LayCare-LLM: Health-Literacy-Aware Medical Reasoning (Team 21)

Shuwei He
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
shuweih@andrew.cmu.edu

Vu Hoang
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
vthoang@andrew.cmu.edu

Lin Park
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA 15213
linp@andrew.cmu.edu

Abstract

Low health literacy remains a major barrier to equitable healthcare access, limiting patients' ability to interpret symptoms, understand prescriptions, and evaluate medical advice. Although Large Language Models (LLMs) show strong performance in medical QA and text simplification, they continue to exhibit safety vulnerabilities, overconfidence, and inconsistent readability. Building on Reflexion Shinn et al. [2023], a framework for verbal reinforcement learning, we investigate the **LayCare-LLMs**, health-literacy–oriented medical reasoning models with two main characteristics (i) *self-reflection with episodic memory* to iteratively learn from prior errors, and (ii) *adversarial training* using automatically generated challenges to identify and correct systematic weaknesses. LayCare-LLMs are designed to produce actionable, comprehensible, and trustworthy patient-facing explanations, advancing both safety and accessibility in medical communication.

1 Motivation

Health literacy—the ability to access, understand, and act on health information—strongly influences outcomes. Nearly half of U.S. adults have limited health literacy, increasing the risk of medication errors, inadequate chronic disease management, and reduced uptake of preventive care Nutbeam [2000]. Existing digital interventions (e.g., portals, chatbots) often fail due to complexity, lack of personalization, or limited safeguards.

Recent LLMs such as Med-PaLM Singhal et al. [2023] and GPT-4 OpenAI [2023] exhibit strong reasoning capabilities, but they are not explicitly aligned with patient literacy needs. They also lack systematic uncertainty estimation and frequently fail under adversarial or ambiguous conditions Ahmad et al. [2023]. This project introduces a *reflective and adversarially-hardened medical*

Preprint. Under review.

reasoning assistant that implements self-reflection and adversarial training to better support patients facing confusing or conflicting health information.

2 Literature review

2.1 LLMs in healthcare

Large language models are increasingly applied in healthcare, from generating patient education materials Aydin et al. [2024] to simplifying medical records Dehkordi et al. [2024]. Frontier LLMs (e.g., Med-PaLM Singhal et al. [2023], BioMistral Labrak et al. [2024], BioMedLM Bolton et al. [2024]) approach expert-level performance on medical QA. However, safety and reliability remain open challenges as LLMs do not guarantee truthfulness, and models may hallucinate or be overconfident, which is particularly salient when outputs can guide health behaviors Ahmad et al. [2023], Huang et al. [2025], Ji et al. [2023]. There are also trade-offs between simplification and fidelity Devaraj et al. [2022], Zaretsky et al. [2024].

2.2 Health literacy and evaluation of patient materials

Health literacy research provides concrete targets for understandability and actionability that go beyond traditional readability formulas. Instruments such as TOFHLA/S-TOFHLA Parker et al. [1995] measure functional literacy, while PEMAT Shoemaker et al. [2014] offers criteria to assess whether materials are easy to understand and prompt clear next steps. Work on physician–patient communication stresses communicating uncertainty clearly Gigerenzer et al. [2008], Trevena et al. [2013]. This motivates a framework that evaluates whether outputs support informed action *and* communicate calibrated uncertainty, not just readability gains.

2.3 Reflexion and reflective learning

A growing body of work studies LLMs that critique their outputs and store experiences to improve future reasoning Madaan et al. [2023], Zhong et al. [2024], Madaan et al. [2022], Shinn et al. [2023]. Related methods (e.g., STaR Zelikman et al. [2022], self-consistency Wang et al. [2023], ReAct Yao et al. [2023]) show gains from decomposing problems, aggregating multiple reasoning paths, and coupling reasoning with tool use. Applications to patient-facing communication remain limited, where reflections must attend to literacy constraints and harm-avoidant advice Busch et al. [2025], Chen et al. [2025]. We couple episodic memory with literacy-aware templates and safety checklists so reflections explicitly target readability, actionability, and clinical caution.

2.4 Adversarial robustness and safety-critical generation

Adversarial training has emerged as a central strategy for strengthening model robustness by exposing LLMs to systematically perturbed or misleading inputs and optimizing them to resist such attacks. In NLP, adversarial prompts reveal vulnerabilities in reasoning, factual consistency, calibration, and instruction adherence Jin et al. [2020], Li et al. [2019], Zhu et al. [2024]. Emerging work in medical AI similarly shows that clinical and biomedical language models are susceptible to adversarial manipulations that can alter diagnostic reasoning, degrade biomedical information extraction, or generate unsafe clinical guidance Yang et al. [2025], Moradi and Samwald [2022], Mozhegova et al. [2025]. Such failures pose concrete clinical safety risks in patient-facing contexts, underscoring the need to treat robustness evaluation as a core requirement for trustworthy deployment rather than an optional extension. Adversarial training also aligns with broader priorities in responsible medical AI by stress-testing how models communicate uncertainty, maintain harm-avoidant decision boundaries, and respond to constrained literacy environments or ambiguous clinical cues. Adversarial methods target the failure modes that arise under distribution shifts, manipulative prompts, and high-stakes ambiguity—conditions that frequently characterize real-world patient communication and safety-critical health settings.

3 Methodology

3.1 Reflexion framework

The key idea behind Reflexion (Shinn et al. [2023]) is self-improvement via verbal feedback. The framework is a simple loop that lets an LLM iteratively improve its next attempt using its own short, natural language reflections. There are three main roles in the framework.

- Actor (LLM): produces an answer (and optionally reasoning)
- Evaluator (signal): scores that attempt with a sparse, task-specific reward (binary correct/incorrect signal)
- Self-reflexion (LLM): reads the attempt and rewards, then writes a brief lesson and plan on what went wrong and what to try next

Then, two memories are kept: a short term trajectory (last attempt) and a long-term memory (the reflection text). The next attempt will condition on both these pieces of information.

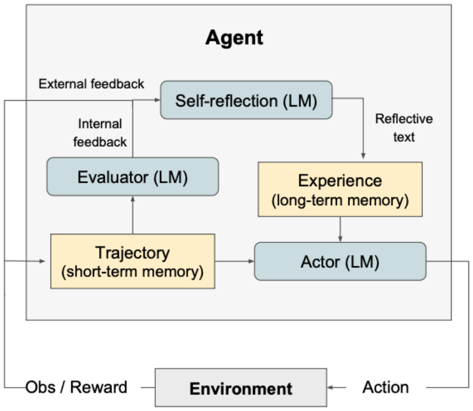
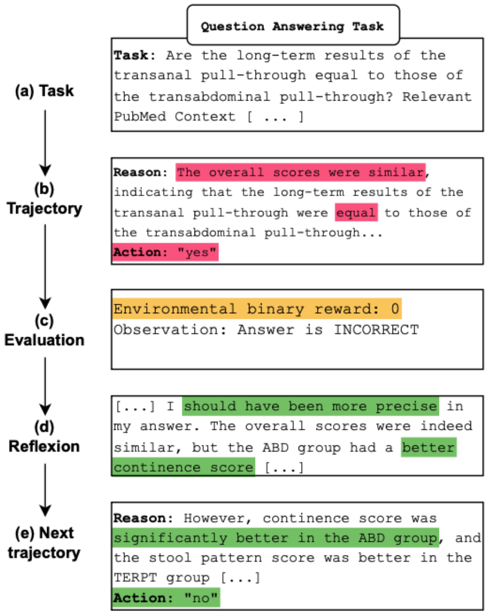


Figure 1: Diagram of reflexion Shinn et al. [2023]



Task: Are the long-term results of the transanal pull-through equal to those of the transabdominal pull-through? Relevant PubMed Context [...]

Reason: The overall scores were similar, indicating that the long-term results of the transanal pull-through were equal to those of the transabdominal pull-through...
Action: "yes"

Environmental binary reward: 0
Observation: Answer is INCORRECT

[...] I should have been more precise in my answer. The overall scores were indeed similar, but the ABD group had a better continence score [...]

Reason: However, continence score was significantly better in the ABD group, and the stool pattern score was better in the TERPT group [...]
Action: "no"

Figure 2: Example of Reflexion pipeline

The reflection process helps to tell the model how to adjust its next attempt without changing the model weights. Compared to well known prompting techniques like CoT (Wei et al. [2023]) or ReAct (Yao et al. [2023]), Reflexion can be thought of as an outer loop that re-used feedback to refine subsequent attempt and can be layered on top of CoT/ReAct.

3.2 Baseline inference pipeline

As a first step, we run an *inference-only* baseline with **Llama 3.1–8B (Instruct)** under a literacy-aware prompting framework. The LLM generates answers using only the provided context, without retrieving information from any external sources. We evaluate on two datasets: **MedRedQA** (∼51K consumer QA pairs; long-form rationales) Nguyen et al. [2023], and **PubMedQA** (1,000 expert-labeled questions with {yes,no,maybe} plus long answers) Jin et al. [2019]. For MedRedQA, we did *score-based* subsampling: we ranked all rows took the top 'Response Score' entries. Each instance is normalized to **query**, **evidence/rationales**, **gold_answer (ground truth label)**, and literacy labels.

3.3 Adversarial training pipeline

To surface robustness failures in biomedical question answering and improve safety under prompt manipulations, we implement an adversarial supervised fine-tuning (SFT) pipeline on PubMedQA that combines literacy-aware response formatting with automatically discovered adversarial examples. The pipeline proceeds through four stages: constructing a literacy-aware training set, generating adversarial prompts and aligned responses, performing adversarial supervised fine-tuning, and evaluating robustness under clean and adversarial prompting conditions.

Stage 1: Dataset construction and literacy-aware supervision. We construct a supervised dataset from the PubMedQA labeled split, formatting each instance as a chat interaction. The system message defines a literacy-aware medical assistant; the user message requests a one-word answer (Yes/No/Maybe) followed by a short lay explanation; and the assistant target includes the gold label and a compressed lay-friendly rationale derived from the PubMedQA long answer. These formatted examples later form one part of the combined training corpus for adversarial fine-tuning.

Stage 2: Automatic adversarial prompt generation. Starting from the base Llama-3.1-8B-Instruct model with LoRA adapters attached (but not fine-tuned), we generate adversarial prompts for a subset of PubMedQA items. For each question–abstract pair, we construct a perturbed prompt that stresses the decision rule by enforcing a structured output format (single-word answer + simplified explanation) and subtly influencing the model's reasoning process. We then query the model with these prompts and use its structured "safe" answers—produced through the controlled generation template—as the supervision signal. Each adversarial training instance therefore consists of the original biomedical content, an adversarially phrased instruction, and the model's aligned, safety-checked response.

Stage 3: Adversarial supervised fine-tuning (the only training stage). We concatenate the clean literacy-aware dataset from Stage 1 with the automatically discovered adversarial dataset from Stage 2, forming a unified training corpus. We then perform a single supervised fine-tuning pass using LoRA adapters while keeping the base model frozen. This produces the adversarial SFT model, explicitly optimized to remain accurate and safety-aligned even when user prompts are adversarially perturbed. In this pipeline, this is the *only* gradient-based training stage.

Stage 4: Robustness evaluation (baseline model vs. adversarial SFT). For evaluation, we compare (i) a baseline model—obtained by loading the base Llama 3.1 model with the same LoRA architecture used in fine-tuning but without any parameter updates—and (ii) the adversarially fine-tuned model. This baseline is intentionally defined to be structurally comparable to the adversarial SFT model so that differences in performance reflect the effect of adversarial training rather than architectural or formatting differences; it does not correspond to the baseline used in the self-reflexion experiments. Both models are evaluated on the PubMedQA labeled split under two prompting conditions: clean prompts requesting a Yes/No/Maybe answer with a brief lay explanation, and adversarial prompts designed to bias the model toward an incorrect label (e.g., by overriding instructions or enforcing a specific answer). We compute clean accuracy, attacked accuracy, the Attack Success Rate (ASR; will explain in metrics section), and the retained-correct rate (the fraction of originally correct predictions that remain correct under attack). This comparison isolates the contribution of adversarial supervised fine-tuning to robustness against prompt-based manipulations.

3.4 Metrics and equations

Accuracy and Classification Balance (PubMedQA). We measure correctness and whether the model performs consistently across the three answer categories (*yes/no/maybe*).

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}.$$

For each class c, define precision $P_c = \frac{\text{TP}_c}{\text{TP}_c+\text{FP}_c}$ and recall $R_c = \frac{\text{TP}_c}{\text{TP}_c+\text{FN}_c}$. The per-class F1 score and macro-averaged F1 are:

$$\text{F1}_c = \frac{2P_cR_c}{P_c + R_c}, \qquad \text{Macro-F1} = \frac{1}{3}\sum_{c\in\{\text{yes,no,maybe}\}} \text{F1}_c.$$

Accuracy measures overall correctness, while Macro-F1 ensures balanced performance and prevents the model from overfitting to the majority class.

Generative Quality We evaluate how well model-generated rationales align with expert reference explanations in person-friendly language using F1-based overlap metrics defined over different tokenizations. Let $\tau(X)$ denote a tokenization function applied to text X, and let $\mathcal{T}(X) = \tau(X)$ be the resulting multiset of tokens. We define the token-level overlap between candidate (C) and reference (R) as

$$\mathrm{overlap}_\tau(C,R) = \sum_{g \in \mathcal{T}(C) \cap \mathcal{T}(R)} \min\big(\mathrm{count}_{\mathcal{T}(C)}(g),\, \mathrm{count}_{\mathcal{T}(R)}(g)\big),$$

with precision and recall

$$P_\tau = \frac{\mathrm{overlap}_\tau}{|\mathcal{T}(C)|}, \qquad R_\tau = \frac{\mathrm{overlap}_\tau}{|\mathcal{T}(R)|},$$

and the unified F1 score

$$F1_\tau = \frac{2P_\tau R_\tau}{P_\tau + R_\tau}.$$

ROUGE-1 F1. Following Lin [2004], ROUGE-1 F1 instantiates this framework with a ROUGE-style unigram tokenizer τ_{rouge} that preserves lexical diversity and repeated words:

$$\text{ROUGE-1 F1} = F1_{\tau_{\text{rouge}}}.$$

Because τ_{rouge} is designed for summarization, ROUGE-1 F1 emphasizes recall and tends to reward longer generations that cover more of the reference wording. In our setting, it serves as a proxy for *completeness* and lexical alignment with expert rationales.

Token F1. Given Rajpurkar et al. [2016], Token F1 uses a QA-style normalization τ_{tok} that lowercases, strips punctuation, and applies whitespace-based tokenization:

$$\text{Token F1} = F1_{\tau_{\text{tok}}}.$$

This normalization collapses superficial lexical variation and makes the metric more sensitive to extraneous or hallucinated content: precision drops whenever the model inserts irrelevant or speculative medical details, even if overall recall is high. As a result, Token F1 captures *focused, faithful reasoning* rather than sheer length or lexical richness.

ROUGE-1 F1 and Token F1 share the same harmonic-mean formula $F1_\tau = 2P_\tau R_\tau/(P_\tau + R_\tau)$, but they differ in the underlying tokenization function τ and thus in what they reward. ROUGE-1 F1, built on τ_{rouge}, primarily reflects lexical recall and can increase when explanations become longer, even if they include redundant or weakly relevant details. Token F1, built on τ_{tok}, penalizes such verbosity because precision sharply decreases when additional tokens do not appear in the expert rationale. In biomedical QA, we therefore interpret ROUGE-1 F1 as measuring *coverage* of the reference explanation, while Token F1 measures how concisely and accurately the model reproduces medically relevant content. Consistent with these goals, our adversarial training experiments on PubMedQA are evaluated using **Token F1**, whereas the self-reflexion experiments, we report **ROUGE-1 F1**.

Readability We assess how accessible the model's rationales are using three complementary metrics: **Flesch Reading Ease (FRE)**, which measures how difficult an English passage is to understand [Flesch, 1948]; **Flesch–Kincaid Grade Level (FKGL)**, which approximates the U.S. school grade level required for comprehension [Kincaid et al., 1975]; and the **SMOG Index**, which estimates years of education needed to understand the text [McLaughlin, 1969]. Using all three metrics could provide a more reliable picture of readability across biomedical rationales.

Calibration To ensure that predicted probabilities reflect real correctness likelihoods, improving transparency and trust, for probabilities $\mathbf{p}_i = (p_{i1}, ..., p_{iK})$ and true one-hot labels $\mathbf{y}_i$:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - y_{ik})^2.$$

We also report Expected Calibration Error (ECE). Partition predictions into bins $\{\mathcal{B}_b\}_{b=1}^{B}$ by $\hat{c}_i = \max_k p_{ik}$:

$$\mathrm{acc}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|} \sum_{i \in \mathcal{B}_b} \mathbf{1}\{\arg\max_k p_{ik} = y_i\}, \quad \mathrm{conf}(\mathcal{B}_b) = \frac{1}{|\mathcal{B}_b|} \sum_{i \in \mathcal{B}_b} \hat{c}_i,$$

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{B}_b|}{N} \big|\mathrm{acc}(\mathcal{B}_b) - \mathrm{conf}(\mathcal{B}_b)\big|.$$

Brier penalizes overconfident errors, while ECE measures how closely stated confidence matches actual correctness. Lower is better.

Robustness A central goal is to assess how well the model maintains accuracy when inputs are adversarially perturbed. *Attacked accuracy*—the model's accuracy on adversarially modified prompts—provides the most direct indicator of robustness, since a reliable model should remain correct even when the prompt attempts to bias or distort its decision. For a LLM f, adversarial inputs $\tilde{x}_i$ with true label y_i, attacked accuracy is defined as:

$$\mathrm{Acc}_{\mathrm{adv}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\, f(\tilde{x}_i) = y_i \,\}.$$

To more precisely quantify degradation under adversarial pressure, we consider a threat model $\mathcal{T}$ that generates perturbed prompts $\tilde{x}_i = \mathcal{T}(x_i)$ from clean inputs x_i. The *Attack Success Rate (ASR)* measures the fraction of originally correct predictions that become incorrect under attack:

$$\mathrm{ASR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\, f(x_i) = y_i \;\wedge\; f(\tilde{x}_i) \neq y_i \,\}.$$

Finally, the *retained-correct rate (RCR)* captures the stability of correct reasoning under perturbation:

$$\mathrm{RCR} = \frac{\sum_{i=1}^{N} \mathbf{1}\{\, f(x_i) = y_i \;\wedge\; f(\tilde{x}_i) = y_i \,\}}{\sum_{i=1}^{N} \mathbf{1}\{\, f(x_i) = y_i \,\}}.$$

High attacked accuracy and retained-correct rate, together with low ASR, indicate that the model preserves correct biomedical reasoning even when prompts are adversarially manipulated. These metrics jointly provide a comprehensive evaluation of robustness in safety-critical biomedical QA.

3.5 Evaluation protocol

Our goal is to maintain competitive task performance while improving *accuracy*, *calibration*, *readability*, and (in later phases) *robustness*.

- **PubMedQA**: (i) **Reflexion+CoT**. Accuracy, Macro-F1; Calibration (Brier, ECE with first-step renormalization). (ii) **Adversarial Training**: Accuracy, Token-F1, and robustness.
- **MedRedQA (long-form)**: ROUGE-1 F1; Confidence (ANLL, entropy); Readability (FRE, FKGL, SMOG).

4 Experiments

We currently use **LLaMA 3.1 8B Instruct** as the baseline: it is accessible, trainable on local/colab resources, and widely used in healthcare-LLM studies. We implement experiments by two parallel ways:

For Reflexion, we then evaluate two settings: (i) **Baseline + Chain-of-Thought (CoT)** and (ii) **Baseline + CoT + Reflexion**. We do not train model weights on either PubMedQA or MedRed; both are used for evaluation. In all settings, we standardize the prompts and ask the model to produce a brief, plain-English explanation.

In the Reflexion framework, the actor, evaluator, and self-reflection modules iterate in a loop across trials until the evaluator deems an attempt correct. Because inference is time-consuming and context windows are limited, we cap the long-term memory size and allow at most three attempts per question in our experiments.

For adversarial supervised fine-tuning, we use **LLaMA-3.1-8B-Instruct** with QLoRA (4-bit NF4 quantization) and LoRA adapters, updating only the adapter weights while keeping all base model parameters frozen. Unlike the Reflexion experiments, which evaluate the unmodified baseline model, this pipeline performs a single fine-tuning stage that combines the literacy-aware PubMedQA examples with the automatically generated adversarial examples. We train using HuggingFace's **SFTTrainer** with the AdamW optimizer, a cosine learning-rate schedule (initial learning rate = 5×10^{-5}), per-device batch size 1, and 16 gradient-accumulation steps (effective batch size 16). All inputs are tokenized to a maximum sequence length of 768 tokens.

For evaluation of adversarial training, we load either (i) the base model with untrained LoRA adapters (baseline) or (ii) the adversarially fine-tuned adapters, and run decoding-only inference with prompts truncated at 1024 tokens and outputs limited to 80–160 tokens. This setup cleanly separates the adversarial-training experiment from the Reflexion-only experiment while preserving a comparable model architecture across conditions.

5 Results

5.1 Reflexion

5.1.1 PubMedQA

Accuracy and macro-F1 both improve when adding reasoning with CoT and reflexion. The Baseline attains 0.755 (accuracy) and 0.532 (F1); with CoT these rise to 0.771 and 0.571; with CoT+Reflexion, they reach 0.795 and 0.601, the best overall. The macro-F1 gains suggest the reasoning steps help on the rarer/ambiguous "maybe" cases, not just the dominant yes/no labels. Calibration also improves as the Brier score decreases from 0.379 (Baseline) to 0.370 with CoT and to 0.372 with Reflexion. By contrast, ROUGE-1 falls from 0.325 (Baseline) to 0.307 with CoT and 0.301 with Reflexion, suggesting more paraphrasing and less direct keyword overlap. Readability becomes harder: FRE drops from 36.70 to 23.82 and 22.05; FKGL increases from 12.65 to 16.29 and 17.17; SMOG increases from 14.34 to 17.01 and 17.16. Overall, multi-turn reasoning increases task performance and improves calibration, at the cost of lower extractive overlap and simpler explanation.

Table 1: PubMedQA results: baseline vs. CoT vs. CoT+Reflexion

Model	**Accuracy/F1** ↑	**ROUGE-1** ↑	**Brier** ↓	**Readability**(FRE ↑ (FKGL,SMOG) ↓)
Baseline	0.755/0.532	**0.325**	0.379	**FRE 36.70, FKGL 12.65, SMOG 14.34**
Baseline+CoT	0.771/0.571	0.307	**0.370**	FRE 23.22, FKGL 16.93, SMOG 17.01
Baseline+CoT+Reflexion	**0.795/0.601**	0.301	0.372	FRE 22.05, FKGL 17.17, SMOG 17.16

5.1.2 MedRedQA

On MedRedQA, the baseline model achieves the highest ROUGE-1 F1 (0.17) but produces very hard-to-read explanations: FRE 25 and grade-level metrics correspond to college-plus level. This is consistent with our qualitative impression that the baseline writes long, jargon-heavy paragraphs that overlap reasonably well with the expert answer but are poorly aligned with our 6th–8th grade health-literacy target. Given the inherent nature of the MedredQA being an open-ended consumer QA setting, Exact-match EM is 0 across all models; free-form answers rarely match the reference verbatim exactly.

Adding CoT (Baseline+CoT) substantially improves readability but hurts ROUGE-1 as the metric drops from 0.17 to 0.12, while FRE jumps from 25.3 to 53.5 and FKGL falls from 16.2 to about 10.1. In other words, CoT encourages the model to explain itself step-by-step in plainer language, which helps readability scores but leads to less literal overlap with the crowd-sourced reference answers. Calibration metrics remains similar. This suggests that prompting alone does not meaningfully fix overconfidence.

Table 2: MedRedQA results: baseline vs. CoT vs. CoT+Reflexion (short) vs. CoT+Reflexion (long)

Model	**ROUGE-1**↑	**Calib. (Brier / ECE)**↓	**Readability**(FRE ↑ (FKGL,SMOG) ↓)
Baseline	**0.17**	0.72 / 0.84	FRE 25.3, FKGL 16.2, SMOG 16.9
Baseline+CoT	0.12	0.71 / 0.84	FRE 53.5, FKGL 10.1, SMOG 12.0
Reflexion (short answer only)	0.10	**0.70 / 0.83**	**FRE 26.6, FKGL 9.7, SMOG 10.1**
Reflexion (long rationale)	0.12	0.72 / 0.84	FRE 29.0, FKGL 13.8, SMOG 13.6

The two Reflexion variants show a similar trade-off but with different emphasis. **Reflexion-short** computes metrics on the short final answer only (EM-style). It attains the lowest ROUGE-1 F1 (0.10) but the most patient-friendly grade levels: FKGL 9.7 and SMOG 10.1, noticeably better than the baseline and slightly better than CoT. FRE, however, remains low (26.6), illustrating the instability of readability heuristics across metrics—short, telegraphic answers can still be penalized by FRE if they contain medical terms. Similar to the CoT model, calibration again remains similar.

Considering the free-form answer type, we test a variant of the Reflexion model from above, and try **Reflexion-long** which evaluates on the final rationale paragraph. Here, ROUGE-1 F1 recovers to 0.12—close to the CoT model and much higher than Reflexion-short—because the longer trajectories tend to quote more of the original question and reference phrasing. This comes at a cost to readability: FKGL rises to 13.8 and SMOG to 13.6, worse than CoT and Reflexion-short but still slightly better than the baseline's very technical outputs. Intuitively, Reflexion-long encourages the model to produce more elaborate, evidence-anchored explanations that echo clinical language, which boosts lexical overlap but drags the reading level back toward "early college."

5.2 Adversarial Training on PubMedQA

The first column of table 3 summarizes clean-performance outcomes on the PubMedQA labeled split. Even without adversarial perturbations, adversarially augmented supervision yields clear gains in both label prediction and rationale quality. The **Baseline Model** attains 0.729 accuracy and 0.256 Token F1, indicating moderate alignment with expert rationales. The **Adversarial SFT** model improves to 0.763 accuracy and 0.316 Token F1, demonstrating that exposure to automatically generated adversarial examples strengthens not only robustness but also standard predictive performance under clean prompts.

Table 3: Robustness evaluation on PubMedQA (clean vs. adversarial prompts).

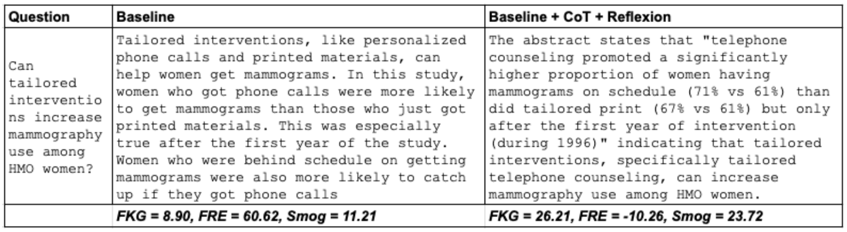
Model	**Clean Acc./Token F1** ↑	**Attacked Acc.** ↑	**ASR** ↓	**Retained-correct Rate** ↑
Baseline Model	0.729/0.256	0.552	**0.270**	**0.730**
Adversarial SFT	**0.763/0.316**	**0.585**	0.279	0.721

Next, we discuss the robustness outcomes under adversarially manipulated prompts. The **Baseline Model** drops from 0.729 clean accuracy to 0.552 under attack (ASR = 0.270), whereas the **Adversarial SFT** model achieves higher clean and attacked accuracy (0.763 and 0.585) but shows a nearly unchanged attack success rate (0.279). This reflects the distinction between unconditional and conditional robustness: attacked accuracy increases partly because adversarial SFT corrects examples the baseline originally mispredicted, whereas ASR isolates the probability that a *clean-correct* answer flips under attack. That conditional vulnerability remains largely unmitigated.

Overall, adversarial SFT improves general predictive ability and unconditional robustness, but only modestly affects the likelihood that targeted prompt manipulations overturn an otherwise correct prediction. This suggests that the perturbations used in our evaluation probe deeper weaknesses in biomedical reasoning or instruction adherence—vulnerabilities that may require targeted adversarial augmentation, consistency regularization, or uncertainty-aware decoding to address effectively.

5.3 Discussion

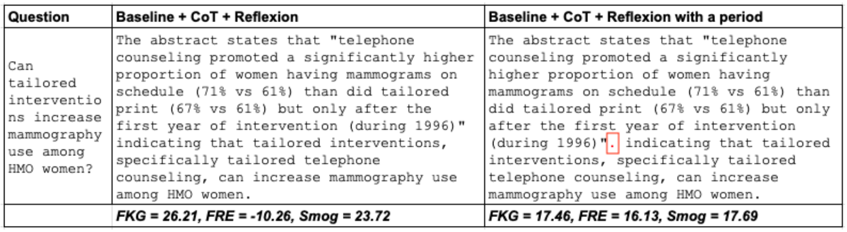
The readability drop on PubMedQA is sizable, so we inspected model outputs from the baseline and the Reflexion setup. Qualitatively, the answers are often similar in content. The main difference is that CoT + Reflexion tends to quote or explicitly cite phrases from the abstract. Moreover, across the chain of thought and reflexion attempts, the model has a tendency to aggregate all of its intermediate thoughts. This makes the rationale more evidence-anchored, but also lengthens sentences.



Question	Baseline	Baseline + CoT + Reflexion
Can tailored interventions increase mammography use among HMO women?	Tailored interventions, like personalized phone calls and printed materials, can help women get mammograms. In this study, women who got phone calls were more likely to get mammograms than those who just got printed materials. This was especially true after the first year of the study. Women who were behind schedule on getting mammograms were also more likely to catch up if they got phone calls	The abstract states that "telephone counseling promoted a significantly higher proportion of women having mammograms on schedule (71% vs 61%) than did tailored print (67% vs 61%) but only after the first year of intervention (during 1996)" indicating that tailored interventions, specifically tailored telephone counseling, can increase mammography use among HMO women.
	FKG = 8.90, FRE = 60.62, Smog = 11.21	***FKG = 26.21, FRE = -10.26, Smog = 23.72***

Figure 3: Comparison of output between baseline and CoT+Reflexion

Closer analysis of the readability scores shows that these metrics heavily penalize long sentences. In the same example in Figure 4, inserting a single period to split one sentence markedly improves all readability scores. Because such metrics (e.g., FRE, FKGL, SMOG) are still widely used to assess LLM outputs, this sensitivity raises questions about how faithfully they reflect true understandability.



Question	Baseline + CoT + Reflexion	Baseline + CoT + Reflexion with a period
Can tailored interventions increase mammography use among HMO women?	The abstract states that "telephone counseling promoted a significantly higher proportion of women having mammograms on schedule (71% vs 61%) than did tailored print (67% vs 61%) but only after the first year of intervention (during 1996)" indicating that tailored interventions, specifically tailored telephone counseling, can increase mammography use among HMO women.	The abstract states that "telephone counseling promoted a significantly higher proportion of women having mammograms on schedule (71% vs 61%) than did tailored print (67% vs 61%) but only after the first year of intervention (during 1996)". indicating that tailored interventions, specifically tailored telephone counseling, can increase mammography use among HMO women.
	FKG = 26.21, FRE = -10.26, Smog = 23.72	***FKG = 17.46, FRE = 16.13, Smog = 17.69***

Figure 4: Readability metrics' sensitivity to punctuation

Based on the MedRedQA results, we find that the model's medical advice is often clinically reasonable but doesn't match the short Reddit-style reference answers. For example, the model might say "He should go to the ER for more extensive heart testing" after an Apple Watch detects a possible left bundle branch block on an ECG, while the reference answer is more like "The Apple Watch isn't a real ECG; wait to see your regular doctor." Similarly, the model may suggest "further testing for an abnormal ECG," whereas the reference simply says "it depends on how it's abnormal." We also see that Reflexion tends to improve framing rather than content: the model adds caveats, more nuanced differentials, and clearer action plans ("see a dermatologist," "consult a neurologist," "seek immediate medical attention for possible meningitis") but still fails the exact-label match. Long-reflection runs produce extremely verbose, instruction-contaminated outputs (lots of meta-text like "Please respond with your answer. . . ") and short-reflection runs are cleaner but show the same pattern: the model self-critiques ("I was too vague," "I didn't give a specific timeline for resuming sex after genital warts treatment," "I should have defined 'severe COVID'") yet its next answer is still scored incorrect. Overall, Reflexion is successfully pushing the model toward more cautious, structured, and self-aware reasoning, but in our setup it doesn't translate into higher exact-match accuracy against the reference answers, and a big failure mode is format/label mismatch rather than purely faulty clinical reasoning.

We attribute the divergence of how reflexion influences readability scores to dataset characteristics. MedRedQA contains consumer Q&A from Reddit: questions are written by laypeople and the supporting context is already in everyday language. When we add CoT/Reflexion, the model tends to paraphrase and structure that lay content into shorter, clearer sentences, which directly boosts FRE and lowers FKGL/SMOG. In contrast, PubMedQA draws on scientific abstracts with domain jargon and numeric results. To answer correctly, the model often must cite trial arms, effect sizes, and qualifiers (e.g., p-values, odds ratios), which lengthens sentences and introduces technical terms. As a result, readability scores worsen even though the medical reasoning improves. Constrained prompts (e.g., "write in sentences under 10 words") can mechanically improve FRE/FKGL/SMOG on PubMedQA, but may hurt fidelity and simply game the metric. A better direction is concept simplification that preserves key evidence. Evaluation should pair readability with clinical faithfulness and expert judgment, rather than optimizing surface-level heuristics alone.

Adversarial fine-tuning yields clear gains in both clean and attacked performance, yet these improvements do not manifest as higher resistance to targeted manipulations. Although the adversarially trained model achieves higher clean accuracy and higher attacked accuracy, its attack success rate remains nearly identical to the baseline. This discrepancy reflects the conditional nature of ASR: it is computed only over the subset of examples that were answered correctly in the clean setting. Adversarial SFT corrects many cases that were previously wrong—raising attacked accuracy—but does not substantially increase stability on clean-correct items. Consequently, the model improves in an *unconditional* sense (more predictions survive perturbation overall) without altering susceptibility on the specific examples where targeted prompt manipulations are most effective.

This pattern indicates that the evaluated adversarial prompts exploit deeper structural weaknesses in biomedical reasoning or instruction following, rather than surface-level linguistic vulnerabilities that additional supervision can easily mitigate. In practice, adversarial SFT strengthens general predictive reliability under distribution shift, but leaves intact a subset of inherently fragile biomedical cases on which both models fail in similar ways. These observations highlight an important methodological nuance: aggregate robustness gains can coexist with persistent conditional vulnerabilities that remain invisible unless ASR is examined directly.

6 Conclusion

Our results show that CoT + Reflexion produces more evidence-rich explanations but also incurs substantial penalties from standard readability metrics, which are highly sensitive to sentence segmentation and may not reliably capture true patient comprehensibility. Adversarial supervised fine-tuning improves both clean and attacked accuracy on PubMedQA, yet leaves the attack success rate effectively unchanged, indicating that robustness gains are largely *unconditional* rather than increased stability on already-correct cases. Taken together, these findings suggest that neither readability-focused prompting nor supervised fine-tuning alone resolves deeper vulnerabilities in biomedical reasoning and grounding. Advancing safety in clinical QA will require robustness-oriented training objectives and human-centered evaluation to ensure that model outputs are not only accurate, but also stable, interpretable, and clinically safe.

Administrative details

Project Github: `https://github.com/tienvu95/laycare_llm`

Contribution:

- Shuwei He: Idea and proposal initialization, adversarial training model implementation, report writing and editing.
- Vu Hoang: Literature review, set up pipieline for PubmedQA experiments (baseline, reflexion), report writing and editing.
- Lin Park: Implementation of MedredQA baseline and reflexion, video presentation recording, report writing and editing.

References

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy LLMs: Dealing with hallucinations in healthcare AI, 2023. URL https://arxiv.org/abs/2311.01463.

S. Aydin, M. Karabacak, V. Vlachos, and K. Margetis. Large language models in patient education: a scoping review of applications in medicine. *Frontiers in Medicine*, 11, 2024. doi: 10.3389/fmed.2024.1477898. URL https://doi.org/10.3389/fmed.2024.1477898.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. BioMedLM: A 2.7b parameter language model trained on biomedical text, 2024. URL https://arxiv.org/abs/2403.18421.

F. Busch, L. Hoffmann, C. Rueger, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5:26, January 2025. doi: 10.1038/s43856-024-00717-2. URL https://doi.org/10.1038/s43856-024-00717-2.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163, 2025. ISSN 2667-1026. doi: 10.1016/j.imed.2025.03.002.

Mahshad Koohi H. Dehkordi, Shuxin Zhou, Yehoshua Perl, Fadi P. Deek, Andrew J. Einstein, Gai Elhanan, Zhe He, and Hao Liu. Enhancing patient comprehension: An effective sequential prompting approach to simplifying EHRs using LLMs. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6370–6377. IEEE, 2024. doi: 10.1109/BIBM62325.2024.10822313.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.506.

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532.

Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and Steven Woloshin. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2):53–96, 2008. doi: 10.1111/j.1539-6053.2008.00033.x.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), January 2025. doi: 10.1145/3703155.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), March 2023. doi: 10.1145/3571730.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025, 2020. doi: 10.1609/aaai.v34i05.6311.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. 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 2567–2577, 2019.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report S6, Institute for Simulation and Training, University of Central Florida, 1975. URL https://stars.library.ucf.edu/istlibrary/56.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024. URL https://arxiv.org/abs/2402.10373.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. In *2019 Network and Distributed System Security Symposium (NDSS)*, 2019. doi: 10.14722/ndss.2019.23318.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memprompt: Memory-assisted prompt editing with user feedback. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2833–2861, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, NIPS '23, 2023.

G. Harry McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

Milad Moradi and Matthias Samwald. Improving the robustness and accuracy of biomedical language models through adversarial training. *Journal of Biomedical Informatics*, 132:104114, 2022. doi: 10.1016/j.jbi.2022.104114.

Ekaterina Mozhegova, Asad Masood Khattak, Adil Khan, Roman Garaev, Bader Rasheed, and Muhammad Shahid Anwar. Assessing the adversarial robustness of multimodal medical ai systems: insights into vulnerabilities and modality interactions. *Frontiers in Medicine*, 12, 2025. doi: 10.3389/fmed.2025.1606238.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.42. URL https://aclanthology.org/2023.ijcnlp-main.42/.

Don Nutbeam. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies. *Health Promotion International*, 15(3):259–267, 2000.

OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Ruth M. Parker, David W. Baker, Mark V. Williams, and Joanne R. Nurss. The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine*, 10(10):537–541, October 1995. doi: 10.1007/BF02640361. URL https://doi.org/10.1007/BF02640361.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Sarah J. Shoemaker, Michael S. Wolf, and Cindy Brach. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Education and Counseling*, 96(3):395–403, 2014. doi: 10.1016/j.pec.2014.05.027. URL https://doi.org/10.1016/j.pec.2014.05.027.

Karan Singhal, Shekoofeh Azizi, Tyna Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Lyndal J. Trevena, Brian J. Zikmund-Fisher, Adrian Edwards, Wolfgang Gaissmaier, Mirta Galesic, Paul K. J. Han, John King, Margaret L. Lawson, Susan K. Linder, Isaac Lipkus, Elissa Ozanne, Ellen Peters, Daniëlle Timmermans, and Steven Woloshin. Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, 13(Suppl 2):S7, 2013. doi: 10.1186/1472-6947-13-S2-S7.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Yifan Yang, Qiao Jin, Furong Huang, and Zhiyong Lu. Adversarial prompt and fine-tuning attacks threaten medical large language models. *Nature Communications*, 16:9011, 2025. doi: 10.1038/s41467-025-39011-4.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Jonah Zaretsky, Jeong Min Kim, Samuel Bashkaroun, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Network Open*, 7(3):e240357, March 2024. doi: 10.1001/jamanetworkopen.2024.0357. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2815615.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Self-taught reasoner bootstrapping reasoning with reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, March 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis (LAMPS '24)*, pages 57–68, 2024. doi: 10.1145/3689217.3690621.