WHEN REWARDS BREAK TRACEABILITY: TRAINING TRACEABLE EVIDENCE PATHS WITH LEDGERRL

Anonymous authors
Paper under double-blind review

ABSTRACT

In high-stakes settings such as medicine and finance, a reader must be able to check the evidence behind an answer: it should be correct, reach the evidence it needs, and expose small units that can be checked. Retrieval-augmented reasoners are usually trained with outcome or citation rewards and judged by accuracy or citation precision; we find that such training admits four shortcuts that break these requirements: outcome rewards leave answers without a cited path, entailment rewards encourage claims that copy their evidence, citation rewards attach many sources to broad assertions, and formats that answer in one context search little. We answer each with a matching design. A typed claim–evidence ledger lets named claims cite evidence and the answer cite claims; staged execution begins with an explicit investigation; and **LedgerRL** trains one shared policy with claim-level rewards that discount copied claims and count only answer-used claims toward quantity credit. On three multi-hop corpora after medical-only training, the staged ledger elicits the search that a flat citation prompt skips, and even untrained it makes more answers traceably correct than a trained citation-RL policy. Unguarded claim credit collapses traceability, whereas the guarded rewards reach 19.7% complete-path correctness, above every baseline under the same prompts, and leave fewer answers uncited than an outcome-only ledger. No arm dominates: against citation RL and a citation cap that searches about as often, LedgerRL leads at claim-level TC@2 and covers more gold evidence than the cap, but not at final-answer check units, in gold precision, or in gold-citation F1; it leaves more answers uncited and trails the cap by 7.9 points in accuracy at 1.9 times its tokens.

1 INTRODUCTION

In medicine, finance, and other high-stakes settings, an answer is acted on only after someone checks where it came from. Retrieval-augmented reasoning systems can answer correctly while leaving this unclear. A reader who wants to check the result needs an explicit path from *answer to claim to evidence*, and that path is useful only if three requirements hold for the same answer. The answer is *correct*; the path is *complete*, reaching the evidence the answer needs; and the path is *resolved*, so that each statement on it can be checked against a small number of sources. Accuracy measures the first requirement, and citation precision or recall measures parts of the second. Neither measures whether all three hold at once.

The usual remedy trains the reasoner to search and cite, with outcome rewards for correct answers (Jin et al., 2025) or rewards for supported citations (Menick et al., 2022; Huang et al., 2024), and evaluates citation precision and recall (Gao et al., 2023b). We find that these training signals and formats each break one of the three requirements (§2). **(S1)** Outcome-rewarded search answers without a path, and an outcome-trained ledger still leaves more than a quarter of its answers uncited. **(S2)** In a claim ledger trained with entailment-based credit, claims increasingly copy their evidence, and out of domain its answers rarely place the gold evidence on a cited path. **(S3)** Sentence-level citation rewards can attach many sources to one broad assertion (*shotgun citation*); a citation cap restores small checks but places less of the retrieved evidence on the answer path. **(S4)** Formats that answer in one context search little: an untrained model given a flat citation prompt skips search on nearly a third of questions, and a ledger trained in a single context settles at one search.

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.

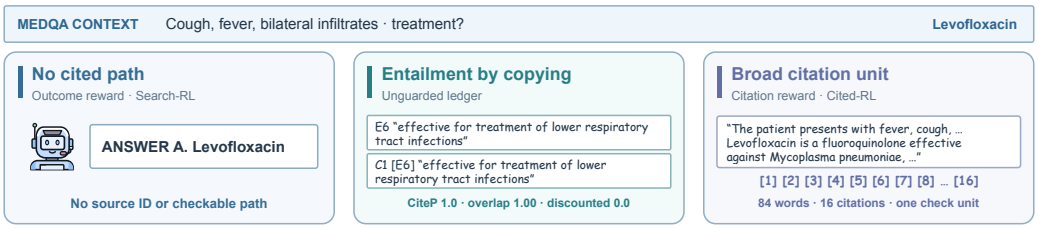

Figure 1: Reward shortcuts. The uncited SEARCH-RL format is schematic; in an audited MedQA case, an unguarded claim copies its evidence and CITED-RL puts 16 citations on one 84-word assertion. Excerpts are shortened here; the text reports the case values and aggregate results.

Figure 1 illustrates S1–S3 with a format schematic and two audited MedQA cases. The failures break different links in the public record: without a source id a correct answer cannot be traced back; a copied claim scores perfectly on entailment without adding a useful intermediate statement; and sixteen citations on one broad assertion leave the reader a wide check. None implies that the answer is false, so support, completeness, and check size need separate measurement (§2).

Each shortcut calls for a change in what the reasoner records and is rewarded for, not only in how its answers are scored. We answer each shortcut with one design choice (§3). A *typed claim–evidence ledger* records evidence rows, named claims that cite them, and an answer that cites claim ids, so the answer–claim–evidence path is explicit and each claim is a check unit (S1, S3). *Staged execution* runs an investigator that searches and writes claims, an assessment pass that annotates them, and a synthesiser that answers by citing claim ids, as separately prompted contexts of one shared policy, so that evidence gathering precedes answering (S4). **LEDGERRL** trains this policy with claim-level rewards that credit supported claims, discount copied ones (S2), and count only claims that the answer uses toward quantity credit (S1). The record does not by itself make the final rationale atomic, so we test completeness and check size separately.

To evaluate these requirements together, we define *traceable correctness* TC@k (§2): the fraction of answers that are correct, reach every gold supporting source through their public path, and require at most k sources for any single check. It uses human evidence annotations that no arm sees during training: all policies are trained on medical questions and evaluated zero-shot on HotpotQA, 2WikiMultihopQA, and MuSiQue. This is a structural, recall-oriented criterion, not a measure of citation precision or semantic validity; we test those limits explicitly.

In the tested configuration, each design counters the shortcut it targets. Untrained, the staged ledger searches three to four times per question and already makes more answers traceably correct than the flat prompt and than trained citation RL of equal accuracy (§5.1). Without guards, claim credit copies evidence and traceability collapses; with them, verbatim overlap halves, and against an outcome-only ledger fewer answers are uncited because more of the retrieved gold reaches the answer path (§5.2). Against flat citation training no arm dominates (§5.3): LEDGERRL leads citation RL and a citation cap that searches about as often at claim-level TC@2 and in complete paths, but not at final-answer units, in gold precision, or in gold-citation F1; it leaves more answers uncited and trails the cap in accuracy at 1.9 times its tokens. Flagging a cited claim makes regenerated answers omit it 20 points more often than a matched control (§6).

Our contributions are:

- **Four shortcuts in attributed reasoning**, each breaking one traceability requirement: uncited outcome answers, copied claims under entailment credit (four seeds, held-out judge), broad multi-source assertions under citation reward, and contracted search in one-context formats.
- **A typed claim–evidence ledger with staged execution** that answers S1, S3, and S4. Without training, it elicits the search a flat prompt does not and yields more traceably correct answers than that prompt and trained citation RL.
- **LEDGERRL**, guarded claim-level rewards that answer S2 and S1: they halve verbatim overlap, avoid the path collapse of unguarded credit, and raise complete-path correctness above every baseline under the same prompts, with a stated trade-off in final-answer granularity, precision, uncited answers, accuracy, and tokens.

2 REWARD SHORTCUTS THAT MOTIVATE THE DESIGN

We find four shortcuts, each breaking one requirement of a traceable answer (Figure 1).

S1: Outcome reward leaves no path. Outcome-rewarded search cites nothing by design. A ledger format alone does not fix this: trained with outcome reward only, it leaves 28.7% of answers without a reachable source (Table 1).

S2: Entailment credit can reward copying. Let assertions $a_1, \ldots, a_m$ cite retrieved snippet sets C_i. Their mean entailment score is

$$\textsc{CiteP} = \frac{1}{m}\sum_{i=1}^{m} g_i, \qquad g_i = \mathbb{1}\big[\mathrm{NLI}\big(\textstyle\bigcup_{c\in C_i} \mathrm{text}(c) \models a_i\big) \geq \tau\big]\,. \tag{1}$$

Quoting a premise satisfies entailment without constructing a useful intermediate claim. In the unguarded ledger, entailment and copying rise together across four seeds (§5.2; Appendix D).

S3: Citation credit can reward broad checks. Attaching many sources to one long assertion can retain coverage while making the support relation hard to locate. We call this *shotgun citation*; Figure 1 shows a case. Across the three evaluation corpora, 24% of CITED-RL's final rationale items reach more than two sources, against 6% for the citation cap (two-seed means; Appendix B). A citation cap narrows checks but places less of the retrieved evidence on the answer path (§5.3).

S4: Answering in one context contracts search. Under a flat citation prompt, the untrained model searches 0.8 times per question and not at all on 29% of questions. A ledger written in one context searches 1.56 times per MedQA question untrained and settles at one search after training (§5.1). These observations motivate measuring the full path and the number of sources needed to check each part, rather than treating entailment or citation count as traceability.

Traceable correctness. A traceable answer is correct, reaches its gold supporting evidence through its public path, and divides that path into checks using few sources. Let $P(a)$ be the source ids reachable from answer a's non-empty rationale items, via claims in a ledger or directly in a flat record. The check units $U(a)$ are cited claims for a ledger and rationale items for a flat record; $E(u)$ is the set of sources cited by unit u. For gold source ids G_q,

$$\mathrm{TC@}k = \frac{1}{N}\sum_{q=1}^{N} \mathbb{1}[\mathrm{EM}(a_q) = 1] \cdot \mathbb{1}[G_q \subseteq P(a_q)] \cdot \mathbb{1}[\forall u \in U(a_q) :\ 1 \leq |E(u)| \leq k]\,. \tag{2}$$

The three factors enforce correctness, gold-source coverage, and bounded check size. We use TC@2 as the headline measure, report other k values, and count uncited SEARCH-RL answers as zero. Gold evidence is unseen in training. Splitting one assertion into several units can reduce check size without improving its support; citing unrelated sources does not reduce gold recall. TC@k therefore omits citation precision and semantic entailment, and its claim-level ledger units are not directly comparable to flat final-rationale units. We therefore also report, post hoc, *answer-level* TC, which uses final rationale items as the units of every format; HotpotQA-style joint EM, a correct answer whose reachable sources equal the gold set (TC@∞ restricted to gold precision one); gold-citation F1, the per-question harmonic mean of gold precision and gold coverage of the path; and the rate of *uncited* answers, which have no reachable source, missing answers included (§5; Appendix B).

3 LEDGERRL

LEDGERRL trains one policy to construct and use a typed claim–evidence record. Its staged implementation separates claim construction from answer construction across investigation, assessment, and synthesis; Figure 2 shows the resulting path and the reward assigned to each stage. Each design choice targets one shortcut: the ledger path with use-gated quantity credit answers S1, the copy discount S2, claims as check units with a quantity cap S3, and staged execution led by investigation S4; §5 tests each against its shortcut. We also test a single-context variant with the same schema and reward terms. A `support` annotation is not grounding credit: in the figure, the copied C2 receives no copy-discounted credit and is unused by the answer.

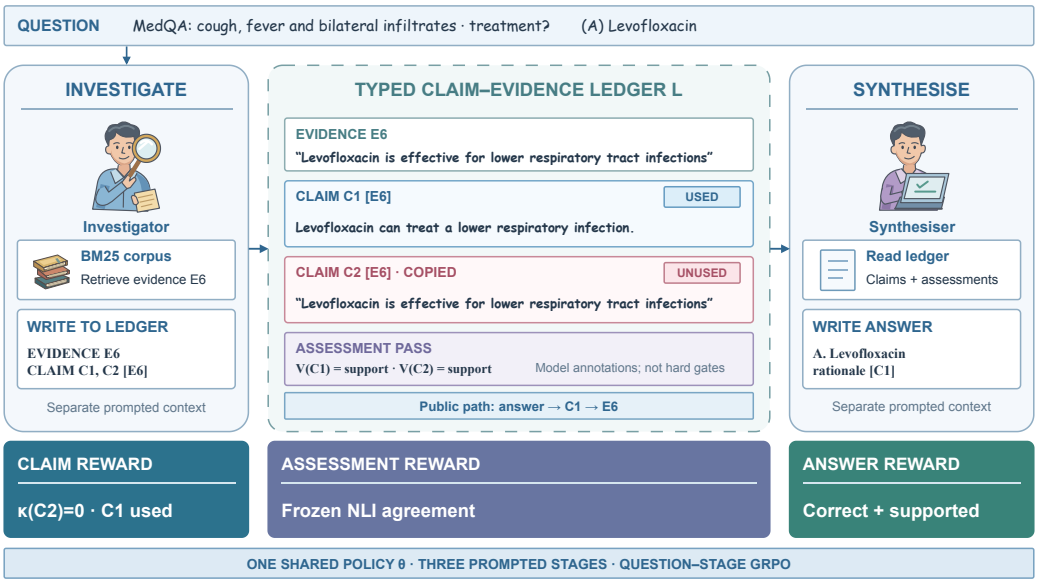

Figure 2: Schematic LedgerRL execution. The investigator writes evidence and claims; the assessment pass adds non-binding verdicts; the synthesiser cites C1, exposing an answer → C1 → E6 path. The copied C2 receives zero copy-discounted grounding credit despite its support annotation. The lower cards map rewards to three prompted stages of one shared policy.

3.1 A TYPED RECORD OF THE EVIDENCE PATH

The ledger L is an append-only sequence of schema-validated rows:

`EVIDENCE`	`{eid, source, snippet_id, text}`
`CLAIM`	`{cid, text, cites:[eid]}`
`VERDICT`	claim id, label ∈ {support, refute, insufficient}, rationale
`ANSWER`	choice or diagnosis; rationale items with text and cited claim ids

An answer cites claim ids, and each claim cites evidence ids that resolve into the retrieval corpus. This names the two support edges separately, so a checker can locate a failed edge or change a claim's verdict before replaying synthesis. Unlike attribution-first generation (Slobodkin et al., 2024), the generated intermediate claims remain individually addressable. The schema guarantees neither truthful nor atomic claims; a `support` verdict is a model judgement, not a certificate.

3.2 CONSTRUCTING AND CONSUMING THE LEDGER

The shared policy θ executes three separately prompted stages in sequence. The **investigator** searches and writes `EVIDENCE` and `CLAIM` rows; the **assessment pass** (*auditor* in the implementation) adds a `VERDICT` per claim; the **synthesiser** writes an `ANSWER` citing claim ids. Generated information passes between contexts through the ledger, not their conversational histories. The synthesiser still has the question and parametric knowledge, so references alone do not force evidence use; we test dependence by intervention (§6). The single-context comparison changes context, protocol, and credit assignment together (§5.1).

3.3 REWARDS FOR SUPPORTED AND ANSWER-USED CLAIMS

Copy-discounted grounding. For claim i, let v_i be the longest contiguous token run shared with its cited evidence, divided by the claim's token count. We define

$$\kappa(v) = \mathrm{clip}\left(\frac{\rho_{\mathrm{hi}} - v}{\rho_{\mathrm{hi}} - \rho_{\mathrm{lo}}}, 0, 1\right), \qquad \textsc{CiteP}_{\mathrm{abs}} = \frac{1}{m}\sum_{i=1}^{m} g_i \kappa(v_i). \tag{3}$$

We set $\rho_{\text{lo}} = 0.40$ and $\rho_{\text{hi}} = 0.80$: shorter shared runs retain full credit, near-verbatim reuse receives none, and the penalty increases linearly between them. This allows some shared terminology while discouraging extended copying; the cutoffs are lexical heuristics, not thresholds of factual correctness. CITEP$_{\text{abs}}$ is a diagnostic held out of the unguarded reward, but an optimisation target for LEDGERRL. Improvements on it therefore require complementary evidence from gold annotations and another judge.

Quantity and use. Let z_i indicate whether the final answer cites claim i, and let n_{good} count claims for which $g_i \kappa(v_i) \geq 1/2$ and $z_i = 1$. We cap quantity credit at $K_c = 3$ good claims and scale grounding credit by $\min(1, m/m_{\min})$, with $m_{\min} = 3$, to discourage a very small ledger. Gating on use prevents unused claims from earning quantity credit. A redundancy term measures the fraction of nonempty content-token claim sets that overlap an earlier claim above Jaccard similarity 0.60. Together, the investigator's guard terms are

$$w_g \, \text{CITEP}_{\text{abs}} \min\left(1, \frac{m}{m_{\min}}\right) + w_n \frac{\min(n_{\text{good}}, K_c)}{K_c} - w_r \, \text{redundancy}(L). \tag{4}$$

The weights are $w_g = 0.5$, $w_n = 0.2$, and $w_r = 0.3$. The full investigator reward also includes outcome and evidence-utility terms, and charges 0.05 per search beyond the first (Appendix A). The quantity gate connects investigation to the claims ultimately selected by the synthesiser.

Assessment, synthesis, and optimisation. The assessment pass learns to match a frozen NLI judge; its stored verdicts are policy outputs, not the judge's training scores. Independently, the synthesiser receives answer, claim–answer support, and citation-coverage rewards. Both support edges thus receive semantic supervision without using policy verdicts as reward labels. GRPO trains the shared policy on three question–stage rows per question. Each of the eight samples in a row reruns the whole protocol, including freshly generated upstream stages; advantages are grouped within that question–stage row, and only the selected stage's generated tokens enter its loss. Thus stage separation increases rollout generation and is not a compute-matched comparison. Omitting assessment input at inference leaves assessment-stage training and both edge rewards intact. Appendix A gives the full objectives, masking, weights, and schema penalties; §5.2 tests their effects.

4 EXPERIMENTAL SETUP

Training and comparisons. We train Qwen3-8B (Yang et al., 2025) with GRPO (Shao et al., 2024) on MedQA training questions without evidence labels for up to 150 steps (batch 24, eight rollouts per row, learning rate 10^{-6}), and evaluate MedQA (Jin et al., 2021) and zero-shot transfer to HotpotQA, 2WikiMultihopQA, and MuSiQue (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). All arms share the policy, training source, learning rate, batch, and maximum search budget. The staged ledger expands $2{,}000$ questions into one row per stage; the flat arms use $6{,}000$ questions so that they train on the same number of rows and rollouts. SEARCH-RL uses outcome reward; CITED-RL adds sentence-level citation rewards, and a citation cap adds an excess-citation penalty. Ledger controls remove the guards or all process rewards, disable one guard at a time, communicate in free prose, or write the same schema in one context. LEDGERRL and the unguarded ledger use seeds 1234, 2025, and 7, and the outcome-only ledger and flat arms seeds 1234 and 2025 (Appendix A, Table 5).

Measurements and statistics. We report MedQA accuracy on all $1{,}273$ test questions and claim behaviour on 200. Out of domain, each corpus contributes its first 200 questions, decoded greedily over a dataset-level BM25 pool (Appendix A); human supporting-fact annotations, unseen in training, define the gold evidence for TC@k, gold citation precision, and the retrieval ceiling (fraction of gold evidence retrieved). A missing final answer counts as incorrect, untraceable, and uncited for every arm. Question-paired bootstrap intervals ($5{,}000$ draws) resample within corpora and keep all runs of a method together; they measure question uncertainty conditional on the trained checkpoints, not training-seed uncertainty; FinQA (Chen et al., 2021) and further evaluations are in the appendix.

5 RESULTS

We test each design choice against the shortcut it targets: search (S4), path completeness and copying (S1, S2), and check size (S3), which also sets the trade-off against flat citation training.

5.1 STAGED EXECUTION RESTORES SEARCH (S4)

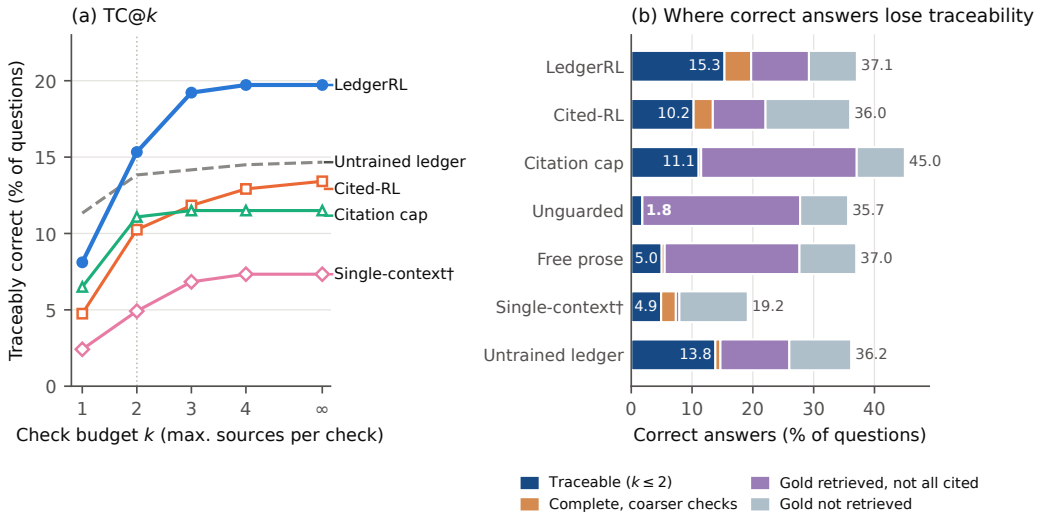

Figure 3: Traceable correctness on HotpotQA, 2Wiki, and MuSiQue (600 questions, zero-shot after medical training). (a) TC@k (Eq. 2): answers that are correct, reach every gold source through their public path, and need at most k sources per check. (b) Correct answers split by the first requirement they miss. LEDGERRL and the unguarded staged ledger average step-150 seeds 1234, 2025, and 7, the other trained step-150 arms seeds 1234 and 2025; †the single-context ledger is compared at step 100 (§5.1). Free-prose stages and the dashed untrained staged ledger are single runs. SEARCH-RL cites nothing and has TC@$k = 0$.

Before any training, the staged protocol already makes more answers traceable: the untrained staged ledger is as accurate as the trained CITED-RL policy but more often traceably correct (Table 1; $+3.6$ TC@2 points, 95% interval $[1.5, 5.8]$), also at the answer level and in joint EM. The gap starts in retrieval. Under the flat prompt the model often answers from memory (S4), whereas the investigation stage searches three to four times per question, so about four times as many questions reach a correct answer with all gold evidence retrieved (Appendix B). Staging thus helps by eliciting search; an external rule forcing four searches is an alternative remedy that makes flat citation more traceable than the ledger (Appendix C). Untrained runs are single runs.

Training keeps this search only when the stages stay separate. SINGLE-CONTEXT LEDGER, with the same row types and rewards in one context, grounds its claims as well as LEDGERRL but converges to one search per question and reaches about a third of LEDGERRL's TC@2 at the same step ($+9.2$ points for LEDGERRL, $[6.7, 11.8]$; Appendix C). Neither restoring the investigator's reward weight nor removing the search cost prevents the contraction, and part of the gap precedes training; the comparison changes contexts, protocol, and credit assignment together.

5.2 GUARDED CLAIM CREDIT COMPLETES PATHS WITHOUT COPYING (S1, S2)

Entailment credit alone is satisfied by quoting (Table 2). Without guards, claims are entailed as often as LEDGERRL's but largely copy their evidence, and copy-discounted grounding falls below the untrained model; the guards keep the entailment rate and halve verbatim overlap, while the outcome-only control copies little but is rarely entailed. The copy discount stops copying and the used-claim gate stops unused claims from inflating the record; the redundancy penalty shows no independent effect (single-seed ablations), and a judge never used in training confirms the ordering (Appendix D). Out of domain, copying costs traceability rather than accuracy: unguarded

Table 1: **Traceability under claim- and answer-level check units.** Values in %; HotpotQA, 2Wiki, MuSiQue, 600 questions, step 150. TC@2 uses each format's own units; answer TC uses final rationale items for every format; joint EM, cite F1, and uncited answers are defined in §2. Two-run means except LEDGERRL and the unguarded ledger (three runs) and the untrained ledger and untrained flat citation (one run each). Paired differences in the text average per-question differences and can differ from differences of rounded means by 0.1. Bold and underline mark the best and second-best arm in each column.

Arm	EM↑	TC@2↑	Answer TC@2↑	TC@∞↑	Joint EM↑	Cite F1↑	Gold P↑	Uncited↓
Flat citation								
Untrained flat citation	20.5	4.8	4.8	5.2	4.5	30.2	39.4	33.5
CITED-RL	36.0	10.2	10.2	13.4	9.8	**49.6**	59.1	**6.2**
Citation cap	**45.0**	11.1	11.1	11.5	10.7	<u>48.1</u>	**63.7**	<u>12.0</u>
Staged ledger								
Untrained ledger	36.2	<u>13.8</u>	**12.5**	<u>14.7</u>	**12.5**	45.7	57.1	21.5
Outcome-only ledger	36.8	12.7	<u>12.0</u>	13.2	<u>12.1</u>	41.7	53.5	28.7
Unguarded ledger	35.7	1.8	1.8	1.8	1.8	40.9	<u>63.0</u>	17.3
LEDGERRL	<u>37.1</u>	**15.3**	10.6	**19.7**	10.8	46.8	53.1	18.2

runs answer and retrieve gold evidence as often as LEDGERRL, yet only 1.8% of their answers are traceable, because retrieved gold rarely reaches the answer path (Figure 3b).

With protocol and search limit fixed, guarded claim credit changes how retrieved evidence is used, not how much is retrieved (Table 3): against the outcome-only and untrained ledgers, a larger share of retrieved gold reaches the answer path (+13.1 and +6.3 points, $[10.6, 15.7]$ and $[3.2, 9.3]$), so fewer answers are uncited and more paths are complete. The paths are not more precise, and because trained rationale items bundle claims, joint EM and answer-level TC@2 do not improve: guarded credit buys completeness, not precision. Removing the copy discount alone lowers TC@2 from 14.2% to 6.0% at step 100, far more than any other single guard (Appendix B); these single-run ablations do not show that copying alone causes the missing paths.

5.3 CHECK SIZE AND THE TRADE-OFF WITH FLAT CITATION TRAINING (S3)

Flat citation training trades check size against evidence use (Figure 3). CITED-RL keeps coverage but loses about a quarter of its complete answers to broad checks. The citation cap keeps checks small and searches about as often as LEDGERRL but uses little of what it retrieves: it retrieves all gold evidence for 82% of its correct answers yet cites all of it for only 26%, against 53% for LEDGERRL.

Against every baseline (Table 3), LEDGERRL leads the flat settings in claim-level TC@2 and in complete paths. Against the cap the gain comes from using retrieved evidence: LEDGERRL retrieves less gold but covers more of it on the answer path (+5.2 points over four corpora, $[3.3, 7.1]$; Appendix E). Claim units do not make the final rationale finer, because final items bundle claims (Appendix B): at final-answer units and in joint EM LEDGERRL ties both flat settings, and it has lower gold precision, more uncited answers, lower gold-citation F1 than CITED-RL, and lower out-of-domain accuracy than the cap at about twice its tokens.

No arm dominates. At the reported checkpoints and under the same prompts, LEDGERRL is ahead of every baseline in complete-path correctness, at the costs above; without a need for intermediate records, the citation cap is the more economical and, out of domain, the more accurate choice.

6 WHAT TC DOES NOT CERTIFY

TC certifies that the gold sources lie on the answer's path and that each check reads few of them, not that those sources entail the statements citing them or that the answer depends on the recorded claims. We test both.

Semantic support. Under frozen BART-MNLI, never used in training, and the DeBERTa training judge, LEDGERRL's final rationale items are not consistently better supported by their cited

Table 2: **Reward components under the staged ledger protocol.** MedQA, $n = 200$, greedy, step 100. Used is the fraction of claims cited by the answer. CITEP$_{\text{abs}}$ is an objective for guarded arms, not an independent success criterion. Unguarded multi-seed results are in Appendix D.

Arm	CITEP	CITEP$_{\text{abs}}$↑	Verbatim↓	Claims/q	Used	Redund.↓	Schema err.↓
UNTRAINED	0.497	0.443	0.229	5.62	0.475	0.016	0.012
OUTCOME-ONLY LEDGER	0.370	0.353	0.181	4.76	0.502	0.019	0.031
LEDGERRL W/O GUARDS	0.886	0.331	0.691	3.60	0.607	0.022	0.000
w/o anti-copy	0.907	0.354	0.682	4.72	0.545	0.021	0.002
w/o used-claim gate	0.863	0.790	0.305	7.92	0.442	0.011	0.004
w/o quantity term	0.912	0.811	0.373	3.25	0.762	0.009	0.001
w/o redundancy penalty	0.805	0.767	0.249	6.25	0.534	0.006	0.003
LEDGERRL	0.886	0.790	0.342	4.54	0.605	0.010	0.002

Table 3: **Paired differences, LEDGERRL minus each baseline.** Points; HotpotQA, 2Wiki, MuSiQue, 600 questions, step 150. Bold: the 95% interval favours LEDGERRL; underlined: it favours the baseline; plain: it includes zero. Intervals are in Table 8. LEDGERRL makes 3.88 searches per question against 1.99 and 3.60 for the flat arms and uses 3.9 and 1.9 times their recorded tokens over four corpora including FinQA (Appendix E); the ledger arms share its protocol and search limit.

LEDGERRL −	TC@2↑	TC@∞↑	Answer TC@2↑	Joint EM↑	Cite F1↑	Gold P↑	Uncited↓	EM↑
CITED-RL	+5.1	+6.3	+0.3	+1.1	−2.8	−6.1	+12.0	+1.1
Citation cap	+4.2	+8.2	−0.5	+0.2	−1.3	−10.7	+6.2	−7.9
Outcome-only ledger	+2.7	+6.5	−1.4	−1.2	+5.1	−0.4	−10.5	+0.3
Untrained ledger	+1.5	+5.1	−1.9	−1.7	+1.1	−4.1	−3.3	+0.9
Unguarded ledger	+13.5	+17.9	+8.7	+9.0	+5.9	−9.9	+0.9	+1.4

evidence than CITED-RL's across corpora, runs, and judges (Appendix F). LEDGERRL's TC advantage therefore concerns which evidence is on the path and how finely it is split, not entailment; requiring both ledger edges to be entailed also lowers support markedly, so grounded claims do not guarantee a supported path.

A local assessment changes the answer's citations. We hold LEDGERRL's question, evidence, and claims fixed, remove the answer, and regenerate with the synthesiser. On questions with both a supported cited claim and a supported uncited claim, we mark exactly one claim `insufficient`: the cited target, or a length-matched uncited control. Success means a valid answer that omits the target.

Flagging one cited claim makes regenerated answers omit it about 20 points more often than flagging the control (Table 4), and on HotpotQA it also lowers retention of the target-linked rationale under both judges. The answer thus depends partly on the claim layer that TC counts, although most flagged targets are kept. The trained assessment pass itself is not a reliable signal: it labels 93–97% of claims `support`, and its verdicts show no EM effect at inference (Appendix G).

7 LIMITATIONS

LEDGERRL has a material in-domain accuracy cost against most flat baselines and uses about twice the tokens of the citation cap. TC@k needs gold evidence annotations and measures completeness and check size, not precision or entailment, which we bound only post hoc; final rationale items stay coarse. Intervals are conditional on the evaluated checkpoints (one to three seeds per arm; single-seed guard ablations). Transfer is zero-shot after medical-only training, protocol comparisons change several factors together, and baselines are our own implementations. We make no claim about faithfulness to hidden computation or reduced verification effort; Appendix J gives the details.

8 RELATED WORK

Attributed generation and its evaluation. Attribution evaluation connects generated assertions to sources (Rashkin et al., 2021; Gao et al., 2023b; Yue et al., 2023); FActScore (Min et al., 2023)

Table 4: **One-claim support withdrawal.** Valid answers omitting the previously cited target after a one-claim support-status intervention. Same investigator outputs, greedy synthesis; the control flags a different, previously uncited claim.

Corpus	Eligible	Flag target	Flag control	Paired difference [95% interval]
HotpotQA	161/200	0.224	0.019	$+0.205$ [+0.143, +0.267]
2Wiki	135/200	0.215	0.015	$+0.200$ [+0.133, +0.267]

further decomposes long text into atomic facts, and citation metrics are surveyed by Schreieder et al. (2026). Training approaches include preference learning for quote-supported answers (Nakano et al., 2021; Menick et al., 2022), fine-grained citation rewards (Huang et al., 2024), and traceability-shaped RAG rewards (Ren et al., 2025). The latter reward selecting each training question's annotated supporting passages before a free-text analysis; our rewards use no evidence labels, credit named intermediate claims through an entailment judge with copy and use guards, and are tested on corpora held out of training. Citation precision and recall are often reported separately; TC@k instead requires correctness, complete gold-source recall, and bounded check size for the same answer. Unlike ALCE's citation-precision and support checks, TC is judge-free but does not test entailment or penalise extra non-gold citations; we report precision-conditioned variants and gold-citation F1 alongside it. TC@∞ relaxes the exact supporting-fact set match of HotpotQA's Joint EM (Yang et al., 2018) to gold-id recall along an anchored answer-reachable path, and TC@k adds a check-size bound; it is related to, not independent of, joint answer–evidence evaluation, and we report Joint EM alongside it. Quotation is appropriate when it is the deliverable, as in quote-supported generation; our copy discount targets only the ledger's intermediate claims.

Attribution before generation and self-assessment. Slobodkin et al. (2024) select source spans and plan sentences before generating locally attributable text, and measure fact-checking effort with people. Self-RAG (Asai et al., 2024) combines adaptive retrieval with reflection tokens that assess relevance and support and can steer decoding. Fine-grained support assessment and controllability are therefore not unique to LEDGERRL. It differs in keeping generated intermediate claims as named record objects with separate source–claim and claim–answer links, trained online with stage-specific rewards; we do not claim head-to-head superiority. EntailmentBank (Dalvi et al., 2021) evaluates multi-step trees from premises through intermediate conclusions to an answer hypothesis; our ledger is an executable attribution record, but its schema and TC score do not certify an entailment tree or the final inference.

Search, roles, and reward misspecification. Our search scaffold follows Jin et al. (2025); agentic RL also refines credit via tool-call entropy (Li et al., 2026) or next-user-turn feedback (Zhao et al., 2026). Multi-agent failure modes are catalogued by Cemri et al. (2025). LEDGERRL uses one shared policy with separate stage contexts and a typed communication record; we do not claim that more agents generally improve reasoning. Reward gaming is an established concern (Amodei et al., 2016; Skalse et al., 2022; Gao et al., 2023a; Pan et al., 2022; Wen et al., 2024), and entailment models rely on lexical overlap (McCoy et al., 2019). We show such shortcuts in attributed reasoning and a guard for each.

9 CONCLUSION

Common training signals and formats each break one requirement of a traceable answer: outcome reward leaves no cited path, entailment credit rewards copied claims, citation credit rewards broad checks, and one-context formats contract search. Answering each shortcut with a matching design, a typed ledger executed in stages elicits search before any training, and guarded claim credit then completes more evidence paths without copying. Against flat citation training no arm dominates: LEDGERRL leads in complete paths at claim-level resolution and trails in final-answer granularity, precision, accuracy, and tokens. Measuring correctness, completeness, and check size jointly is what makes this trade-off visible; accuracy or citation precision alone would hide it. The remaining gaps suggest two directions: rewards that make final rationale units as fine as the claims beneath them, and training on evidence-labelled multi-hop tasks instead of zero-shot transfer.

AI USE STATEMENT

In this work, we used generative AI writing assistants for polishing the English of manuscript drafts. We have not used generative AI tools for generating research ideas, hypotheses, or experimental designs, and no reported number was produced by a generative AI tool: all results come from the deterministic training and evaluation pipelines described in the paper. We have reviewed all AI-assisted work: AI-assisted code was checked against the written specification and is covered by the property tests and the coordinate-reflection audit, and AI-polished text was revised and approved by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendix A specifies the arms, training data, rewards, prompts, optimisation settings, retrieval pools, and cost reconstruction; Appendix B gives the TC protocol, fixed before scoring, and every TC variant. The artifact includes the stage prompts, validation code, reward scorer, corpus-construction scripts, frozen evaluation outputs, and the analysis scripts that produce every table.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370, 2021. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020. arXiv:2011.01060.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.04315.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021.

Zeping Li, Hongru Wang, Yiwen Zhao, Guanhua Chen, Yixia Li, Keyang Chen, Yixin Cao, Guangnan Ye, Hongfeng Chai, and Zhenfei Yin. Rethinking the role of entropy in optimizing tool-use behaviors for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. arXiv:1902.01007.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2201.03544.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*, 2021.

Jingyi Ren, Yekun Xu, Xiaolong Wang, Weitao Li, Ante Wang, Weizhi Ma, and Yang Liu. Towards transparent RAG: Fostering evidence traceability in LLM generation via reinforcement learning. *arXiv preprint arXiv:2505.13258*, 2025.

Tobias Schreieder, Tim Schopf, and Michael Färber. Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. *arXiv preprint arXiv:2508.15396*, 2026.

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. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3344, 2024. doi: 10.18653/v1/2024.acl-long.182. URL https://aclanthology.org/2024.acl-long.182/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics (TACL)*, 2022. arXiv:2108.00573.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. *arXiv preprint arXiv:2409.12822*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, 2023. doi: 10.18653/v1/2023.findings-emnlp.307. URL https://aclanthology.org/2023.findings-emnlp.307/.

Yiwen Zhao, Zhihao Wen, Yuchen Mao, Mingxuan Jiang, Yihao Hu, Pan Wang, Xin Zhang, and Wei Wu. Towards better agents for multi-turn user interaction: The next user turn is more than context. *arXiv preprint arXiv:2608.17499*, 2026.

A REPRODUCIBILITY

Comparison arms. Table 5 lists the ledger variants, internal controls, and external baselines. SEARCH-RL and CITED-RL are our implementations of the search-RL and citation-RL paradigms in the shared setting.

Table 5: **Arms and controls.** Single-guard and single-context comparisons use step 100; the paired OOD comparison uses step 150 with seeds 1234 and 2025, and also seed 7 for LEDGERRL and the unguarded control. The Llama replication is described in Appendix I.

Arm	Comparison tested
LEDGERRL W/O GUARDS	Same protocol; copy, quantity, usage, and redundancy guards disabled.
Single-guard ablations	Remove anti-copy, the used-claim gate, the quantity term, or the redundancy penalty.
OUTCOME-ONLY LEDGER	Same protocol; outcome reward broadcast to all stages, process terms removed.
FREE-PROSE STAGES	Same stages communicating in free prose; provenance links are extracted after generation.
SINGLE-CONTEXT LEDGER	One context writes all row types; same schema and reward terms.
UNTRAINED	Untrained policy, evaluated under the staged, single-context, and flat citation protocols.
SEARCH-RL / CITED-RL	Single-agent search with outcome reward / sentence-level citations with citation rewards.
Citation cap	CITED-RL reward plus a soft cap of 2 citations per rationale item (penalty 0.1).
LEDGERRL	Staged ledger: investigation, assessment, synthesis; full process rewards.

Policy and optimisation. Qwen3-8B, GRPO in `verl` 0.7.1 (FSDP2 with colocated vLLM 0.8.5 rollout), up to 150 steps (checkpoints specified per comparison), batch 24, n=8 rollouts per question, learning rate 10^{-6}, and the same number of rollouts per data row across arms. Each ledger training question is expanded into three rows, one per target stage (the `target_role` field in code); each rollout executes the protocol but only the selected stage's generated tokens enter its loss. The entire sequence, including upstream stages, is sampled afresh for each of the eight continuations; a group does not reuse one fixed upstream record. GRPO groups those eight samples for that question–stage row. Maximum prompt and response lengths are 6,144 and 8,192 tokens, respectively; training temperature is 1.0, with thinking disabled. The actor uses a KL loss of weight 0.001 and no entropy bonus. Equal row and rollout counts do not imply equal total generation tokens across protocols. Retrieval is local BM25 (`bm25s`) over 125,847 medical passages; no arm has web access at train or eval time. Runs use three or four GPUs with the global batch fixed at 24.

Training data. All arms train on MedQA training questions without evidence labels. The staged ledger arms (LEDGERRL, its ablations, and the outcome-only, unguarded, and free-prose controls)

use 2,000 questions, each expanded into one row per stage (6,000 rows); the single-context ledger uses one row per question for the same 2,000. The flat arms (SEARCH-RL, CITED-RL, and the citation cap) use 6,000 distinct questions that include these 2,000. At batch 24 and 150 steps each arm samples 3,600 rows, so the flat arms see more distinct questions.

Baseline rewards. SEARCH-RL is rewarded by answer correctness alone. CITED-RL receives $\text{outcome} + 0.4\,\text{validity} + 0.2\,\text{coverage}$, minus 0.2 when it gives no answer, where validity is the fraction of cited rationale sentences entailed by their cited snippets (the reward judge at threshold 0.5) and coverage the fraction of rationale sentences with a valid citation. The citation cap additionally subtracts the excess-citation penalty below.

Prompts. The prompts that set the untrained protocols differ in who searches. The flat citation prompt lets the model search ("one query per turn, max 4 searches") and asks it to answer "when confident", citing snippet ids for every rationale fact. The staged investigator is told that its job is to "gather evidence" and "post verifiable claims", to "search, then post claims", and to write 3–8 atomic claims citing at most three evidence ids each. The assessment pass labels each claim `support`, `refute`, or `insufficient` from its cited evidence only, and the synthesiser answers from claims marked `support`, citing claim ids in every rationale sentence. The complete prompts are in the experiment repository.

Citation cap. The cap penalty is $\lambda \sum_i \max(0, |C_i| - 2) / \max(1, \sum_i |C_i|)$ over the model's rationale-item citation lists C_i, added to the CITED-RL reward with $\lambda = 0.1$. Cap runs use the medical training data, four-GPU configuration, batch, rollouts, learning rate, and 150 steps of the other arms; the coefficient is stored explicitly in every training and validation record. Evaluation uses the greedy cited-search protocol with at most four searches and four results per search.

OOD retrieval pools and annotation mapping. Each dataset uses one pooled static BM25 index, built from the contexts or filings of its selected evaluation subset. The search endpoint receives only a query and a return count; it does not restrict candidates by question id. The indices contain 20,792 HotpotQA sentence entries from 500 questions, 15,991 2Wiki sentence entries from 500, 9,998 MuSiQue paragraph entries from 500, and 9,392 FinQA text/table entries from 300. The primary profile evaluates the first 200 questions per corpus against these larger pools. BM25 uses English stopword removal and stemming, returns at most four entries per search, and permits at most four searches. The medical passage count above describes training and MedQA, not these OOD search pools.

HotpotQA and 2Wiki construction maps each supporting-fact `(title, sentence-index)` pair to a sentence snippet id while pooling all supplied context sentences. MuSiQue retains paragraph units and maps `is_supporting` annotations to paragraph ids. FinQA indexes `pre_text`, `post_text`, and table rows (joining their cell text), mapping `gold_inds` text/table keys to snippet ids. If a FinQA gold key is absent from the document lists, construction also includes the text supplied by that gold indicator. Thus these are benchmark-derived evidence pools, not unprocessed financial-document search. The numeric answer program is not executed by these policies. Scoring uses the original snippet ids, not equivalence classes of identical text; the 2Wiki index contains 44 duplicate-id rows, all with identical content.

OOD cost reconstruction. For staged runs, each saved system/user/response turn is counted once. For flat runs, each assistant response is a call whose prompt is the entire preceding message history. Prompts are retokenized with the recorded Qwen chat-template settings and thinking disabled; responses use their saved decoded text, so unrecorded terminal tokens are not recovered. BM25 calls are counted separately from model calls. No extra truncation is applied when reconstructing these actual stored requests.

Reward judge. Entailment and contradiction use the following checkpoint, each at threshold 0.5:

`MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`

Reward weights. The investigator uses $w_g = 0.5$, $w_n = 0.2$, $w_r = 0.3$, and gated outcome weight $w_o = 0.4$. Quantity saturates at $K_c = 3$, with $m_{\min} = 3$. Copy thresholds are 0.40 and 0.80;

duplicate claims have content-token Jaccard above 0.60. Search costs 0.05 beyond the first call, with at most four calls. The synthesiser uses outcome weight 1.0, citation validity 0.4, coverage 0.2, and a target of at least two distinct cited claim ids. Auditor agreement has weight 1.0. An exploratory variant that additionally rewards the auditor for flagging a programmatically corrupted evidence row did not learn detection and is not among the reported arms. The validator rejects malformed rows. The reward subtracts 0.2 when the investigator produces no claims, the synthesiser produces no answer, or the auditor does not cover every claim. It does not zero all process rewards merely because another row was rejected. OUTCOME-ONLY LEDGER broadcasts outcome uniformly to the three stages and removes all process terms.

The complete investigator reward. Equation 4 shows the three guard terms. In full, with $\mathcal{Z} = \{i : z_i{=}1\}$,

$$\begin{aligned} r_{\text{inv}} \;&= w_g \Big[\tfrac{1}{m}\textstyle\sum_i g_i\,\kappa(v_i)\Big]\,\min\!\big(1, \tfrac{m}{m_{\min}}\big) \;+\; w_n \tfrac{\min(n_{\text{good}}, K_c)}{K_c} \\ &\quad+\; w_u\,\frac{|\{i \in \mathcal{Z} : g_i\kappa(v_i) \geq \frac{1}{2}\}|}{|\mathcal{Z}|}\cdot \text{outcome} \;+\; w_o\,\text{outcome}\cdot \mathbb{1}[\mathcal{Z} \neq \emptyset] \\ &\quad-\; c\cdot\max(0, n_{\text{search}} - 1) \;-\; w_r\,\text{redundancy}(L), \end{aligned} \tag{5}$$

with $w_u{=}0.3$ the evidence-utility weight and $c{=}0.05$ the per-search cost. Empty denominators contribute zero; an investigator with no accepted claims also receives the missing-claims penalty.

For synthesis, let J index rationale items and J_c those with valid claim citations for which an NLI pair is formed. The semantic citation term is the mean over $j \in J_c$ of the entailment indicator multiplied by the copy discount between the rationale item and its cited claims. Citation coverage is the fraction of rationale items with nonempty citation lists, multiplied by $\min(1, n_c/2)$, where n_c is the number of distinct claim ids cited by the answer. The synthesiser reward is outcome plus 0.4 times semantic citation credit plus 0.2 times coverage. Empty means contribute zero. NLI training premises are capped at $4{,}000$ characters and tokenized to at most $1{,}024$ tokens. Training contexts cap each evidence text at 700 characters; staged evaluation contexts retain the full stored evidence.

Evaluation and measurements. Evaluations use greedy decoding and the first 200 questions per corpus; full MedQA accuracy uses all $1{,}273$ test questions, and larger OOD replications use 500 questions per corpus (300 on FinQA). Outputs with missing or invalid rows are retained rather than replaced, and a missing final answer counts as a failure (incorrect, untraceable, and uncited) for every arm; schema errors and uncited answers are distinct measurements. Claim behaviour on MedQA is measured by CITEP, CITEP$_{\text{abs}}$, verbatim overlap (the longest shared token run divided by the claim length), and claims per question. Out of domain, gold precision is the share of cited sources in the annotated evidence set, R/CEIL is the cited share of the gold evidence present in the retrieved context, and the retrieval ceiling is the share of all gold evidence retrieved; an answer is uncited when no source is reachable from it. LEDGERRL and the unguarded ledger have step-150 runs with seeds 1234, 2025, and 7 (the unguarded ledger also has four seeds at step 75); the outcome-only ledger, SEARCH-RL, CITED-RL, and the citation cap have seeds 1234 and 2025; single-guard ablations, free-prose stages, and the variant without assessment (Appendix G) have one run. Reproducing the runs requires the stage prompts, validation code, scorer, and corpus construction scripts from the experiment repository.

B TRACEABLE CORRECTNESS: FULL RESULTS

Definition and scope. TC@k follows Eq. 2. The definition, the list of k values, the headline $k = 2$, the attrition stages, and the answer-layer sensitivity were fixed in a written protocol before any TC@k value was computed; the untrained single-context, unguarded, and free-prose arms were added afterwards as controls. Records are the first 200 questions of HotpotQA, 2Wiki, and MuSiQue for every run; missing answers count as not traceable. Flat pseudo-claims only resolve citation ids, so a flat check unit is the rationale item with its cited sources. Free-prose provenance is extracted after generation, so its claims serve as check units. The answer-layer variant takes every final rationale item as a check unit, with all sources reachable through it. MuSiQue provides supporting paragraphs (Trivedi et al., 2022), whose coarser units make a fixed k less uniform across corpora.

Intervals are 95% question-paired, corpus-stratified bootstrap intervals (5,000 draws), conditional on the checkpoints; seeds are shown separately.

Table 6: **TC@k (%) for all available arms.** Sources/check is the mean, over all records, of the number of sources cited by the record's largest check unit (zero when a record has no check unit). Step 150 unless noted.

Arm	TC@1↑	TC@2↑	TC@3↑	TC@4↑	TC@∞↑	Answer-layer TC@2↑	Sources/ check
CITED-RL (2 seeds)	4.8	10.2	11.8	12.9	13.4	10.2	2.02
Citation cap (2 seeds)	6.5	11.1	11.5	11.5	11.5	11.1	1.24
LEDGERRL W/O GUARDS (3 seeds)	1.8	1.8	1.8	1.8	1.8	1.8	0.95
Outcome-only ledger (2 seeds)	9.2	12.7	13.2	13.2	13.2	12.0	0.87
Free-prose stages	3.8	5.0	5.2	5.5	5.5	0.3	0.93
UNTRAINED, staged	11.3	13.8	14.2	14.5	14.7	12.5	1.00
SINGLE-CONTEXT LEDGER, step 100 (2 seeds)	2.4	4.9	6.8	7.3	7.3	2.9	1.81
SINGLE-CONTEXT LEDGER, inv.-weighted	3.7	6.3	7.8	8.2	8.2	4.5	1.55
SINGLE-CONTEXT LEDGER, no search cost	1.7	4.8	6.8	7.2	7.2	2.5	2.24
UNTRAINED, single-context	6.3	6.7	7.0	7.0	7.0	6.3	1.04
UNTRAINED, flat citation	3.7	4.8	5.2	5.2	5.2	4.8	0.91
LEDGERRL, step 100	11.2	14.2	16.3	17.2	17.3	12.2	1.48
LEDGERRL (3 seeds)	8.1	15.3	19.2	19.7	19.7	10.6	1.44

Table 7: **Paired differences with 95% intervals.** Arms with several runs enter as the per-question mean of their runs.

Difference (points)	TC@2↑	TC@∞↑	Answer-layer TC@2↑
LEDGERRL − CITED-RL	+5.1 [+3.2, +7.0]	+6.3 [+4.2, +8.4]	+0.3 [-1.5, +2.0]
LEDGERRL − citation cap	+4.2 [+2.1, +6.4]	+8.2 [+5.9, +10.5]	−0.5 [-2.7, +1.4]
LEDGERRL − outcome-only ledger	+2.7 [+0.9, +4.6]	+6.5 [+4.6, +8.4]	−1.4 [-3.2, +0.4]
LEDGERRL − LEDGERRL W/O GUARDS (3 seeds)	+13.5 [+11.3, +15.7]	+17.9 [+15.3, +20.4]	+8.7 [+6.9, +10.6]
LEDGERRL − free-prose stages	+10.3 [+7.7, +12.9]	+14.2 [+11.3, +17.1]	+10.2 [+8.2, +12.3]
LEDGERRL − untrained staged	+1.5 [-0.6, +3.6]	+5.1 [+2.9, +7.2]	−1.9 [-3.9, -0.1]
LEDGERRL − SINGLE-CONTEXT LEDGER (step 100)	+9.2 [+6.7, +11.8]	+10.0 [+7.1, +12.9]	+9.2 [+6.9, +11.7]
Untrained: staged − single	+7.2 [+4.8, +9.7]	+7.7 [+5.2, +10.2]	+6.2 [+4.0, +8.5]
SINGLE-CONTEXT LEDGER: trained − untrained	−1.8 [-3.7, +0.2]	+0.3 [-1.7, +2.5]	−3.4 [-5.2, -1.7]

Table 8: **Intervals for Table 3.** Points; question-paired, corpus-stratified bootstrap, 5000 draws. TC intervals are in Table 7.

LEDGERRL −	Joint EM↑	Cite F1↑	Gold P↑	Uncited↓	EM↑
CITED-RL	+1.1 [-0.7, +2.8]	−2.8 [-4.7, -1.0]	−6.1 [-8.4, -3.8]	+12.0 [+9.8, +14.3]	+1.1 [-1.8, +4.0]
Citation cap	+0.2 [-2.0, +2.1]	−1.3 [-3.4, +0.7]	−10.7 [-13.2, -8.2]	+6.2 [+3.7, +8.6]	−7.9 [-10.8, -5.1]
Outcome-only ledger	−1.2 [-3.0, +0.6]	+5.1 [+3.2, +7.0]	−0.4 [-2.9, +1.9]	−10.5 [-13.1, -7.9]	+0.3 [-2.0, +2.5]
Untrained ledger	−1.7 [-3.7, +0.2]	+1.1 [-1.1, +3.2]	−4.1 [-6.9, -1.3]	−3.3 [-6.2, -0.5]	+0.9 [-1.6, +3.5]
Unguarded ledger	+9.0 [+7.2, +10.8]	+5.9 [+4.2, +7.5]	−9.9 [-12.1, -7.8]	+0.9 [-0.9, +2.8]	+1.4 [-0.4, +3.3]

Table 9: **Attrition from correct to traceably correct answers.** Percent of questions, as in Figure 3(b). The third column requires an anchored complete path ($|E(u)| \geq 1$ for every check unit).

Arm	Correct	+ all gold retrieved	+ all gold on path	+ checks ≤ 2 sources
CITED-RL (2 seeds)	36.0	22.1	13.4	10.2
Citation cap (2 seeds)	45.0	37.1	11.5	11.1
LEDGERRL W/O GUARDS (3 seeds)	35.7	27.8	1.8	1.8
Outcome-only ledger (2 seeds)	36.8	27.0	13.2	12.7
Free-prose stages	37.0	27.7	5.5	5.0
UNTRAINED, staged	36.2	26.0	14.7	13.8
SINGLE-CONTEXT LEDGER, step 100 (2 seeds)	19.2	7.9	7.3	4.9
UNTRAINED, flat citation	20.5	6.7	5.2	4.8
LEDGERRL, step 100	36.2	26.3	17.3	14.2
LEDGERRL (3 seeds)	37.1	29.2	19.7	15.3

Guard ablations at a matched training step. The three step-150 unguarded runs score 1.2, 1.8, and 2.5% at TC@2; the low score is not confined to one seed. The outcome-only ledger reaches 12.7% over two seeds despite having no claim-level entailment credit. At step 100, Table 11 follows

Table 10: **TC@2 (%) by seed and corpus for the step-150 comparison.**

Arm	Seed	HotpotQA	2Wiki	MuSiQue	Mean
CITED-RL	1234	14.0	12.0	7.5	11.2
CITED-RL	2025	14.0	12.0	2.0	9.3
Citation cap	1234	14.5	12.5	5.5	10.8
Citation cap	2025	18.5	12.0	3.5	11.3
LEDGERRL	1234	22.5	21.0	8.0	17.2
LEDGERRL	2025	14.0	17.0	9.5	13.5
LEDGERRL	7	19.5	21.0	5.5	15.3

each single-guard removal from answer correctness through retrieval and onto the answer-reachable citation path. Removing the anti-copy term yields the largest TC@2 decrease (-8.2 points, question-paired interval $[-11.0, -5.5]$). For both it and the fully unguarded ledger, the loss occurs largely before the check-size test: gold sources are retrieved but not placed on the final path. This connects the reward ablation to TC without asserting that copying alone causes every missing citation.

Table 11: **Single-guard ablations at step 100.** OOD percentages over the three TC corpora, one run per arm, 200 questions per corpus. Columns successively impose answer correctness, complete gold retrieval, complete gold path, and bounded checks. The answer-layer variant applies the final check to rationale items.

Step-100 arm	EM↑	Gold retrieved↑	Gold on path↑	TC@2↑	Answer TC@2↑
LEDGERRL W/O GUARDS	35.7	29.5	4.8	4.8	4.8
w/o anti-copy	35.5	28.8	6.0	6.0	5.8
w/o used-claim gate	31.3	24.3	13.3	12.5	11.3
w/o quantity term	37.7	27.2	15.2	13.3	10.7
w/o redundancy penalty	31.3	24.8	15.0	12.8	11.5
LEDGERRL	36.2	26.3	17.3	14.2	12.2

Post-hoc sensitivity to precision, source burden, and support. For each output, define $p_g = |G_q \cap P(a_q)|/|P(a_q)|$, with $p_g = 0$ for an empty path. Starting from the frozen TC@2 indicator, we additionally require $p_g \geq .75$, $p_g = 1$, or $|P(a_q)| \leq 2$. We also replace all check units by final rationale items, or require every final rationale item to have a valid reachable source link and held-out BART-MNLI direct entailment score at least 0.5. The BART condition tests cited rationale text, not whether the public reasoning entails the short final answer. The scores use existing item-level BART outputs matched to the TC rollouts by SHA256; no model was rerun. All variants are post-hoc and use gold snippet ids, without merging content-equivalent occurrence ids. Table 12 reports the observed rates; intervals resample question ids within corpus and keep all runs of an arm for a question together.

Table 12: **Post-hoc sensitivity of TC@2.** Sensitivity rates in percent, with the original TC@2 shown for reference. "+BART" retains the original claim-adjacent check units and adds final-rationale support. The answer-unit and BART variants are separate columns, not jointly imposed.

Arm	TC@2	$+p_g \geq .75$	$+p_g = 1$	$+\lvert P(a)\rvert \leq 2$	Answer unit	+BART
CITED-RL (2 seeds)	10.2	9.6	9.6	9.6	10.2	6.9
Citation cap (2 seeds)	11.1	10.7	10.7	10.6	11.1	8.7
Untrained staged ledger	13.8	12.5	12.5	12.5	12.5	10.2
LEDGERRL (3 seeds)	15.3	11.1	10.8	10.6	10.6	11.7

The LEDGERRL minus citation-cap difference becomes $+0.4$ $[-1.7, 2.4]$ points at $p_g \geq .75$, $+0.2$ $[-2.0, 2.1]$ at $p_g = 1$, 0.0 $[-2.1, 1.9]$ with at most two total sources, and -0.5 $[-2.7, 1.4]$ at the common answer unit. Under the BART condition it is $+3.1$ $[1.1, 5.0]$; this judge-based advantage is sensitive to the entailment model and does not validate the final answer inference. Compared with the untrained staged ledger, the trained model's strict precision-conditioned TC@2 is -1.7 $[-3.7, 0.2]$ points.

Post-hoc units, joint metrics, and evidence use. These measures were added after the TC protocol was fixed. Answer-level TC is the only unit shared by all formats. A two-hop variant that bounds the sources of each cited claim and the claim ids of each rationale item by k lets a ledger rationale item reach up to k^2 sources, so it is not format neutral; at $k = 2$ it gives 13.6% for LEDGERRL,

13.2% for the untrained ledger, and the flat values of Table 1. Uncited answers have no reachable source, with missing answers included, as in Table 22; restricted to answered questions the LEDGERRL, outcome-only, and untrained rates are 15.2%, 21.2%, and 20.8%. Joint EM follows HotpotQA: a correct answer whose answer-reachable source set equals the gold set; joint F1 multiplies answer and supporting-set precision and recall. Gold-citation F1 is the per-question harmonic mean of gold precision and gold coverage of the path. Retrieval share T and cited share of retrieved gold U decompose gold coverage as $V = TU$. All are computed from the same frozen records and gold ids as TC (script and outputs in the experiment archive). Paired differences use the TC bootstrap with question ids sorted within corpus. Relative to the outcome-only ledger, LEDGERRL changes T by $+1.6$ $[0.2, 2.9]$ and U by $+13.1$ $[10.6, 15.7]$, and coverage V by $+10.6$ $[8.7, 12.5]$. Relative to CITED-RL, LEDGERRL's joint EM and joint F1 differences are $+1.1$ $[-0.7, 2.8]$ and $+1.2$ $[-0.8, 3.3]$; relative to the citation cap, $+0.2$ $[-2.0, 2.1]$ and -4.3 $[-6.5, -2.2]$. Relative to the citation cap, LEDGERRL changes T by -9.0 $[-10.8, -7.4]$ and U by $+10.5$ $[7.8, 13.1]$. Table 13 gives the share of final rationale items reaching more than two sources, by run; Table 14 shows the same broad units in domain under other flat citation rewards.

Table 13: **Broad final rationale items.** Final rationale items citing more than two unique sources through any path, pooled over HotpotQA, 2Wiki, and MuSiQue. For ledgers, sources are reached through cited claims.

Arm (run)	Items with > 2 sources (%)	Sources per item
CITED-RL (1234 / 2025)	37.1 / 10.6	2.57 / 1.38
Citation cap (1234 / 2025)	8.6 / 3.5	1.30 / 1.33
Untrained staged ledger	8.9	1.35
Outcome-only ledger (1234 / 2025)	4.6 / 8.6	1.03 / 1.39
LEDGERRL (1234 / 2025 / 7)	26.2 / 43.4 / 18.9	1.85 / 2.50 / 1.47

Broad units under other flat citation rewards. The broad units of S3 are not specific to CITED-RL's coverage term. We also train a flat format with outcome plus entailment credit and no coverage term ($\text{outcome} + 0.4\,\text{cite_valid}$), and a CITED-RL+$\kappa$ arm that adds the copy discount to the standard citation credit. Table 14 reports their MedQA rationale units; copying is measured on model-written text.

Table 14: **Flat citation rewards on MedQA.** $n = 200$, greedy; step 150 unless noted, and the two outcome-plus-entailment rows come from one run. Product is words per rationale unit times citations per unit, a descriptive span statistic rather than reading cost.

Flat reward / run	Acc	Verbatim	Words/unit	Cites/unit	Product
CITED-RL, seed 1234	0.750	0.079	82.8	7.04	583
CITED-RL, seed 2025	0.750	0.165	25.9	1.93	50
Outcome + entailment, step 75	0.775	0.066	100.3	7.14	716
Outcome + entailment, step 150	0.765	0.063	96.5	8.25	797
CITED-RL+κ	0.770	0.054	115.9	5.92	686

Every variant except one CITED-RL seed attaches about six to eight sources to rationale units of 80–120 words, while copying stays low. Removing the coverage term lowers the cited share of retrieved gold out of domain (R/CEIL 0.71–0.88 against 0.79–0.92) and leaves 5–14% of answers uncited against about 2–5%; adding the copy discount does not narrow the units. Entailment credit thus induces copying on ledger claims but not on these flat rationales, whereas broad units recur under several citation rewards; their seed dependence argues against a universal claim.

Segmentation stress tests. We hold each frozen answer and its reachable source set fixed but hypothetically replace its check units by one singleton source unit per existing source link. Under this invalid-as-evidence but informative structural transformation, TC@1 becomes 19.7% for LEDGERRL, 13.4% for CITED-RL, and 11.5% for the citation cap; their original TC@1 values were 8.1%, 4.8%, and 6.5%. If every retrieved source is instead hypothetically cited in its own singleton unit, TC@1 equals correct-answer and complete-gold-retrieval incidence: 29.2%, 22.1%, and 37.1%. Neither transformation creates supported statements or better answers. They expose the score's segmentation and precision incentives rather than new method performance.

Reading the sensitivity analyses. With claims as check units, all three LEDGERRL seeds exceed every seed of the paired CITED-RL and citation-cap baselines at $k = 2$, and the three-seed mean leads on each corpus; in single cells both citation-cap seeds exceed LEDGERRL seed 2025 on HotpotQA (14.5 and 18.5 versus 14.0), and CITED-RL seed 1234 exceeds seed 7 on MuSiQue (7.5 versus 5.5). At the answer layer, LEDGERRL is tied with the flat step-150 baselines and below its own untrained staged ledger: trained rationale items cite several claims. Training the staged ledger increases completeness but also claim size. The unguarded and free-prose controls stay far below LEDGERRL under every variant, including the answer layer.

C SEARCH UNDER STAGED AND SINGLE-CONTEXT EXECUTION

Protocols before training. Table 15 compares the untrained model under the staged and single-context ledger protocols, and the corresponding step-100 policies, on the same questions. Under CITED-RL's flat citation prompt the same untrained model makes 0.80 searches per question over HotpotQA, 2Wiki, and MuSiQue and none on 29.0% of questions (0.69, 1.00, and 0.69 searches and 38, 18, and 31% without search, respectively; FinQA 0.93 and 8.5%). It gives an answer on 99.3% of these questions, so the low search count is not a parsing failure.

Table 15: **Search and evidence use by protocol.** Protocol differences before training and after step 100 (seed 1234 for the single-context ledger), same questions. Searches count calls, not evidence rows. Gold R is unconditional gold recall; the ceiling is the share of gold evidence retrieved.

Protocol / training	Searches/q	Gold P↑	Gold R↑	Ceiling↑	EM↑
HotpotQA					
Staged / untrained	3.18	0.665	0.465	0.704	0.450
Single-context / untrained	1.33	0.621	0.388	0.578	0.380
Single-context / step 100	1.07	0.358	0.424	0.504	0.285
Staged / step 100	3.02	0.699	0.548	0.721	0.440
2Wiki					
Staged / untrained	3.73	0.586	0.437	0.657	0.370
Single-context / untrained	1.72	0.560	0.360	0.563	0.315
Single-context / step 100	1.29	0.348	0.393	0.469	0.140
Staged / step 100	3.58	0.589	0.523	0.653	0.395
MuSiQue					
Staged / untrained	3.39	0.462	0.340	0.635	0.265
Single-context / untrained	1.10	0.487	0.290	0.490	0.125
Single-context / step 100	1.00	0.346	0.390	0.468	0.090
Staged / step 100	2.96	0.525	0.448	0.608	0.250
FinQA					
Staged / untrained	3.47	0.232	0.233	0.375	0.035
Single-context / untrained	2.02	0.211	0.206	0.358	0.025
Single-context / step 100	1.03	0.172	0.320	0.341	0.030
Staged / step 100	3.28	0.251	0.306	0.370	0.035

Single-context training. SINGLE-CONTEXT LEDGER uses one conversation to search, write claims, record verdicts, and answer, with the schema, validator, and reward terms of LEDGERRL. Seeds 1234 and 2025 use $(r_{\text{inv}} + r_{\text{aud}} + r_{\text{syn}})/3$. An investigator-weighted variant uses $r_{\text{inv}} + (r_{\text{aud}} + r_{\text{syn}})/3$, and a no-search-cost variant removes the $0.05 \max(0, n_{\text{search}} - 1)$ charge throughout training (seed 1234 each). All are trained for 100 steps and evaluated with the single-context prompts, which allow one completion pass for a block missing its verdicts or answer. Table 16 gives the two-seed summary used in §5.1; Tables 17 and 18 give every run.

The step-100 TC@2 values are 4.8% and 5.0% for seeds 1234 and 2025, 6.3% for the investigator-weighted run, and 4.8% without the search cost, against 14.2% for the staged reference (Table 6). Out of domain the single-context runs make 1.0–1.3 searches per question. Neither reweighting the investigator nor removing the search charge restores search; the no-cost run also has low CITEP (0.295) and copying (0.130), so strong grounding is not a general property of single-context training. The untrained single-context ledger has higher gold precision than each run in Table 18 on every

Table 16: **Staged versus single-context ledger.** Step 100, greedy; the single-context row averages two seeds. CITEP$_{abs}$ and searches are on MedQA; OOD ceiling and gold precision give ranges over HotpotQA, 2Wiki, and MuSiQue (precision also FinQA). TC@2 pools the three TC corpora.

Method	CITEP$_{abs}$↑	Searches/q	OOD ceiling↑	OOD gold P↑	TC@2 (%)↑
SINGLE-CONTEXT LEDGER	0.877	1.00	0.46–0.50	0.16–0.37	4.9
LEDGERRL (staged)	0.790	2.57	0.61–0.72	0.25–0.70	14.2

Table 17: **Single-context training on MedQA.** MedQA, $n = 200$, greedy; trained arms at step 100. Every trained single-context run settles at about one search per question.

MedQA arm	Searches/q	Acc↑	CITEP$_{abs}$↑	Claims/q
UNTRAINED, staged	3.84	0.695	0.443	5.62
UNTRAINED, single context	1.56	0.725	0.422	3.15
SINGLE-CONTEXT LEDGER, seed 1234	1.00	0.635	0.861	3.08
SINGLE-CONTEXT LEDGER, seed 2025	1.00	0.715	0.892	2.99
SINGLE-CONTEXT LEDGER, inv.-weighted	1.01	0.675	0.918	3.02
SINGLE-CONTEXT LEDGER, no search cost	1.01	0.720	0.294	3.12
LEDGERRL	2.57	0.700	0.790	4.54

Table 18: **Single-context runs out of domain.** $n = 200$ per corpus, greedy, step 100. The staged step-100 reference is in Table 15; every single-context run has lower gold precision, ceiling, and EM than it on every corpus.

SINGLE-CONTEXT LEDGER run	Gold P↑	R/CEIL↑	Ceiling↑	Uncited↓	EM↑
HotpotQA					
seed 1234	0.358	0.842	0.504	0.165	0.285
seed 2025	0.355	0.932	0.493	0.040	0.330
inv.-weighted	0.348	0.848	0.517	0.135	0.295
2Wiki					
seed 1234	0.348	0.837	0.469	0.195	0.140
seed 2025	0.384	0.888	0.469	0.030	0.215
inv.-weighted	0.387	0.817	0.491	0.180	0.195
MuSiQue					
seed 1234	0.346	0.834	0.468	0.090	0.090
seed 2025	0.315	0.906	0.450	0.030	0.090
inv.-weighted	0.345	0.868	0.475	0.035	0.070
FinQA					
seed 1234	0.172	0.937	0.341	0.110	0.030
seed 2025	0.155	0.934	0.341	0.000	0.025
inv.-weighted	0.149	0.938	0.351	0.185	0.020

OOD corpus: training lowers its precision, while the staged protocol already searches more before training.

An external four-search requirement. Under the shared prompts the staged ledger searches more than flat citation. To separate eliciting search from using a fixed search budget, we impose one rule on both formats at inference: exactly four searches before the answer (flat) or before any claim is posted (ledger). An earlier answer or claim is rejected with the same request for one new search; assessment and synthesis are unchanged, and every completed request meets the rule. Table 19 reports each protocol with and without it.

The rule helps the flat protocols most, because they otherwise search less: TC@2 rises from 4.8% to 19.3% for the untrained flat prompt and by 6.0 points for CITED-RL ($[4.0, 8.1]$), while the citation cap, which already searches 3.56 times per question, is unchanged (-0.1, $[-1.5, 1.2]$). The ledger protocols already search 3.4–3.9 times and do not gain; trained LEDGERRL loses 2.6 points ($[-4.1, -1.1]$). With four searches in both formats, flat citation is more often traceably correct than the staged ledger: by 7.0 points untrained ($[4.0, 10.0]$), and CITED-RL exceeds LEDGERRL by 3.5 at TC@2 ($[1.5, 5.5]$) and 5.9 at TC@∞ ($[3.6, 8.4]$) while retrieving 9.2 points more of the gold evidence. LEDGERRL keeps its TC@∞ lead over the citation cap ($+5.6$, $[3.5, 7.7]$) and ties it at TC@2 ($+1.8$, $[-0.2, 3.7]$). The staged ledger's advantage under the shared prompts thus comes

Table 19: **An external four-search requirement.** Each protocol with its native search behaviour and with exactly four searches (%; HotpotQA, 2Wiki, MuSiQue, 600 questions, greedy; trained arms at step 150). Gold retrieved is the retrieved share of gold evidence (T); uncited includes missing answers. Two citation-cap requests under the rule exceed the context window and count as failures.

Arm	Searches	Searches/q	EM↑	TC@2↑	Answer TC@2↑	TC@∞↑	Uncited↓	Gold retrieved↑
Untrained								
UNTRAINED, staged	native	3.44	36.2	13.8	12.5	14.7	21.5	66.5
	four	4.00	35.7	12.3	11.8	13.5	28.7	68.5
UNTRAINED, flat citation	native	0.80	20.5	4.8	4.8	5.2	33.5	35.1
	four	4.00	45.8	19.3	19.3	20.3	0.5	77.7
Trained, step 150								
CITED-RL (2 seeds)	native	1.90	36.0	10.2	10.2	13.4	6.2	62.8
	four	4.00	44.0	16.2	16.2	23.1	3.9	78.6
Citation cap (2 seeds)	native	3.56	45.0	11.1	11.1	11.5	12.0	77.3
	four	3.99	45.4	11.0	11.0	11.6	3.2	79.0
LEDGERRL (3 seeds)	native	3.89	37.1	15.3	10.6	19.7	18.2	68.2
	four	4.00	36.6	12.8	8.9	17.2	19.7	69.3

from eliciting search without an external rule, not from better use of an equal search budget; where such a rule is acceptable, flat citation reaches higher TC with it.

D COPYING AND UNCITED ANSWERS WITHOUT THE GUARDS

Copying across seeds. Table 20 expands the unguarded comparison of Table 2. All four seeds reach step 75; three continue to 150.

Table 20: **The unguarded reward on four seeds.** MedQA, $n = 200$, greedy. Bold marks CITEP$_{abs}$ below the untrained 0.443.

Seed	CITEP	**CITEP$_{abs}$↑**	Verbatim↓	Claims/q	Acc↑
Step 75					
1234	0.871	**0.329**	0.686	5.23	0.705
2025	0.923	0.549	0.556	6.53	0.735
7	0.868	**0.362**	0.652	5.66	0.640
11	0.888	**0.361**	0.674	5.01	0.705
Step 150					
1234	0.852	**0.167**	0.809	3.73	0.650
2025	0.920	0.546	0.566	10.62	0.685
7	0.919	**0.343**	0.702	4.82	0.685

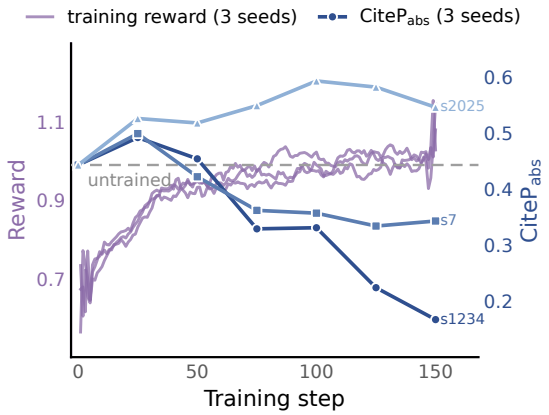

Figure 4: Training reward and copy-discounted claim grounding (CITEP$_{abs}$) for the three unguarded seeds continued to step 150. Two trajectories end below the untrained reference; seed 2025 stays above it while its claim count grows.

At step 75, all four seeds increase entailment and copying. CITEP$_{abs}$ is below its untrained value of 0.443 on three seeds and above it on seed 2025, which instead grows to 10.6 claims per question by step 150. Copying thus replicates, while the copy-discounted score does not decline uniformly; its relation to overlap is partly built into the metric. Figure 4 shows the training trajectories: the logged reward rises in every run while copy-discounted grounding falls below the untrained reference in two of the three runs continued to step 150. The outcome-only control copies little but also has lower CITEP$_{abs}$ than untrained (0.353 versus 0.443), through lower entailment.

A held-out judge. We replace the DeBERTa reward judge with `facebook/bart-large-mnli`, leaving rollouts, thresholds, and the copy discount unchanged. BART differs in architecture and weights and was not used in policy training; both judges include MNLI in their training data, so the supervision is not disjoint. For guarded arms CITEP$_{abs}$ is an optimisation target under the original judge, so this transfer complements the human-annotated evaluation.

Table 21: **Claim grounding under a held-out judge.** MedQA, $n = 200$, greedy. The three unguarded seeds retain their CITEP$_{abs}$ ordering and move by at most 0.020 on that metric; LEDGERRL moves from 0.790 to 0.807. This tests judge transfer, not whether CITEP$_{abs}$ is a complete traceability measure.

	CITEP		**CITEP$_{abs}$↑**	
	reward's NLI	BART	reward's NLI	BART
UNTRAINED	0.497	0.480	0.443	0.423
OUTCOME-ONLY LEDGER	0.370	0.359	0.353	0.341
LEDGERRL W/O GUARDS (s1234, step 100)	0.886	0.903	**0.331**	**0.344**
same, seed 7 (step 100)	0.889	0.916	**0.357**	**0.374**
same, seed 2025 (step 100)	0.931	0.906	0.593	0.573
LEDGERRL	0.886	0.911	**0.790**	**0.807**

Larger samples and a matched step. Table 22 re-evaluates the untrained model, the step-100 ledger checkpoints, and CITED-RL on 500 questions per corpus (300 on FinQA), and compares the guarded and unguarded ledgers at the same seed and step. The unguarded ledger's lower cited share of retrieved gold persists in both settings, while uncited-answer rates vary more across samples.

Table 22: **Larger samples and a matched seed and step.** Greedy decoding. In the matched pair the guarded-minus-unguarded R/CEIL difference ranges from 0.303 to 0.445 across corpora; uncited-answer rates improve less than in the n=200 samples.

Metric	Arm	HotpotQA	2Wiki	MuSiQue	FinQA
		n=500 **per corpus** (300 **FinQA**)			
R/CEIL	UNTRAINED	0.642	0.618	0.436	0.620
R/CEIL	LEDGERRL W/O GUARDS	0.324	0.188	0.248	0.324
R/CEIL	CITED-RL	0.851	0.815	0.800	0.930
R/CEIL	LEDGERRL	0.747	0.748	0.673	0.785
uncited	UNTRAINED	0.200	0.332	0.338	0.537
uncited	LEDGERRL W/O GUARDS	0.408	0.706	0.464	0.560
uncited	CITED-RL	0.012	0.030	0.028	0.020
uncited	LEDGERRL	0.070	0.158	0.082	0.167
	Matched seed and step: seed 1234, step 150, n=200				
R/CEIL	LEDGERRL W/O GUARDS	0.437	0.367	0.314	0.371
R/CEIL	LEDGERRL	0.740	0.702	0.696	0.816
uncited	LEDGERRL W/O GUARDS	0.155	0.310	0.340	0.240
uncited	LEDGERRL	0.130	0.250	0.155	0.205

E COMMON-STEP PROFILE, COVERAGE DECOMPOSITION, AND COST

Table 23 aligns CITED-RL, the citation cap, and LEDGERRL at step 150 on the first 200 questions of HotpotQA, 2Wiki, MuSiQue, and FinQA (seeds 1234 and 2025, and seed 7 for LEDGERRL). Per-corpus means are equally averaged over the four corpora and the runs of each arm; we do not combine metrics into a score. For each final rationale item we resolve citations through any

claims to unique source snippet ids, so claim ids and repeated references are not counted as sources. Support scores native rationale items under frozen BART-MNLI (held out of training) and DeBERTa (the reward-judge family), with a 700-character evidence cap and threshold 0.5. The span proxy is the product of mean words and mean unique sources over final items with a reachable source; it describes item breadth, not reading time. Costs follow Appendix A.

Table 23: **Four-corpus profile.** OOD means at step 150 over four corpora (three runs for LEDGERRL, two for each flat arm). V is unconditional gold coverage, T the retrieved share of gold, and U the cited share of retrieved gold ($U = 0$ when no gold is retrieved); these are question-level means, so TU need not equal V. No source counts questions without a reachable citation, including missing final outputs.

Method	EM↑	V↑	T↑	U↑	Gold P↑	No source↓	BART↑	DeBERTa↑	Span × sources
CITED-RL	.276	.444	.562	.629	.505	6.9%	.361	.566	46.72
Citation cap	.344	.380	.676	.457	.548	12.5%	.481	.643	43.39
LEDGERRL	.285	.433	.609	.542	.459	20.3%	.419	.507	34.81

Table 24: **Cost at the operating points of Table 23.** All 5,600 records, including missing final predictions, are retained; counts include recorded retry calls and repeatedly supplied context. Tokens are decoded-text estimates, not cache-adjusted compute or measured latency.

Method	Model calls/q↓	Searches/q	Prompt tokens/q↓	Generated tokens/q↓
CITED-RL	3.143	1.994	1,871	188
Citation cap	4.727	3.603	3,862	230
LEDGERRL	6.915	3.878	7,044	909

Retrieval versus use of retrieved gold. For gold, retrieved, and answer-reachable source sets G, R, C of a question, $C \subseteq R$ holds for every record and $V = |G \cap C|/|G| = TU$ with $T = |G \cap R|/|G|$ and $U = |G \cap C|/|G \cap R|$. For the ledger L and the citation cap F on the same question, the exact identity

$$V_L - V_F = (T_L - T_F)\frac{U_L + U_F}{2} + (U_L - U_F)\frac{T_L + T_F}{2}$$

splits the coverage difference symmetrically. The averaged retrieval term is -3.08 points and the use term $+8.31$, summing to $+5.23$ ($[3.28, 7.15]$; question-id bootstrap within corpus, 5,000 draws, runs of each arm averaged within question). Every LEDGERRL run has a negative retrieval term and a positive use term: the ledger retrieves less of the gold evidence than the citation cap but places more of what it retrieves on the answer path. This is algebraic accounting, not causal mediation.

Reading the profile. Per run, LEDGERRL minus the citation cap's two-run mean changes EM by -6.7, -4.4, and -6.7 points (seeds 1234, 2025, 7) and coverage by $+6.7$, $+8.6$, and $+0.3$; the three-run coverage difference is positive on every corpus. Over four corpora the citation cap is more accurate (-5.9 for LEDGERRL, $[-8.1, -3.8]$) and has higher gold precision and DeBERTa support, while LEDGERRL uses 1.94 times its tokens, 1.46 times its calls, and 1.08 times its searches. Relative to CITED-RL, LEDGERRL's coverage is tied (-1.1, $[-2.9, 0.7]$) at 3.9 times its tokens. The citation cap's final items cite fewer sources than LEDGERRL's (Table 13) but contain more words, so its span proxy is higher.

F SEMANTIC SUPPORT OF FINAL RATIONALES AND LEDGER EDGES

Gold precision measures membership in a question's annotated evidence set, not support for a generated assertion. For rationale item a_j with reachable cited evidence E_j, we measure whether E_j entails a_j, averaging within each question and then over all 200 questions per corpus; missing answers and invalid citation paths receive zero. The primary judge is frozen BART-MNLI, held out of policy training; the training judge, DeBERTa, is a consistency check. Both use threshold 0.5, no copy discount, at most 700 characters per evidence snippet, and at most 1,024 tokens per pair, truncating the premise and never the hypothesis. Neither the question nor the gold answer enters the premise. Flat pseudo-claims only resolve citations, so every format is scored on its actual rationale text.

Table 25: **Final-rationale support from reachable cited evidence.** Question-macro means including failures; step 150, $n = 200$ per corpus, greedy; CITED-RL has no seed-7 run. These are model-based support judgements, not human verification.

Arm / seed	HotpotQA	2Wiki	MuSiQue	FinQA
BART				
CITED-RL / 1234	0.460	0.312	0.453	0.120
CITED-RL / 2025	0.527	0.338	0.492	0.183
LEDGERRL / 1234	0.530	0.300	0.475	0.155
LEDGERRL / 2025	0.650	0.545	0.765	0.380
LEDGERRL / 7	0.465	0.250	0.395	0.115
DeBERTa				
CITED-RL / 1234	0.620	0.542	0.713	0.560
CITED-RL / 2025	0.618	0.440	0.721	0.313
LEDGERRL / 1234	0.615	0.420	0.590	0.205
LEDGERRL / 2025	0.775	0.670	0.870	0.420
LEDGERRL / 7	0.515	0.335	0.490	0.180

Paired question-bootstrap intervals (10,000 resamples) show that the comparison depends on the run. Under BART, LEDGERRL minus CITED-RL is $+0.070$ $[-0.007, 0.147]$ and $+0.122$ $[0.045, 0.200]$ on HotpotQA for seeds 1234 and 2025, and $+0.022$ $[-0.060, 0.105]$ and $+0.273$ $[0.190, 0.354]$ on MuSiQue. Under DeBERTa, seed 1234 is tied on HotpotQA and lower on 2Wiki, MuSiQue, and FinQA (for example -0.122 $[-0.210, -0.037]$ on MuSiQue), whereas seed 2025 is higher on every corpus. LEDGERRL seed 7 is below the two-run CITED-RL mean on every corpus under both judges. Neither consistently better support nor seed-robust agreement between the judges is established.

Both ledger edges. We separately check whether cited claims support each rationale item and whether each of those claims is supported by its own evidence. For the step-100 staged run, BART's claim-to-rationale support is $0.750/0.615/0.780/0.340$ on HotpotQA/2Wiki/MuSiQue/FinQA; requiring in addition that every cited claim be supported gives $0.460/0.335/0.640/0.135$. On this two-edge diagnostic the untrained ledger reaches $0.245/0.200/0.265/0.065$ and the single-context ledger $0.430/0.400/0.448/0.325$. Grounded intermediate claims therefore do not guarantee a supported path, and the staged protocol does not improve every edge or corpus.

G VERDICT INTERVENTIONS

The step-100 LEDGERRL policy regenerates answers to 200 HotpotQA and 200 2Wiki questions from fixed investigator outputs. Every condition removes the original answer and uses the native synthesiser prompt, a 700-character evidence cap, greedy decoding without thinking, 1,024 output tokens, and one generation, so comparisons are within this experiment. Conditions retain the original verdicts, remove them, replace their labels by `support` with a neutral note, or permute complete verdict records across claim ids.

Table 26: **Inference-time verdict interventions.** $n = 200$ per corpus. Invalid answers count as incorrect; validity uses the native schema, including claim-id resolution.

	HotpotQA		2Wiki	
Condition	EM	Valid answer	EM	Valid answer
No verdicts	0.435	0.955	0.355	0.845
All support	0.430	0.960	0.345	0.830
Permuted verdicts	0.435	0.950	0.350	0.855
Original verdicts	0.435	0.940	0.345	0.835

Removing verdicts changes EM by 0.000 on HotpotQA and $+0.010$ on 2Wiki, and the all-support and permutation contrasts also have paired intervals that include zero. The trained assessment pass labels 92.7–96.9% of its own OOD claims `support`; only nine of 2,935 receive `refute`. The assessment input thus shows no EM benefit at inference, which does not test assessment-stage training.

One-claim support withdrawal. The test in §6 selects a previously supported cited claim and the closest-length supported uncited claim, deterministically by question id. Only the selected verdict becomes `insufficient`, with the same neutral note; evidence and claim text are unchanged. Eligibility is $161/200$ and $135/200$, and mean cited/control lengths are $11.39/11.58$ and $9.90/10.05$ words. The EM differences are $+0.006$ and -0.007, with intervals including zero. Among questions whose regenerated original-verdict answer is valid and cites the target (156 HotpotQA, 121 2Wiki), we test whether the new rationale still entails each original rationale item citing the target. On HotpotQA this retention is 0.673 after target flags versus 0.859 after control flags under BART (-0.186, $[-0.263, -0.115]$) and 0.705 versus 0.885 under DeBERTa (-0.179, $[-0.256, -0.103]$); on 2Wiki both judges give -0.083 $[-0.174, 0.000]$. The withdrawal is synthetic and does not establish that the targeted fact was false.

Removing the assessment stage. A two-stage variant drops the assessment pass from training and evaluation, so the synthesiser answers from unassessed claims (seed 1234, step 150). The change also removes the assessment calls and their training rows, so it tests the stage within the protocol rather than the verdict input alone. Table 27 compares it with the full protocol for the same seed.

Table 27: **LEDGERRL with and without the assessment stage.** Seed 1234, step 150; HotpotQA, 2Wiki, and MuSiQue (600 questions). Both make 3.8 searches per question.

Metric (%)	Full protocol	Without assessment	Difference [95% interval]
EM	36.2	33.8	-2.3 $[-5.5, +0.8]$
TC@2	17.2	12.3	-4.8 $[-7.3, -2.3]$
TC@∞	20.8	15.7	-5.2 $[-8.0, -2.5]$
Answer TC@2	11.7	8.8	-2.8 $[-5.0, -0.7]$
Joint EM	12.0	9.2	-2.8 $[-5.0, -0.7]$
Cite F1	49.6	45.1	-4.5 $[-6.9, -2.0]$
Gold retrieved (T)	68.9	68.0	-0.9 $[-2.6, +0.9]$
Cited share of retrieved gold (U)	62.3	56.1	-6.2 $[-9.7, -2.9]$
Uncited	17.8	19.8	$+2.0$ $[-1.5, +5.5]$

Without the stage, retrieval is unchanged but less of the retrieved gold reaches the answer path, and TC@2 and TC@∞ fall by about five points. The inference-time interventions above leave EM unchanged; this comparison changes training and inference together and does not separate the two.

H ACCURACY

The full MedQA test set is the primary accuracy comparison. LEDGERRL averages 0.635 over three runs, against 0.741 for SEARCH-RL, 0.681 for CITED-RL, and 0.544 for the citation cap over two seeds each; free-prose stages reach 0.717. The citation cap uses the cited-search protocol of CITED-RL, and a missing final answer counts as incorrect. Its seed-1234 run gives no final answer on 590 of the 1,273 questions: it keeps searching past the four-search limit or writes its answer without the required block. This associates an accuracy cost with the ledger configuration here, without establishing a general cost of structured reasoning.

Corpus dependence. Out of domain, the step-100 LEDGERRL checkpoint scores 0.395 on 2Wiki against 0.505 for SEARCH-RL, and 0.250 against 0.260 on MuSiQue, where the untrained model and the outcome-only ledger score 0.265 and 0.275. Attribution improvements need not improve exact match; small gaps are not significance claims.

I SECOND POLICY FAMILY

We repeat the guarded-versus-unguarded comparison on Llama-3.1-8B-Instruct with the same data, optimiser, and prompts, for one seed; both arms are reported at step 75, and a full-budget replication remains open. Untrained Llama fails validation on 15.3% of evaluated rollouts. Without guards it emits one claim, four evidence rows, one verdict, and 199 answer rows across 200 MedQA questions: training reward approaches 0.99 while the claim record disappears. This is not high entailment on copied claims, since the only claim has CITEP 0. The outcome-only control emits 193 claims but no answer-used claims, with schema-error rate 0.569, so structural failure also occurs without process rewards.

Table 28: **Full MedQA accuracy.** All 1,273 MedQA test questions, greedy, step 150 unless noted, with 95% binomial intervals. CITEP$_{abs}$ and overlap use the same rollouts for ledger arms. Missing final answers count as incorrect.

Arm	Seed	Acc↑	95% interval	CITEP$_{abs}$↑	Verbatim↓
SEARCH-RL	2025	**0.749**	[0.726, 0.773]	—	—
SEARCH-RL	1234	0.733	[0.709, 0.757]	—	—
FREE-PROSE STAGES	single run	0.717	[0.693, 0.742]	0.241	—
Citation cap	2025	0.699	[0.674, 0.724]	—	—
CITED-RL	2025	0.691	[0.665, 0.716]	—	—
CITED-RL	1234	0.671	[0.645, 0.697]	—	—
LEDGERRL W/O GUARDS	2025	0.646	[0.619, 0.672]	0.547	0.572
LEDGERRL W/O GUARDS	7	0.635	[0.609, 0.662]	0.346	0.701
OUTCOME-ONLY LEDGER	1234	0.626	[0.600, 0.653]	0.388	0.194
SINGLE-CONTEXT LEDGER (step 100)	1234	0.599	[0.572, 0.625]	0.866	0.321
OUTCOME-ONLY LEDGER	2025	0.592	[0.565, 0.619]	0.420	0.201
UNTRAINED	—	0.589	[0.562, 0.616]	0.419	0.221
LEDGERRL W/O GUARDS	1234	0.531	[0.504, 0.558]	**0.131**	**0.839**
Citation cap	1234	0.388	[0.361, 0.415]	—	—
LEDGERRL	7	0.656	[0.630, 0.682]	0.810	0.313
LEDGERRL	1234	0.638	[0.612, 0.664]	**0.870**	0.293
LEDGERRL	2025	0.610	[0.584, 0.637]	0.824	0.362

Table 29: **Llama on MedQA.** Greedy, one seed. Trained arms: step 75, $n = 200$; untrained: $n = 100$. †The unguarded entailment and overlap values describe its only emitted claim. ‡Strict option-string matching; normalising an initial option letter yields 0.655 for outcome-only, 0.640 for unguarded, and unchanged 0.605/0.290 for guarded/untrained.

Llama / MedQA	Acc↑	CITEP	**CITEP$_{abs}$**↑	Verbatim↓	Claims/q	Schema err↓
untrained	0.290	0.487	0.430	0.214	7.15	0.153
OUTCOME-ONLY LEDGER step 75	0.000‡	0.461	0.415	0.211	0.96	0.569
LEDGERRL W/O GUARDS step 75	0.630	0.000†	**0.000**†	0.308†	**0.01**	**0.576**
LEDGERRL step 75	0.605	0.868	**0.812**	0.280	11.44	**0.007**

Out of domain ($n = 200$ per corpus, step 75), the unguarded arm retrieves nothing on any corpus and leaves every answer uncited. The guarded arm retrieves 7.0–8.6 evidence rows per question and cites 84–96% of the gold evidence it retrieves, but its gold precision is only 0.065–0.177, and it writes more claims than the Qwen operating point. The guards thus restore ledger production in a second family; this is not a cross-family replication of every traceability result.

J LIMITATIONS IN DETAIL

Accuracy and cost. On all 1,273 MedQA test questions, LEDGERRL averages 0.635 over three runs and has a material in-domain accuracy cost against SEARCH-RL, CITED-RL, and the same stages communicating in free prose; only the citation cap, whose mean is lowered by unanswered questions, is less accurate. OOD EM is corpus-dependent (Appendix H). The staged protocol also uses more tokens and calls than the flat baselines, 1.9 times the tokens of the citation cap (§5.3); equal checkpoint steps are not equal inference budgets.

What TC does and does not measure. TC requires exact gold snippet ids and so covers only corpora with evidence annotations; FinQA is excluded because every arm's EM there is at most 0.035. It scores completeness and check size, not citation precision or entailment, which varies across seeds and judges (§6). Its check units follow each format's own statements next to the evidence; splitting statements can improve TC without improving support, and extra non-gold citations do not lower it. The answer-level, joint-EM, gold-citation F1, and precision analyses (Table 1; Appendix B) bound these effects; except for TC@k itself, these measures were added after the TC protocol was fixed. Trained LEDGERRL rationale items bundle several claims: 30% reach more than two sources, the broad-unit pattern we diagnose in citation RL, so a reader who checks the final rationale directly against sources does not find finer units than with the citation cap, and guarded training does not raise joint EM over the untrained ledger. Intervals are conditional on the evaluated checkpoints; LEDGERRL and the unguarded control have three seeds, the other trained arms two, and the free-prose control one.

Scope. The single-guard ablations are single-seed, and the protocol comparison changes contexts, protocol, and credit assignment together. All TC evaluations are zero-shot after medical-only training; we do not establish the same effects after training on evidence-labelled multi-hop tasks. Removing the assessment stage from training and inference lowers TC@2 by 4.8 points and TC@∞ by 5.2 in one seed (Appendix G); this changes the protocol as a whole. Baselines are our implementations. A one-seed Llama-3.1-8B replication restores ledger production under the guards, but with substantial schema failures and low gold precision (Appendix I). We make no claim about faithfulness to hidden computation or reduced human verification effort.