Reducing Citation Hallucinations in Large Language Models

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

Extended Abstract

Large language models (LLMs) routinely fabricate the references they cite. For instance, an untrained LLaMA-3.1-8B-Instruct hallucinates 85.5% of the references it produces on real medical questions (MedAESQA). This is particularly dangerous in scenarios such as medical question-answering (QA) because fabricated citations can lend false authority to possibly-wrong medical claims, are hard for laypeople to disentangle, and turn an answer into a dead end since the reader cannot reach the (nonexistent) source. (The harm persists even when the content itself is correct.)

In this research project, I ask whether an open-weights model can be trained to cite references that *actually exist*, using RL against an automatic citation verifier. Citation existence is objective and externally checkable, which makes it an unusually clean signal for reinforcement learning with verifiable rewards (RLVR): unlike a learned reward model, a deterministic verifier cannot be spoofed by stylistic plausibility. The verifier is a Python port of CheckIfExist (Abbonato, 2026), where each reference is resolved against CrossRef and OpenAlex (with a DOI fast path) and counted valid only when a match clears a confidence threshold. The reward rewards real citations, penalizes fabrications at twice the weight, floors answers that cite nothing, and lightly penalizes uncited sentences:

$$R = \frac{N_{\text{valid}} - 2\,N_{\text{invalid}}}{N} - 0.1\,\frac{k}{|S|},$$

with $R = -1$ when $N = 0$, where k is the number of sentences without any in-text citations and S is the set of all sentences in the model's repsonse. The LLM was trained with GRPO (Shao et al., 2024) and LoRA ($G = 4$ rollouts/prompt) on $1{,}479$ COVID-QA questions and evaluated on 40 held-out MedAESQA questions (160 rollouts/checkpoint).

GRPO measurably reduces hallucination. At the best checkpoint (step 750) the 8B hallucination rate falls from 85.5% to 76.7% (-8.8 pp) and reward rises from -1.56 to -1.31 ($+0.26$, $z = +4.2$, $p < 10^{-4}$). Overall, the model emits 71% more verified references on the same questions. In terms of the full GRPO curve, I see that reward climbs steeply, then plateaus or mildly regresses; the 3B and 8B models trace the same arc and peak at the *same* step even when the learning rate is held constant.

The gain is genuine, not gamed. A six-part audit finds no recycling, duplication, marker mismatch, or cite-stuffing, and positive-reward rollouts keep full citation density (7.5 vs. 7.7 refs) rather than scoring by abstaining. The improvement comes from shrinking the fraction of fully-fabricated answers ($39\% \rightarrow 28\%$), not inflating already-good ones ($R \geq 0$ grows only $1.9\% \rightarrow 2.5\%$). In the full evaluation set, the best model shows improvement on 30 of 40 questions as compared to the base untrained LLM.

The model cites more carefully. Reference count barely moves ($7.2 \rightarrow 7.7$), but references carrying an explicit DOI collapse from 54% to 14% and, interestingly, the policy appears to shifts toward recall-able canonical sources. For instance, NEJM mentions rise $68 \rightarrow 89$, and on an oncology question it learns to cite the actual landmark immunotherapy trials. The same instinct can backfire: on a niche question the policy drifts to famous-but-off-topic papers, a consequence of the fact that my reward rewards existence, not relevance.

In sum, a deterministic citation verifier is a practical, auditable RL reward for factuality (-8.8 pp, $p < 10^{-4}$), supplying the unspoofable signal learned reward models lack. Some limitations of the current approach include a single eval set and seed, a $\sim$7.7% verifier false-negative rate (although, importantly, the false-positive rate was 0.0%), and the focus on existence rather than scientific/medical relevance.

1 Introduction

When a language model answers a medical question, or indeed any question where citations to existing literature and sources are essential, it commonly supports its claims with references that do not exist. On the MedAESQA medical-QA set, an untrained LLaMA-3.1-8B-Instruct hallucinates 85.5% of the references it produces. The harm can be serious, especially in the context of medical QA, and arises from two equally problematic possibilities:

1. There is a risk that the answer includes incorrect information connected to the fabricated citation.

2. Even if the information is correct, the user can hit a "dead end" since they can't explore the attached source further and read the (nonexistent) original scholarly work.

The issue is hard to mitigate on the user's end since a fabricated citation looks stylistically indistinguishable from a real one to a layperson, with correct-looking authors, a plausible title, and a well-formed DOI. Lastly, invented citations cost credibility, and in a clinical setting that can often be the difference between a usable second opinion and a liability.

The standard alignment tool for QA finetuning is RLHF or reinforcement learning from human feedback (Ouyang et al., 2022), but that is a poor fit here because human annotators do not reliably catch fabricated references either: the errors are subtle, numerous, and require database lookups to confirm. The central observation is that citation existence is *objective* and *automatically checkable*. A reference either resolves to a real record in a scholarly index or it does not—an objective, binary result. This makes citation existence an unusually clean target for reinforcement learning with verifiable rewards (RLVR), the paradigm behind recent reasoning models (Shao et al., 2024), but with a reward that is fully deterministic and external rather than a learned proxy.

The traditional tool for addressing citation accuracy in QA settings is 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. RAG systems, however, need a real-time inference-time retrieval index and consume more resources and time, while being vulnerable to information poisoning by less-trustworthy or even outright false sources online.

On the other hand, my approach in this paper rests on the hypothesis that LLMs, within their large parametric knowledge, do contain memory of correct citations, but that current training regimes focused on next-word prediction provide no incentive to the model to produce correct citations that match scholarly databases. I posit that, with reinforcement learning using the correct reward function, LLMs can be trained to cite sources more accurately, which would prove to be useful in medical QA settings.

I 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 its parametric knowledge. My work makes the following contributions:

- I design and train against a deterministic, external citation-existence reward—to my knowledge the first work to reward citation existence directly—and show it reduces 8B hallucination on MedAESQA by 8.8 pp ($85.5\% \to 76.7\%$, $p < 10^{-4}$).

- I document a sharp, reproducible training horizon: both the 3B and 8B models peak at step 750 and then regress, under both linear-decay and constant learning-rate schedules.

- I verify, through a six-part audit, that the gain is genuine learning rather than reward hacking, and rule out two natural mechanistic explanations for the regression (eroded calibration and advantage-noise amplification).

- I characterize *how* the trained model's citing behavior changes—abandoning fabricated DOIs, gravitating to canonical high-impact venues, and recalling named landmark trials—and surface a characteristic failure mode in which existence is improved at the cost of relevance.

2 Related Work and the Gap

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. The 85.5% hallucination rate of the baseline model on MedAESQA sits squarely in this range and motivates an objective, automatically checkable training signal rather than a human- or model-judged one.

RL for factuality. Verifiable-reward RL has reduced hallucination on short factoid QA, where Wei et al. (2025) optimize a ternary truthful/abstain/wrong reward with GRPO, and on long-form generation, where Chen et al. (2025) train reasoning models to be more factual with a precision/detail/relevance reward. But the signal is delicate. Li and Ng (2025) show that reasoning-oriented RL can *increase* hallucination, attributing it to high-variance gradients and entropy-induced randomness, and reward functions built on automatic factuality scorers are prone to gaming, producing less detailed or less relevant responses that nonetheless score well (Chen et al., 2025). These results all reward *claim truthfulness* but none rewards *citation existence* directly, which is the gap I address.

Reward over-optimization. Gao et al. (2023) showed 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, and 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 minor modifications. Figure 1 shows the pipeline. For each extracted reference it (i) checks an in-memory cache keyed 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

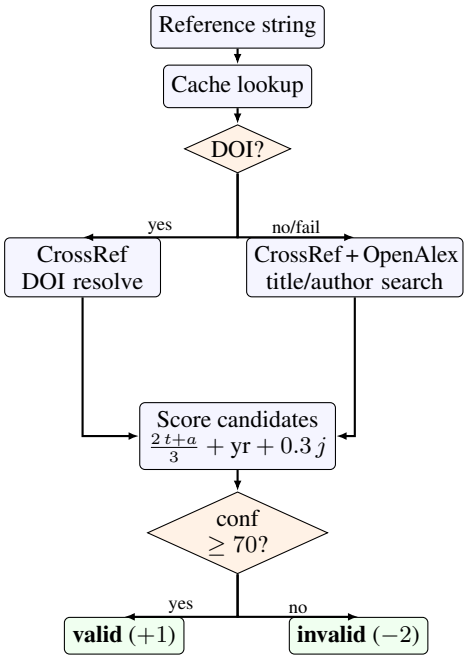

Figure 1: The deterministic citation verifier. A cached, DOI-first pipeline resolves each reference against CrossRef and OpenAlex; a title-dominated score must clear 70 for the citation to count as valid. Invalid citations are penalized at twice the weight of a valid one.

it does not; and (iii) when no DOI is present (or the DOI lookup fails), dispatches concurrent title/author searches to CrossRef and OpenAlex.

Each candidate is scored by a title-dominated function: score $= \frac{2\,t+a}{3} + \text{yr} + 0.3\,j$, where t, a, and j are title, author, and journal string similarities and yr is a year bonus (40 for an exact match, 15 within one year, 5 within two). String similarity takes the maximum of a normalized Levenshtein and a Jaccard word-overlap score.

The best candidate is accepted only if its title similarity exceeds 70; CrossRef and OpenAlex carry confidence ceilings of 100 and 90 respectively, reflecting metadata quality. The original verifier queries a third index (Semantic Scholar) but I removed it after an ablation on a 47-reference ground-truth corpus showed identical verdicts (35/47 valid) at a much higher API rate-limit cost which was becoming a training speed bottleneck. On that ground-truth corpus the verifier had a 0% false-positive rate and a ∼7.7% false-negative rate.

References are extracted from the model output in two passes: a numbered reference-list parser keyed on a "References"/"Bibliography" header, and an inline-marker parser (`[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 the range $[-2.1, +1.0]$. Each term targets a specific failure mode. The core fraction weights invalid citations by two, so a single fabrication outweighs a real citation and a positive reward requires mostly real references rather than a lucky majority. The floor at $N = 0$ removes the trivial exploit of citing nothing to avoid penalties. 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 paper uses a single symbol R throughout to indicate the reward.

3.3 Training and evaluation

I fine-tuned LLaMA-3.1-8B-Instruct and LLaMA-3.2-3B-Instruct with GRPO (Shao et al., 2024). GRPO was found to be appropriate here since it estimates each rollout's advantage from a group baseline rather than a learned critic such as in actor-critic methods like PPO (Schulman et al., 2017). As previously discussed, reward is computed by external API calls rather than training a separate neural head as the critic.

The implementation specifically uses the `GRPOTrainer` class from HuggingFace's TRL library. The reward is a plain Python callable passed as `reward_funcs` such that TRL can hand it the batch of completions and it returns one scalar each from the citation verifier–cum–reward function. Rather than updating all 8B weights, I fine-tuned with LoRA (Low-Rank Adaptation) (Hu et al., 2022), which was taught in CS224N. LoRA is a parameter-efficient method that freezes the pretrained model and trains small inserted low-rank matrices. In my implementation, I use the PEFT library (also from HuggingFace) and specifically the `LoraConfig` class. Lastly, the training was run on an A100 GPU accessed through Modal credits made available in the class. Table 1 lists the full training details.

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 = 160
Hardware	Modal A100-80GB, ~60 s/step

Table 1: Training and evaluation configuration.

The base LLM was trained on 1,479 broad-synthesis questions filtered from COVID-QA (removing yes/no, single-statistic, and date-lookup items that do not call for cited synthesis) and evaluate on the 40 held-out MedAESQA questions with $G = 4$ rollouts each (160 rollouts per checkpoint; $\text{SEM}(R) \approx 0.04$, so reward differences above $\sim$0.08 are significant at $p < 0.05$). To separate the training horizon from the learning-rate schedule, I ran each model under two schedules.

The first run was under linear decay and on a set of 1,000 out of 1,479 questions. Upon observing a plateau after the step-750 checkpoint, I conducted another training run under a constant-LR continuation (the *cont750* run) that resumes from the step-750 checkpoint at the decay schedule's step-750 learning rate (without decaying it further) and trains on the remaining 729 COVID-QA questions. This was done to test whether the eventual plateau in performance was due to the LR schedule: If the plateau were merely the decay schedule approaching zero, the constant-LR continuation should keep improving, but it does not.

In the next section, I report reward R, hallucination rate $N_{\text{invalid}}/(N_{\text{valid}} + N_{\text{invalid}})$, references per rollout, and $z = \Delta/\text{SE}$ for pairwise comparisons, and conduct further investigations and analyses.

4 Results

4.1 GRPO cuts 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%, which is a -8.8 pp reduction over baseline and $\Delta R = +0.26$ ($z = +4.2$, $p < 10^{-4}$). The valid-citation count rises $168 \rightarrow 287$, i.e. the model emits 71% more verified references on the same questions, while references per rollout barely move ($7.24 \rightarrow 7.70$): thus we see that the gain is concentrated in citation

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	$\mathbf{-1.307}$	**76.7%**	7.70
Step 1000 (decay)	-1.356	79.2%	7.64
Cont. $\rightarrow$ 1000	-1.326	77.8%	7.64
Cont. $\rightarrow$ 1250	-1.375	79.2%	7.52
Cont. $\rightarrow$ 1479	-1.408	79.7%	7.72

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

quality, not volume.

Continued training does not help. Under linear decay, step 750 $\rightarrow$ 1000 moves $\Delta R = -0.05$ ($z = -0.76$, n.s.), and the constant-LR continuation drifts down to $R = -1.41$ by step 1479 ($z = -1.6$ vs. the peak). The 3B model behaves roughly identically in shape, peaking at step 750 ($R = -1.70$, $z = +2.5$ vs. baseline) and relaxing toward baseline by step 1250. Interestingly, the peak-at-750 phenomenon is robust across two model scales, two LR schedules, and two independent post-750 data samples (since the step-1000 checkpoint on the decaying-LR schedule and step-1000 checkpoint on the constant-LR schedule are independent perturbations from the same step-750 policy). This supports the hypothesis that the reduction in performance likely originates from unwanted KL divergence after a substantial number of steps (750 in this case), after which further training 'breaks' the base model in subtle ways.

4.2 The win is broad

Another positive signal that was observed, which indicates genuine improvement at the best (step 750) checkpoint, was that the 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). This broad base matters for the reward-hacking question that follows: a gain concentrated on one or two questions would suggest the policy found a question-specific shortcut, whereas a broad improvement is the signature of a general behavioral change.

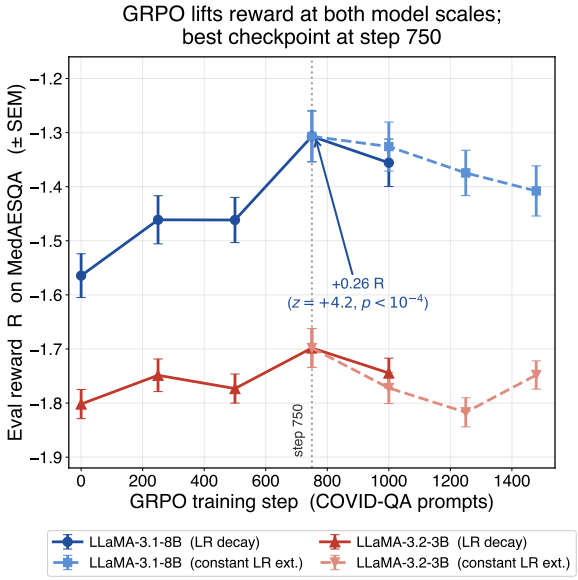

Figure 2: Reward across GRPO training. The 8B and 3B runs trace the same arc and peak at step 750 under both linear-decay and constant-LR schedules, then plateau or mildly regress. Bands are ± 1 SEM over 160 rollouts.

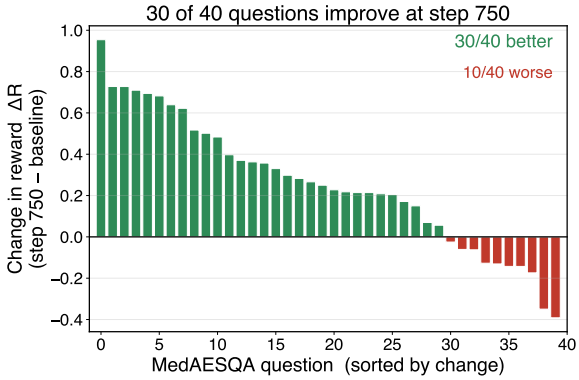

Figure 3: Per-question reward change, baseline $\rightarrow$ step 750 (8B). $30/40$ questions improve; the gain is broad rather than concentrated.

4.3 The gains are not reward hacking

As discussed in multiple lectures, reward-hacking is a common concern in the field of reinforcement learning, so I audited the peak checkpoint along six axes: cross-question reference recycling, within-rollout duplication, inline-marker/reference-list ID mismatch, cite-stuffing, the citation density of positive-reward rollouts, and the shape of the reward distribution. The first four are clean at every checkpoint—the top 5 most-reused references account for 1.6% of citations at both baseline and step 750 with no reference shared across five or more questions, duplicate rates stay near 10%, inline/list ID mismatches are 0%, and the number of

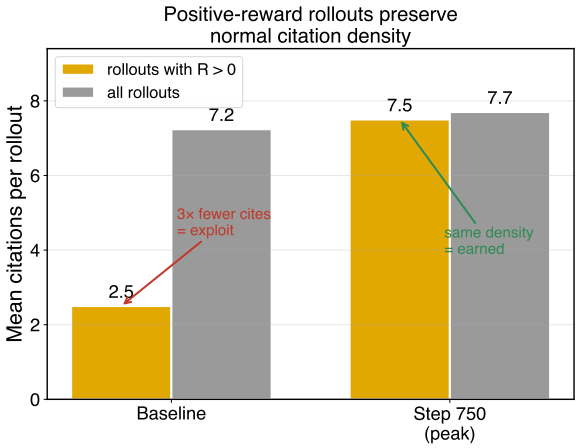

Figure 4: Citation density of positive-reward rollouts. At baseline, $R > 0$ rollouts cite $3\times$ less than average (an exploit); at step 750 they cite at average density, earning reward by citing real papers rather than by abstaining.

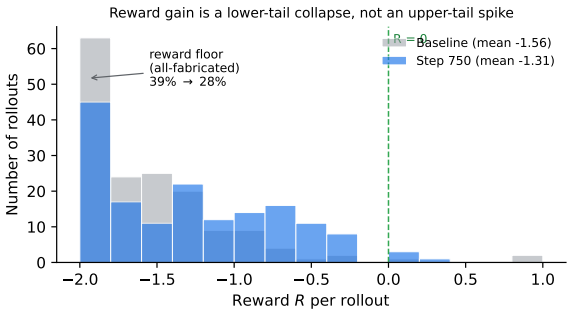

Figure 5: Reward distribution over 160 rollouts. The step-750 gain is a collapse of the lower tail—fewer fully-fabricated answers at the reward floor (39% → 28%)—rather than a spike at the high-reward end.

Checkpoint	Hedges / 1K chars
8B baseline	4.18
8B step 750 (peak)	4.05
8B cont. 1479	4.23
3B baseline	4.26
3B step 750 (peak)	**4.73**
3B cont. 1250	4.34

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

sentences with stuffed (≥ 4) citations stays at 0.

The remaining two tests appear to be even more decisive. First is the question of *which* rollouts earn 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, an exploit of dodging the penalty by citing only a handful citations that all happen to be real. 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 (the hard way). Secondly, I look at the shape of the reward distribution (Figure 5). The gain comes from shrinking the fraction of fully-fabricated answers at the reward floor (39% → 28%), not from inflating already-good ones, as the proportion of rollouts with $R > 0$ grows only 1.9% → 2.5%. In other words, RL is repairing the worst answers rather than gaming the best ones. Thus, the 8.8 pp reduction is genuine improvement in the model's abilities.

4.4 Ruling out eroded caution and advantage noise

Before turning to what the policy learns, I want to share results from two 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 stable or *rises* (see Table 3): the 8B model holds at ≈4.1–4.2 across baseline, peak, and late checkpoints, and 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 → 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% → 2%; 3B 35% → 15%). Neither eroded calibration nor advantage-noise amplification explains the decline.

4.5 The model learns to cite more carefully

Finally, we arrive at the most insightful section—*how* exactly does the trained (and better performing) model cite differently?

Here, only looking at aggregate reward hides the most interesting effects. First, reference count barely moves (7.2 → 7.7), but references carry-

ing 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 the most common hallucination type (Mugaanyi et al., 2024). After the RL training run, the LLM successfully learns not to fabricate DOIs and rely on recall-able title/author matches.

Secondly, the model gravitates more often toward canonical, high-impact sources: mentions of the New England Journal of Medicine (NEJM), one of the longest-running and most prestigious journals in the field, across all 160 rollouts, rise $68 \rightarrow 89$. Thus, the trained policy focuses on easily-recallable source that it might have seen more often during pretraining and thus are stored more prominently in its parametric knowledge—in other words, research articles from prominent journals and other high-impact publications, which are discussed frequently in subsequent literature. This is desirable behavior since such landmark papers are thoroughly vetted and thus highly likely to be both comprehensive and accurate should the user decide to follow the bibliography to learn more.

Figure 6 makes this concrete on a sample question from MedAESQA on the topic of IBS (While this is the question with the single-best improvement between base and best model, the overall results still hold generally.) The baseline produces eight references, of which only 2 resolve correctly (75% hallucinated); the step-750 policy leads with a genuine NEJM review, drops DOIs that might be associated with incorrect papers, and ultimately reaches $5/9$ verified. This shows that the model is recalling specific landmark literature more often instead of inventing plausible-sounding titles.

4.6 When the policy regresses: existence is not relevance

The same instinct to retrive canonical papers that drives broad gains can backfire on some questions. On a different question (which had the largest single-question regression between base and mean), the model was asked how cannabidiol (CBD) affects liver enzymes. This is a niche, recent topic with sparse indexed literature; here, the baseline cites an on-point study ("Effects of cannabidiol on liver enzymes in healthy volunteers"), while the step-750 policy drifts to the more famous CBD literature on anxiety, epilepsy, and schizophrenia. While those papers are more canonical and more recall-able, they are off-topic to the user's questions. This is a limitation of the current method/verifier which only rewards existence and not relevance, and provides room for future work (§6).

5 Discussion

A verifier supplies what learned rewards lack. Whereas RL against learned, RLHF-style rewards can *increase* hallucination (Li and Ng, 2025) or be gamed by exploiting the scorer (Chen et al., 2025), a deterministic external verifier gives an unspoofable factual signal: a citation either resolves in a scholarly index or it does not. The paper's headline result—an 8.8 pp reduction that survives a six-part audit—mirrors, for bibliographic provenance, what RLVR-for-truthfulness work has shown for short-answer factoid QA (Wei et al., 2025). And this follows the broader trend towards verifiable rewards as the 'gold standard' for LLM post-training.

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. The remaining shape resembles reward over-optimization (Gao et al., 2023), but my reward is deterministic, so this is not the learned-reward over-optimization that Gao et al. describe.

Instead, I believe two other hypotheses from recent RLVR theory might fit better. 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. (2025) 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 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.

A third candidate, which emerged repeatedly in

MedAESQA Q18: "what are the symptoms and treatments for IBS?" (✓ verified real, × hallucinated; best rollout per model)

Baseline — $2/8$ verified (75% hallucinated).
Answer: "IBS is a chronic GI disorder of abdominal pain, bloating, and altered bowel habits [1], with symptoms such as cramping [2], gas, and altered movements; causes involve gut bacteria, stress, and diet [3]. Treatment spans lifestyle [4], dietary change [5], medications [6], and a low-FODMAP diet [7,8]."
All eight references carry a DOI; most do not resolve to the cited title:

× [1] Chang & Chey (2017), *Gastroenterology Clinics* — DOI resolves to a different paper. ✓ [2] Mertz (2003), *J. Clin. Gastroenterology*. × [3] Spiller (2013), gut motility — not found. × [4] Chang & Chey (2017), "Stress and IBS" — DOI mismatch. ✓ [5] Spiller (2013), diet in IBS. × [6] Chey & Chang (2017), treatment of IBS — DOI mismatch. × [7] Biesiekierski et al. (2011), gluten/NCGS — DOI mismatch. × [8] Shepherd et al. (2012), low-FODMAP — not found.

Step 750 — $5/9$ verified (44% hallucinated).
Answer: "IBS is a chronic GI disorder with subtypes IBS-D/C/M [1,2], driven by altered motility, hypersensitivity, and the gut microbiome [3]. Treatment combines diet (fiber, low-FODMAP) [4], stress management (meditation, yoga, CBT) [5], and medications [6] — antispasmodics [7], laxatives [8], and SSRIs [9]."
Leads with a landmark NEJM review; DOIs dropped:

✓ [1] Chang & Talley (2018), *NEJM*. ✓ [2] Ford & Talley (2012), *Med. J. Australia*. ✓ [3] Spiller (2003), *Nat. Rev. Gastro. Hepatol.* × [4] Halmos et al. (2014), low-FODMAP — not found as cited. × [5] Grover & Al-Araque (2016) — not found. ✓ [6] Brandt et al. (2009), *Am. J. Gastroenterology*. ✓ [7] Garsed & Houghton (2012), *Clin. Gastro. Hepatol.* × [8] Gastrointestinal Society (2020), web page. × [9] Chang et al. (2018), psychobiotics — not found.

Figure 6: A matched MedAESQA example (best rollout per model). Both answers are clinically fluent; the trained policy improves validity ($2/8 \rightarrow 5/9$), gravitates to a canonical venue (NEJM), and abandons fabricated DOIs.

my discussions with other ML researchers when trying to understand the plateau, is *cumulative policy drift*: GRPO's per-step KL penalty and clipping set a bound on only the single-step deviation from the reference policy, not the total divergence accumulated across $1{,}479$ updates. Shenfeld et al. (2025) show that the KL divergence from the base policy is a reliable predictor of catastrophic forgetting, so as training continues past the peak the policy may drift far enough from the base model to erode the very parametric recall of real references on which the gains depend. This is also the simplest hypothesis and well-grounded in reinforcement learning theory (from CS224R lectures and in general), thus by Occam's Razor it might be the most likely explanation.

I would like to end this section by stressing that the decline is statistically mild ($z = -0.76$ to -1.6), so I offer these as candidate explanations for a possible issue rather than confirmed mechanisms behind a clearly documented phenomenon. The practical implication is that an early stop at the reward peak is the way to go.

6 Limitations

This paper uses a single held-out evaluation set (MedAESQA, 40 questions) and single-seed training runs, so statistical claims can rest on rollout variance and seed variance. 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 of little use to a reader. In the worst case, the LLM could actively print wrong statements but attach technically-correct citations to 'pass' the verifier. However, this would be a broader safety misalignment that is corrected with value alignment programs by the frontier labs.

The verifier also has a $\sim 7.7\%$ false-negative rate (short-title DOIs, proceedings, arXiv-only preprints), which penalizes some legitimate citations.

Scale itself also helps before any RL: doing one last evaluation on MedAESQA reveals that an untrained LLaMA-3.1-70B baseline hallucinates only 68.5% (vs. 76.7% for our 8B model *after* GRPO), reflecting the larger parametric store of real citations that comes with more parameters; compute constraints prevented me from fine-tuning the 70B model, but RL on a larger base to test whether the verifier-reward gain compounds with scale is the natural next step.

7 Conclusion

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 shows the gain is genuine learning rather than reward gaming: the model emits 71% more verified references at unchanged citation density, abandons fabricated 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. The benefit has a sharp, reproducible training horizon: both 3B and 8B peak at step 750 and regress thereafter through a mechanism consistent with catastrophic-forgetting, entropy-collapse and support-shrinkage phenomena in RLVR. Ultimately, the deterministic verifier supplies an objective, factual signal that drives the learning of the LLM and pushes it to be more accurate when citing sources from its own parametric knowledge.

Statement on Generative AI

I wrote code for the core GRPO training with LoRA myself, which is the important reinforcement learning part of the project. To be honest, this was pretty easy since I've worked with these libraries before, and they abstract away a lot of the production-level implementation details.

I did need to use AI assistance from Claude Code to take my initial scripts and Colab notebooks and make them runnable on Modal with checkpointing, remote file management, spinning up parallel GPU containers, which are all CS-Systems topics that were beyond the scope of this class.

However all the design decisions, such as formulating the reward function, made by me based on principles like asymmetric weighting of negatives and having consistent upper and lower bounds to reduce variance, which we learned in class content. I whiteboarded four different reward functions before I settled on the final one used in the paper.

I also used AI to port the original citation verifier, released by previous authors, into Python, since I don't understand TypeScript which the original work was implemented in. Even here, the modifications made to the original design, such as removing the Semantic Scholar API due to rate limiting issues, were entirely conceived by me, in order to adapt the verifier to the needs of my project.

Lastly, the core work of formulating the problem in a way that made it amenable to reinforcement learning in the first place; modifying initial hyperparameters after early failures in experiments; trying different learning-rate schedules to isolate the effect of one hyperparameter; and talking to other researchers to discuss the results I was observing; etc. were all done by me as well.

Note on template

I used the ACL template since I'm thinking of working on the project further after the quarter ends and potentially submitting it to the GroundLM workshop at EMNLP 2026—hope that's fine.

References

Diletta Abbonato. 2026. CheckIfExist: Detecting citation hallucinations in the era of AI-generated content. *arXiv preprint arXiv:2602.15871*.

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.

Xilun Chen and 1 others. 2025. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv 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.

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)*.

Junyi Li and Hwee Tou Ng. 2025. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. *arXiv preprint arXiv:2505.24630*.

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, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*.

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, and 1 others. 2025. TruthRL: Incentivizing truthful LLMs via reinforcement learning. *arXiv preprint arXiv:2509.25760*.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. 2025. The invisible leash? why RLVR may or may not escape its origin. *arXiv preprint arXiv:2507.14843*.