

Does Pedagogy Hurt Truth? Evaluating Educational Rewriting in Medicine

Stanford CS224N Custom Project

Cally Lin

Biomedical Data Science
Stanford University
callylin@stanford.edu

Sasha Simic

Computer Science
Stanford University
ssimic@stanford.edu

Abstract

Large language models are increasingly used to generate medical explanations for patient education and clinical support. Adapting explanations to different audiences can improve clarity but may also introduce distortions or unsupported claims, raising a central question: *does pedagogy hurt truth?* We study evidence-bounded pedagogical rewriting in medicine, where a model rewrites an explanation for either a patient-facing (**Basic**) or clinician-facing (**Technical**) audience while remaining grounded in topic-specific evidence passages. Using 1,015 English MedlinePlus topics and Llama 3.1 8B Instruct, we evaluate prompt-based rewriting, retrieval-augmented generation (RAG), and LoRA-based adaptation using metrics for pedagogical adaptation and evidence-grounded faithfulness. We find that prompt-only audience conditioning produces clear stylistic differentiation but introduces unsupported medical content, particularly in **Technical** rewrites. Unsupported claim rates increase from 3.25% in anchors to 5.59% in **Basic** rewrites and 11.41% in **Technical** rewrites. Retrieval grounding and parameter-efficient fine-tuning reduce unsupported claims, with the largest improvements in the highest-risk technical setting. These results highlight the need to jointly optimize pedagogical clarity and evidence-grounded faithfulness in medical explanation generation.

1 Key Information to include

- Mentor: Caroline Choi
- External Collaborators: No
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
Sasha Simic	5
Cally Lin	4
Total pooled	9
Team size	2
Late days applied for the report	2

2 Introduction

Large language models (LLMs) are increasingly used to generate explanations for diverse audiences in applications such as educational tools and decision support systems. In healthcare, these systems support both patients and clinicians through health information platforms and clinical decision support tools [1]. Because audiences vary in technical knowledge, medical explanations must often be

adapted to match the reader’s expertise. Clinicians may expect precise terminology and step-by-step instructions, while patients typically require simplified language and additional context.

Ensuring that these adaptations remain factually accurate is critical. Even small wording changes can alter clinical interpretation by misrepresenting risks or introducing unsupported claims. Simplifying technical content often requires replacing specialized terminology, restructuring sentences, or adding context, which can introduce factual drift where rewritten explanations include claims not supported by the source evidence [2, 3].

In this work, we study evidence-bounded pedagogical rewriting for medical explanations, addressing a gap in current research on how audience adaptation affects factual faithfulness. Given an anchor explanation, supporting evidence passages, and an audience label (Basic or Technical), a language model rewrites the explanation while remaining grounded in the evidence. We evaluate two objectives: pedagogical adaptation (readability and lexical complexity) and evidence-bounded faithfulness (claim support with respect to the evidence). Our main contribution is an evaluation framework for measuring the tradeoff between pedagogical adaptation and evidence-grounded faithfulness in medical rewriting.

3 Related Work

Pedagogical adaptation and simplification. Recent work shows that large language models can adapt explanations to different audiences, but most evaluations emphasize style and clarity rather than factual grounding. *ELI-Why* [4] frames pedagogical utility as a first-class objective and demonstrates that helpful explanations are not fully captured by generic automatic metrics. Similarly, *Know Your Audience* [5] shows that models respond to audience hints such as vocabulary and sentence complexity, but does not examine whether those adaptations remain faithful to supporting evidence. Research in text simplification provides related tools for controlled adaptation of readability and lexical complexity, including datasets such as Newsela-style rewriting and controllable methods like ACCESS and CUT [6]. Metrics such as SARI [7] evaluate edit behavior (add/delete/keep operations) and remain widely used for simplification tasks. However, these approaches primarily measure surface-level transformations and do not directly detect unsupported factual claims. Our work builds on these lines of research while explicitly evaluating evidence-grounded faithfulness in pedagogical rewriting.

Evidence-grounded faithfulness. A growing body of work shows that fluent LLM outputs may still be unsupported when checked against source evidence. Benchmarks, such as FinanceBench [8], demonstrate that models frequently produce plausible but ungrounded explanations when reasoning over documents. This literature motivates claim-level, evidence-bounded auditing as a more reliable factuality protocol than open-ended plausibility judgments. However, most prior work evaluates question answering or summarization rather than audience-conditioned rewriting. Our work evaluates rewritten explanations using evidence, since adapting explanations for different audiences can sometimes cause the meaning to drift from the original.

Retrieval and lightweight adaptation. Retrieval-augmented generation (RAG) methods, including Atlas and Self-RAG [9], improve factual grounding by conditioning generation on retrieved evidence. At the same time, parameter-efficient fine-tuning approaches, such as LoRA [10], enable efficient specialization of large language models without full model retraining. These methods are typically evaluated for factual question answering or general generation tasks rather than for audience-aware rewriting. Our experiments connect these approaches by comparing prompt-only, RAG, and LoRA configurations under the same audience-conditioned rewriting task and evaluating them using a shared evidence-bounded faithfulness audit.

Taken together, prior work highlights separate advances in pedagogical adaptation, controllable rewriting, and factual grounding. Our work studies the intersection of these areas by evaluating audience-conditioned medical rewriting under both pedagogical and evidence-grounded faithfulness metrics.

4 Approach

Task Let x denote an anchor medical explanation, $E = \{e_1, \dots, e_k\}$ a set of supporting evidence passages, and $a \in \{\text{Basic}, \text{Technical}\}$ the target audience; the task is to generate a rewritten explanation $\hat{x}$ adapted for audience a while remaining grounded in E , formalized as $\hat{x} = f_\theta(x, E, a)$ where f_θ is a language model parameterized by θ . The rewrite is optimized for two objectives: *pedagogical adaptation* (readability and lexical-complexity control) and *evidence-bounded faithfulness* (claims in $\hat{x}$ are supported by passages in E).

Prompt-Based Baselines We establish a baseline using prompt-based rewriting with instruction-tuned language models. Given an anchor explanation x , the model generates an audience-conditioned rewrite while preserving the medical facts in the original text. Audience conditioning is implemented through two instruction variants. The `Basic` condition encourages plain language and simplified explanations, while the `Technical` condition encourages precise clinical terminology and concise phrasing. The full prompt template is provided in Appendix A.1.

Retrieval-Augmented Generation As a technical grounding tool for our baseline, we use sparse lexical retrieval (TF-IDF cosine over unigram/bigram features) to select topic-relevant evidence passages for each anchor. Generation is then performed in an evidence-conditioned prompt that concatenates the anchor (source) text, retrieved top- k passages, and audience instruction. This adds a memory step at inference time, allowing the model to use retrieved text for factual information instead of relying only on what it learned during training.

LoRA Fine-Tuning As a parameter-efficient adaptation tool, we apply PEFT LoRA [10]: base weights are frozen and only low-rank adapter matrices are trained in selected transformer modules (attention-only or attention+MLP, depending on variant). The optimization objective is supervised next-token prediction on audience-conditioned rewrite examples, where targets are GPT-5.1 proxy rewrites (in place of human rewrites) filtered by claim-level truth audits. This pipeline uses LoRA as a controlled adaptation mechanism for style/complexity transfer while constraining training signals toward evidence-consistent rewrites.

We then compare both variations (RAG and LoRa) against the prompt-only baseline on the same splits to analyze how the pedagogy–truthfulness tradeoff changes under retrieval-based grounding and LoRA adaptation.

5 Experiments

5.1 Data

We use the MedlinePlus [11] Topics dataset from the U.S. National Library of Medicine (NLM), snapshot dated March 4, 2026. After filtering to English topics and removing entries with empty summaries, the dataset contains 1,015 topics (from 1,017 raw topics). We split topics deterministically (seed=42) into 812 train, 101 validation, and 102 test topics, with 2,989 / 369 / 360 evidence passages in the three splits. For evidence construction, each topic is divided into passages of about 120 words; if a passage is shorter than 20 words, it is combined with a neighboring passage, and the combined passage can be up to 140 words. This yields 3,718 total evidence passages.

Intended data usage. We use topic-level splits to enforce strict separation across training and evaluation workflows. The `train` split (812 topics) is used to build supervision for LoRA fine-tuning, including filtered rewrite-based training rows. The `val` and `test` splits (101 and 102 topics) are held out for model selection, calibration, and final evaluation, and are never used for LoRA parameter updates. Splits are performed at the topic level to prevent information leakage.

5.2 Evaluation Metrics

We evaluate both pedagogical adaptation and evidence-bounded faithfulness on the full validation set (101 topics, 202 rewrites) and test set (102 topics, 204 rewrites).

Pedagogical adaptation. We measure audience differentiation using readability and lexical complexity proxies comparing `Basic` and `Technical` rewrites. Metrics include total word count, average

words per sentence, Flesch–Kincaid grade level, Flesch Reading Ease, average word length, and lexical diversity measured by type–token ratio. For each metric m , we report audience-level means and the per-topic difference: $\Delta_m = m_{\text{Technical}} - m_{\text{Basic}}$. These measures quantify whether models meaningfully adjust register and complexity for different audiences.

We intentionally exclude SARI from primary pedagogical conclusions. In manual inspection, SARI was unreliable for this setting, particularly when evaluating Technical rewrites.

Evidence-bounded faithfulness (LLM-as-judge baseline) [8, 12]. We evaluate factual consistency using an evidence-bounded claim audit with an LLM-as-judge framework rather than TF-IDF thresholding. Each rewrite is segmented into sentence-level claim candidates (split by sentence punctuation/newlines). Short fragments are excluded in two stages: candidates with fewer than 5 words are dropped, and remaining candidates with fewer than 3 content words (after stopword filtering) are also dropped. Each remaining claim is then evaluated against retrieved topic evidence passages.

Our primary faithfulness metrics are `UnsupportedClaimRate` and `RewriteFlagRate`. For diagnostic analysis, we also report the distribution of judge labels, including `partial_support`, `missing_evidence`, `off_topic`, `contradicted`, `unclear_claim`, and parse-error rates.

$$\begin{aligned}\text{UnsupportedClaimRate}(\mathcal{R}) &= \frac{\text{Total unsupported claims in } \mathcal{R}}{\text{Total claims in } \mathcal{R}}, \\ \text{RewriteFlagRate}(\mathcal{R}) &= \frac{\text{Rewrites in } \mathcal{R} \text{ with } \geq 1 \text{ unsupported claim}}{\text{Total rewrites in } \mathcal{R}}.\end{aligned}$$

We also tested overgeneralization diagnostics, but results were unstable under our baseline. Since our primary target is directional faithfulness (supported vs. unsupported) and human review suggested most overgeneralization-like cases were already captured as `partial_support`, we treat overgeneralization findings as exploratory.

5.3 Experimental Setup

We evaluate on a held-out validation set of 101 topics and a test set of 102 topics drawn from the English MedlinePlus corpus. All generations use Llama 3.1 8B Instruct, served via Ollama on a GCP virtual machine with an NVIDIA T4 GPU. Decoding parameters are fixed across all runs: temperature 0.2, top- p 0.95, maximum 512 output tokens, seed 42, a 120-second timeout, and up to three retries on failure.

We evaluate three rewriting configurations:

- **Prompt-only baseline.** For each topic, the model generates two rewrites: one targeting a lay audience (`Basic`) and one targeting a clinical audience (`Technical`). This produces 202 outputs on the validation set and 204 outputs on the test set.
- **Retrieval-augmented generation (RAG).** Retrieved evidence passages are prepended to the prompt prior to generation. We ablate over two retrievers (BM25 and TF-IDF) and three passage budgets ($k \in \{1, 3, 5\}$), yielding six RAG configurations. To isolate the effect of audience conditioning on faithfulness, we additionally include a *general-audience* ablation that removes the audience split and generates a single neutral rewrite per topic (101 validation outputs and 102 test outputs).
- **LoRA-adapted model.** We augment Llama 3.1 8B Instruct with a parameter-efficient LoRA adapter trained on truth-filtered GPT-5.1 proxy rewrites from the training split. Within this setting, we evaluate two adapter checkpoints (v1/v2) reflecting progressively improved training configurations to assess the effect of adapter quality on downstream rewrite performance.

Experiments use deterministic versioning, topic ordering, fixed decoding parameters, and per-run metadata logging for reproducibility.

Faithfulness Calibration. To select a reliable automated judge, we evaluated several variants, including GPT-4.1-mini with revised prompts, NLI-style labels derived from judge metadata, standalone NLI classifiers, and GPT-5.1 with the original judge prompt. While NLI-style variants increased

sensitivity to partial-support cases, they showed weaker calibration on anchor claims and less stable label behavior. Based on anchor calibration and human spot checks, GPT-5.1 provided the most consistent supported-versus-unsupported decisions and is therefore used as the primary faithfulness evaluator. Table 9 (Appendix D.1.1) summarizes the human–automatic agreement results.

Validation of automated claim evaluation. To validate the automated NLI/Judge based claim classifiers, one annotator reviewed 10 randomly sampled test topics per audience (20 rewrites total) from the prompt-only baseline, segmented rewrites into sentence-level claims, and labeled each as *Entails*, *Partial*, *Contradicts*, or *Neutral*; for binary agreement, *Entails/Partial* were mapped to faithful and *Contradicts/Neutral* to non-faithful. Human and automated labels were aligned by topic ID, audience, and claim index, and 23 human-labeled faithful claims not extracted by the pipeline (segmentation mismatch) were excluded from agreement computation. Table 10 summarizes data we based the decision on.

5.4 Results

5.4.1 Baseline

Faithfulness. We evaluate claim-level faithfulness using Unsupported Claim Rate (UCR) and rewrite-level faithfulness using Rewrite Flag Rate (RFR), defined as the percentage of rewrites containing at least one unsupported claim. We report these in Table 1. Secondary label distributions are reported in Table 11 (Appendix E.1).

Slice	Val UCR	Δ UCR	Test UCR	Δ UCR	Val RFR	Δ RFR	Test RFR	Δ RFR
Anchor	2.50	0.00	3.25	0.00	32.67	0.00	41.18	0.00
Basic	5.91	+3.41	5.59	+2.34	43.56	+10.89	46.53	+5.35
Technical	10.60	+8.10	11.41	+8.16	59.41	+26.74	60.78	+19.60

Table 1: Faithfulness results. Deltas are percentage-point differences vs anchor.

Pedagogical adaptation. Table 2 summarizes baseline pedagogy metrics (means and paired Technical – Basic deltas) on validation and test.

Split	Row	Words	Sentences	Words/Sent	FK Grade	FRE	Avg Word Len	TTR
Val	Basic	217.11	11.73	18.92	10.39	54.28	4.89	0.64
Val	Technical	187.29	9.27	21.06	16.77	12.35	6.02	0.71
Val	Δ (T-B)	-29.82	-2.47	+2.14	+6.38	-41.94	+1.13	+0.07
Test	Basic	210.46	11.54	19.95	10.58	54.76	4.87	0.65
Test	Technical	181.85	9.27	20.76	16.23	15.70	5.93	0.72
Test	Δ (T-B)	-29.65	-2.32	+0.81	+5.68	-39.30	+1.07	+0.07

Table 2: Baseline pedagogy metrics. FK = Flesch–Kincaid grade level; FRE = Flesch Reading Ease; TTR = type-token ratio.

Technical rewrites drive most of the faithfulness gap. Compared with Basic, Technical has much higher unsupported rates in both splits (test UCR: 11.41% vs 5.59%; test RFR: 60.78% vs 46.53%), showing that higher-complexity adaptation is the main source of factual risk.

The dominant failure mode is partial support, not contradiction. The secondary breakdown in Table 11 shows *partial_support* as the largest unsupported bucket and substantially higher for Technical (val: 7.65% vs 3.77%; test: 8.32% vs 4.10%), while *missing_evidence*, *contradicted*, and *off_topic* remain comparatively small.

Pedagogy targets are achieved, but with a complexity–faithfulness tradeoff. Technical rewrites are consistently more complex (higher FK, lower FRE, longer words, higher TTR, more words per sentence) even though they are shorter in length. This implies a technical rewrite relied primarily on increased density, likely sacrificing truthfulness.

5.4.2 LoRA Ablations

Table 3 summarizes LoRA V1 and V2 adapter configurations, including optimization settings and final training and validation losses.

LoRA	Epochs	LR	r	α	Dropout	Target modules	Max len	Train loss	Val loss
V1	3	2e-4	16	32	0.05	attn-only (q,k,v,o)	1024	0.6340	0.7087
V2	8	8e-5	32	64	0.10	attn + MLP (q,k,v,o,gate,up,down)	1536	0.1698	0.5336

Table 3: LoRA training config and results.

Faithfulness. For the LoRA experiments, we report only test-set results in the main text because validation data was used for checkpoint selection.

Audience	Metric	Baseline	LoRA V1	Δ V1 (pp)	LoRA V2	Δ V2 (pp)
Basic	UCR	5.59	6.31	+0.72	7.21	+1.62
Basic	RFR	46.53	40.20	-6.33	53.92	+7.39
Technical	UCR	11.41	7.22	-4.18	9.81	-1.60
Technical	RFR	60.78	48.04	-12.74	62.75	+1.97
Overall	UCR	8.21	6.72	-1.49	8.37	+0.16
Overall	RFR	53.69	44.12	-9.57	58.33	+4.64

Table 4: Baseline vs. LoRA variants. UCR = Unsupported Claim Rate; RFR = Rewrite Flag Rate.

Model	Audience	partial_support	missing_evidence	off_topic	contradicted	unclear_claim	parse_error
LoRA V1	Overall	3.72% (-2.28)	2.18% (+1.32)	0.09% (+0.04)	0.00% (-0.24)	0.04% (-0.10)	0.68% (-0.23)
LoRA V1	Basic	3.43% (-0.68)	2.26% (+1.56)	0.08% (+0.08)	0.00% (-0.09)	0.08% (-0.18)	0.47% (+0.03)
LoRA V1	Technical	4.09% (-4.23)	2.09% (+1.03)	0.10% (-0.01)	0.00% (-0.43)	0.00% (+0.00)	0.95% (-0.54)
LoRA V2	Overall	5.43% (-0.57)	1.65% (+0.79)	0.08% (+0.03)	0.12% (-0.12)	0.24% (+0.10)	0.85% (-0.07)
LoRA V2	Basic	4.59% (+0.48)	1.75% (+1.05)	0.07% (+0.07)	0.15% (+0.06)	0.15% (-0.12)	0.51% (+0.07)
LoRA V2	Technical	6.48% (-1.83)	1.53% (+0.46)	0.09% (-0.02)	0.09% (-0.34)	0.36% (+0.36)	1.26% (-0.23)

Table 5: Secondary LLM-judge diagnostic labels and parse-error rates for LoRA runs on test claims. Parenthetical values denote percentage-point differences relative to the baseline within the same audience.

Pedagogical adaptation Table 6 summarizes LoRAs impact on pedagogy. For pedagogy analysis, we focus on LoRA V1 only as LoRA V2 showed very weak held-out faithfulness.

Metric	Basic			Technical		
	Baseline	V1	Δ	Baseline	V1	Δ
Word count	210.46	185.71	-24.75	181.85	180.62	-1.24
Avg words/sentence	19.95	15.02	-4.92	20.76	18.54	-2.21
FK grade	10.58	8.24	-2.34	16.23	14.25	-1.98
Flesch ease	54.76	62.75	+7.98	15.70	25.94	+10.25
Avg word length	4.87	4.71	-0.15	5.93	5.65	-0.28
Type-token ratio	0.65	0.61	-0.04	0.72	0.67	-0.04

Table 6: Test-set pedagogy metrics for prompt-only baseline vs. LoRA V1. Δ = V1 - baseline.

LoRA V1 is the only variant that improves held-out faithfulness overall: compared with baseline, V1 lowers test UCR (8.21% $\rightarrow$ 6.72%, -1.49 pp) and test RFR (53.69% $\rightarrow$ 44.12%, -9.57 pp), with the largest gains in Technical rewrites (UCR -4.18 pp, RFR -12.74 pp), whereas V2 does not

generalize as well (overall UCR +0.16 pp, RFR +4.64 pp) despite much lower training/validation loss (train 0.1698 vs 0.6340; val 0.5336 vs 0.7087), suggesting overfitting and lack of generalization.

Pedagogy (plain-language takeaway): LoRA V1 makes rewrites easier to read for both audiences, with the biggest effect in Basic. In practice, sentences get shorter and readability improves, so outputs are easier to follow at a glance. The same trend appears in Technical, but less strongly, which suggests V1 improves clarity without fully minimizing the audience distinction.

Truthfulness: LoRA V1 improves directional faithfulness overall, with non-uniform error shifts. V1 reduces test UCR (8.21%→6.72%) and RFR (53.69%→44.12%), with the largest gains in Technical rewrites (UCR −4.18 pp, RFR −12.74 pp); however, secondary labels show composition changes (notably lower `partial_support` but higher `missing_evidence` vs. baseline), indicating a shift in unsupported-claim composition rather than a uniform reduction across error types.

5.4.3 Retrieval-Augmented Generation

Config	k	Audience	Val Unsupported (%)	Val Flag (%)	Test Unsupported (%)	Test Flag (%)
Baseline	–	Basic	5.91	43.56	5.59	46.53
		Technical	10.60	59.41	11.41	60.78
		General	10.90	48.50	5.50	42.20
BM25	1	Basic	7.70 (+1.79)	55.40 (+11.84)	7.20 (+1.61)	52.90 (+6.37)
		Technical	12.70 (+2.10)	60.40 (+0.99)	10.90 (-0.51)	54.90 (-5.88)
		General	5.20 (-5.70)	42.60 (-5.90)	4.50 (-1.00)	37.30 (-4.90)
BM25	3	Basic	7.20 (+1.29)	49.50 (+5.94)	6.10 (+0.51)	51.00 (+4.47)
		Technical	10.70 (+0.10)	55.40 (-4.01)	9.90 (-1.51)	51.00 (-9.78)
		General	5.10 (-5.80)	39.60 (-8.90)	3.80 (-1.70)	33.30 (-8.90)
BM25	5	Basic	7.20 (+1.29)	50.50 (+6.94)	6.00 (+0.41)	46.10 (-0.43)
		Technical	10.80 (+0.20)	53.50 (-5.91)	8.40 (-3.01)	47.10 (-13.68)
		General	5.00 (-5.90)	41.60 (-6.90)	3.50 (-2.00)	32.40 (-9.80)
TF-IDF	1	Basic	7.50 (+1.59)	51.50 (+7.94)	6.80 (+1.21)	51.00 (+4.47)
		Technical	12.90 (+2.30)	61.40 (+1.99)	10.40 (-1.01)	55.90 (-4.88)
		General	5.00 (-5.90)	40.60 (-7.90)	4.90 (-0.60)	38.20 (-4.00)
TF-IDF	3	Basic	6.80 (+0.89)	49.50 (+5.94)	5.80 (+0.21)	49.00 (+2.47)
		Technical	10.90 (+0.30)	48.50 (-10.91)	9.90 (-1.51)	52.00 (-8.78)
		General	4.90 (-6.00)	37.60 (-10.90)	4.60 (-0.90)	37.30 (-4.90)
TF-IDF	5	Basic	7.70 (+1.79)	54.50 (+10.94)	5.90 (+0.31)	46.10 (-0.43)
		Technical	10.00 (-0.60)	47.50 (-11.91)	7.80 (-3.61)	43.10 (-17.68)
		General	5.00 (-5.90)	41.60 (-6.90)	4.30 (-1.20)	35.30 (-6.90)

Table 7: Retriever ablation for RAG across retriever type (BM25 vs. TF-IDF) and passage budget k . Values show absolute rates with percentage-point change relative to the prompt-only baseline in parentheses.

Table 7 reports the retriever ablation across sparse retrieval methods (BM25 and TF-IDF) and passage budgets $k \in \{1, 3, 5\}$. Retrieval improves faithfulness in several configurations, with the strongest gains appearing in Technical rewrites and in the general-audience setting.

The best audience-conditioned configuration is TF-IDF with $k = 5$, reducing the Technical test unsupported claim rate to 7.8% and the flag rate to 43.1% (−3.61 and −17.68 pp vs baseline). Increasing k generally benefits Technical rewrites, suggesting that additional evidence helps maintain grounding when generating more specialized explanations.

6 Analysis

Beyond aggregate metrics, we perform qualitative and slice-based analysis to understand why the model succeeds or fails under different conditions.

Behavioral Comparison Across Methods Across methods, we observe a consistent tradeoff between audience adaptation and faithfulness: prompt-only gives the clear audience separation but is

most vulnerable to unsupported technical elaboration, retrieval improves grounding by anchoring generation to explicit evidence (especially in technical outputs), and LoRA V1 gives the strongest aggregate unsupported-claim reduction by making rewriting behavior more conservative. The broader insight is that retrieval and LoRA improve faithfulness through different levers: retrieval mostly lowers unsupportedness in higher-risk technical rewrites by adding explicit evidence at inference time, while LoRA mainly reshapes generation behavior (with readability gains and shifted error composition), so the preferred method depends on whether the goal is stronger audience differentiation or safer evidence-bounded outputs.

Audience-Specific Behavior Across settings, Technical rewrites are consistently more complex and less faithful than Basic rewrites. The likely driver is that technical prompts push denser terminology and greater specificity, which improves audience differentiation but increases risk of evidence drift. Notably, technical outputs are usually shorter, not longer, indicating adaptation through compressed specialization (higher density/complexity, lower readability) rather than length expansion.

LoRA behavior. LoRA V1 appears to make rewrites easier to read in both Basic and Technical modes while still preserving a meaningful pedagogy gap between audiences. A plausible explanation for its faithfulness gains is that it reduces over-dense, over-elaborated technical phrasing, which helps technical rewrites stay closer to evidence instead of adding unsupported detail. At the same time, the secondary buckets show a redistribution of errors (lower `partial_support`, higher `missing_evidence`), indicating a sharper supported/unsupported boundary rather than a uniform improvement across all failure types. This pattern is consistent with increased polarization and may reflect fewer borderline false positives/false negatives, though that is only an assumption and not based in data.

Retrieval Effects Across Audiences Retrieval provides its clearest benefits in the Technical condition, where the prompt-only baseline exhibits the highest unsupported claim rates. As shown in Table 7, grounding generation with retrieved evidence substantially reduces unsupported technical claims. The ablation results (Table 7) also show that increasing the retrieval budget generally improves performance in the more difficult settings, particularly for Technical and general-audience rewrites. This pattern suggests that retrieval mainly helps by constraining unsupported specificity rather than uniformly improving all forms of rewriting. In contrast, the Basic condition benefits less from retrieval. This is likely because the prompt-only baseline already produces relatively stable rewrites in this configuration.

Judge Calibration A key observation is that *judge anchoring* is critical for reliable faithfulness analysis. When calibrated on anchor claims and checked against lightweight human labels, GPT-5.1 produced the most stable supported-vs-unsupported decisions (over 90% binary agreement in both Basic and Technical) and fewer parse/schema failures than GPT-4.1-mini, especially on harder claim forms (long composite or question-like claims). This stability is important because our downstream comparisons depend on consistent directional judgments. One limitation remains: overgeneralization was not cleanly separable from `partial_support`; since these cases were still counted as unsupported in the binary decision, we report UCR/RFR as primary outcomes and treat finer-grained buckets as diagnostic.

Error Composition Aggregate unsupported rate does not fully capture how systems fail. As shown in Table 11, `partial_support` is the dominant unsupported label, especially for Technical rewrites, while `contradicted` and `off_topic` remain rare. This indicates that most failures are near-supported claims or incompletely grounded claims rather than outright false statements. It also suggests boundary spillover between supported and unsupported decisions (particularly around `partial_support` and `missing_evidence`), so further improvements in better sub-classification of partially supported claims may be important.

Scope broadening. Another failure pattern involves expanding the scope of a claim. Instead of restating the evidence-supported statement, the model generalizes it to a broader population or condition category. Although these claims are often medically reasonable, they exceed the information contained in the provided evidence and therefore count as unsupported under the evidence-bounded

evaluation protocol. This overgeneralization boundary is not always clear and may need some special consideration and human training.

7 Conclusion

We study evidence-bounded pedagogical rewriting in medicine: rewriting medical explanations for different audiences while preserving factual grounding in source evidence. Our experiments show that audience-conditioned rewriting introduces a measurable tradeoff between pedagogical adaptation and evidence-bounded faithfulness. Across settings, `Technical` rewrites consistently produce higher unsupported-claim rates than `Basic` rewrites, indicating that increased explanatory detail and clinical specificity increase the risk of semantic drift.

We also find that evaluation design is as critical as generation design. After comparing multiple judge configurations and diagnostic strategies, GPT-5.1 provided the most stable calibration for evidence-bounded evaluation when validated against anchor references and human spot checks.

Our conclusion is that parameter-efficient adaptation can improve faithfulness in high-risk settings. In particular, LoRA V1 reduced both claim-level and rewrite-level unsupported rates relative to the prompt-only baseline, with the largest gains in the `Technical` condition. However, improvements were configuration-sensitive: LoRA V2 achieved lower training and validation loss but worse test faithfulness, indicating that optimization loss did not reliably predict evidence-grounded generalization. This is consistent with recent findings that LoRA placement (which modules are adapted), batch size, and learning rate can materially change downstream behavior and must be tuned carefully.

The primary limitations of this work are (1) reliance on automated judging rather than large-scale human annotation, (2) sensitivity of results to LoRA configuration and proxy-training data quality, and (3) the limited reliability of overgeneralization as a standalone diagnostic metric. Because many overgeneralization cases collapse into `partial_support`, we focus reporting on directional faithfulness (supported vs. unsupported), UCR/RFR, and stable judge diagnostic categories. However, we also acknowledge that many claims lie close to the supported/unsupported boundary.

Overall, these results highlight the importance of jointly evaluating pedagogical quality and evidence-grounded faithfulness when deploying LLMs for medical communication. Future work should explore tighter claim-to-evidence supervision, larger human-annotated calibration sets, and retrieval-aware objectives that explicitly penalize unsupported claim additions. It should also test prompting strategies that discourage unnecessary complexity increases when they are not evidence-supported, so audience adaptation does not come at the expense of grounding.

Team Contributions

Sasha Simic set up the project infrastructure, including the GCP environment, data pipelines, and baseline scripts. He also implemented the LoRA experiments and contributed to generating baseline data.

Cally Lin contributed to generating baseline data, implemented the RAG experiments, and conducted human evaluations. Both authors collaborated on experiment design, analysis of results, and writing the paper.

Acknowledgment: In our codebase, we used AI coding tools to assist with README preparation, versioning workflows, GCP pipeline development, and improvements to observability (logging).

References

- [1] Li Zhui, Nina Yhap, Liu Liping, Wang Zhengjie, Xiong Zhonghao, Yuan Xiaoshu, Cui Hong, Liu Xuexiu, and Ren Wei. Impact of large language models on medical education and teaching adaptations. *JMIR Med Inform*, 12:e55933, Jul 2024.
- [2] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization, 2020.

- [3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629, 2022.
- [4] Brihi Joshi, Keyu He, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, and Xiang Ren. Eli-why: Evaluating the pedagogical utility of language model explanations, 2025.
- [5] Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. Know your audience: Do llms adapt to different age and education levels?, 2023.
- [6] Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. Controllable sentence simplification, 2020.
- [7] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.
- [8] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering, 2023.
- [9] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
- [10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [11] National Library of Medicine (US). Medlineplus topics xml file. https://medlineplus.gov/xml/mplus_topics_2026-02-21.xml, 2026. National Institutes of Health. Accessed February 21, 2026.
- [12] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022.

A Prompt Templates

This section provides the prompt templates used.

A.1 Baseline

[System]

You are a careful medical rewriting assistant.

Your job is to rewrite text for a target audience while preserving factual content from the source.

Never introduce new unsupported medical information.

[User]

Audience: {{AUDIENCE}}

{{AUDIENCE_INSTRUCTION}}

Task:

- Rewrite the SOURCE TEXT for this audience.
- Preserve source medical facts and meaning.
- Do not add new causes, risks, treatments, numbers, or medical claims not in SOURCE TEXT.
- Keep length close to source intent.

Topic title:

{{TOPIC_TITLE}}

SOURCE TEXT:
{{SOURCE_TEXT}}

Output:
- Return only the rewritten explanation text.

The audience-conditioned baseline inserts one of two instruction blocks depending on the target audience. The Basic instruction directs the model to use plain language, define medical terms when necessary, and target a length of 120-220 words. The Technical instruction directs the model to use precise clinical terminology and concise phrasing, with a target length of 140-260 words.

For comparison, the general ablation replaces both audience-specific instructions with a single neutral instruction that asks the model to preserve medical accuracy while writing accessibly, with a target length of 130-240 words. This setting produces one rewrite per topic without explicit audience conditioning.

A.2 Retrieval-Augmented Generation

[System]
You are a careful medical rewriting assistant.
Your job is to rewrite text for a target audience while preserving factual content.
You are given EVIDENCE PASSAGES from the same medical topic to help you stay grounded.
Use the evidence to verify and support your rewrite, but do not introduce any claims not present in the SOURCE TEXT or EVIDENCE.
Never fabricate new medical information.

[User]
Audience: {{AUDIENCE}}

{{AUDIENCE_INSTRUCTION}}

Task:
- Rewrite the SOURCE TEXT for this audience.
- Preserve source medical facts and meaning.
- Use the EVIDENCE PASSAGES below to support accuracy, but do not add claims absent from SOURCE TEXT.
- Do not add new causes, risks, treatments, numbers, or medical claims not in SOURCE TEXT or EVIDENCE.
- If audience is Basic: use plain language, shorter sentences, and briefly define necessary medical terms.
- If audience is technical: use precise medical terminology and concise clinical phrasing.
- Keep length close to source intent.

Topic title:
{{TOPIC_TITLE}}

SOURCE TEXT:
{{SOURCE_TEXT}}

EVIDENCE PASSAGES:
{{EVIDENCE}}

Output:
- Return only the rewritten explanation text.

B Effective Number of Retrieved Passages

Retriever	k	Avg Passages	Range
BM25	1	1.0	1–1
BM25	3	2.2	1–3
BM25	5	2.8	1–5
TF-IDF	1	1.0	1–1
TF-IDF	3	2.2	1–3
TF-IDF	5	2.8	1–5

Table 8: The number of retrieved passages per topic on the test set. The actual number of retrieved passages can be smaller than k when fewer than k evidence passages are available for a topic.

C Milestone Pivots

C.1 Base Model

During milestone development we evaluated multiple instruction-tuned models (Llama 3.2 3B [13] and Mistral 7B [14]) to establish baseline behavior. For the final experiments we standardize on Llama 3.1 8B Instruct [13] to enable controlled comparisons across prompt-only, RAG, and LoRA configurations.

D Calibration

D.1 Judge Selection

Table 9 summarizes our anchor-based calibration comparison across candidate faithfulness evaluators. We report raw unsupported rate, parse-adjusted unsupported rate, and parse/missing-metadata incidence to separate true unsupported behavior from evaluator instability. This comparison motivates our final choice of GPT-5.1 + prompt v1 as the primary faithfulness baseline.

Anchor Eval Variant	Unsupported (%)	Adj. Unsupported (%)	Parse / Missing-meta (%)	Notes
GPT-5.1 + prompt v1	2.50	1.99	0.52	Most reliable anchor calibration; selected as the main faithfulness baseline.
GPT-4.1-mini + prompt v2	7.22	3.48	3.87	
GPT-4.1-mini + NLI-derived labels from judge metadata	6.37	6.37	0.09	Uses <code>nli_proxy</code> metadata as a decision proxy; useful as a diagnostic signal.
Standalone NLI threshold sweeps	8.26	8.26	0.00	Useful as a threshold-based diagnostic, but less reliable for fine-grained subtype analysis.

Table 9: Faithfulness calibration summary. Adjusted unsupported excludes parse/missing-meta cases from the denominator where applicable.

D.1.1 Human Evaluation

Table 10 compares human NLI-style judgments with automated classification on claims from the prompt-only baseline test split. Binary agreement is 90.4% for Basic rewrites and 90.3% for Technical rewrites, indicating strong alignment between manual and automated labels.

We use this data mostly just to get confidence in judge selection, but in future it could be used directly for validation and any threshold adjustment.

Human verdict	Basic ($n = 83$)			Technical ($n = 72$)		
	Auto:supported	Auto:unsupported	Total	Auto:supported	Auto:unsupported	Total
Entails	73	4	77	60	4	64
Partial	2	0	2	5	0	5
Contradicts	1	0	1	0	0	0
Neutral	3	0	3	3	0	3
All	79	4	83	68	4	72
Binary agreement	90.4%			90.3%		

Table 10: Agreement between human NLI judgments and automated classification on claims from the prompt-only baseline test split.

Most claims are labeled as supported by both annotators. The automated classifier marks 79/83 Basic claims and 68/72 Technical claims as supported. Disagreements occur in both directions: in each audience, four claims classified as unsupported by the automated system were labeled as entailing by the human annotator, while a smaller number of claims labeled non-faithful by the human annotator were classified as supported by the automated system.

E Secondary Metrics

E.1 Baseline Faithfulness

Table 11 reports the secondary LLM-judge label breakdown. Across both splits, `partial_support` is the most frequent non-supported label, with higher rates for Technical rewrites than for Basic rewrites. Other categories, including `missing_evidence`, `contradicted`, and `off_topic`, remain relatively infrequent.

Split	Audience	partial_support	missing_evidence	off_topic	contradicted	unclear_claim	parse_error
Val	Overall	5.48%	0.91%	0.05%	0.10%	0.29%	1.15%
Val	Basic	3.77%	1.03%	0.00%	0.00%	0.43%	0.69%
Val	Technical	7.65%	0.77%	0.11%	0.22%	0.11%	1.75%
Test	Overall	6.00%	0.86%	0.05%	0.24%	0.14%	0.91%
Test	Basic	4.10%	0.70%	0.00%	0.09%	0.26%	0.44%
Test	Technical	8.32%	1.07%	0.11%	0.43%	0.00%	1.49%

Table 11: Secondary LLM-judge diagnostic labels and parse-error rates at the claim level, reported as percentages by split and audience.