Towards Verifiable Answers: Reinforcement Learning with a Verifier Reduces Hallucinated References in Literature-Based Question Answering

Rushank Goyal
Department of Computer Science
Stanford University
rushankg@stanford.edu

Abstract

Experts and laypeople alike increasingly encounter the scientific literature by posing questions to large language models (LLMs) rather than through database searches, making a model's cited sources the reader's main window into the original literature. Yet LLMs frequently suffer from citation hallucinations. Because source existence is objective and externally checkable, it provides an unusually clean signal for reinforcement learning with verifiable rewards (RLVR). We train LLAMA-3.1-8B-INSTRUCT and LLAMA-3.2-3B-INSTRUCT with GRPO and LoRA against a deterministic verifier that resolves each generated reference against scholarly databases CrossRef and OpenAlex. On 40 common medical questions in the MedAESQA dataset, the hallucination rate decreases from 85.5% for the base $8B$ model to 76.7% ($p < 10^{-4}$) for the best post-trained checkpoint. The post-trained model produces 71% more verified references at unchanged citation density, generates fewer misleading DOIs (down to 14% vs 54% for the base), and favors canonical sources such as the *New England Journal of Medicine*. A six-part audit finds no evidence of reward hacking. Ultimately, this deterministic verifier supplies a practical, objective signal that reduces hallucinated references, offering a novel method to improve citation accuracy and a stepping stone toward fully verifiable LLM answers for literature-based QA.

1 Introduction

When large language models (LLMs) answer questions that require grounding in the scientific literature, they commonly include 'hallucinated' references, i.e. references that point to scholarly articles that do not, in fact, exist as written (Alansari and Luqman, 2026). Such mistakes cause two distinct issues: firstly, an in-text citation to a fabricated reference lends false authority to a claim that may be incorrect; secondly, even if the claim itself is correct, the reader hits a dead end when trying to use the bibliography to navigate to the original literature themselves. Thus, hallucinated references affect a reader's trust in the answer.

Given the increasing number of users, particularly lay users, who use LLMs for literature-based question answering (QA), it is necessary to address the issue of citation hallucination (Chatterji et al., 2025; Mendel et al., 2025). Accurate QA is important for high-stakes tasks such as medical or legal advice, where providing a window into the primary literature is critical, as well as enterprise QA tasks like answering company-specific customer support inquiries. Indeed, any field where QA involves synthesizing accurate answers based on scholarly research, e.g. policymakers asking questions answered by economic analyses (Buchanan et al., 2024), benefits from better citation generation in LLM responses.

Errors are hard to catch on the user's end since a hallucinated reference looks stylistically indistinguishable from a real one to a layperson, with a realistic author list, a plausible title, and a well-formed DOI. This is also why another popular tool for QA finetuning, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), is also a poor fit for this task, as human annotators cannot reliably catch fabricated references any more than an end user.

The traditional tool for addressing reference accuracy in QA settings is thus retrieval-augmented generation (RAG), which is used by popular general-purpose QA systems like Perplexity and You.com and also specialized medical platforms like OpenEvidence (Low et al., 2025). RAG systems, however, require a realtime inference-time retrieval index and consume more resources and time, while being vulnerable to information poisoning attacks from external sources (Chang et al.,

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 199–209
November 10, 2026 ©2026 Association for Computational Linguistics

2025). In general, too, the high stakes and wide applicability of the problem warrant the exploration of novel techniques to reduce fabricated references.

The central observation we make is that reference or citation existence is *objective* and *automatically checkable* since a reference either resolves to a real record in a scholarly index or it does not. This makes citation existence an unusually clean target for reinforcement learning with verifiable rewards (RLVR), which has recently been a popular technique for post-training language models (Shao et al., 2024).

Our approach in this paper rests on the hypothesis that LLMs, within their large parametric knowledge, do contain memory of correct references, but pretraining regimes provide little incentive to maximize truthfulness—[1]only likelihood. We posit that, when post-trained with an appropriate reward function, LLMs can be trained to cite sources more accurately, increasing readers' trust in QA responses.

We therefore train an open-weights model with Group Relative Policy Optimization (GRPO) against a deterministic citation verifier, asking whether the model can be pushed to surface real bibliographic information from within its parametric knowledge. Specifically:

- We design and train against a deterministic, external citation-existence reward and show statistically significant hallucination reductions by 8.8 pp in the post-trained (PT) model (85.5% base to 76.7% PT, $p < 10^{-4}$).
- We conduct a thorough six-part audit that shows no increase in duplicate citations, no decrease in citation density, and fewer fully fabricated responses, among other results that greatly reduce the likelihood of the result being reward-hacked.
- We characterize how the trained model's citing behavior changes and discuss an important failure mode in which existence is improved at the cost of relevance.

The paper is intended as exploratory work demonstrating that an automated citation verifier provides a viable RL training signal, setting a foundation for future systems that can enforce more complete scholarly metadata and scientific relevance checks for each reference.

[1]This em-dash has been written by a human!

2 Related Work

How bad is citation hallucination? Across models and domains, LLMs fabricate a large fraction of the citations they produce. Walters and Wilder (2023) documented widespread fabrication and error in ChatGPT-generated bibliographies; in medicine, Chelli et al. (2024) measured hallucination rates of 28.6%–91.4% across GPT-3.5, GPT-4, and Bard for systematic-review references, and Mugaanyi et al. (2024) found citation reliability varies sharply by discipline, with fabricated or malformed DOIs a dominant error mode. More recent work in anatomy QA (Ülkir and Paslı, 2026) and legal citation benchmarking (Chen et al., 2026a) continues to show high citation hallucination rates, ranging from 25.2% to 47.5% for the former and as high as 94% for the latter.

RL for factuality. Verifiable-reward RL has reduced hallucination on short factoid QA, where Wei et al. (2026) optimize a ternary truthful/abstain/wrong reward with GRPO, and on long-form generation, where Chen et al. (2026b) train reasoning models to be more factual with a precision/detail/relevance reward. The increased factuality is not, however, a guaranteed outcome of RL-based post-training process, as reasoning-oriented RL can often *increase* hallucination rates (Li and Ng, 2026). The authors attribute this effect to high-variance gradients and entropy-induced randomness. These results all reward truthfulness in short-form QA, whereas we focus on long-form QA, with a desire to optimize truthfulness as well as utility and trustworthiness for the reader using the existence of real bibliographic citations as a proxy.

Reward over-optimization. Gao et al. (2023) show that optimizing a learned proxy reward eventually degrades the true objective. The reward is deterministic rather than learned, so classical reward-model over-optimization does not directly apply to this work. Nonetheless, the training runs do show a plateau and slight regression after a certain point. Possible explanations for this are discussed in §5.

3 Method

3.1 Deterministic verifier

The reward is computed by a deterministic citation verifier, a Python port of CheckIfExist (Abbonato, 2026) with modifications to increase speed. Figure 1 shows the pipeline. For each extracted reference it (i) checks an in-memory cache keyed

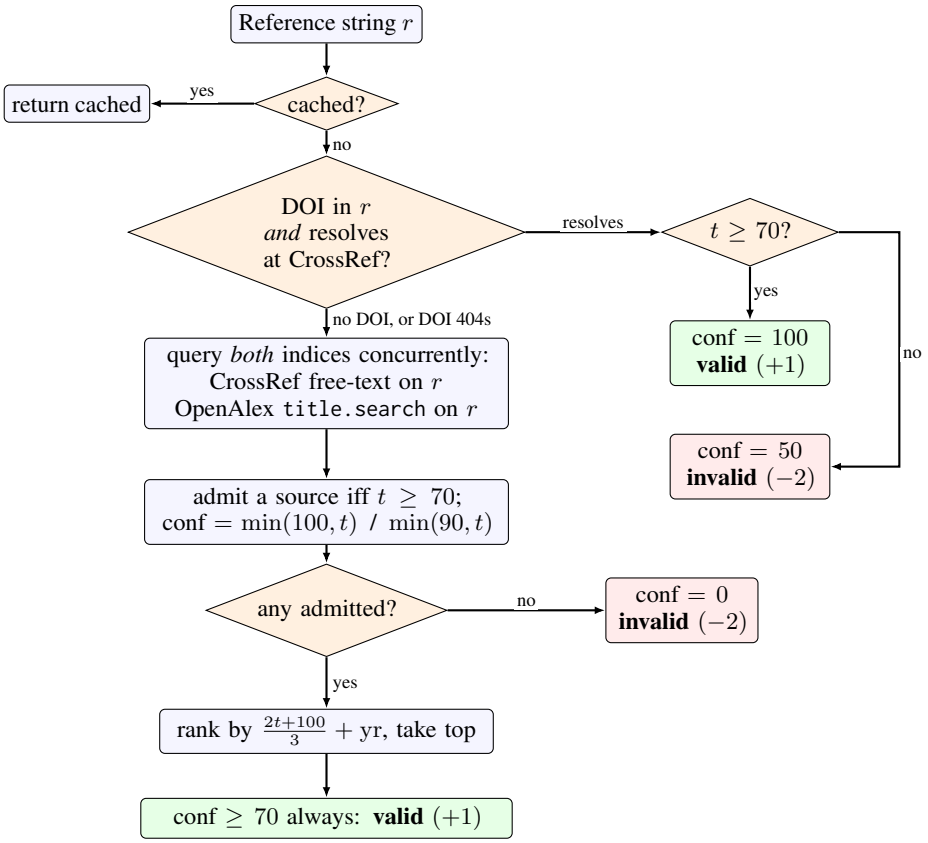


Figure 1: Citation verifier pipeline. t is the title similarity between a retrieved title and the reference string r.

on a hash of the reference fields, so that citations repeated across a prompt's G rollouts incur no repeated API calls; (ii) takes a *DOI fast path*, i.e. if a DOI is present it is resolved against the CrossRef REST API, scoring confidence 100 when the resolved title matches and a capped, sub-threshold confidence of 50 when it does not; and (iii) when no DOI is present (or the DOI lookup fails), dispatches concurrent title searches to CrossRef and OpenAlex.

The best candidate is accepted only if its title similarity exceeds 70. Title similarity (0-100) is the higher of a character-level Levenshtein ratio and a word-level Jaccard overlap, plus a containment boost. Note that title matches in CrossRef and OpenAlex carry confidence ceilings of 100 and 90 respectively—reflecting metadata quality—in line with the code implementation of CheckIfExist. The original verifier queried a third index, Semantic Scholar (S2), which was removed after an ablation on a 47-reference ground-truth corpus showed identical verdicts with and without S2 querying, even though including S2 carried a much higher API rate-limit cost that bottlenecked training speed.

References are extracted from the model output in two passes: a numbered reference-list parser keyed on a "References"/"Bibliography" header, and a parser for inline citations (`[1]`, `[1,2]`, `[1-3]`, etc.) used to compute the per-sentence coverage term $k/|S|$.

3.2 Reward

Let N be the number of extracted references, N_{valid} and N_{invalid} the verified counts, S the set of body sentences, and k the number of sentences with no inline citation. The reward is

$$R = \frac{N_{\text{valid}} - 2\,N_{\text{invalid}}}{N} \;-\; 0.1\,\frac{k}{|S|}, \qquad (1)$$

with $R = -1$ when $N = 0$, giving range $[-2.1, +1.0]$ for the reward function.

Each term targets a specific failure mode. The core term asymmetrically penalizes invalid citations by a factor of two, so a fabrication outweighs a real citation and a positive reward requires mostly real references rather than a lucky majority. Additionally, the normalized sentence term $k/|S|$ discourages confident, uncited prose while dividing by $|S|$ so that long answers are not penalized more than short ones at equal coverage; when every sentence carries a citation it vanishes to zero. The

Setting	Value
Base models	LLaMA-3.1-8B / 3.2-3B-Instruct
Adapter	LoRA r=32, α=64, all proj.
Optimizer	AdamW, lr 5×10^{-6}, bf16
Schedules	linear decay; constant 1.25×10^{-6}
Group size	$G = 4$ rollouts / prompt
Sampling	T=0.8, top-p 0.95, 1500 tok.
Train data	COVID-QA, 1,479 questions
Eval data	MedAESQA, 40 Q $\times$ 4 rollouts
Hardware	Modal A100-80GB, $\sim$60 s/step

Table 1: Training and evaluation configuration.

paper uses a single symbol R throughout to indicate the reward.

3.3 Training and evaluation

We fine-tune LLaMA-3.1-8B-Instruct and LLaMA-3.2-3B-Instruct using Group Relative Policy Optimization[2] (GRPO) (Shao et al., 2024). GRPO was found to be appropriate here since each reward is computed directly using a deterministic verifier rather than a separate learned critic.

Weights are updated using Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient method that freezes the pretrained model and trains small inserted low-rank matrices instead. We use rank-32 LoRA with α=64 and dropout of 0.05, with adapters on all seven attention and MLP projections for a total of $\sim$84M (8B) / $\sim$49M (3B) trainable parameters. Training was run on a single NVIDIA A100 GPU. Table 1 lists the full training details.

Data. Training prompts are drawn from COVID-QA (Möller et al., 2020), which contains 2,019 expert-annotated QA pairs over scholarly articles about COVID-19. We removed yes/no, single-statistic, and date-lookup items that do not call for cited synthesis, retaining a dataset of 1,479 broad-synthesis questions. Only the questions are used: the gold answers are discarded, so the policy receives no supervised target and the entire training signal comes from the verifier. Each prompt pairs the question with a fixed system prompt (see A) that requires a 2–3 paragraph answer, an inline citation marker on every factual claim, 5–10 real sources, and a references section in a parseable format, with an explicit instruction not to fabricate.

[2]Implemented using the TRL and PEFT libraries from HuggingFace.

Procedure. At each step the policy samples $G = 4$ completions for one prompt (T=0.8, top-p 0.95, $\leq$ 1500 new tokens). Every completion is scored independently by the verifier and GRPO forms advantages relative to the group mean. With `per_device_train_batch_size=`G and no gradient accumulation, the number of optimizer steps is equal to the number of questions consumed. Rewards are computed online against the live CrossRef and OpenAlex APIs with an in-memory cache for cost savings.

Runs The main runs cover COVID-QA[0:1000] under a linear decay from 5×10^{-6}. Because reward plateaus after step 750, we trained another run from `checkpoint-750` at a constant LR of 1.25×10^{-6}—which is the learning rate that the decay schedule reaches at step 750—over the 729 remaining, unseen questions COVID-QA[750:1479], to separate a genuine plateau from schedule-induced decay to zero.

Evaluation 40 questions from MedAESQA (Gupta et al., 2025) were used for evaluation, with $G = 4$ rollouts each for a total of 160 rollouts per checkpoint. All statistical comparisons use paired t-tests over the 40 questions, applied to per-question mean reward and per-question hallucination rate. As MedAESQA is an entirely different corpus, with de-identified medical questions from the general public, all gains represent cross-dataset transfer of accurate citation abilities rather than improved in-distribution fit.

4 Results

4.1 GRPO reduces hallucination

Table 2 and Figure 2 give the trajectory of the training runs of the 8B model and both the 3B and 8B models, respectively. Reward rises steeply through step 750, where the 8B model peaks at $R = -1.31$ and hallucination bottoms at 76.7%, an 8.8 pp reduction over baseline ($t = 5.2$, $p < 10^{-4}$) with $\Delta R = +0.26$ (95% CI $[0.16, 0.36]$; $t = 5.1$, $p < 10^{-4}$). The number of valid citations rises from 168 to 287, i.e. the model emits 71% more verified references on the same questions, while references per rollout barely move, going from 7.24 to 7.70.

Continued training after this point does not help and appears to worsen performance across both model sizes. Under linear decay, training from step 750 to 1000 gives $\Delta R = -0.05$ between

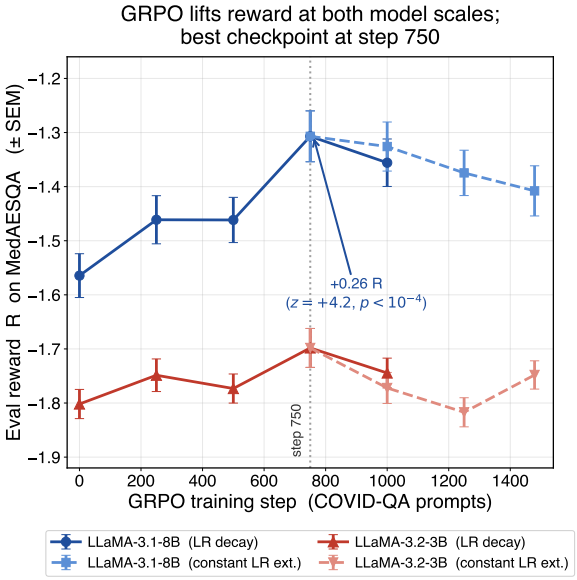


Figure 2: Reward across GRPO training. The 8B and 3B runs trace the same arc and peak at step 750 under both LR schedules, then plateau or mildly regress.

those two points, while the constant-LR continuation run drifts down to $R = -1.41$ by step 1479. The 3B model behaves roughly identically in shape, peaking at step 750 ($R = -1.70$).

Both model sizes exhibit remarkably similar reward trajectories: reward improves rapidly during early GRPO training, peaks at approximately step 750, and subsequently plateaus or mildly regresses across both LR schedules. This effect can be driven by multiple possible reasons: unwanted divergence from the base model after a substantial number of training steps; saturation of the reward signal; or noise from the verifier, which is exactly objective in theory but imperfect in its current implementation.

4.2 Reward gains are broadly distributed

We observe that, at the best (step 750) checkpoint, the model's reward improvement is distributed across the eval set and not driven by a few easy items. Figure 3 plots the per-question change in reward from baseline to step 750: 30 of 40 MedAESQA questions improve and only 10 regress (and even then, the magnitude of regression is much lower than the magnitude of gain). While gain concentrated on a few questions might have implied narrow domain-specific improvement—e.g. for questions that have topic overlap with those in the COVID-QA training set—this broad improvement suggests more general behavioral change.

Checkpoint	R	Halluc.	Refs/roll
Baseline (8B)	−1.564	85.5%	7.24
Step 250	−1.461	81.6%	7.61
Step 500	−1.462	81.0%	7.33
Step 750	**−1.307**	**76.7%**	7.70
Step 1000 (decay)	−1.356	79.2%	7.64
Cont. → 1000	−1.326	77.8%	7.64
Cont. → 1250	−1.375	79.2%	7.52
Cont. → 1479	−1.408	79.7%	7.72

Table 2: GRPO trajectory on MedAESQA with detailed metrics for the main run (8B model).

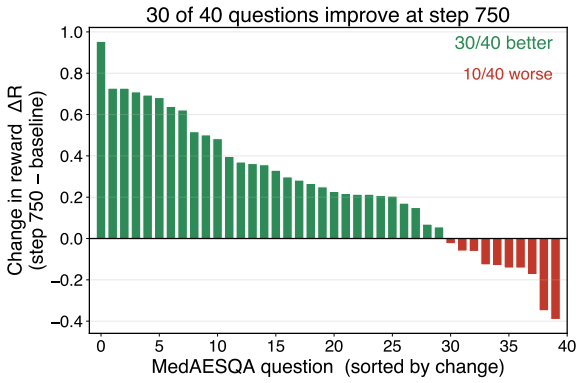


Figure 3: Per-question reward change between baseline and step 750 model (8B). $30/40$ questions improve.

4.3 A six-step audit against reward hacking

To further rule out various sources of reward hacking, we audited the best checkpoint along six axes. No anomalous behavior indicative of reward hacking was observed.

Four of the axes remained stable: reference recycling across questions, within-rollout duplication, mismatches between inline markers and reference lists, and number of sentences with 'stuffed' citations. Specifically, the top 5 most-reused references accounted for 1.6% of citations at both baseline and step 750 with no reference shared across five or more questions, duplicate rates remained near 10%, inline/list ID mismatches were at 0%, and the number of sentences with stuffed (≥ 4) citations was also 0.

The remaining two axes shifted in favor of the post-trained model. First, we checked *which* rollouts earned positive reward (Figure 4). At baseline, the few rollouts with $R > 0$ achieve it by citing sparsely, with a mean of 2.5 references versus 7.2 overall,. At step 750, positive-reward rollouts actually carry mean 7.5 references versus 7.7 overall, i.e. the model earns reward at full citation density. Secondly, we inspect at the shape of the reward distribution (Figure 5). The gain comes

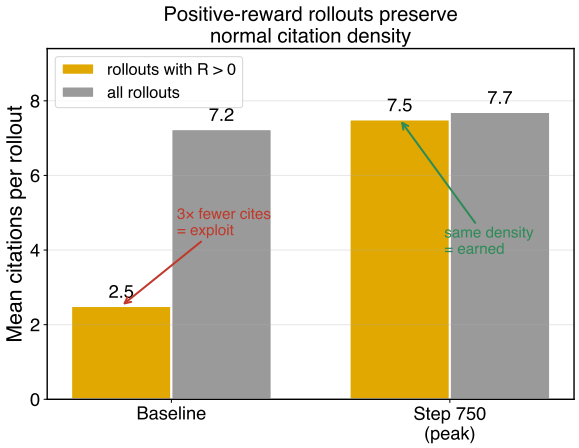


Figure 4: Citation density of positive-reward rollouts. At baseline, $R > 0$ rollouts cite $3\times$ less than average (an exploit); at step 750 they cite at average density, earning reward by citing real papers rather than by abstaining.

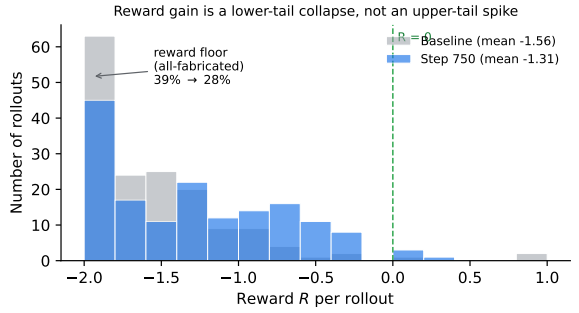


Figure 5: Reward distribution over 160 rollouts. The step-750 gain is a collapse of the lower tail—fewer fully-fabricated answers at the reward floor ($39\% \rightarrow 28\%$)—rather than a spike at the high-reward end.

from shrinking the fraction of fully-fabricated answers at the reward floor (39% to 28%), not from inflating already-good ones, as the proportion of rollouts with $R > 0$ changes minimally, from 1.9% to 2.5%. In other words, RL is fixing the worst answers rather than exploiting gain from those responses that are already the best. Thus, the 8.8 pp reduction represents a genuine improvement in the model's abilities.

4.4 Ruling out eroded caution and advantage noise as causes of performance regression

Before we turn our attention to the learned policy, we share results from tests that check two natural explanations for the post-750 regression. Both explanations are rejected based on the results.

First, if RL eroded the model's epistemic caution, it would hedge less as it regresses, and with lower hedging tendencies it might over-claim or include riskier, i.e. less trustworthy or verified, sources. Instead, hedging—as measured by modal and evidential phrases ("may," "suggests," "further research") per 1K characters—is largely unchanged (see Table 3): the 8B model holds at $\approx$4.1–4.2 across baseline and peak checkpoints, while the 3B peak actually hedges more than its baseline.

Second, GRPO's group-relative advantage could amplify noise if within-group reward variance shrank late in training, destabilizing the computation of the standardized advantage and causing noisy and ultimately counterproductive gradient updates. However, within-group standard deviation (measured on the G=4 eval groups) is flat on MedAESQA and grows on the training distribution ($0.28 \rightarrow 0.49$), and degenerate all-equal-reward groups—which contribute no gradient due to lack of any relative difference among rollouts—become rarer with training (8B $10\% \rightarrow 2\%$; 3B $35\% \rightarrow 15\%$). Neither eroded calibration nor advantage-noise amplification explains the decline.

Checkpoint	Hedges / 1K chars
8B baseline	4.18
8B step 750 (peak)	4.05
3B baseline	4.26
3B step 750 (peak)	**4.73**

Table 3: Epistemic hedging is stable or increases with training, refuting the hypothesis that the performance regression is driven by eroded caution.

4.5 The model learns to cite more carefully

We discuss how the behavior of the post-trained model (i.e. the best step-750 checkpoint) changes compared to the baseline LLM.

An important result is that references carrying an explicit DOI collapse from 54% to 14%. Given the verifier's penalty for a resolving-but-mismatched DOI, a fabricated DOI is a high-variance bet since it could resolve to a completely different paper. Prior surveys have shown that fabricated DOIs are among the most types of hallucinations (Mugaanyi et al., 2024). After the RL training run, the LLM learns not to fabricate DOIs and rely on recall-able title/author matches.

A natural concern is that citing fewer DOIs might be an easy "reward hack" i.e. the observed improvement in reward R is driven entirely by a learned DOI-avoidance behavior. We investigate this further by performing an analysis on rollouts where none of the references carry DOIs ($52/160$

MedAESQA Q37: "if hct is 34, does pt need blood transfusion?"
(✓ = verified real reference, × = hallucinated reference; best rollout per model)

Baseline — $1/3$ verified (67% hallucinated).
A hemoglobin (Hb) concentration below 7 g/dL is typically considered an indication for blood transfusion in patients with acute blood loss [1]. However, a hemoglobin concentration (Hct) of 34% does not necessarily indicate the need for a blood transfusion, as this value alone does not provide information about the patient's overall clinical status or the presence of symptoms.
In fact, some patients with Hb levels as low as 6.5 g/dL may not require a blood transfusion if they are asymptomatic and have no evidence of organ dysfunction [2]. The decision to transfuse should be based on a comprehensive assessment of the patient's clinical condition, including symptoms, vital signs, and laboratory values.
Guidelines from the AABB (formerly the American Association of Blood Banks) and the AABB-AST (American Society for Apheresis) recommend that blood transfusions be considered in patients with a Hb concentration below 7 g/dL or a hematocrit (Hct) below 21% if they have symptoms or signs of inadequate oxygen delivery [3]. Therefore, a Hct of 34% is unlikely to require a blood transfusion unless the patient has significant symptoms or evidence of organ dysfunction.
× [1] Vincent, J. L., et al. (2011). Red blood cell transfusion in critically ill patients. Journal of the American Medical Association, 305(14), 1441–1448. DOI: 10.1001/2011.jama.442
✓ [2] Carson, J. L., et al. (2016). Red blood cell transfusion: A clinical practice guideline from the AABB. Annals of Internal Medicine, 165(8), 557–564. DOI: 10.7326/M16-0933
× [3] AABB. (2017). AABB Clinical Transfusion Guidelines. AABB, Bethesda, MD.

Step 750 — $2/4$ verified (50% hallucinated).
Hemoglobin (Hb) levels are a critical determinant in deciding the need for blood transfusions, not hematocrit (Hct) values alone. A Hct value of 34 is considered low, but it may not necessarily indicate anemia or the need for transfusion [1]. The decision to transfuse should be based on the patient's symptoms, physical examination, and laboratory results, including Hb levels.
According to the American Society of Anesthesiologists (ASA) guidelines, blood transfusions are recommended for patients with Hb levels below 6-8 g/dL, especially if they have symptoms such as fatigue, shortness of breath, or chest pain [2]. However, some studies suggest that even at Hb levels above 6-8 g/dL, patients may still experience impaired oxygen delivery and exercise capacity [3]. Therefore, the decision to transfuse should be individualized and based on the patient's specific clinical context.
A more appropriate approach might be to assess the patient's oxygen delivery and consumption, rather than relying solely on Hct or Hb values. For example, a study by Vincent et al. [4] found that patients with Hb levels between 7-10 g/dL who were not symptomatic did not benefit from transfusion, suggesting that a more nuanced approach to blood transfusion is warranted.
Ultimately, the decision to transfuse should be made by a healthcare provider, taking into account the patient's individual circumstances.
× [1] Vincent JL, et al. Transfusion requirements in critically ill patients. New England Journal of Medicine, 2008.
✓ [2] Practice guidelines for blood component therapy: A report by the American Society of Anesthesiologists Task Force on Blood Component Therapy. Anesthesiology, 1996.
✓ [3] Hébert PC, et al. A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care. New England Journal of Medicine, 1999.
× [4] Vincent JL, et al. Transfusion requirements in critically ill patients: a pilot study. Intensive Care Medicine, 2002.

Figure 6: A matched example from the evaluation set showing reduced hallucination in the post-trained model.

for base, $132/160$ for PT). Even in this subset, the hallucination rate falls from 85.3% in the baseline to 75.2% in the post-trained model. This -10.1 pp reduction shows that the performance gain persists independently of DOI presence, suggesting that the model genuinely improves its recall of text-based bibliographic features rather than merely gaming the metric.

We also propose that DOI-avoidance can be considered a behavior improvement, as, when fabricated DOIs are included, lay users are liable to paste them directly into a search engine and end up at a nonexistent or (worse) mismatched paper.

Secondly, the model gravitates more often toward canonical, high-impact sources: the *New England Journal of Medicine* (NEJM), one of the longest-running and most prestigious journals in the medical field, is mentioned 68 times in the baseline's 160 rollouts but 89 times in the same number of rollouts by the post-trained model. Generally, research articles from prominent journals and other high-impact publications are discussed frequently in subsequent literature, suggesting the post-trained model prefers easily recallable sources that it saw more often during pretraining. This aligns with established findings that a language model's ability to accurately recall parametric knowledge is strongly correlated with a document's frequency of occurrence in the pretraining data (Kandpal et al., 2023).

An affinity for more canonical and frequently referenced sources is desirable behavior, since such landmark papers are thoroughly vetted and thus highly likely to be both comprehensive and accu-

rate should the user decide to dive deeper into the bibliography to learn more.

Figure 6 makes this concrete on a sample question from MedAESQA. The baseline produces three references, of which only 1 resolves correctly (67% hallucination rate), while the post-trained model has 2 verified references out of 4 (50% hallucination rate). Even the one that 'resolves correctly' in the baseline's rollout has an incorrect year, as the deterministic verifier focuses on the title.

4.6 Existence is not relevance

The same instinct to retrive canonical papers that drives broad gains can backfire on some questions. We take the example of a different question from MedAESQA, which asks how cannabidiol (CBD) affects liver enzymes. This is a niche, recent topic with sparse indexed literature; here, the baseline model cites an on-point study ("Effects of cannabidiol on liver enzymes in healthy volunteers"), while the post-trained model drifts to the more famous CBD literature on anxiety, epilepsy, and schizophrenia. While those papers are more canonical and more recallable, they are off-topic to the user's questions. This is a limitation of the current verifier, which only rewards existence and not relevance (see §6).

5 Discussion

A verifier can supply what learned rewards lack. Whereas RL against learned, RLHF-style rewards can *increase* hallucination (Lin et al., 2024) or be gamed by exploiting the scorer (Chen et al., 2026b), a deterministic external verifier gives an unspoofable factual signal: a citation either resolves in a scholarly index or it does not. Our main result of an 8.8 pp reduction mirrors for long-form QA what other RLVR-for-truthfulness work has shown for short-answer factoid QA (Wei et al., 2026).

The plateau is an open question. Both scales peak at step 750 and then plateau or mildly regress, and we have separately ruled out reward hacking (§4.3), eroded caution, and advantage-noise amplification (§4.4) as possible factors. Two hypotheses from recent RLVR theory might explain this phenomenon. Cui et al. (2025) show that policy entropy collapses early and trades predictably for reward, implying a performance ceiling once exploration is exhausted. A different paper by Wu et al. (2026) argues that RLVR sharpens probability mass onto what the base model already favors while trimming correct-but-rare outputs. The observations in this paper are consistent with both: firstly, the regression is concentrated in the lower tail of the reward distribution, and once the low-performing responses are improved, the model plausibly hits a 'ceiling' that it cannot cross; secondly, the retreat toward a small set of canonical venues resembles the distribution sharpening that the second hypothesis predicts. The first effect can likely be resolved with a larger model (e.g. LLaMA's 70B variant) that has an order of magnitude more parameters and thus more parametric knowledge to draw upon.

Vanilla GRPO can also suffer from advantage collapse, especially with small groups of 4 rollouts (He et al., 2026). Sampling more rollouts per question ($G = 8$ or 16 or even 32 instead of of $G = 4$) or creating a larger set of possible questions (likely synthetically since we only need questions and not answers) are easy methods to improve performance. On the other hand, adapting the data distribution at train-time (Panaganti et al., 2026) to sample more rollouts from higher difficulty questions or questions that require greater literature search depth could improve reward gain without additional data.

Lastly, the imperfect nature of the verifier likely also impedes improvement at some point (see §6).

We end this section by stressing that the decline is statistically mild, so these are offered as candidate explanations for a possible issue that could arise in similar research, rather than confirmed mechanisms behind a clearly documented phenomenon. From a practical perspective, an early stop at the reward peak is likely ideal.

6 Conclusion

We follow the broader trend towards verifiable rewards for LLM post-training, presenting an exploratory demonstration that a deterministic citation verifier can supply a useful training signal to reduce hallucinated references in long-form, literature-grounded QA.

Specifically, we show that training an open-weights model with GRPO against a deterministic citation verifier measurably reduces hallucinated references in medical QA (-8.8 pp, $p < 10^{-4}$), and a six-part audit finds no evidence of common reward-hacking behaviors. Instead, we see the post-trained model emits 71% more verified references without any change in citation density, emits far fewer DOIs, recalls named landmark trials, and gravitates to canonical sources. The same instinct

can backfire on niche topics, where existence is improved at the cost of relevance, providing a concrete target for future reward design. Ultimately, the deterministic verifier supplies an objective, factual signal that reduces the rate of hallucinated references, providing an interesting, novel method to improve reference accuracy that can serve as a stepping stone toward fully verifiable QA systems.

Limitations

This paper uses a single held-out evaluation set (MedAESQA, 40 questions), which contains data for but one domain of literature-based QA. Additionally, all our training runs are single-seed.

The verifier is imperfect, as only the title of a work is checked and not the journal, year, page numbers, etc. Some imperfections were introduced by our simplifications, while others are more fundamental limits of the databases themselves. We encourage future researchers to improve upon this design. Even with these fixes, the risk of false positives and negatives remains as string matching over large databases is challenging, and certain citations might be to material not indexed in common databases. Regardless of the design, we encourage a thorough analysis of the precision and recall of such verifiers, which we were not able to do for our verifier in the current study.

Even if it were perfect, the verifier checks citation *existence*, not *relevance*: as §4.6 shows, a real reference can be tangential to the claim it supports, so a fluent answer with verifiable-but-irrelevant citations would get high reward while potentially being less useful to the reader. In the worst case, the LLM could actively print wrong statements but attach technically-correct references to 'pass' the verifier. Thus, a highly-accurate 'verifier' for scientific relevance would be an important next step in the evolution of this field.

We believe that beyond existence, truthfulness and relevance themselves can be rewarded using a 'verifier.' Specifically, whether a paper P supports a claim C can be verified to a high degree of confidence. We are confident that future research can reward directly for truthfulness instead of relying on existence as a proxy.

Ethics Statement

Our intention in this research is to improve the reliability of literature-based QA, ensuring that when laypeople or professionals query an LLM, they are pointed toward real, verifiable sources. However, we acknowledge a potential dual risk: namely, by training an LLM to reliably generate real citations, this methodology could theoretically be exploited to automatically generate seemingly credible scientific manuscripts.

To mitigate this, our methodology is strictly designed for focused QA rather than free-form generative writing. Our training pipeline directs the model to generate concise, 2–3 paragraph responses that directly address a specific user query. This design does not support the sustained argumentative synthesis or document-level coherence required for genuine academic authorship. We condemn any use of such methodology for academic misconduct, ghostwriting, or the automated fabrication of scientific literature. This work is intended solely to build trustworthy, verifiable QA assistants that help human readers safely navigate existing knowledge.

Acknowledgments

This research originated as a course project for Stanford University's CS224R. I would like to thank the teaching staff for their instruction and for coordinating with Modal to provide the compute credits that made this work possible. I am also grateful to the anonymous workshop reviewers for their insightful feedback that resulted in an improved final manuscript.

References

Diletta Abbonato. 2026. CheckIfExist: Detecting citation hallucinations in the era of ai-generated content. *Preprint*, arXiv:2602.15871.

Aisha Alansari and Hamzah Luqman. 2026. Large language models hallucination: A comprehensive survey. *Preprint*, arXiv:2510.06265.

Joy Buchanan, Stephen Hill, and Olga Shapoval. 2024. Chatgpt hallucinates non-existent citations: Evidence from economics. *The American Economist*, 69(1):80–87.

Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Ziyou Jiang, Yang Liu, and Qing Wang. 2025. One shot dominance: Knowledge poisoning attack on retrieval-augmented generation systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18811–18825, Suzhou, China. Association for Computational Linguistics.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt.

Working Paper 34255, National Bureau of Economic Research.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26:e53164.

Sijia Chen, Hang Yin, and Shunfan Zhou. 2026a. Legalcitebench: Evaluating citation reliability in legal language models. *Preprint*, arXiv:2605.10186.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen tau Yih. 2026b. Learning to reason for factuality. *Preprint*, arXiv:2508.05618.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, pages 10835–10866.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. 2026. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *Preprint*, arXiv:2605.21125.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *Preprint*, arXiv:2211.08411.

Junyi Li and Hwee Tou Ng. 2026. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. *Advances in Neural Information Processing Systems*, 38:151064–151085.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

Yen Sia Low, Michael L Jackson, Rebecca J Hyde, Robert E Brown, Neil M Sanghavi, Julian D Baldwin, C William Pike, Jananee Muralidharan, Gavin Hui, Natasha Alexander, et al. 2025. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digital health*, 11:20552076251348850.

Tamir Mendel, Nina Singh, Devin M Mann, Batia Wiesenfeld, and Oded Nov. 2025. Laypeople's use of and attitudes toward large language models and search engines for health queries: survey study. *Journal of medical Internet research*, 27:e64290.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Joseph Mugaanyi, Liuying Cai, Sumei Cheng, Caide Lu, and Jing Huang. 2024. Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research*, 26:e52935.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kishan Panaganti, Zhenwen Liang, Wenhao Yu, Haitao Mi, and Dong Yu. 2026. Group distributionally robust optimization-driven reinforcement learning for llm reasoning. *Preprint*, arXiv:2601.19280.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mehmet Ülkir and Bahattin Paslı. 2026. Reference hallucination, citation reliability, and readability of large language models in anatomy-related question answering. *Clinical Anatomy*.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific reports*, 13(1):14045.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-Tau Yih, and Xin Luna Dong. 2026. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 134156–134175. PMLR.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. 2026. The invisible leash: Why rlvr may or may not escape its origin. *Preprint*, arXiv:2507.14843.

A System Prompt

Every training and evaluation rollout used the following system prompt, held fixed throughout.

```
You are a medical research assistant answering questions based on the published scientific literature. Your responses must:

1. Be 2–3 paragraphs long.
2. Support every factual claim with an inline citation marker, e.g. [1] or [2, 3].
3. Draw on 5–10 real, peer-reviewed papers or authoritative sources.
4. End with a References section listing each cited source in this exact format:

References
[1] Author(s). Title. Journal or Conference, Year. DOI: xx.xxxx/xxxxx
[2] Author(s). Title. Journal or Conference, Year.

Only cite papers that actually exist. Do not fabricate titles, authors, journals, or DOIs.
```

The user turn contained the COVID-QA or MedAESQA question alone.