EvidenceRL: Reinforcing Evidence Consistency for Trustworthy Language Models

J. Ben Tamo[*♠], **Yuxing Lu**[*♠♡], **Benoit L. Marteau**[♠], **Micky C. Nnamdi**[♠], and **May D. Wang**[†♠]

[♠]Georgia Institute of Technology
[♡]Peking University

{jtamo3, yxlu, bmarteau3, mnnamdi3, maywang}@gatech.edu

Abstract

Large Language Models (LLMs) are fluent but prone to hallucinations, producing answers that appear plausible yet are unsupported by available evidence. This failure is especially problematic in high-stakes domains where decisions must be justified by verifiable information. We introduce **EvidenceRL**, a reinforcement learning framework that enforces evidence adherence during training. EvidenceRL scores candidate responses for grounding (entailment with retrieved evidence and context) and correctness (agreement with reference answers) and optimizes the generator using Group Relative Policy Optimization (GRPO). We evaluate across two high-stakes domains, cardiac diagnosis and legal reasoning, where EvidenceRL consistently improves evidence grounding and faithfulness without sacrificing task accuracy. On cardiac diagnosis, F1@3 increases from 37.0 to 54.5 on Llama-3.2-3B while grounding ($G_{\max}$@3) rises from 47.6 to 78.2; hallucinations drop nearly 5× and evidence-supported diagnoses increase from 31.8% to 61.6%. On legal reasoning, EvidenceRL raises Faithfulness from 32.8% to 67.6% on Llama-3.1-8B, demonstrating consistent behavioral change across domains. Our code is open-sourced at https://github.com/Wizaaard/EvidenceRL.git.

1 Introduction

The integration of Large Language Models (LLMs) into safety-critical domains such as healthcare, law, and finance has exposed reliability failures that limit real-world deployment (Gallagher et al., 2024). A key driver of these failures is the lack of evidence grounding: models often produce plausible outputs that are not supported by the available evidence (Gallagher et al., 2024; Magesh et al., 2025; NANDA, 2025; Sarvari et al., 2025). These findings indicate that without mechanisms to enforce evidence adherence, LLMs remain unreliable for high-stakes decision support.

Retrieval-Augmented Generation (RAG) attempts to mitigate hallucinations by conditioning generation on retrieved documents (Lewis et al., 2020). However, retrieval alone does not guarantee that outputs are derived from the provided evidence, and models frequently produce unsupported answers even when relevant context is available (Gao et al., 2023b; Zhang et al., 2025; Wallat et al., 2025). This failure mode, often termed context-memory conflict, persists even when high-quality evidence is provided, revealing a core limitation of existing RAG pipelines: retrieval is treated as a soft prompt rather than a constraint on generation.

Existing approaches address reliability through either inference-time filtering or training-time alignment. Post-hoc verification methods detect unsupported statements after generation (Manakul et al., 2023; Farquhar et al., 2024), while alignment techniques such as instruction tuning and reinforcement learning from human feedback (RLHF) shape model behavior but do not explicitly optimize evidence grounding (Peng et al., 2023; Bai et al., 2022). Hybrid approaches introduce retrieval or verification during reasoning (Asai et al., 2023; Dhuliawala et al., 2024), yet evidence usage remains loosely coupled to the training objective. As a result, current systems lack mechanisms that directly enforce evidence-consistent generation during learning.

To address these limitations, we introduce EvidenceRL, a reinforcement learning framework that optimizes evidence adherence during training rather than enforcing it post hoc. Unlike standard alignment methods that rely on sparse or noisy holistic rewards, EvidenceRL integrates two complementary signals: (1) a fine-grained grounding reward using a Focus-Then-Verify architecture to compute sentence-level entailment against the con-

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 894–914
November 6-10, 2026 ©2026 Association for Computational Linguistics

text and prevent signal dilution; and (2) a semantic correctness reward using an LLM judge to verify domain-specific answer equivalence. Across multiple model families (Llama, Gemma, GPT-oss) and scales (3B–120B), EvidenceRL consistently improves both task accuracy and evidential grounding. In the medical domain (MIMIC-IV-Ext), diagnostic F1@3 improves by up to 17 points while grounding increases substantially (e.g., on Llama-3.2-3B, $G_{\max}$@3 rises from 47.6 to 78.2). In the legal domain (BarExam MBE), EvidenceRL similarly shifts predictions toward evidence-supported reasoning, increasing the Evidence-Based rate from 18.8% to 41.0% and Faithfulness from 32.8% to 67.6% on Llama-3.1-8B. Behavioral analysis across both domains shows a clear redistribution toward evidence-based predictions: hallucinations drop by nearly $5\times$, while evidence-supported answers increase substantially.

Our contributions are as follows:

- We propose **EvidenceRL**, a reinforcement learning framework that **enforces** evidence grounding as a differentiable training objective using automated NLI and LLM-judge rewards, removing the dependency on human preference annotations.
- We introduce the **Focus-Then-Verify** reward architecture, which targets **sentence-level entailment** within focused context windows to resolve the signal dilution inherent in document-level scoring.
- We show that EvidenceRL improves accuracy and evidence grounding across model scales, reducing hallucinations and increasing evidence-supported predictions on high-stakes medical and legal domains.

2 Related Work

2.1 Evidence-Grounded Generation

RAG conditions LLM outputs on retrieved documents (Lewis et al., 2020), yet hallucinations remain common even in production systems (Huang et al., 2025). A central failure mode is *unconstrained generation*: models often produce answers that are correct in isolation but unsupported by the supplied evidence, a phenomenon termed *post-rationalization* (Wallat et al., 2024). Citation-based metrics frequently fail to detect this behavior because models can attach superficially related passages without relying on them during reasoning (Liu et al., 2023a). Similar issues appear in explanation and chain-of-thought generation, where produced rationales do not reflect the information actually used by the model (Turpin et al., 2023).

Recent work attributes these failures to competition between *parametric knowledge* and *contextual evidence*. When the two conflict, models often default to memorized associations rather than retrieved information (Xu et al., 2024; Longpre et al., 2021). Mechanistic studies link this behavior to interactions between feedforward knowledge circuits and attention-based copying mechanisms, where hallucinations arise when parametric pathways dominate contextual signals (Sun et al., 2025). Long-contexts exacerbate the problem, as relevant evidence is often attenuated or ignored (Liu et al., 2024), and models struggle to reason over conflicting retrieved sources (Cattan et al., 2025).

Existing solutions largely operate at inference time. Approaches such as Corrective-RAG (Yan et al., 2024), Self-RAG (Asai et al., 2024), CiteGuard (Choi et al., 2025), and RARE (Tran et al., 2025) introduce critique, filtering, or iterative retrieval, while FaithfulRAG (Zhang et al., 2025) explicitly models factual conflicts during reasoning. Other work targets transparency or mechanistic intervention (Ye et al., 2026; Shi et al., 2024). While these methods reduce hallucinations, they treat evidence grounding as a *post-hoc correction problem*. As a result, the underlying generation policy remains weakly constrained by evidence.

2.2 Verification and Post-hoc Fact Checking

A parallel line of work focuses on detecting unsupported statements after generation. Early evaluation metrics such as ROUGE fail to capture factual inconsistency, motivating NLI-based verification methods such as FactCC (Kryściński et al., 2020). Subsequent approaches adopt question-answering or claim-level verification formulations (QAGS, QuestEval, SummaC) to better capture semantic grounding (Wang et al., 2020; Scialom et al., 2021; Laban et al., 2022). More recent systems rely on LLM-based judges or self-consistency signals, including SelfCheckGPT, semantic entropy, and rubric-guided evaluation (Manakul et al., 2023; Farquhar et al., 2024; Liu et al., 2023b). Tool-based pipelines such as FacTool and FactScore further decompose generations into atomic facts for verification (Chern et al., 2023; Min et al., 2023).

While these methods improve hallucination de-

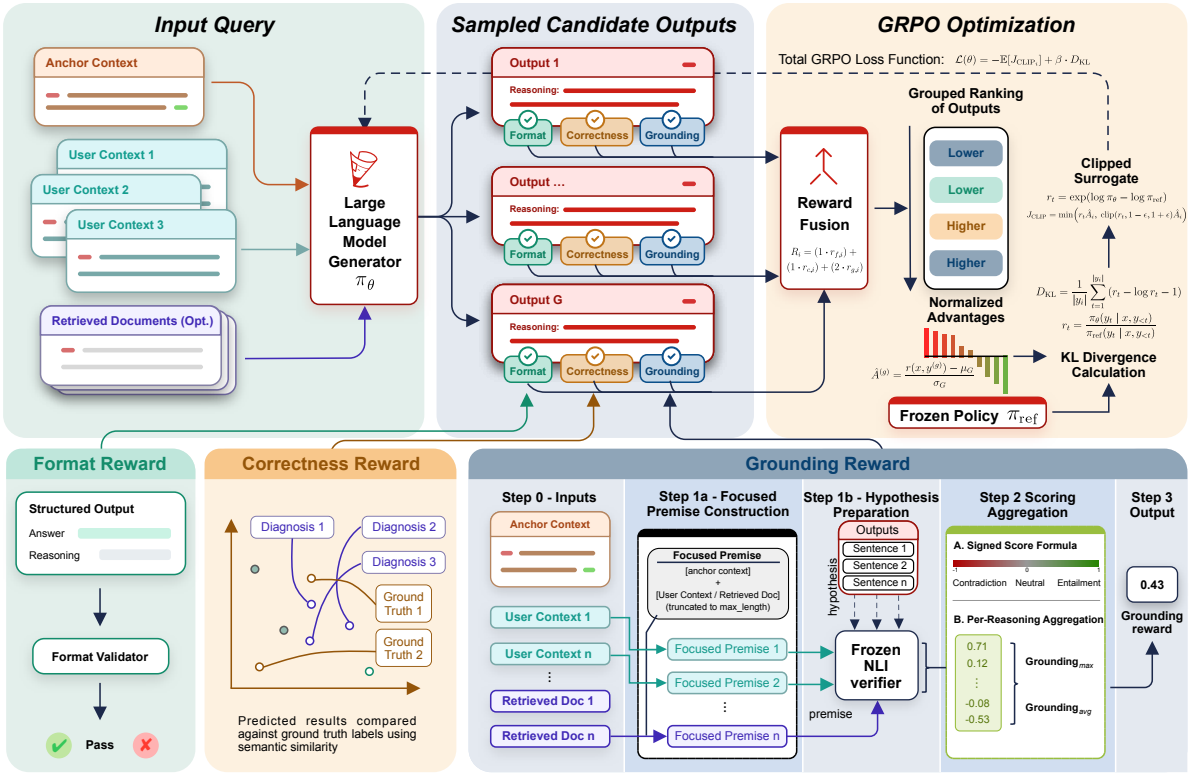


Figure 1: EvidenceRL aligns task accuracy with faithful evidence use across domains. Training uses GRPO with rewards for correctness (r_c), format (r_f), and evidence grounding (r_g). Grounding is computed via a *Focus–Then–Verify* procedure: (1) focused (premise, hypothesis) pairs are constructed by combining an anchor context with individual evidence sections, and (2) each pair is scored by a frozen NLI cross-encoder.

tection, verification alone does not alter generation behavior. Models can produce answers from parametric memory and attach superficially supportive evidence afterward (Gao et al., 2023a). Attribution benchmarks such as ALCE formalize this gap between correctness and groundedness. While recent work begins to optimize against verification signals (Tang et al., 2025; Xu et al., 2025), most approaches remain evaluation frameworks rather than directly enforcing evidence-consistent generation.

2.3 RL for Truthfulness and Grounding

Reinforcement learning has emerged as a scalable mechanism for aligning LLM behavior. RLHF improves helpfulness and safety (Ouyang et al., 2022), but human feedback is poorly calibrated for factual accuracy and uncertainty, often rewarding plausibility over verifiability (Augenstein et al., 2024; Casper et al., 2023). This misalignment leads to overconfident answers and weak grounding in evidence (Xiao et al., 2025). Methods such as TruthRL address this by explicitly rewarding epistemic humility and boundary awareness (Wei et al., 2025).

More recent work replaces human feedback with automated factuality signals. Preference optimization using factuality metrics such as FactScore has been shown to outperform human-labeled rewards (Tian et al., 2023), while aggregated verifier ensembles improve reward stability (Ye et al., 2025). In retrieval-augmented settings, RL has been used to encourage citation use and evidence-grounded answers, beginning with WebGPT and GopherCite (Nakano et al., 2021; Menick et al., 2022) and extending to modern RAG systems.

However, most existing methods optimize *outcome correctness* rather than *evidence consistency*. Context-DPO (Bi et al., 2025) improves robustness under conflicting contexts and PA-RAG (Wu et al., 2025) introduces sequential preference optimization for citation quality. Despite this progress, reward signals are typically coarse (document-level or answer-level), which allows models to remain correct while relying on unsupported reasoning. EvidenceRL builds on this line of work but targets a more fundamental objective: enforcing that model outputs remain *consistent with the provided evidence at the level of individual claims*.

3 Methodology: EvidenceRL

3.1 Setup and Notation

Let $\mathcal{D}$ be a dataset of pairs $(x, y^\star)$, where x is the input context and $y^\star$ is the reference answer. The generator is an autoregressive policy π_θ:

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T} \pi_\theta(y_t \mid y_{<t}, x). \tag{1}$$

Each sampled output y contains (i) an explicit reasoning segment $r(y)$ and (ii) final predictions.

3.2 Reward

We decompose the training reward into three components that jointly encourage well-structured, evidence-grounded, and diagnostically accurate outputs.

Format Reward. Because downstream grounding verification requires parseable structured reasoning, we include a binary format signal:

$$r_f(y) = \begin{cases} 1 & \text{if } y \text{ is valid JSON ,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Grounding Reward: Focus–Then–Verify Decomposition. Assessing whether reasoning is *grounded* is challenging in multi-source settings, as concatenating all evidence into a single NLI premise causes context overflow and signal dilution. We therefore adopt a **focus–then–verify** strategy that performs targeted natural language inference (NLI) checks against individual evidence sources. Let $a(x)$ denote an *anchor* context capturing the core framing of the input (chief complaint and history of present illness). For each supplementary section $s_i \in \mathcal{S}(x)$ (e.g., physical exam, imaging reports), we construct a focused premise:

$$\mathcal{P}(x) = \{a(x) \oplus s_i\}_{i=1}^{|\mathcal{S}|}. \tag{3}$$

When retrieval-augmented generation is used, retrieved documents $\{e_j\}_{j=1}^{k}$ are added as additional premises:

$$\mathcal{P}(x) \leftarrow \mathcal{P}(x) \cup \{a(x) \oplus e_j\}_{j=1}^{k}. \tag{4}$$

For each premise $p \in \mathcal{P}(x)$ we compute a support signal using a frozen NLI model:

$$\begin{aligned} \Delta_{\text{NLI}}(p, r(y)) = P(\text{entail} \mid p, r(y)) \\ - P(\text{contradict} \mid p, r(y)). \end{aligned} \tag{5}$$

We aggregate over premises by taking the score with the largest magnitude (sign preserved), capturing the strongest supporting or contradicting evidence:

$$r_g^{\max}(x, y) = \Delta_{\text{NLI}}(p^\star, r(y)), \tag{6}$$

$$p^\star = \text{argmax}_{p \in \mathcal{P}(x)} \left|\Delta_{\text{NLI}}(p, r(y))\right|. \tag{7}$$

We also report a complementary *average grounding score* measuring overall consistency across sources:

$$r_g^{\text{avg}}(x, y) = \frac{1}{|\mathcal{P}(x)|} \sum_{p \in \mathcal{P}(x)} \Delta_{\text{NLI}}(p, r(y)). \tag{8}$$

Correctness Reward. During training, correctness is estimated using a lightweight proxy rather than an LLM judge to keep online RL efficient. For open-ended tasks (e.g., clinical diagnosis), predicted answers are embedded with a domain-specific encoder and compared to references via cosine similarity:

$$r_c(x, y) = \frac{1}{|\hat{\mathcal{Y}}|} \sum_{\hat{y} \in \hat{\mathcal{Y}}} \mathbf{1}\Big[\max_{y^\star \in \mathcal{Y}^\star} s(\hat{y}, y^\star) > \tau\Big], \tag{9}$$

where $\hat{\mathcal{Y}}$ denotes the predicted answers, $\mathcal{Y}^\star$ the reference set, $\phi(\cdot)$ the embedding function, $s(\hat{y}, y^\star) = \cos(\phi(\hat{y}), \phi(y^\star))$ the cosine similarity between a prediction and a reference, and τ a similarity threshold. For tasks with discrete answer sets (e.g., multiple-choice legal reasoning), correctness reduces to exact matching with the reference.

Combined Reward. The final training reward combines all three components with scalar weights:

$$r(x, y) = w_f \cdot r_f(y) + w_c \cdot r_c(x, y) + w_g \cdot \tilde{r}_g(x, y), \tag{10}$$

where $\tilde{r}_g = (r_g + 1)/2$ is the normalized grounding score, and we set $w_f = w_c = 1$, $w_g = 2$ to emphasize evidence grounding, our primary objective.

3.3 Optimization with GRPO

We optimize π_θ using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each input x, we sample a group of G completions from the current policy, $\{y^{(1)}, \ldots, y^{(G)}\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$. Each completion is scored using the reward in Eq. (10). Rewards are normalized within the group to obtain advantages:

$$\hat{A}^{(g)} = \frac{r(x, y^{(g)}) - \mu_G}{\sigma_G}, \tag{11}$$

where μ_G and σ_G are the mean and standard deviation across the group. Following standard GRPO, the policy is updated using a clipped policy-gradient objective with KL regularization against a frozen reference policy π_{ref}, with advantages applied uniformly across tokens of each sampled completion.

4 Experiments

4.1 Task and Dataset

We evaluate EvidenceRL in two high-stakes domains, medicine and law. Medical diagnosis requires synthesizing clinical findings into ranked hypotheses, whereas legal analysis involves applying statutory rules to case facts.

MIMIC-IV-Ext (Cardiac). We use de-identified ICU cases from MIMIC-IV-Ext (Johnson et al., 2023). Ground-truth labels are derived from ICD-10 cardiac codes (prefix "I"). Given patient context x and optional retrieved evidence E_{pre}, the model predicts five ranked diagnoses with supporting reasoning. We split the dataset into 3,700 training and 1,000 held-out cases (Appendix A.2).

BarExam MBE (Legal). We use the Multistate Bar Examination benchmark (Zheng et al., 2025), where each instance contains a fact pattern, four answer choices, and a gold legal passage as evidence. Given context x and retrieved authority E_{pre}, the model selects the correct answer with grounded reasoning. The dataset includes 954 training and 173 test cases across six legal subjects.

Knowledge Source. For both tasks, models can receive retrieved domain-specific evidence E_{pre} to support grounded reasoning (Appendix A.3).

4.2 Models and Baselines

We evaluate eight backbones: Gemma-3 (4B, 12B, 27B), Llama (3.2-3B, 3.1-8B, 3.3-70B), and GPT-oss (20B, 120B). All methods use the same structured reasoning prompt (Appendix D).

We compare six approaches. Reasoning Only performs direct diagnosis generation from patient context without retrieval. Self-RAG performs adaptive retrieval and evidence critique during generation (Asai et al., 2024). Self-Consistency samples N completions and aggregates via semantic clustering and majority voting (Wang et al., 2022). SFT applies supervised fine-tuning on curated diagnosis–reasoning pairs (Appendix A.4). Faithful DPO (fDPO) constructs cross-model preference pairs based purely on NLI grounding quality, without any correctness signal, in the spirit of context-faithful preference optimization (Bi et al., 2025). EvidenceRL applies GRPO training with the combined reward in Eq. 10, jointly optimizing both grounding and correctness. All fine-tuning uses LoRA (Hu et al., 2022) (Appendix B).

4.3 Metrics

Task Correctness. We report F1@k or Accuracy. For MIMIC, correctness is determined by an LLM judge evaluating semantic equivalence between predicted and ground-truth diagnoses, Judge$(\hat{y}_i, y_j^\star)$.

Evidence Grounding. We report Grounding@k using the NLI scores $r_g^{\max}$ and r_g^{avg} (Eqs. 7–8), averaged over the top-k predictions.

Diagnostic Taxonomy. We classify each prediction into a 3×2 taxonomy based on correctness and grounding strength ($r_g^{\max}$). Grounded ($r_g^{\max} > 0.5$), Weak ($r_g^{\max} \in [-0.5, 0.5]$) and Contradicted ($r_g^{\max} < -0.5$):

	Correct Answer	Incorrect Answer
Grounded	Evidence-Based	Grounded Error
Weak	Weakly Supported	Unsupported Error
Contradicted	Lucky Guess	Hallucination

We report key rates: EB%, H%, LG%, and Weak%. We additionally compute faithfulness, which measures how often correct predictions are genuinely evidence-supported.

$$\text{Faithfulness} = \frac{\text{EB}}{\text{EB} + \text{WS} + \text{LG}}, \qquad (12)$$

5 Results

We evaluate EvidenceRL across two structurally distinct high-stakes domains, clinical diagnosis and legal reasoning, where ungrounded predictions carry real-world consequences, measuring both diagnostic accuracy and evidential reliability.

5.1 EvidenceRL Reduces the Faithfulness Gap and Narrows the Scale Gap

A persistent challenge in high-stakes domains is the *faithfulness gap*: models achieve high task accuracy through parametric recall rather than evidence-grounded reasoning. This pattern appears in both domains. In medicine (Table 1), Llama-3.1-8B achieves 56.6% F1@3 but only 19.4 G_{avg}@3, indicating that many correct diagnoses are unsupported

Table 1: Diagnostic performance and evidential reliability (±95% bootstrap CI). We report F1@3, average grounding (G_{avg}@3), and a diagnostic taxonomy decomposing predictions into evidence-based correctness (EB), hallucinations (H), and lucky guesses (LG). Faithfulness (F) measures the fraction of correct predictions supported by evidence (Eq. 12). Green marks the best backbone within each method; yellow marks the overall best.

Method	Backbone	F1@3 (↑)	G_{avg}@3 (↑)	G_{max}@3 (↑)	EB (↑)	H (↓)	W (↓)	LG (↓)	F (↑)
Reasoning Only	Llama-3.2-3B	37.0 ± 1.6	45.3 ± 3.5	47.6 ± 3.6	31.8%	11.6%	16.5%	**6.2%**	**72.5%**
	Llama-3.1-8B	38.2 ± 1.3	19.4 ± 3.6	21.3 ± 3.7	27.4%	17.7%	20.2%	11.5%	56.6%
	Llama-3.3-70B	51.3 ± 1.3	**46.9 ± 3.0**	**49.4 ± 3.1**	**43.8%**	**8.5%**	**14.3%**	9.3%	71.9%
	Gemma-3-4B	44.2 ± 1.4	34.3 ± 3.3	36.3 ± 3.4	35.1%	13.4%	19.2%	8.9%	66.0%
	Gemma-3-12B	46.5 ± 1.3	21.5 ± 3.2	23.5 ± 3.4	31.0%	14.4%	19.2%	14.3%	56.3%
	Gemma-3-27B	**52.2 ± 1.3**	37.1 ± 3.2	39.0 ± 3.4	39.1%	**8.5%**	15.5%	13.7%	63.1%
	GPT-oss-20B	46.0 ± 1.3	-2.2 ± 3.4	-0.9 ± 3.6	24.9%	15.6%	17.7%	26.0%	40.2%
	GPT-oss-120B	43.6 ± 1.3	18.3 ± 3.4	19.7 ± 3.5	30.8%	14.4%	15.5%	17.9%	53.3%
Self-RAG	Llama-3.2-3B	36.5 ± 1.5	24.4 ± 4.4	25.6 ± 4.6	26.0%	17.3%	18.5%	10.4%	59.6%
	Llama-3.1-8B	39.4 ± 1.3	30.0 ± 3.3	33.1 ± 3.5	31.5%	13.6%	19.9%	9.6%	62.7%
	Llama-3.3-70B	51.6 ± 1.3	39.9 ± 3.1	42.1 ± 3.3	**41.0%**	9.0%	**16.3%**	11.6%	66.3%
	Gemma-3-4B	46.5 ± 1.3	**46.2 ± 3.1**	**48.7 ± 3.2**	39.2%	**8.2%**	18.1%	**7.8%**	**70.4%**
	Gemma-3-12B	47.0 ± 1.3	17.5 ± 3.2	19.0 ± 3.4	29.5%	14.4%	19.3%	16.2%	52.9%
	Gemma-3-27B	**53.3 ± 1.3**	25.8 ± 3.4	27.4 ± 3.5	34.5%	9.5%	17.5%	17.7%	54.3%
	GPT-oss-20B	36.5 ± 1.6	0.2 ± 3.0	1.7 ± 3.1	25.7%	16.3%	17.6%	23.7%	42.7%
	GPT-oss-120B	37.1 ± 1.6	11.4 ± 3.1	12.9 ± 3.2	29.9%	15.7%	16.4%	18.2%	52.6%
Self Consistency	Llama-3.2-3B	39.3 ± 2.2	48.2 ± 4.7	50.3 ± 4.8	33.6%	10.7%	15.6%	**5.9%**	71.8%
	Llama-3.1-8B	34.8 ± 1.3	25.8 ± 3.1	27.6 ± 3.3	26.8%	17.6%	19.7%	8.7%	60.7%
	Llama-3.3-70B	46.8 ± 1.3	**51.2 ± 3.0**	**53.2 ± 3.1**	**41.2%**	9.2%	**12.1%**	7.8%	74.5%
	Gemma-3-4B	41.9 ± 1.4	48.3 ± 3.1	50.5 ± 3.2	37.3%	10.6%	14.9%	6.2%	**75.3%**
	Gemma-3-12B	41.0 ± 1.3	24.4 ± 3.3	26.0 ± 3.5	28.7%	16.4%	17.0%	12.0%	59.7%
	Gemma-3-27B	**48.6 ± 1.3**	45.9 ± 3.1	48.1 ± 3.2	39.6%	**8.7%**	13.9%	10.1%	68.7%
	GPT-oss-20B	44.0 ± 1.3	0.4 ± 3.5	1.6 ± 3.6	26.9%	16.5%	15.7%	24.6%	43.5%
	GPT-oss-120B	41.2 ± 1.3	25.1 ± 3.3	26.7 ± 3.4	32.3%	15.2%	16.1%	13.0%	59.2%
SFT	Llama-3.2-3B	**49.1 ± 1.3**	**34.8 ± 3.2**	**36.5 ± 3.4**	**41.3%**	**11.5%**	**15.5%**	**12.2%**	**67.0%**
	Llama-3.1-8B	47.2 ± 1.4	24.9 ± 3.3	26.5 ± 3.5	37.9%	13.3%	17.5%	13.8%	61.4%
	Llama-3.3-70B	35.2 ± 1.8	8.9 ± 3.0	10.0 ± 3.2	32.5%	14.3%	18.1%	19.5%	52.1%
	Gemma-3-4B	39.9 ± 1.6	2.1 ± 3.3	1.8 ± 3.4	28.4%	18.2%	18.6%	21.1%	47.3%
	Gemma-3-12B	33.8 ± 1.8	0.3 ± 3.1	0.6 ± 3.3	28.5%	16.9%	18.8%	23.4%	45.3%
	Gemma-3-27B	36.6 ± 1.7	1.2 ± 3.1	1.2 ± 3.3	27.8%	18.0%	19.7%	21.5%	45.7%
	GPT-oss-20B	47.7 ± 1.4	31.8 ± 3.4	33.1 ± 3.6	38.0%	12.1%	15.7%	12.8%	63.5%
fDPO	Llama-3.2-3B	38.7 ± 1.4	62.4 ± 2.7	63.6 ± 2.8	37.9%	7.9%	11.4%	3.6%	82.2%
	Llama-3.1-8B	39.7 ± 1.4	51.4 ± 3.4	53.0 ± 3.5	39.5%	9.0%	11.2%	7.3%	76.7%
	Gemma-3-4B	46.3 ± 1.3	**68.9 ± 2.6**	**70.8 ± 2.6**	**47.6%**	**5.9%**	**9.8%**	**3.1%**	**86.4%**
	Gemma-3-12B	47.9 ± 1.3	37.6 ± 3.1	39.9 ± 3.3	37.0%	10.2%	16.5%	11.0%	65.6%
	Gemma-3-27B	**53.1 ± 1.3**	62.5 ± 2.7	64.6 ± 2.8	48.5%	4.4%	11.6%	7.0%	77.7%
	GPT-oss-20B	46.0 ± 1.3	14.5 ± 3.4	15.7 ± 3.6	29.9%	13.1%	16.5%	20.8%	48.4%
EvidenceRL with GRPO (Ours)	Llama-3.2-3B	**54.5 ± 1.3**	**77.0 ± 2.2**	**78.7 ± 2.2**	**61.6%**	**2.4%**	**8.6%**	**3.4%**	**87.5%**
	Llama-3.1-8B	53.9 ± 1.4	**58.9 ± 2.9**	**61.1 ± 3.0**	**55.3%**	**4.3%**	14.2%	7.2%	**76.1%**
	Gemma-3-4B	46.8 ± 1.4	54.0 ± 3.0	56.0 ± 3.1	42.4%	7.8%	**13.9%**	**6.6%**	76.4%
	Gemma-3-12B	49.3 ± 1.3	38.0 ± 3.1	40.5 ± 3.3	38.0%	10.6%	17.2%	10.3%	65.9%
	Gemma-3-27B	**54.9 ± 1.3**	46.2 ± 3.1	48.4 ± 3.2	44.4%	7.0%	14.1%	11.4%	69.0%

Table 2: BarExam MBE: Accuracy, Evidence Grounding, and Diagnostic Taxonomy across backbone models. Wider confidence intervals reflect the smaller test(n=173) set compared to the medical domain (n=1,000).

Method	Backbone	Acc.(↑)	G_{avg}(↑)	G_{max}(↑)	EB(↑)	H(↓)	W(↓)	LG(↓)	F(↑)
Reasoning Only	Llama-3.2-3B	42.7 ± 9.0	-3.8 ± 7.2	5.3 ± 13.6	13.7%	15.4%	46.2%	9.4%	32.0%
	Llama-3.1-8B	57.3 ± 9.4	-1.7 ± 8.6	3.3 ± 12.2	18.8%	10.3%	51.3%	12.8%	32.8%
	Llama-3.3-70B	**82.1 ± 6.8**	**13.9 ± 4.8**	**49.5 ± 12.4**	**50.4%**	**1.7%**	**30.8%**	**6.0%**	**61.5%**
	Gemma-3-4B	47.9 ± 8.6	5.3 ± 6.6	14.5 ± 14.8	19.7%	10.3%	44.4%	6.8%	41.1%
	Gemma-3-12B	60.7 ± 8.5	4.2 ± 4.6	21.0 ± 15.3	22.2%	5.1%	43.6%	12.0%	36.6%
	Gemma-3-27B	65.8 ± 8.5	4.5 ± 4.4	20.2 ± 15.7	28.2%	7.7%	45.3%	7.7%	42.9%
	GPT-oss-20B	54.7 ± 8.5	-8.3 ± 8.1	-1.5 ± 14.7	15.4%	12.8%	53.8%	9.4%	28.1%
	GPT-oss-120B	77.8 ± 7.7	1.0 ± 6.0	7.7 ± 15.8	33.3%	9.4%	31.6%	17.9%	42.9%
EvidenceRL with GRPO (Ours)	Llama-3.2-3B	50.4 ± 9.0	7.7 ± 5.3	29.3 ± 14.4	26.5%	10.3%	36.8%	6.0%	55.9%
	Llama-3.1-8B	**60.7 ± 9.0**	**18.9 ± 6.2**	**51.9 ± 12.5**	**41.0%**	**6.0%**	26.5%	**4.3%**	**67.6%**
	Gemma-3-4B	48.7 ± 8.5	7.8 ± 4.4	33.8 ± 15.2	26.5%	**6.0%**	39.3%	6.8%	54.4%
	Gemma-3-12B	**60.7 ± 8.5**	11.0 ± 5.9	33.6 ± 15.2	37.6%	6.8%	**23.1%**	9.4%	62.0%

by patient evidence. The legal domain (Table 2) shows an even stronger version of this effect: under reasoning-only inference, grounding is negative for several models (e.g., $G_{\text{avg}} = -3.8$ for Llama-3.2-3B and -8.3 for GPT-oss-20B). Even the strongest baseline, Llama-3.3-70B (Acc 82.1%, F 61.5%),

exhibits only moderate grounding ($G_{\text{avg}} = 13.9$) with an EB rate of 50.4%, meaning nearly half of correct predictions remain non-evidential.

EvidenceRL training fundamentally shifts the *basis* of correct predictions. In the medical domain, GRPO increases Llama-3.2-3B's EB correctness from 31.8% to 61.6% while reducing hallucinations and lucky guesses to 2.4% and 3.4%, and raising F1@3 from 37.0 to 54.5. A similar redistribution appears in the legal domain: Llama-3.1-8B's EB rate rises from 18.8% to 41.0%, hallucinations fall from 10.3% to 6.0%, lucky guesses from 12.8% to 4.3%, and faithfulness increases from 32.8% to 67.6% alongside G_{avg} improving from -1.7 to 18.9. Across both domains, EvidenceRL provides a training signal that penalizes unsupported correctness, moving correct predictions into the explicitly evidence-grounded category.

This shift also narrows the traditional scaling advantage of larger models. Under zero-shot reasoning, larger models are not consistently more evidence-grounded (e.g., Gemma-3-4B achieves higher grounding than Gemma-3-12B in both domains). After EvidenceRL training, smaller models can surpass much larger baselines while remaining more grounded: in medicine, EvidenceRL Llama-3.2-3B (F1@3 = 54.5) exceeds reasoning-only Llama-3.3-70B (51.3) and Gemma-3-27B (52.2); in law, EvidenceRL Llama-3.2-3B (F = 55.9%) surpasses several larger reasoning-only models. These results indicate that improving *how* models use evidence during training can partially substitute for parameter scale, offering a compute-efficient path to reliable reasoning in high-stakes settings.

5.2 When the Objective Is Not Aligned: SFT and Inference-Time Controls

Methods that do not modify the learning objective fail to reliably align diagnostic accuracy with evidence grounding. Supervised Fine-Tuning (SFT) preserves plausible predictions but collapses grounding: across the Gemma family, $G_{\text{avg}}@3$ falls nearly to zero (e.g., 37.1 $\rightarrow$ 1.2 on Gemma-3-27B) while F1@3 remains similar (36.6). This indicates that SFT teaches models to imitate answer format and cite evidence without learning the semantic relationship between diagnoses and supporting text.

Inference-time controls show a different but similarly limited effect. Techniques such as Self-RAG and Self-Consistency occasionally improve grounding but fail to consistently improve both accuracy and evidence use. For example, on Llama-3.2-3B, Self-RAG slightly reduces F1@3 (37.0 $\rightarrow$ 36.5) while substantially lowering grounding (45.3 $\rightarrow$ 24.4), while Self-Consistency improves grounding through aggregation but retains a higher fraction of reasoning failures. These patterns suggest that inference-time methods primarily redistribute outputs within the model's existing preference structure, whereas EvidenceRL alters the learning objective itself through reward-based credit assignment, enabling simultaneous gains in diagnostic accuracy and evidential reliability.

5.3 Comparing Alignment Objectives: fDPO vs. EvidenceRL

Both fDPO and EvidenceRL improve evidence grounding over SFT and inference-time baselines, but they optimize the accuracy–grounding trade-off through different mechanisms.

fDPO trains on cross-model preference pairs, where chosen responses come from the most grounded model outputs and rejected responses from the least grounded. This provides a strong offline grounding signal (mean grounding +0.75 vs. −0.55), but also introduces implicit cross-model distillation: weaker backbones may be trained to imitate responses generated by stronger models. On Llama-3.2-3B (Table 1) this results in high Faithfulness (82.2%) but little accuracy improvement (F1@3 = 38.7 vs. 37.0 baseline). The Evidence-Based rate increases only modestly (31.8%$\rightarrow$37.9%), while Weak and Lucky Guess predictions shrink (16.5%$\rightarrow$11.4%, 6.2%$\rightarrow$3.6%). Faithfulness rises primarily because fewer predictions remain correct rather than because more grounded correct predictions are produced.

EvidenceRL avoids this confound by computing rewards on-policy. GRPO evaluates grounding and correctness directly on the model's own rollouts, rewarding improvements relative to its own candidate generations. As a result, grounded correctness expands rather than contracts. On Llama-3.2-3B, the Evidence-Based rate nearly doubles (31.8%$\rightarrow$61.6%) while diagnostic accuracy also increases (F1@3 = 54.5), yielding the highest Faithfulness (87.5%). Here, improvements arise from generating more predictions that are both correct and evidence-supported.

5.4 Reward Proxy Alignment Analysis

Figure 3 compares the BioLORD-2023 similarity threshold ($\tau = 0.80$) against a judge's binary verdicts. Across 24,896 pairs from our GRPO-trained

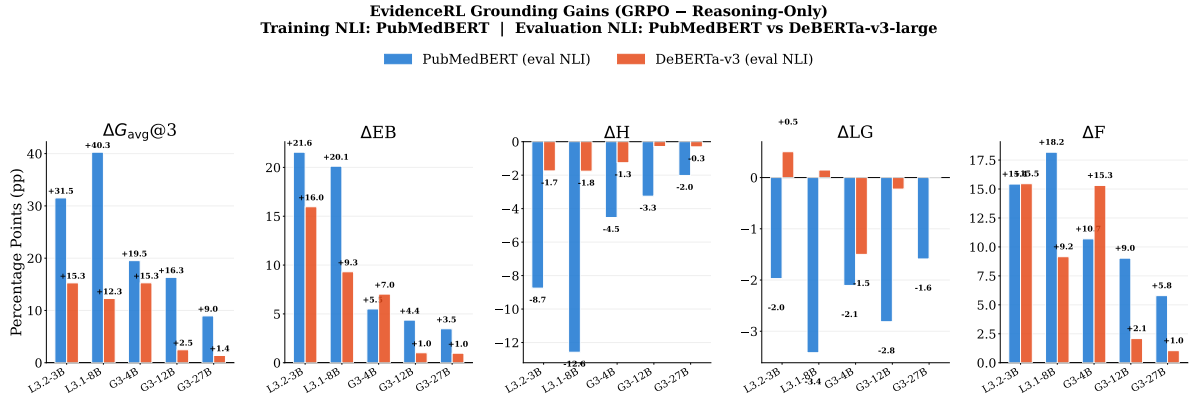


Figure 2: All five metrics shift in the same direction under both evaluators, confirming that grounding improvements reflect genuine evidence use rather than reward model overfitting.

models, the embedding proxy achieved 98.0% precision, preventing the RL policy from being falsely rewarded for incorrect predictions. The rare false positives (145 cases, or 0.58%) were legitimate clinical near-misses (e.g., "Left Bundle Branch Block" vs. "Atrioventricular and left bundle-branch block") rather than systematic exploitation. Furthermore, untrained reasoning baselines demonstrated a nearly identical precision (97.2% across 38,629 pairs), confirming that the optimization process does not learn to hack the proxy metric. Although the conservative $\tau = 0.80$ threshold limits recall to 58.0% by under-crediting non-literal synonyms, this asymmetry safely prioritizes strict error penalization over rewarding diverse phrasing.

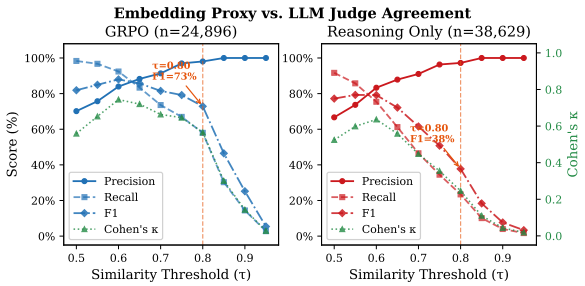


Figure 3: At $\tau = 0.80$, precision remains high for both approaches. Stable recall and high Cohen's κ indicate a conservative reward signal, with no evidence of proxy hacking by GRPO models.

To ensure grounding improvements reflect genuine evidence use rather than overfitting to the training NLI reward model, we re-evaluated all models using an independent evaluator (DeBERTa-v3-large). As shown in Figure 2, grounding gains persist across all five backbones: Evidence-Based predictions increased by 1.0–16.0 pp, Hallucinations decreased by 0.3–1.8 pp, and Faithfulness improved by 1.0–15.5 pp. Although absolute magnitudes are smaller, likely due to DeBERTa-v3-large lacking biomedical pretraining, the consistent direction of improvement confirms that GRPO enhances true evidence-grounded reasoning rather than exploiting reward model artifacts.

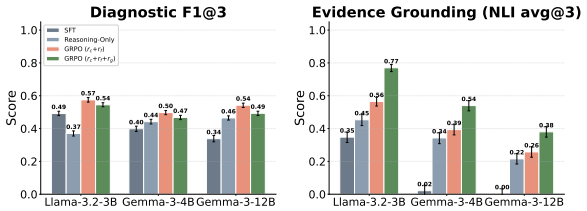


Figure 4: SFT yields reasonable accuracy but weak grounding. GRPO with correctness reward (r_c+r_f) maximizes F1, while adding the grounding reward (r_g) substantially improves evidence attribution with only minor accuracy trade-offs.

5.5 Ablation Study

Figure 4 isolates the contribution of each reward component across three backbones on the medical domain. SFT produces plausible but ungrounded outputs: while diagnostic precision remains competitive (e.g., P@3=0.614 on Llama-3.2-3B), grounding collapses on the Gemma models (G∗avg@3 = 0.021 for Gemma-3-4B and 0.003 for Gemma-3-12B), indicating that demonstration tuning teaches citation format without enforcing semantic evidence use.

RL with correctness reward (r_c+r_f) yields the highest diagnostic accuracy across all models (F1@3: 0.575, 0.497, 0.542) while moderately improving grounding (0.564, 0.394, 0.258). Adding the grounding reward (r_g) substantially increases evidence grounding (0.770, 0.540, 0.380)

with only a small reduction in diagnostic accuracy (e.g., 0.575→0.545 on Llama-3.2-3B). Overall, the full reward produces the best accuracy–grounding trade-off, demonstrating that explicit grounding supervision is necessary to align diagnoses with supporting evidence.

6 Conclusion

Large language models can produce correct answers while ignoring the evidence provided in context. In high-stakes domains such as clinical diagnosis and legal reasoning, this leads to plausible but unsupported outputs that undermine trust. EvidenceRL addresses this by making evidence adherence a training objective. Across multiple model families and two domains, it consistently shifts model behavior toward predictions that are both accurate and evidence-grounded while reducing hallucinations and lucky guesses, without sacrificing task performance. These results highlight a fundamental limitation of inference-time controls: retrieval, self-consistency, and post-hoc verification can filter outputs but cannot change how models reason. Grounding-aware reinforcement learning does, reshaping the model's decision policy to prefer evidence-supported reasoning.

7 Limitation

EvidenceRL relies on automated proxies to scale training and evaluation. Grounding is assessed using a frozen biomedical NLI cross-encoder, which provides a domain-appropriate signal but remains an approximation of true evidential reasoning. We mitigate this through sentence-level verification and by evaluating all methods with the same grounding model, ensuring any systematic bias affects approaches equally. Training cost is higher than SFT because GRPO requires sampling and scoring multiple completions per iteration; inference cost remains unchanged. As with retrieval-augmented systems generally, the method assumes retrieved evidence is reliable, so errors in the knowledge source may propagate into otherwise well-grounded outputs. The fDPO baseline uses cross-model preference pairs, introducing a mild knowledge distillation component alongside the grounding signal; within-model grounding-only optimization would provide a cleaner ablation and is a natural extension of this work. Finally, while we demonstrate consistent improvements across two high-stakes domains, cardiac diagnosis and legal reasoning, evaluating EvidenceRL across additional domains and evidence sources remains an important direction for future study.

8 Potential risks

EvidenceRL is a research contribution toward more trustworthy AI reasoning and is not intended for autonomous deployment in clinical or legal settings. Responsible use in these domains would require prospective validation, regulatory evaluation, and human-in-the-loop oversight. While EvidenceRL improves evidence grounding, well-cited reasoning may appear authoritative even when the underlying conclusion is incorrect. The framework therefore aims to make model reasoning *verifiable* rather than automatically trustworthy. In practical deployments, evidence citations should support human review rather than replace it. The grounding reward enforces consistency with retrieved evidence but does not verify the correctness of that evidence itself. Errors or outdated information in the knowledge source may therefore propagate into otherwise well-grounded outputs, making evidence curation and source quality essential. Finally, training at the scales explored here requires substantial computational resources. Ensuring that evidence-grounded AI tools remain accessible across institutions will require continued work on efficient training and adaptation to smaller models.

Acknowledgments

The authors acknowledge the AI Makerspace of the College of Engineering (RRID:SCR_028058), provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA (RRID:SCR_027619), for computational resources used in this work.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott

Hale, Alon Halevy, and 1 others. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, and 1 others. 2025. Context-dpo: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300.

Jiawei Cao and Sendong Zhao. 2025. MIMIC-IV-Ext Cardiac Disease. *PhysioNet*. Version 1.0.0.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *arXiv preprint arXiv:2506.08500*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Yee Man Choi, Xuehang Guo, Yi R Fung, and Qingyun Wang. 2025. Citeguard: Faithful citation attribution for llms via retrieval-augmented validation. *arXiv preprint arXiv:2510.17853*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Shannon K Gallagher, Jasmine Ratchford, Tyler Brooks, Bryan P Brown, Eric Heim, William R Nichols, Scott Mcmillan, Swati Rallapalli, Carol J Smith, Nathan VanHoudnos, and 1 others. 2024. Assessing llms for high stakes applications. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 103–105.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and 1 others. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

MIT NANDA. 2025. State of ai in business 2025. *Preprint at https://www. artificialintelligence-news. com/wp-content/uploads/2025/08/ai_report_2025. pdf*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1586–1596.

Peter Sarvari, Zaid Al-Fagih, Alexander Abou-Chedid, Paul Jewell, Rosie Taylor, and Arouba Imtiaz. 2025. Challenges and solutions in applying large language models to guideline-based management planning and automated medical coding in health care: Algorithm development and validation. *JMIR Biomedical Engineering*, 10(1):e66691.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6594–6604.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Xiaqiang Tang, Yi Wang, Keyu Hu, Rui Xu, Chuang Li, Weigao Sun, Jian Li, and Sihong Xie. 2025. Ssfo: Self-supervised faithfulness optimization for retrieval-augmented generation. *arXiv preprint arXiv:2508.17225*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025. Rare: Retrieval-augmented reasoning enhancement for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18305–18330.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not faithfulness in rag attributions. *arXiv preprint arXiv:2412.18004*.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, and 1 others. 2025. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. 2025. Pa-rag: Rag alignment via multi-perspective preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9091–9112.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. *arXiv preprint arXiv:2505.01997*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Zhichao Xu, Zongyu Wu, Yun Zhou, Aosong Feng, Kang Zhou, Sangmin Woo, Kiran Ramnath, Yijun Tian, Xuan Qi, Weikang Qiu, and 1 others. 2025. Beyond correctness: Rewarding faithful reasoning in retrieval-augmented generation. *arXiv preprint arXiv:2510.13272*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. 2026. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. *arXiv preprint arXiv:2601.06842*.

Yuxuan Ye, Raul Santos-Rodriguez, and Edwin Simpson. 2025. Optimising factual consistency in summarisation via preference learning from multiple imperfect metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17342–17355.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193.

A Experimental Setup

A.1 LLMs Usage

Large Language Models (LLMs) were used solely as general-purpose assistive tools to help polish the manuscript's language and to refine instructions within our prompt templates. Specifically, LLMs aided in improving grammar, clarity, and style, and in suggesting alternative phrasings for prompt templates. All scientific ideas, experimental design, and key arguments were conceived and written by the authors, and all factual statements were independently verified.

A.2 Datasets

A.2.1 MIMIC-IV-Ext Cardiac Dataset Details

We use the MIMIC-IV-Ext Cardiac Disease dataset (Cao and Zhao, 2025), a collection of de-identified ICU hospitalizations with structured clinical notes and gold-standard cardiac diagnoses. Each case includes the following note sections:

- **Chief complaint**: primary reason for admission (e.g., dyspnea, orthopnea).
- **History of present illness (HPI)**: narrative clinical history preceding presentation.
- **Physical exam**: vital signs and examination findings.
- **Imaging**: radiology reports (e.g., X-ray, CT, MRI).
- **Catheterization (CATH)**: invasive hemodynamic and procedural findings.
- **ECG / ECG machine report**: electrocardiographic interpretation.
- **Invasions**: invasive procedure documentation.

Ground-truth diagnoses are derived from ICD-10 codes with the cardiac prefix ("I"). To prevent label leakage and standardize note structure, we apply the following preprocessing steps:

1. Normalize line breaks across heterogeneous note formats.
2. Remove placeholders and template artifacts (e.g., list markers, boilerplate fragments).
3. Remove diagnostic summary sections (`IMPRESSION`, `FINAL DIAGNOSIS`) so the model must infer diagnoses from the clinical evidence.
4. Deduplicate ECG machine text to eliminate repeated autogenerated phrases.
5. Normalize whitespace by collapsing repeated spaces and empty lines.

A.2.2 BarExam Dataset Details

We evaluate legal reasoning using the BarExam QA benchmark (Zheng et al., 2025), which consists of multiple-choice questions derived from the Multistate Bar Examination (MBE). Each instance includes:

- **Fact pattern**: narrative describing the legal scenario, including events, actors, and circumstances.
- **Question stem**: the specific legal issue to resolve.
- **Four answer choices** (A–D): one correct option and three distractors.
- **Gold legal passage**: an authoritative legal excerpt (statute, rule, or case law) supporting the correct answer.

Questions span six core MBE subjects: Constitutional Law, Contracts, Criminal Law, Evidence, Real Property, and Torts. The dataset draws from historical MBE administrations between 1972 and 1998.

A.3 Retrieval Pipeline

Both domains use an optional retrieval pipeline to provide external evidence when required. We encode text using `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) (384-d embeddings) and index passages with FAISS (Douze et al., 2025) for approximate nearest-neighbor search via cosine similarity.

In the medical domain, we index cardiovascular clinical knowledge from the `ilyassacha/cardiologyChunks` dataset (3.2M records), chunked into 320-token segments with 64-token overlap, yielding 822,861 indexed chunks. Retrieval queries are constructed from the chief complaint and history of present illness, and the top-$k = 3$ passages are inserted into the prompt as evidence.

In the legal domain, we index the BarExam legal passage corpus (856,835 passages) comprising case law paragraphs, Wex encyclopedia entries, and MBE explanations. Passages are pre-segmented and indexed directly. Depending on the experiment, models receive either the gold supporting passage or the top-$k = 3$ retrieved passages using the question text as the retrieval query.

A.4 Model Backbones and Inference Configuration

We evaluate EvidenceRL across multiple open-weight instruction-tuned model families spanning

Table 3: Retrieval pipeline configuration.

Parameter	Medical	BarExam
Embedding model	`all-MiniLM-L6-v2`	
Embedding dim.	384	
Index type	FAISS HNSW	
Chunk size	320 tokens	N/A (pre-segmented)
Chunk overlap	64 tokens	N/A
Top-k	3	1 (gold) / 3 (FAISS)

a range of parameter scales, including Llama 3.x, Gemma 3, and GPT-OSS variants. The evaluated models range from 3B to 120B parameters and represent diverse architectures and training recipes. Table 4 lists all backbones used in our experiments.

Table 4: Model backbones used in experiments. All models run in bf16 via vLLM for inference.

Model	Family	Params	Inference
Llama-3.2-3B-It	Llama 3.2	3B	bf16, TP=2
Llama-3.1-8B-It	Llama 3.1	8B	bf16, TP=2
Llama-3.3-70B-It	Llama 3.3	70B	bf16, TP=4
Gemma-3-4B-IT	Gemma 3	4B	bf16, TP=2
Gemma-3-12B-IT	Gemma 3	12B	bf16, TP=2
Gemma-3-27B-IT	Gemma 3	27B	bf16, TP=2
GPT-OSS-20B	GPT-OSS	20B	bf16, TP=2
GPT-OSS-120B	GPT-OSS	120B	bf16, TP=4

All inference is performed using `vLLM` with bfloat16 precision, which provides efficient batched decoding and tensor-parallel serving. Models up to 27B parameters use tensor parallelism with TP=2 across two NVIDIA H100 GPUs (80 GB each); the 70B and 120B models use TP=4 across four NVIDIA H200 GPUs (141 GB each). GPT-OSS generation uses a maximum generation length of 4,092 tokens. Detailed inference parameters are summarized in Table 5.

Table 5: vLLM inference configuration.

Parameter	Standard	Self-Consistency
Temperature	0.7	0.9
Top-p	0.9	0.9
Top-k	50	50
Repetition penalty	1.15	1.15
Max tokens	2,048 / 1,024*	2,048
n (samples per prompt)	1	10
Tensor parallelism	2–4	2
GPU memory utilization	0.90	0.90

*2,048 for medical, 1,024 for BarExam.

A.4.1 Self-Consistency Pipeline

Our self-consistency (SC) implementation follows Wang et al. (2022) but adapts majority voting for free-form clinical diagnosis text using embedding-based semantic clustering rather than exact-match voting.

1. **Diverse generation**: Generate $N = 10$ completions per patient using higher temperature ($T = 0.9$) via vLLM's batched n-parameter sampling.
2. **Parse**: Extract 5 diagnoses from each completion (up to 50 diagnoses per patient).
3. **Pool**: Collect all diagnosis names across all N samples.
4. **Cluster**: Embed all unique diagnosis names with `BioLORD-2023` (Remy et al., 2024) (768-dim, MPNet-based). Apply greedy agglomerative clustering: process names sorted by frequency (most common first); for each name, merge into the first existing cluster whose centroid has cosine similarity ≥ 0.85, or start a new cluster.
5. **Rank**: Rank clusters by a composite score: score $=$ vote_count $\times$ 100 $-$ avg_position, where vote count is the number of distinct samples containing the diagnosis and avg_position is its average rank across samples.
6. **Select reasoning**: For the top-5 clusters, select the best reasoning paragraph (longest reasoning among votes ranked in top-2 within their respective samples).

A.4.2 Self-RAG Pipeline

Our Self-RAG (Asai et al., 2023) implementation adapts the retrieve-then-read paradigm with an explicit self-critique step that makes retrieval *adaptive*, only patients with uncertain diagnoses trigger evidence retrieval.

1. **Zero-shot generation**: Generate an initial set of 5 diagnoses with reasoning from patient context alone (no evidence).
2. **Self-critique**: The same model evaluates each diagnosis, assigning a confidence level (`HIGH`/`LOW`) and a boolean `needs_evidence` flag.
3. **Conditional retrieval**: If any diagnosis is flagged as uncertain, a retrieval query is constructed from the uncertain diagnoses' names and reasoning. The top-k=3 chunks are retrieved via FAISS. Patients with all high-confidence diagnoses skip retrieval entirely.
4. **Selective refinement**: For patients with retrieved evidence, the model regenerates all 5 diagnoses given the original patient context, its initial diagnoses (annotated with confidence tags), and the retrieved evidence. Patients without re-

trieval retain their zero-shot output.

B Training and Reward Implementation

B.1 SFT, fDPO, and GRPO Datasets

We construct separate training datasets for supervised fine-tuning (SFT), faithfulness DPO (fDPO), and GRPO reinforcement learning. All datasets are derived from the same 3,700 training patients in the MIMIC-IV-Ext cardiac dataset or 954 training cases across six legal subjects in BarExam.

SFT Dataset. SFT examples are generated from candidate outputs produced by all eight model backbones on the training cases. Each model predicts five ranked diagnoses with reasoning, yielding up to 29,600 candidate outputs. Outputs are filtered using the diagnostic taxonomy (Section 4.3). An output is retained if at least two of the top-3 diagnoses are *Evidence-Based (EB)*, meaning the diagnosis is both correct and grounded ($r_g^{\max} > 0.5$). Qualified outputs are ranked by (i) the number of EB diagnoses in the top-3 and (ii) the mean $r_g^{\max}$ across the top-3 as a tiebreaker. After deduplication, we cap at two examples per patient to avoid over-representation. The final SFT dataset contains 4,534 examples from 2,634 patients.

fDPO Dataset. The fDPO dataset provides preference pairs that encourage grounded reasoning. For each training patient, all eight backbones generate responses under the no-retrieval setting. Each response is scored by the NLI grounding evaluator, and we compute the mean grounding score across the five predicted diagnoses. For each patient, the highest-scoring response is selected as the *chosen* example and the lowest-scoring as the *rejected* example. We apply the following filtering criteria:

1. grounding gap ≥ 0.7 between chosen and rejected responses,
2. chosen response grounding ≥ 0.1,
3. rejected response grounding ≤ -0.1.

This yields 2,292 preference pairs with a mean grounding gap of 1.30 (chosen mean = 0.75, rejected mean = -0.55). No correctness filtering is applied, making fDPO a pure grounding preference objective.

GRPO Dataset. GRPO training uses all 3,700 training patients (954 in the legal domain) without filtering. Each entry contains the patient context, ground-truth diagnoses, and optional retrieved evidence (for RAG variants). Unlike SFT and fDPO, GRPO does not rely on pre-generated outputs; completions are generated on-policy during training and rewards are computed dynamically.

B.2 Grounding NLI Models

Medical Domain. We use PubMedBERT-MNLI-MedNLI[1] as the frozen NLI model for grounding evaluation. This is a PubMedBERT-based (Gu et al., 2021) cross-encoder fine-tuned on MultiNLI (Williams et al., 2018) and MedNLI (Romanov and Shivade, 2018), providing domain-specific natural language inference for clinical text.

Legal Domain. We use nli-deberta-v3-large[2], a DeBERTa-v3-Large cross-encoder fine-tuned on SNLI and MultiNLI. As no legal-domain NLI model is publicly available, we use this general-purpose model with token-aware premise truncation to ensure the hypothesis is never clipped at the 512-token input limit.

B.3 Training Hyperparameters

All models are fine-tuned using LoRA adapters (Hu et al., 2022) applied to all transformer projection layers.

Table 6: SFT training hyperparameters.

Parameter	Value
Epochs	3
Learning rate	2×10^{-5}
Scheduler	Cosine with 10% warmup
Per-device batch size	2
Gradient accumulation	8
Max sequence length	2,048 tokens
Loss masking	Completion-only
Precision	bfloat16

During GRPO, each of the $G{=}8$ completions per prompt is scored by three independent reward functions that share the same structure across domains but differ in their domain-specific implementations:

1. R_{format}: Binary format compliance. Returns $\{0, 1\}$.
   - *Medical*: Valid JSON with exactly 5 diagnoses, each containing non-empty name and reasoning fields.
   - *Legal*: Valid JSON with an answer letter (A/B/C/D) and non-empty reasoning.
2. $R_{\text{correctness}}$: Task correctness. Returns $[0, 1]$.

[1] https://huggingface.co/pritamdeka/PubMedBERT-MNLI-MedNLI

[2] https://huggingface.co/cross-encoder/nli-deberta-v3-large

Table 7: fDPO training hyperparameters. Uses TRL `DPOTrainer` with sigmoid loss on faithfulness preference pairs.

Parameter	Value
Epochs	2
Learning rate	5×10^{-6}
Scheduler	Cosine with 10% warmup
Per-device batch size	1
Gradient accumulation	8
DPO β	0.1
Loss type	Sigmoid
Max sequence length	8,192 tokens
Max prompt length	6,144 tokens
Validation split	5%
Weight decay	0.01
Precision	bfloat16

Table 8: GRPO training hyperparameters. Implemented using TRL `GRPOTrainer` with vLLM colocate mode for on-policy generation.

Parameter	Value
Epochs	2
Learning rate	5×10^{-6}
Scheduler	Cosine with 5% warmup
Gradient accumulation	8
Group size G	8 completions
Temperature (medical)	0.7
Temperature (legal)	0.7–1.0†
Max prompt length	6,144 tokens
Max completion length	4,096 tokens
KL coefficient β	0.04
Clipping ϵ	0.2
Policy updates per generation	2
Reward weights (w_f, w_c, w_g)	(1, 1, 2)
Precision	bfloat16

- *Medical*: Embedding-based matching using `BioLORD-2023` (Remy et al., 2024). For the top-3 predicted diagnoses, computes cosine similarity against all ground-truth diagnoses; a prediction is correct if max similarity exceeds τ=0.80. Returns the fraction correct.
- *Legal*: Exact match on the answer letter. Returns $\{0, 1\}$.

3. $R_{\text{grounding}}$: NLI-based evidence grounding, normalized to $[0, 1]$.
   - *Medical*: Uses `PubMedBERT-MNLI-MedNLI`. Averages the max-grounding score across the top-3 diagnoses using the focus-then-verify architecture (Section 3.2).
   - *Legal*: Uses `nli-deberta-v3-large`. Splits reasoning into sentences, scores each against the gold legal passage with token-aware premise truncation, and averages per-sentence scores.

Training curves (Figure 5) reveal a consistent learning pattern across models. Format compliance converges quickly, with near-perfect structured outputs within the first few training steps. Accuracy reward improves rapidly during the first epoch, while grounding reward increases more gradually throughout training.

Table 9: LoRA adapter configuration (shared across SFT, fDPO, and GRPO).

Parameter	Value
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Target modules	`q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`
Bias	None
Task type	`CAUSAL_LM`

The largest improvements appear on the Llama models, while gains on Gemma models are more modest, suggesting that model families differ in how effectively reinforcement learning reshapes reasoning behavior. Llama models reach higher final reward values under GRPO, indicating that instruction-tuned models may retain additional capacity that reinforcement learning can unlock through targeted reward signals.

Overall, the results reveal a consistent pattern: inference-time techniques (retrieval and self-consistency) can influence output selection but rarely change the model's underlying reasoning behavior. EvidenceRL, in contrast, alters how models generate diagnoses. The taxonomy analysis shows fewer hallucinations, fewer lucky guesses, and substantially higher evidence-based reasoning rates, suggesting that reinforcement learning shifts models from parametric reasoning toward genuine use of the clinical evidence provided in the prompt.

C Extended Results on MIMIC-IV-Ext

This appendix provides additional analyses supporting the results presented in Section 5. We examine the trade-off between diagnostic accuracy and evidence grounding, analyze the distribution of prediction types under the diagnostic taxonomy, inspect grounding distributions across patients, present a representative clinical case study, and audit the behavior of the grounding reward.

C.1 Accuracy–Grounding Trade-Off

To better understand the interaction between diagnostic performance and evidential reliability, we

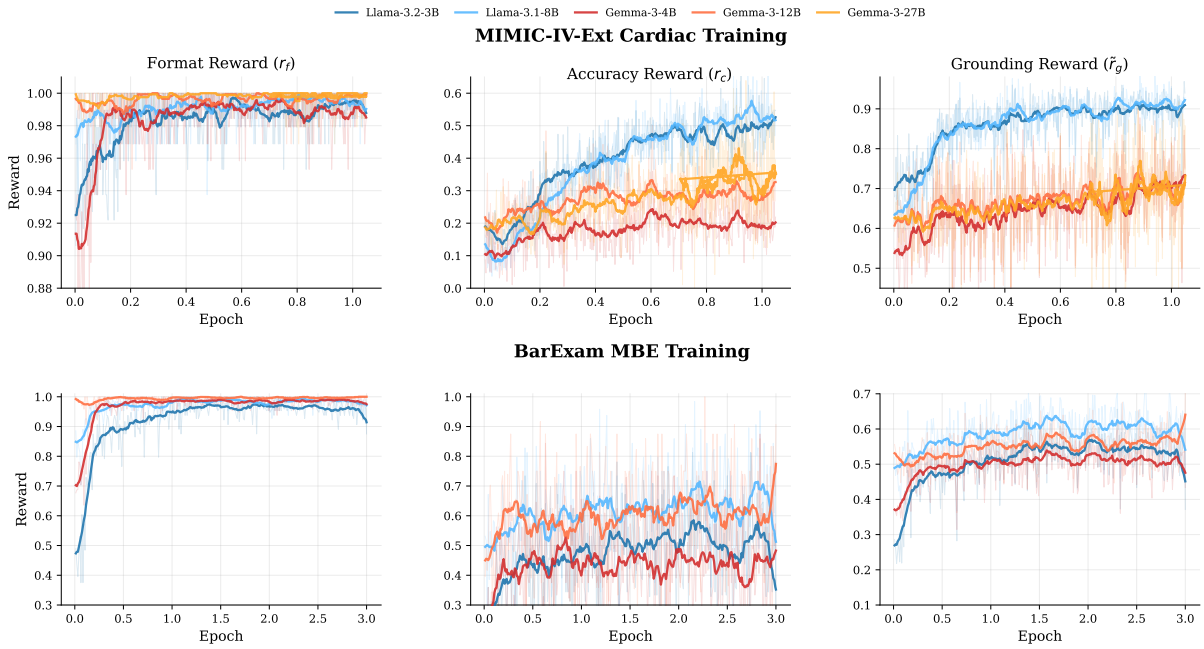


Figure 5: Training reward dynamics across model scales and objectives using MIMIC. We illustrate the training progress for the Llama-3.1-8B, Llama-3.2-3B, and the Gemma-3 series (4B, 12B, and 27B) Performance is evaluated across three primary reward components: (left) Format Reward (r_f) measuring adherence to structural constraints; (center) Accuracy Reward (r_c) assessing the correctness of generated responses; and (right) Grounding Reward ($\tilde{r}_g$) quantifying the extent to which outputs are supported by provided context. Larger model scales generally exhibit higher reward ceilings and more stable convergence across all metrics.

plot model performance across the two axes of diagnostic accuracy (F1@3) and evidence grounding ($G_{\max}$@3). Figure 6 shows the resulting Pareto landscape across all evaluated backbones and training methods.

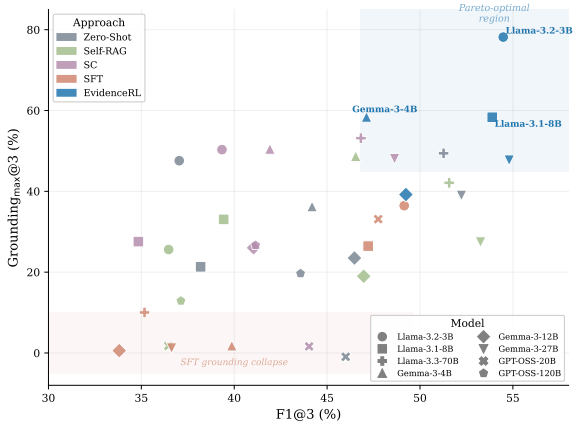


Figure 6: F1@3 vs. $G_{\max}$@3 across all model backbones and approaches. EvidenceRL points (blue) occupy the Pareto-optimal region—high accuracy *and* high grounding. SFT points (orange) collapse to near-zero grounding despite moderate accuracy. Self-RAG (green) clusters with or below zero-shot baselines.

EvidenceRL consistently occupies the Pareto-optimal region, achieving both higher diagnostic accuracy and stronger grounding than baseline approaches. In contrast, SFT models cluster in a region of moderate accuracy but near-zero grounding, indicating that the model learns to imitate plausible answers without grounding them in clinical evidence. Inference-time interventions such as Self-RAG and Self-Consistency remain near the zero-shot baseline, suggesting that retrieval or sampling alone does not reliably improve both axes simultaneously.

These results highlight a central property of EvidenceRL: by directly rewarding grounded correctness during training, the method shifts models toward solutions that jointly optimize accuracy and evidential reliability rather than trading one for the other.

C.2 Taxonomy Analysis

We further analyze model behavior using the diagnostic taxonomy introduced in Section 4.3, which categorizes predictions into Evidence-Based (EB), Weakly Supported (WS), Lucky Guess (LG), Hallucination (H), and Reasoning Failure (RF).

Figure 7 shows the distribution of prediction types for Llama-3.2-3B across all evaluated ap-

proaches. EvidenceRL produces the largest proportion of Evidence-Based predictions, nearly doubling the EB rate relative to the zero-shot baseline. At the same time, hallucinations are substantially reduced.

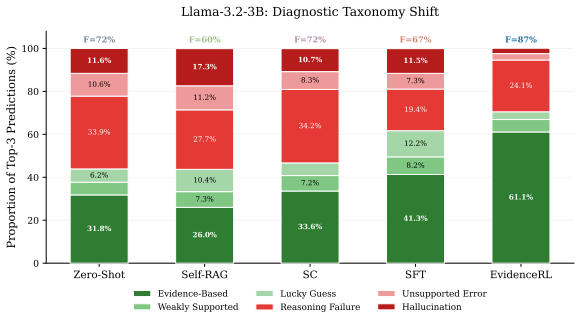


Figure 7: Diagnostic taxonomy distribution for Llama-3.2-3B across all five approaches. Each bar decomposes top-3 predictions into six categories. Faithfulness (F) is annotated above each bar. EvidenceRL achieves the highest Evidence-Based rate (61.1%) and Faithfulness (87%), while Self-RAG has the lowest Faithfulness (60%) due to elevated Lucky Guess and Hallucination rates.

In contrast, SFT improves the overall number of correct predictions but does not improve the evidential basis of those predictions. Much of its performance gain comes from increases in Lucky Guesses, indicating correct diagnoses that are unsupported by the evidence cited in the reasoning.

Inference-time interventions exhibit a different failure pattern. Self-RAG increases both hallucinations and lucky guesses, suggesting that retrieved documents can introduce confounding information when the model is not trained to critically evaluate evidence.

C.3 Per-Patient Grounding Distributions

Aggregate grounding scores can obscure important differences in reliability across individual patients. To examine this, we plot the distribution of per-patient grounding scores ($G_{\max}@3$) for each approach (Figure 8).

EvidenceRL produces a markedly tighter distribution, with the majority of patients exhibiting strong grounding scores. The median grounding score approaches the upper end of the scale, and a large fraction of patients exceed the grounding threshold used in our taxonomy.

Baseline approaches show much broader distributions. Zero-shot reasoning and Self-Consistency exhibit bimodal patterns, where grounding is strong

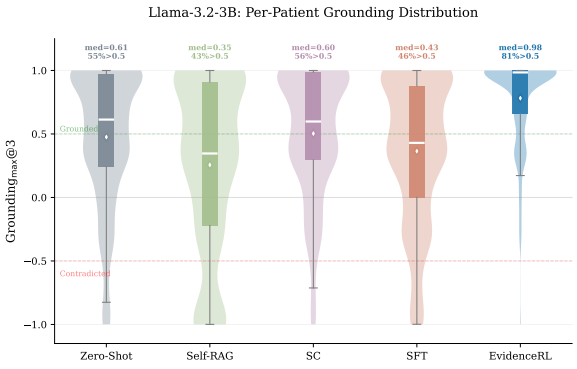


Figure 8: Per-patient grounding distributions (Llama-3.2-3B). Each violin shows the full distribution of $G_{\max}@3$ across patients. Diamonds indicate means; white lines indicate medians. EvidenceRL concentrates mass near 1.0 (median=0.98, 81% of patients above the grounding threshold), while other approaches show broad, bimodal distributions.

for some patients but weak or contradictory for others. Self-RAG displays the most variability, reflecting sensitivity to retrieval quality: when retrieved passages are relevant grounding improves, but when retrieval is noisy grounding deteriorates substantially.

These results suggest that EvidenceRL not only improves average grounding scores but also stabilizes evidence use across patients, an important property for clinical decision support.

C.4 Case Study: Patient #29017807

To illustrate how different approaches reason about the same patient, we present a detailed case study. Table 11 summarizes diagnostic performance; Table 10 shows the actual reasoning text.

Clinical presentation. A male patient with diabetes mellitus and chronic kidney disease (baseline creatinine 2.0) presented with chest pain, dyspnea, orthopnea, and paroxysmal nocturnal dyspnea. Key findings: stress test with anteroapical perfusion defect; echocardiogram showing apical thrombi, LVEF 42–45%, mild mitral regurgitation, mild pulmonary hypertension; troponin of 4.67; cardiac catheterization revealing 90% proximal LAD lesion and 80% distal LCx disease, with successful PCI of the LAD.

Ground truth. (1) Acute myocardial infarction, (2) Heart failure, (3) Chronic ischemic heart disease, (4) Hypertensive chronic kidney disease.

Table 10: Reasoning comparison for the top-ranked diagnosis of patient #29017807 across all five approaches. Clinical values cited from the patient record are underlined.

Approach	Top-1 Dx	$G_{\max}$	Reasoning (excerpted)
EvidenceRL	Myocardial Ischemia	**+0.93**	"The presence of a significant LAD lesion with 90% proximal occlusion, combined with the patient's recent stress test showing antero-apical perfusion defect, strongly suggests intermittent myocardial ischemia. The patient's elevated troponin level (4.67) and reduced ejection fraction (LVEF = 42%) support the diagnosis."
SFT	NSTEMI	−0.35	"The initial presentation included elevated Troponins (4.67 ng/mL), indicative of myocardial necrosis. . . a previously documented stress test showing an apical perfusion defect followed by development of apical thrombi. . . warranting immediate evaluation."
Zero-Shot	CAD	+0.14	"The patient presents with a history of NSTEMI and subsequent stress test showing ant/apical perfusion defects. . . The presence of 90% LAD lesion, 50% CX, and successful stenting of the LAD support this diagnosis."
Self-RAG	CAD	−0.34	"The presence of a 90% proximal lesion in the LAD explains the patient's angina. . . The physical exam shows elevated blood pressure (SBP 153 mmHg). . . Criteria from Evidence 1 and 2 support the presence of coronary artery disease."
SC	CAD	+0.08	(Consensus from N=10 samples; reasoning from the highest-ranked generation.)

Table 11: Case study summary for patient #29017807 (Llama-3.2-3B).

Approach	Correct	G_{avg}@5	Combined
EvidenceRL	**5/5**	**+0.758**	**0.879**
SFT	4/5	−0.869	−0.034
SC	4/5	+0.084	0.442
Zero-Shot	3/5	−0.694	−0.047
Self-RAG	3/5	−0.861	−0.130

D Prompt Templates

Both domains share a common prompting architecture: a domain-expert system role, structured JSON output, and domain-specific *synthesis rules* that guide the model to ground its reasoning in the provided evidence. We present the full prompt templates below.

Despite operating in different domains, both prompt families enforce the same structural principles: (1) a domain-expert persona, (2) structured JSON output for reliable parsing, (3) three synthesis rules that require explicit citation of evidence rather than generic summaries, and (4) critical instructions constraining the model to begin its response with the opening brace. The RAG→no-RAG adaptation follows a parallel pattern in both domains: the evidence-citing rule (*Guideline Alignment* in medical, *Authority Anchoring* in legal) is replaced with a knowledge-based alternative (*Avoid generic summaries* and *Rule Statement*, respectively).

D.0.1 Medical Domain (Cardiac Diagnosis)

Medical RAG Prompt

```
You are an expert cardiology clinical assistant. Based on the patient information and retrieved clinical evidence, provide exactly 5 cardiac diagnoses ranked from most to least likely.

For EACH diagnosis, you MUST provide:
1. The diagnosis name (concise clinical term)
2. A reasoning paragraph following these "Clinical Synthesis" rules:
- Pathophysiological Link: Explain how the specific symptoms (e.g., dyspnea) are directly explained by the clinical findings (e.g., the mitral regurgitation seen on ultrasound).
- Evidence Integration: Use exact values from the Physical Exam (BP, RR) and Imaging (LVEF, PA pressures) as "anchors" for your argument.
- Guideline Alignment: Explicitly state which criteria from the Retrieved Clinical Evidence are met by this specific patient's data.

IMPORTANT: Your response MUST be valid JSON in exactly this format:
{
"diagnoses": [
{"name": "Diagnosis 1 name", "reasoning": "..."},
...
{"name": "Diagnosis 5 name", "reasoning": "..."}
]
}

Patient Information:
{patient_context}
```

```
Retrieved Clinical Evidence:
[Evidence 1]: {passage_1}
[Evidence 2]: {passage_2}
[Evidence 3]: {passage_3}

CRITICAL INSTRUCTIONS FOR YOUR RESPONSE:
- Begin your response IMMEDIATELY with the opening brace {
- Do NOT include any thinking, explanation, preamble, or commentary before the JSON
- Do NOT show your reasoning process outside the JSON - all reasoning goes in the "reasoning" fields
- Output ONLY valid JSON, nothing else
- Start your response with {
```

Medical No-RAG Prompt

```
You are an expert cardiology clinical assistant. Based on the patient information below, provide exactly 5 cardiac diagnoses ranked from most to least likely.

For EACH diagnosis, you MUST provide:
1. The diagnosis name (concise clinical term)
2. A reasoning paragraph following these "Clinical Synthesis" rules:
- Pathophysiological Link: Explain how the specific symptoms (e.g., dyspnea) are directly explained by the clinical findings (e.g., the mitral regurgitation seen on ultrasound).
- Evidence Integration: Use exact values from the Physical Exam (BP, RR) and Imaging (LVEF, PA pressures) as "anchors" for your argument.
- Avoid generic summaries: Do not just list facts; explain the "why" behind the diagnosis using the patient's unique data.

IMPORTANT: Your response MUST be valid JSON in exactly this format:
{
"diagnoses": [
{"name": "Diagnosis 1 name", "reasoning": "..."},
...
{"name": "Diagnosis 5 name", "reasoning": "..."}
]
}

Patient Information:
{patient_context}

CRITICAL INSTRUCTIONS FOR YOUR RESPONSE:
- Begin your response IMMEDIATELY with the opening brace {
- Do NOT include any thinking, explanation, preamble, or commentary before the JSON
- Do NOT show your reasoning process outside the JSON - all reasoning goes in the "reasoning" fields
- Output ONLY valid JSON, nothing else
- Start your response with {
```

Medical LLM-as-Judge Prompt

```
You are evaluating the diagnosis prediction of a clinical model.

CANDIDATE ANSWER: "{predicted_diagnosis}"

ACCEPTED GROUND TRUTHS:
- {ground_truth_1}
- {ground_truth_2}
- ...

TASK: Does the CANDIDATE ANSWER semantically match ANY of the ACCEPTED GROUND TRUTHS?
Respond 'TRUE' if the candidate refers to the same underlying clinical concept as any item in the list (allowing for synonyms, abbreviations, or minor wording differences).
Respond 'FALSE' if it represents a different clinical concept, severity, or anatomical location.

Verdict:
```

D.0.2 Legal Domain (BarExam QA)

Legal RAG Prompt

```
You are an expert legal analyst taking the Multistate Bar Examination (MBE). Based on the question and retrieved legal authorities, select the best answer.

For your answer, you MUST provide:
1. The answer letter (A, B, C, or D)
2. A reasoning paragraph following these "Legal Synthesis" rules:
- Factual Application: Identify the key facts from the question and explain how they map to the legal rule or doctrine at issue.
- Authority Anchoring: Cite specific language, holdings, or principles from the Retrieved Legal Authorities as direct support for your conclusion. Do NOT reason beyond what the authorities state.
- Rule-to-Fact Bridge: Explicitly connect the legal rule from the authority to the specific facts given, showing why the rule applies (or does not apply) to each answer choice.
- Avoid generic summaries: Do not just restate the rule; explain the "why" using the specific facts of the question.

IMPORTANT: Your response MUST be valid JSON in exactly this format:
{
"answer": "X",
"reasoning": "Detailed legal reasoning..."
}

Question:
{question_text_with_choices}

Retrieved Legal Authorities:
[Legal Authority 1]: {passage_1}
[Legal Authority 2]: {passage_2}
[Legal Authority 3]: {passage_3}

CRITICAL INSTRUCTIONS FOR YOUR RESPONSE:
```

```
- Begin your response IMMEDIATELY with the opening brace {
- The "answer" field MUST be exactly one letter: A, B, C, or D
- All reasoning goes in the "reasoning" field
- Output ONLY valid JSON, nothing else
- Start your response with {
```

Legal No-RAG Prompt

```
You are an expert legal analyst taking the Multistate Bar Examination (MBE). Based on your legal knowledge, select the best answer.

For your answer, you MUST provide:
1. The answer letter (A, B, C, or D)
2. A reasoning paragraph following these "Legal Synthesis" rules:
- Factual Application: Identify the key facts from the question and explain how they map to the legal rule or doctrine at issue.
- Rule Statement: State the applicable legal rule or doctrine from your knowledge.
- Rule-to-Fact Bridge: Explicitly connect the legal rule to the specific facts given, showing why the rule applies (or does not apply) to each answer choice.
- Avoid generic summaries: Do not just restate the rule; explain the "why" using the specific facts of the question.

IMPORTANT: Your response MUST be valid JSON in exactly this format:
{
"answer": "X",
"reasoning": "Detailed legal reasoning..."
}

Question:
{question_text_with_choices}

CRITICAL INSTRUCTIONS FOR YOUR RESPONSE:
- Begin your response IMMEDIATELY with the opening brace {
- The "answer" field MUST be exactly one letter: A, B, C, or D
- All reasoning goes in the "reasoning" field
- Output ONLY valid JSON, nothing else
- Start your response with {
```

Legal Accuracy Evaluation. Unlike the medical domain, BarExam accuracy is evaluated via **exact match** on the answer letter (A/B/C/D), making an LLM-as-judge unnecessary.