264 host machines have cases in more than one split (48 cases).

OOD test set. 83 reports (NTSB highway 39, NTSB pipeline 25, TSB Canada pipeline 19, mostly determined), 20 medical cases relabelled by whether the record contains the confirming finding (8 determined, 12 not), 18 microservice faults (one per system and fault type), and 20 Tennessee Eastman runs relabelled by whether the data can identify the fault; 109 should close and 32 should not. Domain predicts label: answering each domain's majority label, read from the test answers, gives 90.3 balanced accuracy, so we report domains separately (Table 14).

Counterfactual versions. The grounds of a case are the evidence items the teacher's final answer cites as support for its conclusion. Versions remove 40%, 75% or 100% of the grounds; the control removes as many items never cited by the teacher. E_1 is never removed, identifiers are not renumbered, and the brief is regenerated with the removed items dropped from the index. A leak check flags removed items whose content is likely restated elsewhere; no edits are made, and the 20 cases without a flagged item in either the 100% or the control version form the leakage-clean subset.

Reference audit. Of 247 references checked, about 10 (4%) contain an extraction error, such as a conclusion field holding a safety notice. We kept them unchanged, since every stage of the pipeline used the same references.

B TRAINING DETAILS

SFT. Qwen3.5-9B, LoRA rank 16, $\alpha = 32$, dropout 0, all linear layers; AdamW, peak learning rate 5×10^{-5} in epoch 1 and 2×10^{-5} in epoch 2 (cosine, 8 warm-up steps), gradient clipping 1.0; batch size 1 per GPU on two H100 GPUs, no packing; 32,768-token sequences in bf16; 273 steps per epoch. The training split has 6.0M tokens, 1.4M of them supervised. Validation cross-entropy after epochs 1–4 is 0.554, 0.537, 0.532 and 0.532; we chose epoch 2 before any test-set use for the reasons in §5. We also computed token loss on a 33-case pilot set of teacher trajectories from the OOD domains (lowest at epoch 3, 0.548 vs. 0.549 at epoch 2). All 33 cases belong to the OOD test set, so we do not count this as a selection criterion and report the overlap here. Inference uses vLLM with the unmerged adapter; we verified that the adapter is active by comparing token losses with the training framework.

RLVR. Synchronous GRPO: both GPUs generate with vLLM, which then sleeps while two-GPU data-parallel training runs, and the new adapter is hot-swapped with a loss check. Each step draws 22 cases (8 rollouts each, at least 18 usable groups) and makes 4 AdamW updates; advantages are standardised within groups and groups with identical rewards are dropped; the loss is averaged over sampled assistant tokens. Rollouts use temperature 1.0 in non-thinking mode, 15 turns, 4,096 tokens per turn and 12,000 per case. The training pool has 544 cases (one training case lacked a complete evidence store). Format penalties: no citation -0.25; citing unread or non-existent items -0.5; each turn beyond six -0.1 (capped at -1); each repeated item beyond ten -0.1 (capped at -1); each malformed call -0.2 (capped at -1); missing marker -2.

Reward variants. The two judged variants add $(2q_{\text{conc}} - 1)$ for closed answers and $\frac{1}{2}(2q_{\text{conc}} - 1) + \frac{1}{2}(2q_{\text{gap}} - 1)$ for unclosed ones, where q_{conc} and q_{gap} come from the checklist of §3 answered by GPT-6 Luna during training; "always counted" adds this to every rollout, "only if decision right" only when $D = +1$. Other runs not in the main tables: a first design that also rewarded evidence coverage and judged hypothesis updates (stopped at step 17: validation accuracy unchanged, and the hypothesis judge rewarded not updating); the always-counted variant with source-weighted

sampling (stopped at step 15: closure drifted along source priors); and a run on non-host cases plus evidence-removed variants (stopped at step 28; only screened).

Checkpoint selection. Checkpoints at steps 10–50 were screened on 80 cases $\times$ 3 samples (30 validation, 30 test, 20 OOD) and a 20-case counterfactual set; step 30 was chosen for accuracy with a counterfactual effect on par with Nautil-SFT on that set. Table 6 removes the screening cases. Because 9 of the 10 non-host test cases the source-only rule gets wrong were among them, the rule scores 98.3 on the rest; these numbers are therefore not evidence of robustness.

Table 6: Test non-host cases not used in checkpoint screening (42 cases: 12 should close, 30 should not). Balanced accuracy.

Model	Balanced
Nautil-SFT	68.6
RL: checklist always counted (step 10)	57.2
RL: checklist if decision right (step 30)	72.5
Nautil-RLVR	86.4
GPT-6 Sol (teacher, 1 sample)	85.0
Source-only rule	98.3

C Evaluation protocol and judges

Protocol. Local models run in vLLM (Qwen3.5-9B models with the non-thinking sampling recommended by the model card), 3 samples per case, 15 turns, 4,096 tokens per turn, 12,000 generated tokens per case and a 32k context. API models run through OpenRouter with provider-default sampling and reasoning, reasoning content replayed across turns, 64,000 output tokens per turn and no per-case cap, so that reasoning is not truncated. A run without CASE CLOSED or CASE NOT CLOSED counts as wrong on both sides. Confidence intervals are 95% case bootstraps (5,000 draws) stratified by label (and by source $\times$ label for within-source accuracy), shared across conditions so that differences are paired.

Judges. GPT-6 Sol judges the first sample of every condition, seeing the final answer, the titles of read and unread items, the text of cited items and the reference (hidden from the investigator), but not the condition name. Hedged statements ("X was not established") are not counted as claims. When an answer makes no causal claim, the supported-claims item does not apply and the conclusion score averages the other four items. Because Sol is also the teacher, its own investigations are judged by Claude Sonnet 5 (Table 11), and 50 stratified requests were re-judged by Sonnet 5 and Grok 4.7 (Table 12): conclusion scores correlate at 0.71–0.82, agreement on the cause item is 88% for every pair, and although absolute overstatement rates differ across judges (70%, 28% and 72% overall), all three judges find trained models overstate far less than untrained ones.

D Additional results

Tables 7–14 repeat the main results with confidence intervals. On boards, paired differences in balanced accuracy are +8.7 [−12.2, +31.1] for Nautil-RLVR minus Gemini and +8.5 [−1.3, +18.3] for Nautil-RLVR minus Nautil-SFT. The tables also give counterfactual results for all conditions and versions (Tables 8, 9), conclusion quality with all columns (Table 10), host and all-112 results (Table 13), and OOD per domain (Table 14).

E Own fetching and long case files

Nautil-SFT fetches about two thirds of the items in cases with up to 50 items, but only 30–40 items (7–23%) in cases with 100 to over 1,000 items (Figure 4). Table 15 compares own fetching with giving the whole file in the first message: on the 20 largest files the untrained model closes every case when given everything (an always-close pattern), while Nautil-RLVR closes more of the

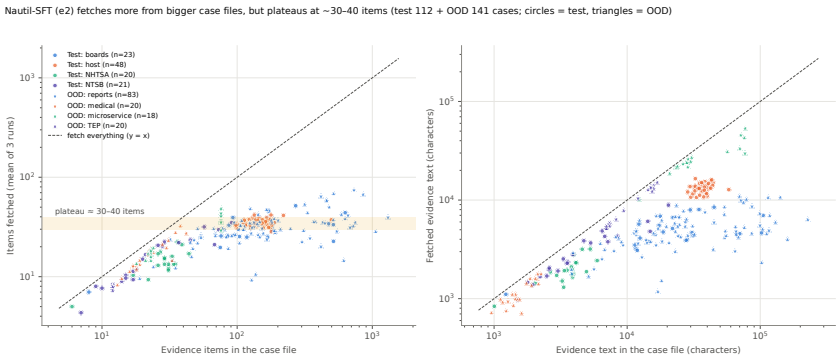

Figure 4: Evidence fetched by Nautil-SFT against case size (test and OOD, mean of 3 samples). Fetching grows with the case but levels off at 30–40 items beyond about 100 items per case.

should-close cases with its own fetching (93% vs. 61%); the main cause is matched more often when everything is given, because own fetching reads only a small share of the largest files. The v2 comparison on 16 medium-long cases does not separate the two modes.

Table 7: Main results. Test set, non-host cases (23 should close / 41 should not), and OOD (109 / 32). Close→C: share of should-close cases the model closed; Not→N: share of should-not-close cases it left open; Bal.: their mean. Within-src: balanced accuracy computed inside each source (boards, NHTSA, NTSB) and averaged, which is exactly 50 for any rule that only looks at the source. No-mark / Fetch-err / Turns refer to the test non-host runs. A run with no closure marker counts as wrong on both sides. 9B models: mean of 3 samples per case; [1]one sample per case (API models, provider-default sampling, no reasoning cap).

	Test, non-host				OOD		Runs (test non-host)		
Model	Close→C	Not→N	Bal.	Within-src	Bal.	Within-dom	No-mark	Fetch-err	Turns
Qwen3.5-9B (base)	81.2	18.7	49.9 [43.7, 56.1]	48.2 [42.3, 54.4]	50.6 [45.6, 56.3]	46.7	9.4	23.4	5.4
+ ReAct prompt	63.8	31.7	47.7 [39.0, 56.3]	47.9 [39.5, 56.0]	44.7 [38.2, 51.7]	45.9	19.8	20.3	8.4
+ ledger prompt	92.8	17.9	55.3 [49.8, 61.2]	56.9 [52.2, 62.2]	56.9 [50.0, 64.3]	56.8	3.6	10.4	5.5
Abstain-R1 (3B)	8.7	47.2	27.9 [22.6, 33.7]	25.3 [20.0, 30.8]	39.1 [34.3, 43.8]	37.4	43.8	15.6	3.7
Qwen2.5-3B-Instruct	10.1	68.3	39.2 [33.4, 45.4]	40.9 [32.9, 49.6]	31.9 [25.7, 38.1]	27.7	32.3	12.5	2.4
Gemini 3.8 Flash[1]	100.0	58.5	79.3 [72.0, 86.6]	77.0 [68.5, 86.1]	68.9 [59.9, 77.9]	59.1	0.0	0.0	4.4
Nautil-SFT	44.9	93.5	69.2 [62.4, 76.1]	60.4 [51.7, 68.8]	61.7 [56.4, 66.6]	52.1	5.2	9.4	5.5
Nautil-RLVR	81.2	85.4	83.3 [74.8, 90.5]	74.1 [67.3, 80.4]	67.0 [60.0, 73.7]	59.3	0.0	0.5	4.6
GPT-6 Sol (teacher ref.)[1]	78.3	85.4	81.8 [71.6, 91.5]	73.0 [56.3, 89.3]	n/a	n/a	0.0	21.9	5.6
Source-only rule	78.3	87.8	83.0 [73.1, 92.0]	50.0 [50.0, 50.0]	90.3 [83.9, 95.4][†]	50.0[†]	–	–	–

GPT-6 Sol generated the SFT training trajectories and its original verdicts define the non-host test labels; its row is a teacher reference, not a competitor, and it was not run on OOD. Source-only rule: boards → close, NHTSA/NTSB → not close (training-set majority label of each source); it reads no evidence. [†]On OOD the rule row uses each domain's majority label among the test answers: an oracle upper bound, not a usable baseline.

F Case studies

G Release

We will release the Nautil-SFT and Nautil-RLVR adapters, the Nautil cases (including the anonymised textual host cases) with their counterfactual versions and OOD set, the teacher trajectories, the evaluation code and judge prompts, and all evaluation trajectories. Licences follow the sources listed in Appendix A.

Table 8: Counterfactual grounding test (V2). Closure rate (%) on the intact file, on a control that removes as many non-grounding items as the 100% version removes, and after removing 100% of the evidence the teacher's conclusion rests on. Removed−control: closure-rate difference (more negative = the verdict follows the grounds). CF-Bal.: balanced accuracy over the two versions, per case (closed on control + explicitly not closed on removed)/2; a run without a closure marker counts as wrong on both sides, as in Table 7. vs SFT: paired difference against Nautil-SFT. Leakage-clean: neither the 100% nor the control version removes an item flagged as restated elsewhere.

Model	Intact	Control	−100%	Removed−control	vs SFT	CF-Bal.	vs SFT
Non-host cases (23; headline)							
Qwen3.5-9B (base)	87.0	92.8	87.0	−5.8 [−15.9, +4.3]	+20.3 [0.0, +39.1]	47.8 [44.9, 50.7]	−14.5 [−23.9, −5.1]
Gemini 3.8 Flash‡	97.1	100.0	92.8	−7.2 [−17.4, 0.0]	+18.8 [−1.4, +37.7]	53.6 [50.0, 58.7]	−8.7 [−18.8, +2.9]
Nautil-SFT	50.7	52.2	26.1	−26.1 [−43.5, −8.7]	–	62.3 [53.6, 71.0]	–
Nautil-RLVR	75.4	76.8	60.9	−15.9 [−30.4, −1.4]	+10.1 [−10.1, +29.0]	58.0 [50.7, 65.2]	−4.3 [−14.5, +5.8]
GPT-6 Sol (teacher ref.)[1]	87.0	82.6	34.8	−47.8 [−69.6, −26.1]	−21.7 [−44.9, +2.9]	71.7 [60.9, 82.6]	+9.4 [−2.2, +21.0]
All teacher-closed test cases (62, incl. 39 host)							
Qwen3.5-9B (base)	90.9	94.1	92.5	−1.6 [−6.5, +3.2]	+15.1 [+5.4, +25.3]	48.1 [46.2, 49.7]	−7.8 [−12.9, −3.0]
Gemini 3.8 Flash‡	95.7	96.8	92.5	−4.3 [−12.4, +3.2]	+12.4 [−0.5, +24.2]	52.2 [48.4, 56.2]	−3.8 [−9.9, +3.0]
Nautil-SFT	72.0	69.4	52.7	−16.7 [−25.8, −7.5]	–	55.9 [51.1, 60.8]	–
Nautil-RLVR	88.7	88.2	79.0	−9.1 [−16.1, −2.2]	+7.5 [−2.7, +17.2]	54.6 [51.1, 58.1]	−1.3 [−6.7, +4.0]
GPT-6 Sol (teacher ref.)[1]	40.3	40.3	14.5	−25.8 [−37.1, −16.1]	−9.1 [−22.0, +3.2]	62.1 [57.3, 67.7]	+6.2 [0.0, +12.6]
Leakage-clean subset (20)							
Qwen3.5-9B (base)	90.0	96.7	88.3	−8.3 [−16.7, −1.7]	+15.0 [−5.0, +33.3]	49.2 [47.5, 50.0]	−12.5 [−20.8, −3.3]
Gemini 3.8 Flash‡	88.3	90.0	96.7	+6.7 [−5.0, +23.3]	+30.0 [+10.0, +46.7]	46.7 [38.3, 52.5]	−15.0 [−23.4, −5.0]
Nautil-SFT	61.7	70.0	46.7	−23.3 [−40.0, −5.0]	–	61.7 [52.5, 70.0]	–
Nautil-RLVR	83.3	83.3	78.3	−5.0 [−16.7, +6.7]	+18.3 [−0.0, +36.7]	52.5 [46.7, 58.3]	−9.2 [−18.3, +0.0]
GPT-6 Sol (teacher ref.)[1]	55.0	55.0	20.0	−35.0 [−55.0, −15.0]	−11.7 [−33.3, +10.0]	67.5 [57.5, 77.5]	+5.8 [−5.0, +16.7]
Leakage-clean, non-host (10)							
Qwen3.5-9B (base)	86.7	96.7	86.7	−10.0 [−23.3, 0.0]	+16.7 [−16.7, +46.7]	48.3 [45.0, 50.0]	−15.0 [−30.0, +1.7]
Gemini 3.8 Flash‡	96.7	100.0	93.3	−6.7 [−16.7, 0.0]	+20.0 [−10.0, +46.7]	53.3 [50.0, 58.3]	−10.0 [−23.4, +5.0]
Nautil-SFT	46.7	56.7	30.0	−26.7 [−53.3, +3.3]	–	63.3 [48.3, 76.7]	–
Nautil-RLVR	73.3	80.0	70.0	−10.0 [−30.0, +10.0]	+16.7 [−13.3, +43.3]	55.0 [45.0, 65.0]	−8.3 [−21.7, +6.7]
GPT-6 Sol (teacher ref.)[1]	100.0	90.0	40.0	−50.0 [−80.0, −20.0]	−23.3 [−50.0, +6.7]	75.0 [60.0, 90.0]	+11.7 [−3.3, +25.0]

How to read: Intact and Control should be high, −100% should be low; the more negative Removed−control, the more closure follows the grounds. [1]One sample per case and version. ‡Three samples per case and version on the 23 non-host cases and one on the 39 host cases (per-case mean, then mean over cases), so the all-62 and leakage-clean-20 blocks mix the two; the one-sample Gemini row is in Table 9. GPT-6 Sol is the teacher: the removed grounds are the items its own teacher answer cited, and it closes only 40% of intact files (13% of host files), which limits the room to drop.

Table 9: Appendix. Counterfactual closure rates for all versions on the 23 non-host cases, and removed−control on the 23 non-host / all 62 / 20 leakage-clean cases, for all conditions. The two 3B models close only 9–14% of intact files and often produce no marker, so their rows carry little information.

	Closure rate, non-host 23					Removed−control		
Model	Intact	−40%	−75%	−100%	Control	Non-host 23	All 62	Clean 20
Qwen3.5-9B (base)	87.0	88.4	79.7	87.0	92.8	−5.8 [−15.9, +4.3]	−1.6 [−6.5, +3.2]	−8.3 [−16.7, −1.7]
Gemini 3.8 Flash‡	97.1	97.1	92.8	92.8	100.0	−7.2 [−17.4, 0.0]	−4.3 [−12.4, +3.2]	+6.7 [−5.0, +23.3]
Nautil-SFT	50.7	39.1	29.0	26.1	52.2	−26.1 [−43.5, −8.7]	−16.7 [−25.8, −7.5]	−23.3 [−40.0, −5.0]
Nautil-RLVR	75.4	82.6	60.9	60.9	76.8	−15.9 [−30.4, −1.4]	−9.1 [−16.1, −2.2]	−5.0 [−16.7, +6.7]
GPT-6 Sol (teacher ref.)[1]	87.0	78.3	52.2	34.8	82.6	−47.8 [−69.6, −26.1]	−25.8 [−37.1, −16.1]	−35.0 [−55.0, −15.0]
No hypothesis ledger	63.8	55.1	50.7	53.6	62.3	−8.7 [−21.7, +4.3]	−6.5 [−12.9, 0.0]	−8.3 [−21.7, +5.0]
No not-closed examples	87.0	87.0	92.8	85.5	94.2	−8.7 [−17.4, −1.4]	−2.2 [−6.5, +1.6]	−5.0 [−11.7, +1.7]
Gated checklist (s30)	55.1	44.9	36.2	37.7	58.0	−20.3 [−30.4, −8.7]	−8.6 [−15.6, −1.1]	−11.7 [−25.0, +1.7]
Full checklist (s10)	17.4	21.7	11.6	8.7	24.6	−15.9 [−24.6, −7.2]	−15.6 [−23.1, −8.1]	−20.0 [−35.0, −6.7]
Abstain-R1 (3B)	8.7	13.0	13.0	11.6	14.5	−2.9 [−17.4, +11.6]	−4.8 [−12.9, +3.8]	−11.7 [−26.7, +5.0]
Qwen2.5-3B-Instruct	8.7	5.8	5.8	5.8	4.3	+1.4 [−5.8, +8.7]	0.0 [−5.9, +5.9]	+5.0 [−6.7, +15.0]
Gemini 3.8 Flash[1]	100.0	100.0	91.3	91.3	100.0	−8.7 [−21.7, 0.0]	−4.8 [−12.9, +3.2]	+5.0 [−10.0, +25.0]

[1]One sample per case and version. ‡Three samples per case and version on the 23 non-host cases, one on the 39 host cases (the row used in Table 8); the last row is the same model with sample 0 only, for reference. Gemini's counterfactual V1 run (different data build) gave −11.3 on all 62; on the 49 cases whose inputs are byte-identical in V1 and V2 it gave −14.3 vs. −4.1: single-sample variance, not the data build.

Table 10: Conclusion content on the test non-host cases, judged by GPT-6 Sol (sample 0 of each condition). Concl.: mean of the checklist items (cause, chain with partial = 0.5, not overstated, alternatives addressed, share of supported claims); when an answer states no causal claim, the supported-claims item is not applicable and the score averages the other four. Overst.: % of answers stating as established what the evidence only suggests (lower is better). Cause: initiating cause matches the reference. Cause∧¬Overst.: both at once. Supp.: supported share of causal claims (answers with $\geq$1 claim). Alt.: addresses the main competing explanation. 0-claim: % answers with no causal claim. #NC: number of not-closed answers; Gap, G1 (names specific missing evidence), G2 (it is really absent from the file), G3 (matches why the reference is open) are computed on those. Answers without a closure marker were not judged.

Model	Scored	Concl.	Overst.	Cause	Cause∧¬Overst.	Chain	Supp.	Alt.	0-claim	#NC	Gap	G1	G2	G3
Qwen3.5-9B (base)	59/64	0.30 [0.26, 0.34]	97 [91, 100]	54 [42, 67]	3 [0, 9]	0.48	26	19	3	10	0.80 [0.65, 0.93]	100	80	60
+ ReAct prompt	54/64	0.35 [0.29, 0.42]	94 [88, 100]	56 [42, 69]	4 [0, 9]	0.56	29	26	4	12	0.70 [0.56, 0.84]	92	64	58
+ ledger prompt	62/64	0.28 [0.23, 0.32]	95 [89, 100]	55 [42, 67]	3 [0, 8]	0.42	21	16	2	9	0.66 [0.49, 0.82]	100	56	44
Abstain-R1 (3B)	38/64	0.16 [0.11, 0.21]	95 [87, 100]	29 [15, 43]	3 [0, 8]	0.33	10	3	11	31	0.33 [0.26, 0.42]	84	8	16
Qwen2.5-3B-Instruct	43/64	0.28 [0.20, 0.36]	77 [64, 89]	40 [26, 54]	14 [5, 25]	0.38	19	16	7	40	0.39 [0.32, 0.46]	88	14	22
Gemini 3.8 Flash[1]	64/64	0.55 [0.50, 0.60]	91 [83, 97]	84 [75, 92]	9 [3, 17]	0.71	56	53	0	24	0.96 [0.91, 1.00]	100	96	92
Nautil-SFT	63/64	0.58 [0.52, 0.64]	35 [25, 45]	56 [44, 67]	43 [31, 55]	0.59	78	38	27	47	0.38 [0.28, 0.48]	32	80	43
Nautil-RLVR	64/64	0.53 [0.47, 0.58]	38 [27, 48]	62 [50, 73]	36 [25, 47]	0.51	76	17	31	43	0.39 [0.30, 0.48]	30	54	51

[1]One sample per case. GPT-6 Sol does not judge its own investigations; see Table 11.

Table 11: Conclusion content under a different judge (Claude Sonnet 5), test non-host cases, five conditions; bottom rows: paired differences against Nautil-SFT. GPT-6 Sol is scored only here, because it cannot judge itself. The judge's reference includes Sol's own original teacher answer, which favours Sol on Cause, Chain and G3. Sonnet 5 is more lenient than GPT-6 Sol on overstatement; compare rows within this table only.

Model	Scored	Concl.	Overst.	Cause	Cause∧¬Overst.	Chain	Supp.	Alt.	0-claim	#NC	Gap	G1	G2	G3
Qwen3.5-9B (base)	59/64	0.38 [0.32, 0.44]	81 [71, 90]	47 [34, 60]	10 [3, 19]	0.33	62	27	3	10	0.70 [0.51, 0.88]	100	70	40
Gemini 3.8 Flash[1]	64/64	0.75 [0.69, 0.80]	55 [42, 67]	89 [81, 95]	42 [31, 55]	0.76	86	77	5	24	0.97 [0.94, 1.00]	100	100	92
Nautil-SFT	63/64	0.64 [0.58, 0.69]	11 [5, 19]	68 [57, 79]	63 [52, 75]	0.50	96	30	40	47	0.45 [0.34, 0.56]	40	79	49
Nautil-RLVR	64/64	0.55 [0.50, 0.60]	20 [12, 30]	61 [48, 72]	50 [39, 61]	0.43	89	14	36	43	0.38 [0.28, 0.50]	33	64	49
GPT-6 Sol (teacher ref.)[1]	64/64	0.87 [0.83, 0.90]	5 [0, 11]	91 [83, 97]	86 [77, 94]	0.71	97	78	8	40	0.91 [0.85, 0.96]	98	95	80
Qwen3.5-9B (base) − SFT	58	−0.25 [−0.33, −0.18]	+72 [+59, +84]	−19 [−36, −2]	−53 [−66, −41]		−28 [−38, −17]	−2 [−16, +14]			+0.23 [+0.02, +0.45]			
Gemini 3.8 Flash[1] − SFT	63	+0.11 [+0.04, +0.18]	+43 [+28, +56]	+21 [+8, +33]	−21 [−36, −5]		−9 [−17, −0]	+46 [+32, +60]			+0.43 [+0.29, +0.58]			
Nautil-RLVR − SFT	63	−0.09 [−0.14, −0.03]	+10 [0, +19]	−6 [−18, +5]	−13 [−25, 0]		−9 [−19, +1]	−16 [−28, −5]			−0.10 [−0.22, +0.04]			
GPT-6 Sol (teacher ref.)[1] − SFT	63	+0.23 [+0.17, +0.28]	−6 [−15, +2]	+22 [+10, +35]	+22 [+9, +35]		+2 [−3, +10]	+48 [+33, +61]			+0.48 [+0.36, +0.60]			

Table 12: Agreement between judges on the same 50 requests (stratified sample; no Gemini answers). Item agreement in % with Cohen's κ; conclusion-score correlation (Pearson / Spearman) and mean difference (first minus second judge). Bottom: overstatement rate by group. Untrained: not trained on our data (base, two prompts, Abstain-R1, Qwen2.5-3B); trained: all SFT/RLVR variants. The absolute overstatement rate depends strongly on the judge, the direction does not.

Judges	Cause	Chain	Overst.	Alt.	Concl. r / ρ	Mean diff.
GPT-6 Sol vs Claude Sonnet 5	88 (0.75)	70 (0.45)	50 (0.15)	78 (0.49)	0.73 / 0.76	−0.14
GPT-6 Sol vs Grok 4.7	88 (0.74)	74 (0.52)	86 (0.66)	80 (0.48)	0.82 / 0.82	+0.06
Claude Sonnet 5 vs Grok 4.7	88 (0.74)	88 (0.78)	56 (0.26)	74 (0.38)	0.71 / 0.74	+0.20
Overstatement rate (%) on the same 50 answers						
Judge	Untrained (19)	Untrained 9B (11)	Trained (31)	All 50	Mean Concl.	
GPT-6 Sol	100	100	52	70	0.38	
Claude Sonnet 5	58	45	10	28	0.51	
Grok 4.7	95	91	58	72	0.31	

Table 13: Appendix. Left: host cases of the test set (39 / 9), labelled by the host teacher's closure marker (no ground-truth root cause); the rule is "always close". These labels do not reproduce: GPT-6 Sol, re-investigating under our protocol, closed only 5 of the 39 host cases labelled should-close. Right: all 112 test cases (non-host and host labels mixed).

	Host (teacher labels)				All 112		
Model	Close→C	Not→N	Bal.	No-mark	Close→C	Not→N	Bal.
Qwen3.5-9B (base)	95.7	3.7	49.7 [46.6, 53.8]	2.8	90.3	16.0	53.2 [49.2, 57.6]
+ ReAct prompt	89.7	3.7	46.7 [42.7, 51.6]	5.6	80.1	26.7	53.4 [47.8, 59.3]
+ ledger prompt	90.6	0.0	45.3 [41.9, 48.3]	4.9	91.4	14.7	53.0 [48.4, 58.0]
Abstain-R1 (3B)	22.2	33.3	27.8 [19.2, 36.5]	38.2	17.2	44.7	30.9 [26.1, 35.9]
Qwen2.5-3B-Instruct	19.7	44.4	32.1 [21.5, 43.0]	27.8	16.1	64.0	40.1 [35.2, 44.9]
Gemini 3.8 Flash[1]	94.9	0.0	47.4 [43.6, 50.0]	0.0	96.8	48.0	72.4 [65.4, 79.6]
Nautil-SFT	79.5	18.5	49.0 [40.3, 58.1]	1.4	66.7	80.0	73.3 [66.9, 79.6]
Nautil-RLVR	95.7	7.4	51.6 [46.2, 59.8]	0.0	90.3	71.3	80.8 [74.8, 86.8]
GPT-6 Sol (teacher ref.)[1]	12.8	55.6	34.2 [17.5, 50.9]	0.0	37.1	80.0	58.5 [50.5, 66.8]
Source-only rule	100.0	0.0	50.0 [50.0, 50.0]	–	91.9	72.0	82.0 [74.7, 88.6]

Table 14: Appendix. OOD by domain: share of correct verdicts per cell (should-close cells: closed; should-not cells: left open). Case counts in parentheses. Domain and label are tightly coupled; only the medical domain has enough cases on both sides.

	Reports		Medical		TEP		Microservice
Model	C (80)	N (3)	C (8)	N (12)	C (3)	N (17)	C (18)
Qwen3.5-9B (base)	83.8	0.0	87.5	5.6	77.8	25.5	94.4
+ ReAct prompt	63.7	0.0	91.7	2.8	77.8	39.2	72.2
+ ledger prompt	85.4	11.1	95.8	5.6	100.0	43.1	92.6
Abstain-R1 (3B)	11.7	66.7	12.5	61.1	0.0	72.5	5.6
Qwen2.5-3B-Instruct	5.8	33.3	8.3	50.0	0.0	68.6	3.7
Gemini 3.8 Flash[1]	90.0	0.0	100.0	33.3	66.7	64.7	94.4
Nautil-SFT	40.4	44.4	25.0	91.7	11.1	100.0	0.0
Nautil-RLVR	66.2	33.3	33.3	75.0	55.6	92.2	7.4
Source-only rule[†]	100.0	0.0	0.0	100.0	0.0	100.0	100.0

[†]Oracle domain-majority rule (peeks at the label distribution). GPT-6 Sol was not run on OOD.

Table 15: Appendix. Long cases: the model fetching evidence itself vs. receiving all evidence in the first message (no tools, one turn). Judge score: GPT-6 Sol, 1–5 long-case rubric (no final answer = 1). Main cause: match / partial / miss / contradicts among judged answers. Δ: paired all-at-once minus own-fetching judge score. The two modes also differ in prompt wording, turns, per-turn and context limits.

Model	Mode	Close→C	Not→N	Bal.	No-mark	Score	Main cause	Δ score
Medium-long v2 (16; 8 / 8)								
Qwen3.5-9B	own fetching	79.2	0.0	39.6 [31.2, 47.9]	7	2.12	8/28/2/5	
Qwen3.5-9B	all at once	100.0	0.0	50.0 [50.0, 50.0]	0	2.42	17/25/0/6	$+0.29$ [-0.06, $+0.67$]
Nautil-SFT	own fetching	58.3	50.0	54.2 [35.4, 72.9]	3	2.58	13/20/11/1	
Nautil-SFT	all at once	62.5	58.3	60.4 [47.9, 72.9]	1	2.75	14/32/1/0	$+0.17$ [-0.25, $+0.58$]
Abstain-R1	own fetching	0.0	41.7	20.8 [12.5, 31.2]	30	1.54	0/5/27/10	
Abstain-R1	all at once	33.3	33.3	33.3 [20.8, 45.8]	8	2.02	1/32/5/2	$+0.48$ [$+0.25$, $+0.71$]
Qwen2.5-3B	own fetching	4.2	37.5	20.8 [10.4, 31.2]	23	1.75	0/12/27/6	
Qwen2.5-3B	all at once	50.0	29.2	39.6 [27.1, 52.1]	5	2.06	4/30/6/3	$+0.31$ [$+0.10$, $+0.54$]
Longest 20 (v3; 18 / 2)								
Qwen3.5-9B	own fetching	70.4	0.0	35.2 [27.8, 42.6]	16	1.72	10/19/2/14	
Qwen3.5-9B	all at once	100.0	0.0	50.0 [50.0, 50.0]	0	2.62	36/20/0/4	$+0.90$ [$+0.67$, $+1.12$]
Nautil-RLVR	own fetching	92.6	33.3	63.0 [44.4, 81.5]	0	2.47	8/41/6/5	
Nautil-RLVR	all at once	61.1	33.3	47.2 [39.8, 54.6]	6	2.65	16/38/2/0	$+0.18$ [0.00, $+0.38$]

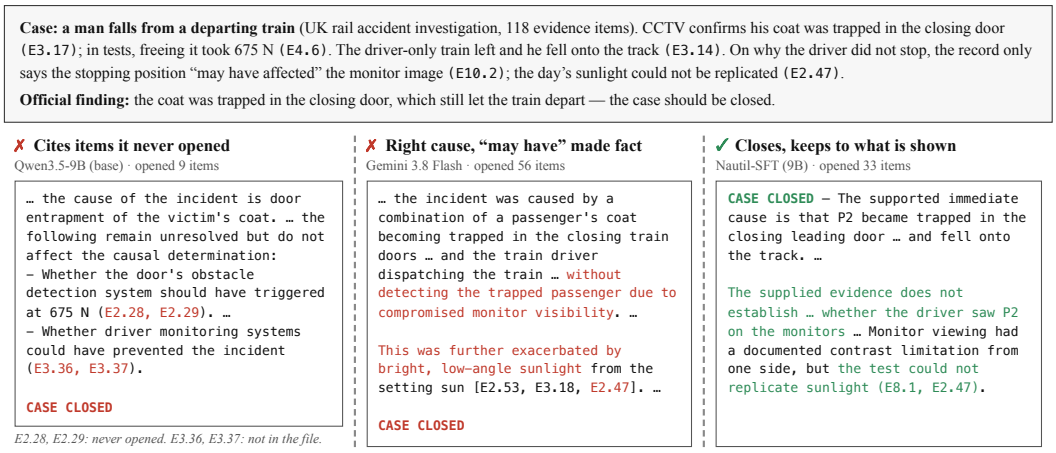

Figure 5: A case that should be closed (a passenger's coat was trapped in a closing train door and the train departed); all three investigators name the trapped coat. The base model (left) cites six items it never opened, two of which do not exist; the frontier model (middle) turns hedged records of monitor contrast and sunlight into the established reason the driver missed the passenger. Nautil-SFT (right) closes on what the CCTV and tests show and leaves what the driver saw open. Excerpts shortened (. . .); one sample per model.

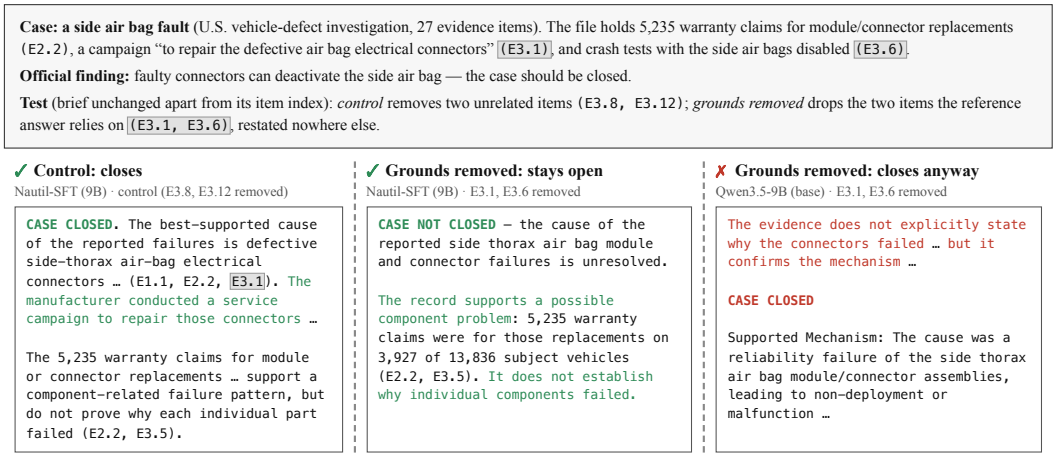

Figure 6: The same case with different evidence removed; the brief is unchanged except that removed items drop out of its index. With two unrelated items removed (left), Nautil-SFT closes on the manufacturer's campaign to repair defective connectors (E3.1); with E3.1 and the crash tests (E3.6) removed (middle), it treats the remaining warranty counts as only a possible component problem and leaves the case open. The base model (right) closes the reduced file on the warranty counts alone. Excerpts shortened (. . .); one sample per model and version.

Case: engine loss after a fuel-tank switch (U.S. aviation safety investigation, 28 evidence items). The engine ran smoothly until the pilot selected the left tank, then lost power; he landed in the water (`E3.4–E3.6`). Preflight draining showed no water (`E3.1`). An added "FERRY TANK" valve in the left fuel line was found ON, with no guard or safety wire (`E4.8–E4.10`); the left tank connected to the engine only with that valve OFF (`E4.9`).

Official finding: inadvertent activation of the unguarded ferry-tank valve caused fuel starvation — the case should be closed.

✓ **Reads the valve test, closes**
Gemini 3.8 Flash

The total loss of engine power was caused by fuel starvation … due to an unapproved, unguarded ferry-tank selector valve that had been left in the "ON" position …

Continuity from the left tank to the engine was only established when the ferry-tank valve was moved to "OFF" [E4.9]. …

CASE CLOSED

✗ **Misreads the test, stays open**
Nautil-SFT (9B)

CASE NOT CLOSED. … the pilot reported no visible water in either tank at preflight draining … continuity through the ferry-tank selector was absent in both observed positions (E4.9). …

What is not established: whether the left tank contained fuel contaminated with water … Whether the undocumented ferry-tank arrangement caused the interruption is likewise unproven.

✗ **Reads the test, still stays open**
Nautil-RLVR (9B)

CASE NOT CLOSED — the target cause, a fuel-path interruption …, is supported but not conclusively identified. …

no continuity could be established from the left selector position to the engine except when the ferry-tank selector was OFF (E4.4, E4.7, E4.9). … Its role in the interruption is plausible but unproven by this evidence.

Figure 7: A failure of our models on a case that should be closed: the file shows that an unguarded ferry-tank valve set to ON blocked fuel from the left tank, and the frontier model (left) reads this and closes. Nautil-SFT (middle) misreads the continuity test (E4.9) and keeps water contamination open, although its own answer notes that preflight draining showed no water. Nautil-RLVR (right) reads the test correctly but still calls the valve's role unproven. Excerpts shortened (. . .); one sample per model.