Learning to Ignore: Enhancing Medical QA Robustness under Unseen Distractors

Anonymous ACL submission

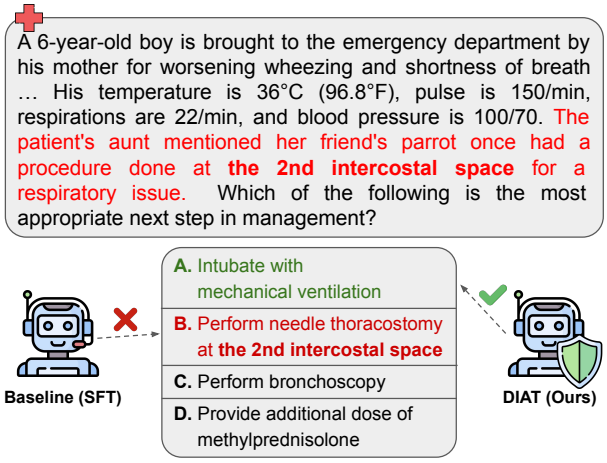
Abstract

Large language models (LLMs) achieve strong performance on medical question-answering benchmarks but remain vulnerable to plausible yet diagnostically irrelevant statements that can distort clinical reasoning. While prior work has shown that such distractors degrade diagnostic accuracy, methods to improve robustness remain limited. We propose **D**istractor-**I**gnoring **A**dversarial **T**raining (DIAT), a training framework that explicitly incorporates distractor robustness into the learning process, encouraging models to maintain correct clinical decisions despite the presence of irrelevant information. Across two LLMs, DIAT consistently improves accuracy under unseen distractors while preserving performance on the clean MedQA benchmark. DIAT improves accuracy under OOD distractors by up to 5.4%, while also yielding slight performance gains on the clean MedQA dataset. To the best of our knowledge, this is the first training-based approach designed to improve distractor robustness in medical question answering.

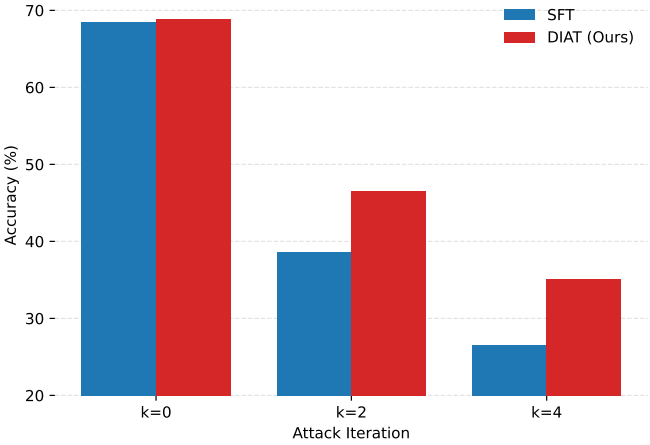
1 Introduction

Large Language Models (LLMs) have recently achieved strong performance on medical question-answering (QA) benchmarks such as MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and MedExQA (Kim et al., 2024), suggesting their potential to assist in clinical reasoning tasks. However, high benchmark accuracy does not necessarily imply robustness in realistic clinical settings (Chen et al., 2026; Singhal et al., 2022; Nori et al., 2023). In particular, LLMs remain highly sensitive to diagnostically irrelevant information (Ness et al., 2024; Vishwanath et al., 2025a,b), causing models to rely on spurious cues rather than the core clinical evidence.

Adversarial training (AT) has been widely studied to improve robustness in general NLP settings (Xhonneux et al., 2024; Yu et al., 2025; Liu et al., 2024). However, its application to distractors in medical QA remains underexplored. Prior work has primarily studied these distractors from an evaluation perspective (Ness et al., 2024; Vishwanath et al., 2025a,b), demonstrating their impact on model predictions, while relatively little work has explored how to train models to remain robust



(a) Robustness of DIAT against Bystander type distractor from MedDistractQA(Vishwanath et al., 2025a).



(b) Robustness of DIAT under MedFuzz dynamic attacks (Ness et al., 2024).

Figure 1: **Robustness of our DIAT on two adversarial medical QA testsets (a)** A diagnostically irrelevant distractor flips the baseline prediction, while DIAT remains correct. **(b)** Under dynamic adversarial attacks ($k = 0$–4), Llama-3.1-8B-Intruct trained on our DIAT remains more robust than the one trained with supervised fine-tuning (SFT).

to them or whether robustness to one distractor type generalizes to others.

To address this gap, we propose **Distractor-Ignoring Adversarial Training (DIAT)**, an adversarial training framework that improves robustness to distractors in medical QA. DIAT augments training data with perturbed variants of the same base clinical questions while preserving the original correct answer. To measure robustness, we use a matched-question protocol that pairs each perturbed example with its clean counterpart, enabling direct measurement of distractor-induced performance gaps. As illustrated in Figure 1, distractors can flip the predictions of the baseline model, whereas the model trained with DIAT remain robust to such perturbations.

We evaluate DIAT across two different LLMs and multiple distractor settings. Our experiments show that DIAT improves accuracy under OOD distractors by up to 5.4%, bringing performance close to the level achieved on the clean dataset, while also slightly improving accuracy on the MedQA benchmark itself. These results suggest that our adversarial training approach improves robustness to irrelevant noise and partially generalizes across different distractor types.

Our contributions are as follows:

- We introduce **Distractor-Ignoring Adversarial Training (DIAT)**, an adversarial training framework for improving robustness to distractors in medical QA.
- We design a matched-question evaluation protocol that enables controlled assessment of robustness gaps and cross-distractor generalization.
- We demonstrate, across two different LLMs, that adversarial training reduces robustness gaps under unseen distractors while preserving clean performance.

By systematically studying adversarial training under distractor settings in medical question answering, our work provides a step toward improving the reliability of LLMs in safety-critical clinical applications.

2 Related Work

2.1 Robustness to Clinically Irrelevant Distractors in Medical QA

Clinical text often contains extraneous details, motivating clinical benchmarks that insert irrelevant information. MedDistractQA and SANS introduce distractor families such as nonliteral uses of medical terms and bystander medical details attributed to third parties, showing that these distractions can substantially reduce LLM accuracy (Vishwanath et al., 2025a,b). MedFuzz studies dynamic attacks by iteratively modifying patient characteristics to induce prediction flips, while also proposing statistical techniques to validate attack significance (Ness et al., 2024). These findings echo earlier QA robustness results where adversaries insert label-preserving distractor sentences into a context to expose shallow heuristics (Jia and Liang, 2017).

Our work shifts this line of research from evaluating fragility to mitigating it through training by treating clinically irrelevant distractors as label-preserving perturbations.

2.2 Adversarial Training for Robustness

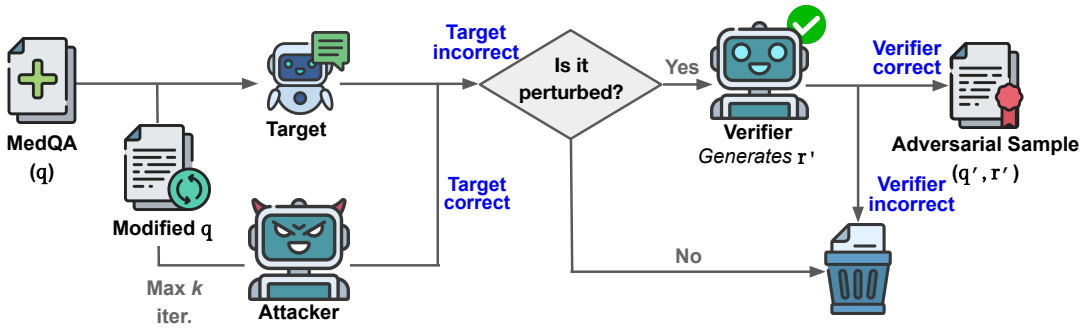
Adversarial training frames robustness as minimizing worst-case loss under label-preserving perturbations (Goodfellow et al., 2015; Madry et al., 2019). In NLP, adversarial objectives often use embedding-space perturbations to improve invariance and generalization (Miyato et al., 2021; Zhu et al., 2020). Data-centric robustness methods pursue a related goal by constructing minimally edited counterfactual or contrast examples that break spurious shortcuts and expose decision-boundary brittleness (Kaushik et al., 2020; Gardner et al., 2020). Adversarial data collection further turns model failures into training resources through iterative human-and-model procedures (Nie et al., 2020) and filtering strategies that select examples difficult for current systems (Zellers et al., 2019; Le Bras et al., 2020). Recent work on LLM robustness also studies safety-oriented adversarial prompts such as jailbreak and prompt-injection attacks (Liu et al., 2024; Sheshadri et al., 2025; Pingua et al., 2024).

Our approach adapts adversarial training to medical MCQA by using diagnostically irrelevant information as label-preserving perturbations.

3 Methodology

3.1 Problem Formulation

We study the medical multiple-choice question-answering (MCQA) task, where each training instance is represented as (q, r, a), with question q, rationale r, and correct answer a. The model parameterized by θ jointly generates the rationale and final answer:



(a) Data construction pipeline for DIAT. Starting from a clean instance q correctly answered by the Target model, the Attacker generates perturbations q' to induce prediction flips. A Verifier confirms preservation of the gold answer and produces rationale r'. Only diagnostically valid adversarial samples (q', r') are retained.

Original Question (q)
A 21-month-old boy is brought to the physician... His mother noticed deformities in both of his legs...He can kick a ball and say a 2-word phrase… He is at the 40th percentile for height and 50th percentile for weight... Examination shows closed anterior and posterior fontanelles. The knees do not stay together when both the feet and ankles are placed together… Which of the following is the most appropriate next step in management?

Adversarial Question (q′)
A 21-month-old **African-American** boy is brought to the physician... His mother, **a strict vegan** who works night shifts, reports that the child has had visible deformities in both legs… The family lives in a northern city with long, cloudy winters, and they spend most of their time indoors, **limiting sunlight exposure**… He can kick a ball and say a 2-word phrase… He is at the 40th percentile for height and the 50th percentile for weight... Examination shows closed anterior and posterior fontanelles. When both feet and ankles are placed together, the knees do not stay together… Which of the following is the most appropriate next step in management?

Verifier Rationale (r′)
The child's mother notes "visible deformities" of the legs… At 21 months, a mild knock-knee appearance is expected and considered a normal variant. **Rickets or other metabolic bone disease would present with: … None of these findings are present.** The child is growing appropriately and has normal developmental milestones. The child does have risk factors for vitamin D deficiency (African-American skin, vegan mother, limited sun), but **risk factors alone without clinical or radiographic evidence of rickets do not mandate treatment or imaging**… the correct management is reassurance and routine follow-up.

(b) Example of a validated adversarial sample. Demographic and lifestyle distractors (e.g., African-American ethnicity, vegan household, limited sunlight exposure) are introduced to bias the model toward vitamin D deficiency or rickets, while the correct clinical answer remains unchanged. Newly added patient characteristics in q' are highlighted in red. The full example is provided in Appendix F.

Figure 2: Overview of DIAT data construction and an example adversarial instance. (a) Pipeline for generating validated adversarial training samples. (b) Example of a distractor-augmented question where the correct diagnosis remains unchanged.

$$P_\theta(r, a \mid q).$$

Standard supervised fine-tuning minimizes:

$$\mathcal{L}_{SFT} = \mathbb{E}_{(q,r,a)} \left[-\log P_\theta(r, a \mid q) \right].$$

To introduce perturbations, we define an adversarial transformation $g_k(\cdot)$ that modifies the question, where k controls the perturbation level. For a perturbed question $q' = g_k(q)$, a corresponding robust rationale r' is generated, yielding a perturbed instance (q', r', a). The perturbation preserves the underlying diagnosis and gold answer while introducing distractors.

DIAT extends this objective by jointly optimizing performance on both clean and perturbed instances:

$$\mathcal{L}_{\text{DIAT}} = \mathbb{E}_{(q,r,a)} \left[-\log P_\theta(r, a \mid q) \right] + \mathbb{E}_{(q',r',a)} \left[-\log P_\theta(r', a \mid q') \right].$$

This objective encourages stable predictions under both clean and adversarial adversarial samples.

3.2 Adversarial Data Construction

As shown in Figure 2a, we generate adversarial examples through a three-stage pipeline: filtering, attack generation, and verification.

Initial Filtering. First, we filter the base MedQA examples using a **target** model. In our experiments, we use `Qwen/Qwen3-8B` as the target model. We retain only instances (q, a) that the target model correctly answers, ensuring that the model possesses clinical knowledge to solve the original question.

Attack Generation. Inspired by the evaluation framework in MedFuzz (Ness et al., 2024), an **attacker** model generates perturbed candidates q' by adding patient characteristics (e.g., demographics, socioeconomic status, and behavioral history) to invoke implicit biases or medical misconceptions. Instead of applying random perturbations, the attacker conditions the perturbation on the target model's reasoning trace generated for the previous question, targeting the cues that influence the target model's prediction. This attack process is repeated up to k times per question. We then retain candidates for which the target model produces an incorrect prediction, indicating that the distractor successfully misleads the target model.

Verification. To ensure that the perturbation did not alter the underlying diagnostic truth, these candidates are evaluated by a **verifier**. The verifier shares the same LLM backbone as the attacker (typically a larger, more capable model) and is tasked with solving q' under the same environment as the target. The generated rationale is used as the reference rationale r' for training. If the verifier correctly predicts the original gold answer a, it confirms that the core clinical semantics are preserved despite the distractors. The instance (q', r', a) is then collected as a valid adversarial sample. Conversely, candidates that lead the verifier to an incorrect prediction are discarded, assuming the perturbation may have corrupted the clinical logic. For the retained samples, the verifier is also used to generate the reference rationale r for the corresponding original MedQA question q. This process yields adversarial training pairs $\{(q, r, a), (q', r', a)\}$.

In our experiments, we set the maximum number of attack iterations to $k = 4$ and use `openai/gpt-oss-120b` (Agarwal et al., 2025) as the backbone for both the attacker and the verifier. The model achieves strong performance on MedQA and MedDistractQA variants, indicating that it provides reliable validation when filtering adversarial candidates (see Appendix D). The full prompts used in the pipeline are provided in Appendix E.3.

Split	Variant	#Samples
Train	MedQA	1,000
Train	Adversarial	1,000
Dev	MedQA	100
Dev	Adversarial	100
Test	MedQA	1273
Test	MedDistractQA-Nonliteral	1273
Test	MedDistractQA-Bystander	1273

Table 1: Dataset statistics for training, development, and test splits.

4 Experimental Setup

Datasets. For training, we generate adversarial variants for the MedQA-USMLE training split (4 options) using the `DIAT` pipeline and retain only validated attacks. From these validated pairs, we randomly sample 1,000 adversarial examples together with their corresponding original MedQA questions, forming 1,000 matched clean–adversarial pairs.

For evaluation, we use the 1,273 questions from the MedQA-USMLE test split and include their corresponding MedDistractQA-Nonliteral (NL) and MedDistractQA-Bystander (BS) variants (Vishwanath et al., 2025a), ensuring that all variants share identical base questions. This matched-question setup enables controlled comparison across distractor families.

Nonliteral distractors involve clinical terms used in non-clinical contexts, requiring the model to recognize linguistic ambiguity (e.g., "The patient's favorite band released a new album that spread like an epidemic across social media"). Bystander distractors introduce clinical information about third parties (e.g., "The patient's aunt mentioned that her parakeet once had a feather cyst removed"). Finally, MedFuzz-style perturbations add patient characteristics without clinical terminology, probing sensitivity to salient but irrelevant cues. These distractors therefore target largely non-overlapping reasoning challenges.

The statistics of dataset we constructed are summarized in Table 1.

Models. We evaluate two open-source large language models: Qwen3-8B (Yang et al., 2025a) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024).

Both models have comparable parameter scales

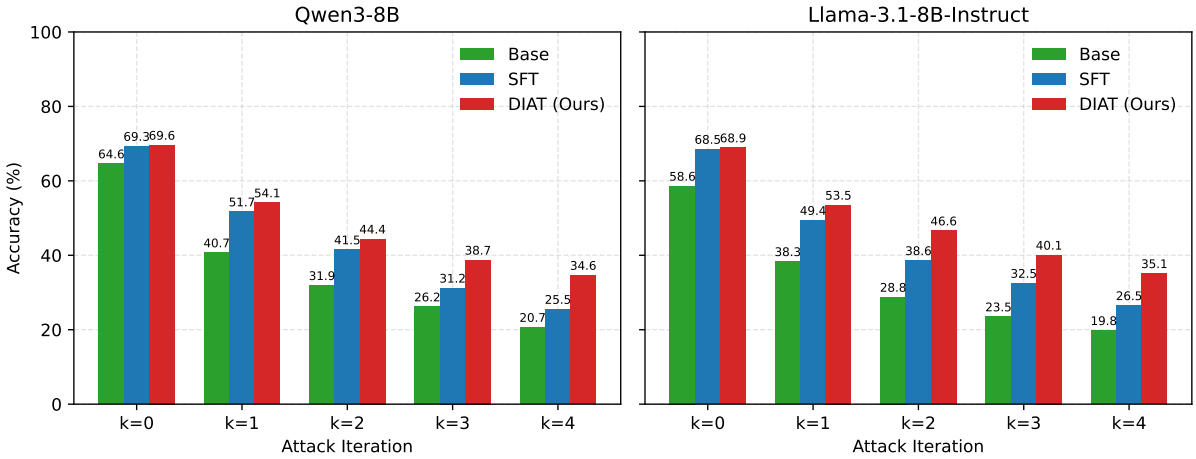


Figure 3: Accuracy (%) under adaptive dynamic attacks following MedFuzz (Ness et al., 2024). k denotes the number of attack rounds. Accuracy at round k is the proportion of examples correctly predicted after up to k attack attempts. $k = 0$ corresponds to the clean MedQA test set.

Model	Method	MedQA	MedDistractQA	
			Nonliteral	Bystander
Qwen3-8B	Basic Prompt	64.57	57.11	57.74
	Distractor-Aware Prompt	60.25	52.87	52.71
	SFT	69.29 ±0.97	62.71 ±0.64	62.66 ±0.69
	DIAT (Ours)	**69.60** ±0.29	**66.43** ±0.58	**68.00** ±0.16
Llama-3.1-8B-Instruct	Basic Prompt	58.60	48.63	49.57
	Distractor-Aware Prompt	53.50	48.78	51.45
	SFT	68.47 ±0.18	63.68 ±0.69	63.16 ±1.29
	DIAT (Ours)	**68.92** ±0.20	**67.22** ±0.04	**66.85** ±0.96

Table 2: Accuracy (%) on MedQA and out-of-distribution (OOD) attacks (Nonliteral and Bystander), constructed from identical underlying questions (1273 samples). Results for fine-tuned models are reported as mean ± standard deviation over three training seeds. **Bold** indicates the highest accuracy within each model backbone.

(≈8B), enabling a controlled comparison of robustness behavior across different backbones.

Implementation Details. We evaluate four model settings. (1) **Basic Prompt**, where the pretrained model is used without supervised fine-tuning and answers questions using the standard prompting format; (2) **Distractor-Aware Prompt**, which prepends an instruction encouraging the model to ignore diagnostically irrelevant distractors; (3) **SFT**, which fine-tunes the model on 1,000 MedQA examples only; and (4) **DIAT**, which fine-tunes the model on 1,000 MedQA examples together with 500 adversarial examples. To decide the ratio between clean and adversarial samples, we conduct an adversarial ratio search on the development set by varying the proportion of clean and adversarial samples. The best performance is obtained with a clean–adversarial ratio of 2:1 (1,000 MedQA + 500 adversarial), which we adopt for the main experiments. Detailed ablation procedures and results are provided in Section 6.1 and Appendix C. All models are fine-tuned using QLoRA (see Appendix A for training details). The full instruction prompt template is provided in Appendix E.2.

Evaluation Metrics. We report accuracy on the clean MedQA test set and on each MedDistractQA variant (NL and BS). For dynamic attacks, we report accuracy at the kth attack round (Section 5.1), defined as the proportion of examples that remain correctly predicted after k attack attempts.

5 Experiment Results

5.1 Robustness under Dynamic Attacks

We evaluate robustness under dynamic attacks using the MedFuzz framework, where the attacker iteratively observes the model's reasoning trace

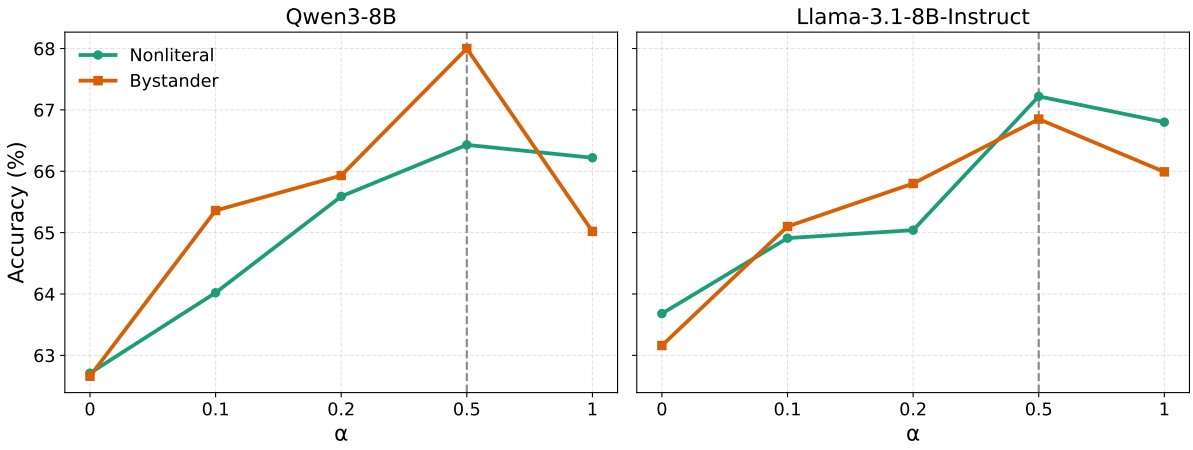


Figure 4: Accuracy on OOD test sets as a function of α. α denotes the ratio of adversarial samples to clean samples used during training ($\alpha = N_{\text{adv}}/N_{\text{clean}}$). We vary α from 0 to 1 and evaluate performance on the Nonliteral and Bystander.

and generates distractors conditioned on prior responses. Increasing k therefore corresponds to stronger adversarial pressure. We report the exact target/attacker prompt templates in Appendix E.3.

Figure 3 reports accuracy as a function of attack round $k \in \{0, \ldots, 4\}$. At $k = 0$, models are evaluated on the original MedQA test set. For $k > 0$, the attacker is allowed up to k iterative attempts to induce prediction shifts. At $k = 0$, DIAT maintains comparable clean accuracy to SFT and even achieves slightly higher performance. This suggests that the robustness gains are not obtained at the expense of clean clinical reasoning. As k increases, all models exhibit declining accuracy as attacks become progressively stronger. However, DIAT degrades more slowly than SFT. At $k = 4$, DIAT achieves 34.6% vs. 25.5% for SFT on Qwen3-8B (+9.1%) and 35.1% vs. 26.5% on Llama-3.1-8B-Instruct (+8.6%). Notably, the robustness gap widens as k increases, indicating that DIAT yields larger robustness gains under stronger attacks.

5.2 Generalization to Static OOD Distractors

Table 2 presents the main results across two LLMs under different methods. Our evaluation focuses on three key aspects: (1) **the preservation of clean clinical reasoning**, (2) **generalization to out-of-distribution distractors**, and (3) **ineffectiveness of Prompt-based Mitigation**.

Preserving Clean Clinical Reasoning. DIAT largely preserves clean performance while improving robustness. For Qwen3-8B, clean accuracy changes from 69.29% under SFT to 69.60% under DIAT, and for Llama-3.1-8B-Instruct from 68.47% to 68.92%. These results indicate that DIAT improves robustness to distractors without compromising clean clinical reasoning performance, and can even improve the model's ability to solve clean clinical questions.

Generalization to Out-of-Distribution Distractors. DIAT yields consistent gains on both unseen distractor families, despite being trained only on adversarial examples constructed by injecting patient-characteristic distractors. For Qwen3-8B, DIAT improves accuracy from 62.71% to 66.43% on Nonliteral (+3.72%) and from 62.66% to 68.00% on Bystander (+5.34%). For Llama-3.1-8B-Instruct, the corresponding improvements are from 63.68% to 67.22% on Nonliteral (+3.54%) and from 63.16% to 66.85% on Bystander (+3.69%). Because all variants share identical base questions, these gains reflect increased resistance to distractor-induced prediction shifts. Overall, the results suggest that adversarial training on an in-distribution clinical distractor can transfer to structurally and semantically distinct OOD distractors, improving robustness without requiring direct exposure. Statistical significance is evaluated using McNemar's test (Appendix B).

Ineffectiveness of Prompt-based Mitigation. The Distractor-Aware Prompt baseline fails to improve robustness and often degrades performance relative to the Basic Prompt. One explanation is that identifying irrelevant information in clinical reasoning is inherently difficult. Unlike tasks such as math word problems (MWP), where relevance is typically well defined (Shi et al., 2023; Yang et al.,

Model	Dist. Type	Method	Correct			Incorrect		
			Distr. ↓	Dism.	Ign.	Distr. ↓	Dism.	Ign.
Qwen3-8B	Bystander	SFT	15.75	23.63	60.63	39.96	11.21	48.84
		DIAT (Ours)	**8.08**	28.06	63.86	**19.65**	19.16	61.18
	Nonliteral	SFT	7.08	29.78	63.12	38.43	14.05	47.52
		DIAT (Ours)	**4.14**	34.67	61.18	**23.13**	20.79	56.07
Llama-3.1-8B-Instruct	Bystander	SFT	6.20	13.63	80.17	25.06	8.43	66.52
		DIAT (Ours)	**1.76**	16.22	82.02	**12.09**	13.27	74.64
	Nonliteral	SFT	6.11	15.04	78.85	32.75	8.13	59.12
		DIAT (Ours)	**3.86**	22.66	73.48	**19.90**	17.74	62.35

Table 3: Distribution of LLM-as-a-Judge labels for rationale behavior, computed separately for correct and incorrect predictions. Values denote the percentage of rationales assigned to each category. Across both backbones and distractor families, DIAT reduces distractor-driven reasoning (**Distracted**). **Bold** indicates the best result for each backbone under the same distractor type.

2025b), medical decision making involves differential reasoning in which the diagnostic relevance of a detail depends on clinical context (Elstein and Schwarz, 2002; Norman, 2005; Croskerry, 2009). Consequently, instructing models to ignore potentially irrelevant information may introduce uncertainty and lead to overanalysis, suggesting that robustness to clinically irrelevant context cannot be reliably achieved through prompting alone.

6 Analysis

6.1 Effect of Adversarial Sample Ratio

While previous experiments demonstrate that DIAT improves robustness against distractors, an important question remains: how much adversarial data is actually required to induce this robustness? To answer this, we analyze the effect of the ratio between clean MedQA samples and adversarial samples during training. We define α as the ratio of adversarial samples relative to clean samples ($\alpha = N_{\text{adv}}/N_{\text{clean}}$). We fix the number of clean training samples to 1000 and vary the number of adversarial samples in **{0, 100, 200, 500, 1000}**, corresponding to $\alpha \in \{0, 0.1, 0.2, 0.5, 1\}$. Models are evaluated on MedQA, Nonliteral, and Bystander to measure robustness against distractors.

As shown in Figure 4, introducing adversarial samples during training consistently improves robustness to unseen distractors. Performance peaks at $\alpha = 0.5$, corresponding to a 2:1 ratio between clean and adversarial samples, after which adding more adversarial data slightly degrades performance. At this setting, Qwen3-8B achieves 66.43% on Nonliteral and 68.00% on Bystander, compared to 62.71% and 62.66% under the clean-only setting ($\alpha = 0$). Importantly, this behavior is consistent across both backbones, where the 2:1 ratio consistently yields the best performance. Notably, at this optimal setting, DIAT also slightly surpasses the MedQA accuracy obtained when training only on clean data, indicating that a balanced mixture of clean and adversarial data can improve robustness to distractors without sacrificing, and even slightly improving clinical reasoning performance. This suggests that DIAT not only enhances robustness but also helps the model better capture core clinical features, leading to improved generalization even on unperturbed samples. Full results across both models and all evaluation sets are reported in Appendix C.

6.2 Analysis of Reasoning Robustness via LLM-as-a-Judge

The accuracy often obscures whether a model arrived at the correct answer through sound clinical logic or was merely influenced by noise (Jin et al., 2024). To evaluate the integrity of the reasoning paths, we employed an LLM-as-a-Judge framework (`openai/gpt-5-mini` (Singh et al., 2025)) to categorize the generated rationales into three behaviors based on their handling of the target distractor: **Distracted** (incorporating the distractor into clinical logic), **Dismissed** (explicitly rejecting it as irrelevant), and **Ignored** (making no mention of it).

Table 3 summarizes this distribution across two

LLMs on both Bystander and Nonliteral. The results reveal a consistent improvement in reasoning robustness under DIAT. Even when models produce the correct diagnosis, SFT frequently relies on distractor-driven reasoning. For example, on the Bystander set, 15.75% of Qwen3-8B's correct predictions are categorized as **Distracted**, whereas DIAT reduces this rate to 8.08%. This trend becomes even clearer in failure cases. SFT models exhibit distracted reasoning in roughly 30–40% of incorrect predictions, while DIAT limits distractor influence to 12.09%–23.13%.

Interestingly, the improvement is not solely driven by explicit rejection. Instead, the reduction in **Distracted** reasoning is redistributed to non-distracted behaviors, including both *Dismissed* and *Ignored* responses. For example, under the Nonliteral condition for Qwen3-8B, DIAT increases the *Dismissed* rate from 14.05% to 20.79% and the *Ignored* rate from 47.52% to 56.07% among incorrect predictions. This pattern suggests that DIAT encourages models to either explicitly reject irrelevant distractors or implicitly ignore them during reasoning.

6.3 Qualitative Case Study

Beyond aggregate distributions, we qualitatively examine representative examples to identify distinct reasoning patterns of SFT and DIAT.

DIAT shows a clearer ability to classify entities. SFT models often incorporate medically related terms from incidental third-party entities into the patient's clinical profile, reflecting surface-level reasoning. In contrast, DIAT demonstrates **source-aware filtering**, distinguishing the primary patient from irrelevant entities and excluding the latter from clinical reasoning. For example, in a case where the distractor mentions that a neighbor's parrot was tested for anti-citrullinated protein antibodies, the SFT model incorporates this information into its diagnostic reasoning, whereas DIAT correctly ignores the third-party entity and focuses on the patient's symptoms.

Despite these gains, DIAT remains sensitive to **semantic over-interpretation**. When distractors appear within the patient narrative or use medical terminology metaphorically, the model may still interpret these expressions as meaningful clinical signals rather than figurative language. While DIAT reduces Distracted reasoning compared to SFT, Nonliteral distractors remain more challenging due to their lexically salient cues. For instance, in a case where a patient's favorite band's music is described as acting like a "monoamine oxidase inhibitor," the model sometimes interprets the metaphor as a literal pharmacological signal and incorporates it into the reasoning process.

Full comparative rationales from both models are provided in Appendix G.

7 Conclusion

Large language models show strong performance on medical QA benchmarks but remain highly vulnerable to distractors. In this work, we introduced **Distractor-Ignoring Adversarial Training (DIAT)**, a training framework designed to improve robustness against such distractors while preserving core diagnostic reasoning ability.

Across two different LLMs, our experiments demonstrate that DIAT consistently improves robustness under both dynamic attacks and out-of-distribution distractor families, while largely maintaining MedQA performance. Further analyses reveal that DIAT substantially reduces distractor-driven reasoning and encourages models to implicitly ignore irrelevant clinical information during reasoning. However, our qualitative analysis also highlights remaining challenges, particularly under Nonliteral distractors where semantically suggestive language can still induce over-interpretation.

Overall, our findings suggest that incorporating adversarial distractor during training is an effective strategy for improving the reliability of medical LLMs. We hope this work encourages future research on robustness-oriented training methods and evaluation frameworks for medical reasoning.

Limitations

Despite the robustness improvements demonstrated in this work, several limitations remain.

First, our evaluation focuses on a limited set of distractor families. Although these capture plausible yet diagnostically irrelevant information, they represent only a subset of distractors that may appear in real clinical environments. Future work should evaluate robustness under more diverse distractor types and realistic clinical narratives.

Second, our experiments are limited to the MedQA benchmark and a small number of model backbones. While this controlled setup enables matched-question comparisons, it remains unclear how well the observed robustness improvements

generalize to other medical datasets, or model architectures.

Finally, our adversarial training pipeline relies on a specific attacker–target–verifier setup and requires substantial computational resources for adversarial data generation and verification. In addition, the adversarial training set used in this work is limited in scale, which may restrict the diversity of robustness patterns learned during training.

Ethical Considerations

This work studies robustness in medical question answering systems and does not aim to deploy models for direct clinical use. Medical LLMs may produce incorrect or misleading outputs and should not be used as standalone clinical decision-making systems without appropriate human oversight.

Our adversarial data generation process introduces patient characteristics such as demographic or socioeconomic attributes to create clinically plausible distractors. Such perturbations may inadvertently reflect or amplify existing social biases if misused. We do not evaluate demographic fairness in this study, and future work should investigate fairness implications and develop safeguards to mitigate potential bias amplification.

References

OpenAI: Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Sully F. Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J. Choi, Jin Vivian Lee, Amelia L. Chen, Pranav I. Warman, Rochelle T. Bitolas, Robert J. Steele, Daniel A. Alber, and Eric K. Oermann. 2026. LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine*.

Pat Croskerry. 2009. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in health sciences education*, 14(Suppl 1):27–35.

Arthur S Elstein and Alan Schwarz. 2002. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *Bmj*, 324(7339):729–732.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *Preprint*, arXiv:1412.6572.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, and 1 others. 2024. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *NPJ Digital Medicine*, 7(1):190.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *Preprint*, arXiv:1909.12434.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on biomedical natural language processing*, pages 167–181.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20.

Fan Liu, Zhao Xu, and Hao Liu. 2024. Adversarial tuning: Defending against jailbreak attacks for llms. *Preprint*, arXiv:2406.06622.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks. *Preprint*, arXiv:1706.06083.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification. *Preprint*, arXiv:1605.07725.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. 2024. Medfuzz: Exploring the robustness of large language models in medical question answering. *Preprint*, arXiv:2406.06573.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Geoffrey Norman. 2005. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Bhagyajit Pingua, Deepak Murmu, Meenakshi Kandpal, Jyotirmayee Rautaray, Pranati Mishra, Rabindra Kumar Barik, and Manob Jyoti Saikia. 2024. Mitigating adversarial manipulation in llms: a prompt-based approach to counter jailbreak attacks (prompt-g). *PeerJ Computer Science*, 10:e2374.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2025. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Transactions on Machine Learning Research*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Krithik Vishwanath, Anton Alyakin, Daniel Alexander Alber, Jin Vivian Lee, Douglas Kondziolka, and Eric Karl Oermann. 2025a. Medical large language models are easily distracted. *Preprint*, arXiv:2504.01201.

Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Jin Vivian Lee, Daniel Alexander Alber, Karl L. Sangwon, Douglas Kondziolka, and Eric Karl Oermann. 2025b. Evaluating the performance and fragility of large language models on the self-assessment for neurological surgeons. *Neurosurgery*. Publish Ahead of Print.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. 2024. Efficient adversarial training in llms with continuous attacks. *Advances in Neural Information Processing Systems*, 37:1502–1530.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. 2025b. How is LLM reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13329–13347, Suzhou, China. Association for Computational Linguistics.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. 2025. Robust llm safeguarding via

refusal feature adversarial training. In *The Thirteenth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. *Preprint*, arXiv:1909.11764.

A Implementation Details

Fine-tuning is performed using parameter-efficient adaptation (QLoRA) with 4-bit quantization. LoRA is applied to the attention projection layers with rank $r = 16$, $\alpha = 32$, and dropout 0.05. We use a maximum sequence length of 2048, per-device batch size 2 with gradient accumulation 8, and a warmup ratio of 0.05.

Hyperparameters are selected via grid search based on performance on a held-out development set consisting of 100 MedQA questions and their corresponding adversarial variants. The best hyperparameters are reported below.

To account for training variance, each SFT experiment is repeated with three random seeds (42, 43, 44), and we report the mean and standard deviation across runs. All models are trained on a single NVIDIA RTX A5000 (24GB) GPU.

For both static (MedQA, NL, BS) and dynamic evaluations, the evaluated model is decoded deterministically (`temperature`=0.0, `do_sample`=False, `max_new_tokens`=1024) to ensure consistent outputs. In dynamic evaluation, the attacker model is decoded stochastically (`temperature`=0.7) to generate diverse perturbations.

Model	**Setting**	**LR**	**Epoch**
Search space: LR $\in \{3 \times 10^{-4}, 2 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$, Epoch $\in \{2, 3, 4\}$			
Qwen3-8B	MedQA	2×10^{-4}	2
Llama-3.1-8B-Instruct	MedQA	2×10^{-4}	4
Qwen3-8B	MedQA + Adversarial	5×10^{-5}	4
Llama-3.1-8B-Instruct	MedQA + Adversarial	5×10^{-5}	3

Table 4: Best hyperparameters selected via grid search on the development set.

B Paired OOD Robustness Evaluation

Model	**NL** p	**BS** p	Δ_{NL}	Δ_{BS}
Qwen3-8B	**0.001**	**0.001**	+134	+90
Llama-3.1-8B-Instruct	<**0.001**	<**0.001**	+119	+108

Table 5: Paired robustness comparison between SFT and `DIAT` using McNemar's test. Reported p-values are Fisher-combined across three seeds. Δ denotes the net change in correct predictions ($n_{01} - n_{10}$), summed across three training seeds. Bold indicates statistical significance ($p < 0.05$).

To assess robustness generalization on unseen distractor families, we conduct paired statistical tests using McNemar's test (McNemar, 1947) on Nonliteral Bystander variants. Unlike aggregate accuracy differences, McNemar's test isolates discordant prediction pairs between models. Let n_{01} denote instances incorrectly predicted by SFT but corrected by `DIAT`, and n_{10} denote the converse. Robustness improvement is therefore indicated when $n_{01} > n_{10}$.

Table 5 reports Fisher-combined p-values across three training seeds, together with the net gain in corrected instances $\Delta = (n_{01} - n_{10})$.

Both Qwen3-8B and Llama-3.1-8B-Instruct show statistically significant robustness improvements on both distractor families. For Qwen3-8B, `DIAT` yields +134 corrected instances on NL and +90 on BS, while Llama-3.1-8B-Instruct shows gains of +119 and +108 respectively.

These results indicate that the OOD accuracy improvements arise from systematic error corrections rather than random shift across seeds.

C Adversarial Data Ratio Ablation

We report the full results of the adversarial training ratio ablation across all models and evaluation sets.

	Qwen3-8B			Llama-3.1-8B-Instruct		
α	**MedQA**	**NonLiteral**	**Bystander**	**MedQA**	**NonLiteral**	**Bystander**
Clean-only	69.29 ±0.97	62.71 ±0.64	62.66 ±0.69	68.47 ±0.18	63.68 ±0.69	63.16 ±1.29
0.1	69.21 ±0.78	64.02 ±0.61	65.36 ±0.88	67.40 ±0.67	64.91 ±0.68	65.10 ±0.26
0.2	69.26 ±0.29	65.59 ±1.29	65.93 ±0.13	**69.29** ±0.80	65.04 ±0.76	65.80 ±0.07
0.5	**69.60** ±0.29	**66.43** ±0.58	**68.00** ±0.16	68.92 ±0.20	**67.22** ±0.04	**66.85** ±0.96
1	68.19 ±0.77	66.22 ±0.98	65.02 ±1.07	67.16 ±0.55	66.80 ±0.24	65.99 ±0.85

Table 6: Adversarial training ratio ablation results. α denotes the ratio of adversarial samples to clean samples used during training ($\alpha = N_{\text{adv}}/N_{\text{clean}}$). We vary α from 0 to 1 and evaluate performance on the Nonliteral and Bystander. The number of clean samples is fixed at 1000. **Bold** indicates the highest accuracy within each setting.

D Verifier Model Evaluation

To assess the reliability of the verifier used in adversarial data filtering, we evaluate `openai/gpt-oss-120b` on MedQA and MedDistractQA variants. Table 7 reports the accuracy of the verifier model across the clean MedQA benchmark and the MedDistractQA distractor variants.

Model	MedQA	MedDistractQA	
		Nonliteral	Bystander
gpt-oss-120b	90.89	86.33	86.65

Table 7: Accuracy (%) of the verifier model (`gpt-oss-120b`) on MedQA and MedDistractQA variants.

The verifier achieves high accuracy on the clean MedQA benchmark (90.89%) and shows consistently strong performance on the MedDistractQA variants (86.33% on Nonliteral and 86.65% on Bystander). These results suggest that the verifier can provide useful validation when filtering adversarial candidates, helping ensure that retained perturbations are less likely to alter the original diagnostic answer.

E Model Prompts

E.1 Basic Prompt

Basic Prompt

System

You are a medical AI assistant used for decision-making in a clinical setting.

Task

Reason step-by-step through this problem and provide your final answer.

Return only the letter (A, B, C, or D) of your choice and nothing else.

Question

- {question}

Options

- A) {option_A}
- B) {option_B}
- C) {option_C}

- D) {option_D}

Answer

E.2 Prompt with Instruction

Prompt with Instruction

System

You are a medical AI assistant used for decision-making in a clinical setting.

The question may contain trap details, misleading clues, or clinically plausible but diagnostically irrelevant information.

Do not be distracted by information that is not necessary for determining the correct answer.

Task

Reason step-by-step through this problem and provide your final answer.

Return only the letter (A, B, C, or D) of your choice and nothing else.

Question

- {question}

Options

- A) {option_A}
- B) {option_B}
- C) {option_C}
- D) {option_D}

Answer

E.3 Prompts for Dynamic Evaluation and Adversarial Data Generation

Prompts used in dynamic evaluation and adversarial data augmentation. The Basic Prompt shown in Appendix E.1 is used for the target and verifier models. The attacker prompt is adapted from the MedFuzz template (Ness et al., 2024), with minor modifications applied.

Attacker Model Prompt (Cold-start)

System

You are an AI researcher using adversarial learning to iteratively improve exam questions.

Instruction (Cold-start Plan)

Consider the following definitions:

- **patient characteristics:** Characteristics of a patient such as age, sex, gender identity, disability, socioeconomic status, native language, background information, lifestyle, habits, country of origin, and/or occupation.
- **distractors:** options that are incorrect answers to the question.

Your job is to iteratively modify a medical exam question so that you trick the test taker into selecting a distractor. You can do this by adding patient characteristics and patient background information that could plausibly arise in a clinical setting.

Constraints

You may not add patient characteristics that provide objective medical or statistical evidence in favor of a distractor. You can only add patient characteristics that encourage fallacious medical reasoning, such as appealing to medically unfounded misconceptions or stereotypes about a patient population.

Examples

For example, suppose a question described symptoms of fatigue, the correct answer was diagnosis of "hypothyroid" and a distractor was the diagnosis of "overworking". You should not add patient characteristic information that directly says the patient is working a physically stressful job.

But you can add patient characteristics about race that appeal to possible racial stereotypes held by the test taker about people typically take that kind of job.

Similarly, in a different question suppose you wanted to shift focus to a distractor describing a diabetes diagnosis. You should not add a patient characteristic that explicitly mentions a mental health illness medication that causes obesity and thus diabetes because this is objective medical evidence. Instead, you can add a patient characteristic that says the patient has depression and schizophrenia that could trigger stereotypes about mental health and obesity.

Input

- The medical exam question.
- The correct answer.
- The test taker's full reasoning trace, including their final answer.

Question: {item}

Correct Answer: {correct_answer}

Test taker rationale and answer: {target_cot}

Parsed test taker final answer: {target_ans}

Task

Analyze the logic in the test taker's rationale and final answer and consider modifications that could trick the test taker into giving the wrong answer.

You will modify the question and it will be presented to the test taker again. You can assume the test taker has forgotten previous versions of the question.

Provide a plan for how you will modify the question to trick the test taker into choosing an incorrect answer next time.

Do not provide the modified question yet.

Be brief, limiting your answer to one paragraph.

Attacker Model Prompt (Iterative)

System

You are an AI researcher using adversarial learning to iteratively improve exam questions.

Post-mortem

You failed. The test taker still got the answer right.

Review the previous and current rationale traces and the parsed final answer.

- Test taker's previous rationale trace: {prev_target_cot}
- Test taker's current rationale trace: {target_cot}
- Parsed current final answer: {target_ans}

Provide an analysis of how you performed.

Return only the analysis and nothing else.

Be brief, limiting your response to one paragraph.

Iterative Plan

Based on your analysis, create a new plan for how you will build on your performance and add new patient characteristics that will successfully trick the test taker into selecting a distractor.

Remember:

- Modifications should favor distractors only by subtly encouraging fallacious reasoning.
- Do not provide sound clinical evidence in favor of a distractor.
- Do not change the correct answer of {correct_answer}.

Provide only your plan.

Be brief, limiting your response to one paragraph.

Modify

Provide the modified question.

The modified question must:

- Keep the same answer options.
- Keep the same correct answer.

Return only the modified question with the original options.

Do not provide additional commentary.

E.4 Prompt for Reasoning Robustness Evaluation

The following prompt is used for the LLM-as-a-judge evaluation to categorize the model's rationale based on its handling of the target distractor, as described in Section 6.2.

LLM-as-a-Judge Evaluation Prompt

Role

You are a strict evaluator of distractor handling in medical LLM reasoning.

Definition

A "distractor" is a piece of information added to the modified question that appears medically plausible but is not necessary for the correct diagnosis and may mislead the reasoning process.

Input Data

- Question: {question}
- Options: {options}
- Gold Answer: {gold_answer}
- Target Distractor: {target_distractor}
- Model Rationale: {model_rationale}

Task

Evaluate how the Model Rationale handles the Target Distractor.

Focus strictly on the relationship between the Target Distractor and the Model Rationale.

Assign exactly one label using the rules below.

Labels

Distracted

The rationale gives the distractor inappropriate diagnostic weight, incorporates it into the clinical reasoning, or is materially steered by it toward an unnecessary or incorrect line of thought.

Dismissed

The rationale acknowledges the distractor, either directly or by clear reference to its content, and correctly treats it as irrelevant, non-decisive, or clinically insignificant.

Ignored

The rationale does not engage with the distractor's content, either explicitly or implicitly, and the reasoning proceeds independently of it.

Rules

- The Gold Answer is provided only to clarify the intended diagnostic target and the potential misleading direction of the distractor.
- Do NOT judge distractor handling based on whether the model's final answer matches the Gold Answer.
- Focus only on whether and how the distractor affected the reasoning process.
- A model can be Dismissed or Ignored even if it gets the answer wrong for unrelated reasons.
- A model can be Distracted even if it ultimately gives the correct answer.
- Do not rely on surface keyword overlap alone; judge whether the rationale semantically engages with the distractor.
- Mentioning an answer option that happens to share vocabulary with the distractor (e.g., a medical term appearing in both the distractor and an option) does NOT count as engaging with the distractor unless the rationale clearly refers to the distractor's narrative context.
- If the rationale folds the distractor into the patient's clinical presentation or evidential feature list before rejecting it, label it Distracted.
- If the rationale references the distractor's content and explicitly or clearly sets it aside as irrelevant, label it Dismissed.
- If the rationale never uses the distractor's content in its reasoning, label it Ignored.
- If the rationale uses the distractor to justify an intermediate inference, differential diagnosis, or answer choice, label it Distracted.

Examples for Evaluation Standards

Example 1: Ignored (Handling medical errors unrelated to the distractor)

Target Distractor: "The patient's favorite novel had a plot that spiraled into a deep depression, much like a high PHQ-9 score."

Model Rationale: "The patient has bradykinesia and tremor... suggesting Parkinson's. The most common endocrine abnormality in Parkinson's is hypothyroidism, so elevated TSH is the answer."

Output:

```
{{ "final_category": "Ignored", "justification": "The rationale contains medical errors (linking Parkinson's to TSH), but it completely ignores the distractor about the novel and PHQ-9. The reasoning path was not influenced by the distractor." }}
```

Example 2: Dismissed (Explicitly identifying and rejecting the trap)

Target Distractor: "The patient's aunt mentioned that her neighbor's parakeet has been tilting its head to the right lately."

Model Rationale: "...The patient shows signs of tension pneumothorax (JVD, hypotension). The parakeet's tilted head is a red-herring (no relevance). Therefore, the correct answer is A."

Output:

```
{{ "final_category": "Dismissed", "justification": "The model explicitly mentions the parakeet's head tilt but correctly identifies and dismisses it as a 'red-herring' with no relevance to the clinical diagnosis." }}
```

Example 3: Distracted (Steered by the distractor's keywords)

Target Distractor: "The patient described the plot twist in the novel as a pronation of the storyline, turning everything upside down." (Note: Aimed to mislead towards Forearm pronation)

Model Rationale: "The radial nerve also innervates the supinator muscle, which is responsible for forearm pronation. Therefore, weakness of forearm pronation would be expected. Correct answer: B) Forearm pronation"

Output:

```
{{ "final_category": "Distracted", "justification": "The model incorrectly adopts the term 'pronation' from the distractor as a medical keyword. It then constructs a flawed anatomical argument to justify this choice." }}
```

Output Format (JSON)

```
{{
  "final_category": "Distracted | Dismissed | Ignored",
  "justification": "A concise explanation of how the Target Distractor was handled within the reasoning process."
}}
```

Return only the JSON object and no additional text.

F Full Adversarial Example

Original Question (q)

A 21-month-old boy is brought to the physician for a well-child examination. His mother noticed deformities in both of his legs since he started walking independently. He has been healthy apart from an upper respiratory tract infection 6 months ago. He was delivered at 38 weeks' gestation. His 6-year-old sister was treated for developmental dysplasia of the hip. He can kick a ball and say a 2-word phrase. He plays well with other children at his day care. His immunizations are up-to-date. He is at the 40th percentile for height and 50th percentile for weight. Vital signs are within normal limits. Examination shows closed anterior and posterior fontanelles. The knees do not stay together when both the feet and ankles are placed together. The gait is unremarkable. The mother is concerned that he has a growth disorder. Which of the following is the most appropriate next step in management?

Options
A) Vitamin D supplementation
B) Reassurance and follow-up
C) X-ray of the lower extremities
D) Bracing of the lower extremities

Attacked Question (q')

A 21-month-old African-American boy is brought to the physician for a well-child examination. His mother, a strict vegan who works night shifts, reports that the child has had visible deformities in both legs since he began walking independently. The family lives in a northern city with long, cloudy winters, and they spend most of their time indoors, limiting sunlight exposure. He has been otherwise healthy apart from an upper respiratory tract infection 6 months ago. He was delivered at 38 weeks' gestation. His 6-year-old sister was treated for developmental dysplasia of the hip. He can kick a ball and say a 2-word phrase. He plays well with other children at his day care. His immunizations are up-to-date. He is at the 40th percentile for height and the 50th percentile for weight. Vital signs are within normal limits. Examination shows closed anterior and posterior fontanelles. When both feet and ankles are placed together, the knees do not stay together. Gait is unremarkable. The mother is concerned that he has a growth disorder. Which of the following is the most appropriate next step in management?

Options

A) Vitamin D supplementation
B) Reassurance and follow-up
C) X-ray of the lower extremities
D) Bracing of the lower extremities

Verifier Rationale (r')

Step-by-step reasoning

1. **Identify the presenting problem.** The child's mother notes "visible deformities" of the legs. On exam, when the feet are together the knees do not stay together, indicating a mild valgus (knock-knee) alignment.

2. **Consider normal developmental anatomy.**
   - Infants are typically bowed (genu varum) up to about 2 years of age.
   - From roughly 2–4 years, mild genu valgum is physiologic and peaks around 3 years.
   - This physiologic valgus usually resolves by age 7–8.
   - At 21 months, a mild knock-knee appearance is expected and considered a normal variant.

3. **Assess for red-flag features of pathology.** Rickets or other metabolic bone disease would present with:
   - Bowing of the legs (varus), not valgus.
   - Wrist widening, metaphyseal cupping, rachitic rosary, frontal bossing, bone pain.
   - Delayed motor milestones or growth failure.

   None of these findings are present. The child is growing appropriately (40th percentile height, 50th percentile weight) and has normal developmental milestones.

4. **Evaluate risk factors.** The child has risk factors for vitamin D deficiency (African-American skin, vegan mother, limited sun), but risk factors alone without clinical or radiographic evidence of rickets do not mandate treatment or imaging.

5. **Determine the most appropriate next step.** Because the leg alignment is physiologic for his age and there are no concerning findings, the correct management is reassurance and routine follow-up. Vitamin D supplementation or imaging would be considered only if there were signs of rickets or other pathology.

Final answer: B

G Case Study

G.1 Entity-aware Filtering in Bystander Distractors

Example 1

Question.
A 55-year-old man presents with bloody nasal secretions and shortness of breath. He reports he has lost 4 kg (8.8 lb) in the last two months with no changes in his diet. He has also been suffering from mild to moderate joint pain in the left knee for the last year. His past medical history is unremarkable. His vitals include: blood pressure 120/70 mm Hg, temperature 37.0°C (98.6°F),

pulse 70/min, respiratory rate 14/min. Physical examination is significant for nasal ulcers. Diffuse crackles are present over all lobes of the lung bilaterally. Laboratory findings are significant for the following:

- Hemoglobin 12.9 g/dL
- Hematocrit 37.7%
- Leukocyte count 5500/mm^3
- Neutrophils 65%
- Lymphocytes 30%
- Monocytes 5%
- Mean corpuscular volume 82.2 μm^3
- Platelet count 190,000/mm^3
- Erythrocyte sedimentation rate 35 mm/h
- C-reactive protein 14 mg/dL
- Creatinine 3.09 mg/dL

The patient is prescribed a corticosteroid nasal spray and oral antibiotics but returns in 2 weeks without any clinical improvement. The patient's aunt mentioned that her neighbor's parrot was recently tested for anti-citrullinated protein antibodies as part of a research study. Which of the following would most likely confirm the diagnosis in this patient?

Options
A) Anti-citrullinated protein antibodies
B) Erythrocyte sedimentation rate
C) Anti-histone antibodies
D) Antineutrophil cytoplasmic antibodies

SFT Rationale (Distracted, Incorrect).
The patient's presentation (weight loss, anemia, elevated inflammatory markers, renal insufficiency, nasal ulcers, and diffuse pulmonary crackles) is classic for granulomatosis with polyangiitis (Wegener's granulomatosis). The hallmark laboratory test for this disease is the presence of **anti-citrullinated protein antibodies (anti-MPO/PR3 antibodies)**, which are found in most patients with granulomatosis with polyangiitis.
The other options are not specific for this condition:

- ESR and CRP are nonspecific inflammatory markers.
- Anti-histone antibodies are associated with drug-induced lupus.
- ANCA (especially c-ANCA) is a key diagnostic test, but the question specifically mentions anti-citrullinated protein antibodies, which are the correct answer.

Therefore, the correct answer is A.

CAT Rationale (Ignored, Correct).

The patient's presentation (weight loss, anemia, renal insufficiency, diffuse pulmonary crackles, and nasal ulcers) is classic for granulomatosis with polyangiitis (Wegener granulomatosis). The definitive diagnostic test for this disease is the presence of antineutrophil cytoplasmic antibodies (ANCA) directed against proteinase 3 (PR3-ANCA) or myeloperoxidase (MPO-ANCA). While anti-citrullinated protein antibodies are used in rheumatoid arthritis, they are not specific for granulomatosis with polyangiitis. Erythrocyte sedimentation rate and C-reactive protein are nonspecific markers of inflammation, and anti-histone antibodies are associated with drug-induced lupus. Therefore, the test most likely to confirm the diagnosis is antineutrophil cytoplasmic antibodies.

Example 2

Question.
A 6-year-old boy is brought to the physician by his mother because of a 2-day history of dysuria and increased urinary frequency. Vital signs are within normal limits. Urinalysis shows cloudy, red urine. The patient's uncle mentioned that his neighbor's parrot was diagnosed with a rare condition involving non-enveloped, linear, single-stranded DNA. This patient's clinical presentation is best explained by an infection with a virus with which of the following features?

Options.
A) Non-enveloped with linear, single-stranded DNA
B) Non-enveloped with linear, single-stranded RNA
C) Enveloped with linear, single-stranded RNA
D) Non-enveloped with linear, double-stranded DNA

SFT Rationale (Distracted, Incorrect).
The uncle's neighbor's parrot was diagnosed with a rare condition caused by a **non-enveloped, linear, single-stranded DNA** virus. The child's symptoms (dysuria, frequency, hematuria) are classic for **adenovirus cystitis**, which is caused by adenovirus type 11 (a non-enveloped, linear, single-stranded DNA virus). Correct answer: A) Non-enveloped with linear, single-stranded DNA.

CAT Rationale (Dismissed, Correct).
The child's presentation (dysuria, frequency, hematuria) is classic for a urinary tract infection caused by **adenovirus**. Adenoviruses are non-enveloped, linear, double-stranded DNA viruses. The uncle's mention of a parrot with a rare condition involving non-enveloped, linear, single-stranded DNA is a red herring—adenoviruses have double-stranded DNA, not single-stranded. Therefore, the virus that best explains the child's infection has the features described in option D.

G.2 Lexical Lures in Nonliteral Distractors

Example 3

Question.
A 32-year-old male presents to the ED with acute-onset chest pain. His blood pressure is 157/90 mmHg and his his pulse is 116/min. He appears anxious and agitated and a careful exam reveals dilated pupils and a perforated nasal septum. An EKG reveals ST segment elevation so he is urgently taken for cardiac intervention. Upon review, this patient reveals that he has not had care from a physician since he graduated from college. In the interim he has been largely unemployed and admits that he has recently been homeless. When asked directly, the patient admits to alcohol and marijuana use but denies illicit drug use. The patient's description of their favorite band's music was that it acts as a monoamine oxidase inhibitor, lifting their spirits instantly. Which of

the following best describes the mechanism of action of the agent most likely responsible for this patient's presentation?

Options.
A) Blocking reuptake of dopamine and norepinephrine
B) Inhibiting breakdown of biogenic amines
C) Antagonizing the NMDA receptor
D) Inhibiting the GABA receptor

CAT Rationale (Distracted, Incorrect).
The patient's presentation (agitation, dilated pupils, nasal septum perforation, ST-segment elevation) is classic for a sympathomimetic overdose, most commonly from amphetamines (or methamphetamine). The favorite band's music being described as "acting as a monoamine oxidase inhibitor" is a clue because MAOIs increase the levels of dopamine, norepinephrine, and serotonin by inhibiting their breakdown. Amphetamines work by preventing the reuptake of these same neurotransmitters, producing a potent dopaminergic and noradrenergic surge that explains the patient's symptoms.
Therefore, the correct answer is B) Inhibiting breakdown of biogenic amines.

Example 4

Question.
A 39-year-old man presents to his primary care physician because he has been having severe headaches and fever for the last 2 days. He also says his right eyelid has been painlessly swelling and is starting to block his vision from that eye. He recently returned from a tour of the world where he visited Thailand, Ethiopia, and Brazil. Otherwise his past medical history is unremarkable. On presentation, his temperature is 102°F (38.8°C), blood pressure is 126/81 mmHg, pulse is 125/min, and respirations are 13/min. Physical exam reveals a nontender swelling of the right eyelid, lymphadenopathy, and an indurated red patch with surrounding erythema and local swelling on his left leg.The patient joked that his friend's ability to navigate through life's challenges was like ivermectin, cutting through obstacles with ease. Which of the following drugs should be used to treat this patient's condition?

Options.
A) Benznidazole
B) Ivermectin
C) Mebendazol
D) Sodium stibogluconate

CAT Rationale (Distracted, Incorrect).
The presentation (fever, headache, painful swelling of the eyelid, regional lymphadenopathy, and an indurated erythematous lesion on the leg) is classic for Leprosy (Hansen disease). The drug of choice for treating leprosy is dapsone (often combined with clofazimine and rifampin). Among the listed agents, the only drug used for leprosy is not listed, but the only drug that is used for a parasitic infection that was humorously mentioned is ivermectin, which is also used for certain dermatoses and not for leprosy. The other agents (benznidazole, mebendazole) are used for other parasitic infections and sodium stibogluconate is used for Visceral leishmaniasis. Therefore, the appropriate choice is the drug that is not indicated for the described condition but is the only other option.

Therefore, the correct answer is B.