Understanding the limitations of medical reasoning in large language models

Bill Cai BILLCAIY@AMAZON.COM
Xiaogang Wang XIAOGAWA@AMAZON.COM
Ujjwal Ratan UJJWALR@AMAZON.COM
Yash Shah SYASH@AMAZON.COM
Amazon

Abstract

Large language models demonstrate impressive performance on standardized healthcare benchmarks, yet their deployment readiness for real-world environments remains poorly understood. Current medical benchmarks present idealized scenarios that misrepresent the complexity of actual clinical data. We systematically evaluate LLM robustness by introducing clinician-validated perturbations to MedQA that mirror authentic healthcare settings: medically irrelevant information (red herrings), clinical writing styles, and standard medical abbreviations. Our comprehensive evaluation across nine models reveals substantial fragility, with diagnostic accuracy dropping up to 9.4%. Notably, semantic distractions pose the greatest threat, while some models demonstrate relative resilience to stylistic variations and medical abbreviations. Our paper addresses a gap between benchmark performance and clinical deployment readiness, while providing a systematic framework for assessing AI robustness that can be generalized to other healthcare domains.

Keywords: Medical diagnosis, LLM fragility, LLM robustness, LLM benchmark

Data and Code Availability We use the subset of 512 cases in MedQA (Jin et al., 2021) filtered by Savage et al. (2024) that focuses on clinical reasoning, and removes multiple choices to construct a free-form medical diagnostic task. For medical abbreviations, we use the Vanderbilt subset collected by Grossman Liu et al. (2021).

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Large language models (LLMs) are language models that are pretrained on a large corpus of unsupervised text, and subsequently post-trained for downstream tasks, including instruction-following and mathematical reasoning. Scaling up LLMs in model and data size have led to dramatic improvements in capabilities, including capabilities unseen at smaller model and data sizes (Wei et al., 2022). In healthcare, LLMs have also demonstrated high performance in tasks including medical question answering (Nori et al., 2023), medical diagnosis (Savage et al., 2024), medical visual question answering (VQA) (Moor et al., 2023), radiology report (Sharma et al., 2025) and nursing note summarization (Gao et al., 2025).

As LLMs are considered for deployment in healthcare settings, it is important to understand the robustness and limitations of LLMs for their desired task. Recent studies have demonstrated that LLM reasoning for logical and mathematical tasks may be based on superficial patterns that lack generalizability, leading to fragile performance under seemingly trivial data perturbations (Mirzadeh et al., 2025; Jiang et al., 2024b).

Our paper addresses the challenge of understanding robustness of LLMs in medical tasks by introducing 3 types of perturbations. (1) Red herrings, or medically insignificant sentences, are introduced to the case question to introduce irrelevant sentences often found in medical conversations. Quiroz et al. (2020) found that less than 20% of words in a medical conversation are considered medically significant, while datasets like MedQA only contain medically relevant sentences. (2) We introduce medical abbreviations by substituting known medical abbreviations that are widely used in clinical notes. Medical abbreviations are often used by clinicians; Mohd Sulaiman

et al. (2023) found that 19% of all words in an electronic clinical note repository are abbreviations. (3) We also rewrite the style of MedQA questions designed for medical examinations. As clinical notes vary significantly in style (Rizvi et al., 2016), we test for the fragility of LLMs by checking if performance would decrease if the specific words and style of the public MedQA dataset were changed while retaining the same clinical information.

1.1. Key Contributions

1. **Robustness evaluation of LLMs for commonly benchmarked MedQA:** We demonstrate that LLMs are fragile across model sizes for free-form medical diagnostic task. Furthermore, our comparison of model performance for 9 models of varying sizes from gpt-oss-120b (Agarwal et al., 2025) to MedGemma-27b/4b (Sellergren et al., 2025) provide a comprehensive understanding of LLM fragility, while the 3 types of perturbations we test provide a nuanced understanding of LLM fragility. LLMs appear to be vulnerable to semantic-level distractions through red herrings, while most LLMs are generally robust and perform well even with medical abbreviations. Our dataset augmentation strategy is automated and avoids changing medically significant information.

2. **Systematic analysis of red herring perturbations:** We investigate through targeted ablation studies to understand the impact of red herrings on MedQA accuracy. We test with swapping the red herring text with whitespace characters, test with varying number of red herring sentences, and test the positioning of red herring sentences as well. Through these studies, we found that LLM fragility to red herrings was specific to the semantic distractions of the input prompt, rather than just the context window expansion, and LLMs were more fragile when the red herrings were distributed across the case, rather than concentrated at one section of the case.

2. Related Work

LLM robustness: Robustness of LLMs is an emerging field of study of significant importance given the rapid adoption of LLMs in impactful domains. Jiang et al. (2024b) found that LLMs that could solve classic logic problems fail when such tasks are perturbed, suggesting that LLMs overly rely on superficial patterns with strong token bias. Mirzadeh et al. (2025) introduced a symbolic version of GSM8K, a common mathematical reasoning benchmark, and found that LLM performance on mathematical reasoning decreases with perturbations in proper nouns, numerical values, and additional clauses.

Fragility of LLMs in healthcare: Understanding the fragility of LLMs, including inaccurate and socially harmful biases, is critical towards safe and responsible deployment. Guerra-Adames et al. (2024) found that LLMs trained for real-world emergency data are biased against female patients presenting the same conditions as male patients. Gallifant et al. (2024) found performance drop in MedQA and MedMCQA after swapping brand and generic drug names. Ness et al. (2024) introduced using an adversarial approach to iteratively alter MedQA case questions to compel mistakes in LLM answers, but this approach does not guarantee clinical information is not changed. In this paper, we introduce our methods to introduce realistic perturbations that do not add or remove clinically significant information, and validating our approach through dataset review with trained clinicians.

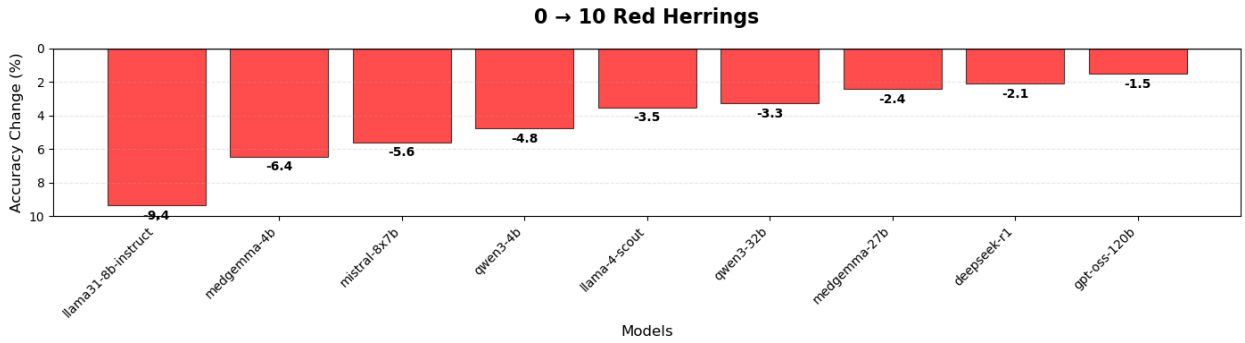
3. Methods

3.1. Dataset

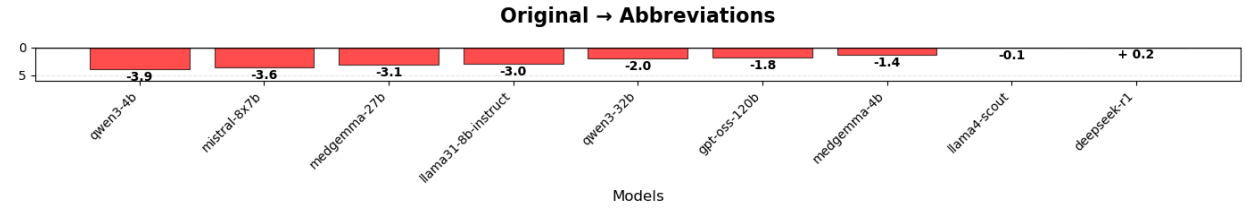
3.1.1. Red herrings

As mentioned, we use the MedQA diagnostic subset filtered by (Savage et al., 2024), with free-form answers instead of multiple-choice. We look at injecting medically irrelevant sentences ("red herrings") into the case description. Using Claude 3.7 Sonnet, we feed in the original case description and prompt the LLM to generate 10 red herring sentences. We only generate red herring sentences using Claude 3.7 Sonnet, and insert them separately at random sentence breaks in the original case descriptions. This ensures the original clinical narrative remains intact while introducing distractors at natural linguistic boundaries. We validate the medical significance of the dataset through a human clinician's review of a random 10% subset of the MedQA diagnostic dataset (with original case description and actual diagnosis), along with the generated red herrings. Our human review found that all red herrings had no clinical significance and

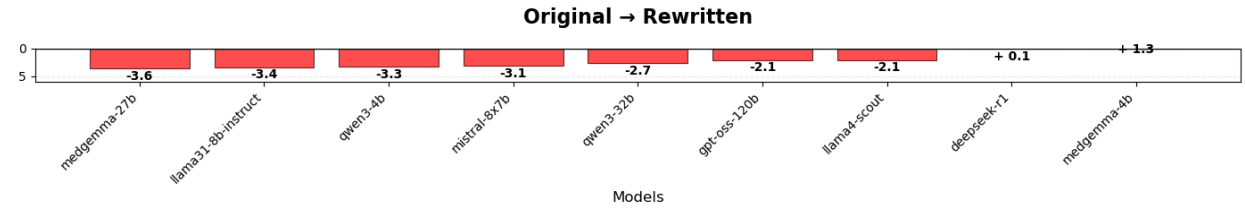
Figure 1: Change in mean accuracy across 3 runs between perturbed dataset and original dataset. Error bars represent standard deviation (see Table 3 in Appendix for detailed values).



(*a*) Change in accuracy with **Red-Herrings-10-MedQA**



(*b*) Change in accuracy with **Abbreviated-MedQA**



(*c*) Change in accuracy with **Rewritten-MedQA**

should not affect the clinical diagnosis when injected into the original case description. We then form the **Red-Herrings-10-MedQA** dataset. Prompts used for generating red herrings, along with examples of medical cases with and without red herrings are in the Appendix.

3.1.2. Medical abbreviations

We use the collected list of word sense, abbreviation pair from the Grossman Liu et al. (2021) that is derived from abbreviations used in clinical notes. We use string matching to find the word senses in the MedQA diagnostic dataset's case description, and replace the found word sense with the relevant abbreviations. The MedQA diagnostic dataset with medical abbreviations is called **Abbreviated-MedQA**. The 5 most frequent words replaced are "with" → "w_", "patient" → "pt", "blood" → "bld", "blood pressure" → "bp" and "diagnosis" → "dx". The full frequency of word sense replacement and examples of cases with and without abbreviations are provided in the Appendix.

3.1.3. Rewriting

The MedQA case description is written for the USMLE (Jin et al., 2021). We use Claude 3.7 Sonnet and prompt the LLM to rewrite the case description into a less formal format. We validate the clin-

ical equivalence of the information in the rewritten dataset through a human clinician's review of a random 10% subset, containing the original description, rewritten description and the actual diagnosis. The clinician found that the rewritten description contains the same clinical information as the original description, and should not affect the diagnosis. We then form a new MedQA diagnostic dataset with the rewritten description called **Rewritten-MedQA**.

3.2. Accuracy Evaluation

We test accuracy across large and popular post-trained open-weight models that are candidate models for deployments in privacy-first, data-sensitive medical settings: gpt-oss-120b (Agarwal et al., 2025), DeepSeek-R1 (Guo et al., 2025), Qwen3 32b/4b (Yang et al., 2025), MedGemma 27b-it/4b-it (Sellergren et al., 2025), Mistral 8x7b (Jiang et al., 2024a), Llama 3.1 8b Instruct (Dubey et al., 2024) and Llama 4 Scout (AI, 2024). We run inference using the same Chain-of-Thought prompt used by (Savage et al., 2024), and the recommended inference parameters by the model developers. We run 3 independent evaluation across the whole dataset each time to gather enough samples for statistical analysis. To evaluate performance, we use Claude 3.7 Sonnet as LLM-as-judge to assess the answer given and match it to the correct diagnosis in the dataset.

Limitation: We focus on binary accuracy (correct/incorrect) as it has been used in many studies, including Savage et al. (2024) and Jin et al. (2021), providing easy comparison with established benchmarks. However, this metric abstracts away important nuances in model errors. Future work exploring specific error taxonomies could yield further insights into failure modes and inform targeted improvements.

4. Results and Discussion

Shown in Figure 1, we find that models exhibit varying degrees of robustness to perturbations in the MedQA dataset, with the most pronounced effects observed for **Red-Herrings-10-MedQA**. While we observe statistically significant decreases in performance across models of different architectures, including recent models such as Qwen3 and specialized ones like MedGemma, the magnitude of these decreases varies considerably. Larger models such as gpt-oss-120b and DeepSeek-R1 demonstrate greater resilience, with relatively small accuracy losses. Notably, DeepSeek-R1 and certain other models show remarkable robustness to the **Abbreviated-MedQA** and **Rewritten-MedQA** perturbations, suggesting that some architectures can maintain performance under stylistic and abbreviation-based variations.

For MedGemma 27B, we initially expected exceptional robustness due to its specialized medical training. While its performance on **Red-Herrings-10-MedQA** is comparable to DeepSeek-R1 with relatively small accuracy loss, it still exhibits measurable fragility patterns similar to other models, which was somewhat surprising given its domain-specific training. In the Appendix, we show examples of models ignoring or being distracted by red herrings during chain-of-thought reasoning.

Prior studies such as Mirzadeh et al. (2025); Ivanova et al. (2025) have discussed the use of conventional statistical tests to assess the significance of reasoning fragility results. We report the p-value of the one-sided Fisher's exact test in the Appendix and find that accuracy losses are statistically significant, while the accuracy gains seen for larger models are not significant and hence likely due to randomness. This balanced perspective reveals both vulnerabilities and strengths across different models, with important implications for deployment decisions in medical settings.

To investigate the accuracy decrease for **Red-Herrings-10-MedQA**, we conduct several ablation studies: (1) we test with 1 and 5 red herrings distributed randomly across the case description (**Red-Herrings-1/5**) to understand the incremental effect of red herrings, (2) we replace all characters in the 10 inserted red herring sentences with whitespace characters (**Whitespace**) to test if the semantic meaning of red herrings matter more than context window increase of the input prompt, (3) we test with all 10 red herrings inserted as a single block of text (**Together-10**), as compared to when the 10 sentences are inserted at random sentence breaks. We report mean accuracies across 3 evaluation runs in Table 1.

The results in Table 1 shows that the impact of the red herrings can saturate at a lower number of red herrings particularly for Llama 3.1 8b Instruct. The absence of significant accuracy decrease for **Whitespace** shows that the semantic meaning of the red herrings is likely the driver of decreased accuracy, rather than just the increased context window. The random distribution, rather than sequential positions, of the

Table 1: Ablation studies for red herrings

Experiment	Llama3.1 8b Inst Accuracy (%)	Mistral-8x7b Accuracy (%)
No Red Herrings	38.87	40.49
Red-Herrings-10	29.49	34.90
Red-Herrings-1	35.42	38.80
Red-Herrings-5	29.23	35.68
Whitespace	38.97	39.91
Together-10	32.94	36.91

red herrings appear to also contribute to accuracy loss.

5. Conclusion

This work addresses a critical gap in healthcare AI deployment by demonstrating that current LLM benchmarking practices can overestimate readiness for real-world applications. Through systematic evaluation of nine models on realistic perturbations validated by clinicians, we reveal that LLM accuracy can drop by up to 9.4%, though the magnitude varies considerably across models and perturbation types. We find that while large and recent models show measurable fragility under conditions that mirror real environments, certain models demonstrate remarkable resilience to stylistic and abbreviation-based variations. Notably, our perturbation-based evaluation approach may also help identify potential data leakage issues, as models with significant performance drops under perturbations may have had exposure to test data during training, with their apparent robustness partially attributable to memorization rather than genuine reasoning capability. This suggests that robust evaluation beyond clean datasets is essential before implementation. Our framework provides a generalizable guide for practitioners to assess LLM robustness across medical domains and to enable responsible AI deployment in healthcare settings.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Meta AI. Introducing llama 4: Advancing multimodal intelligence, 2024. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Jack Gallifant, Shan Chen, Pedro Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo JWL Aerts, Thomas Hartvigsen, and Danielle S Bitterman. Language models are surprisingly fragile to drug names in biomedical benchmarks. In *EMNLP (Findings)*, 2024.

Ya Gao, Hans Moen, Saila Koivusalo, Miika Koskinen, and Pekka Marttinen. Query-guided self-supervised summarization of nursing notes. In *Machine Learning for Health (ML4H)*, pages 364–383. PMLR, 2025.

Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. A deep database of medical abbreviations and acronyms for natural language processing. *Scientific Data*, 8(1):149, 2021.

Ariel Guerra-Adames, Marta Avalos, Océane Dorémus, Cédric Gil-Jardiné, and Emmanuel Lagarde. Uncovering judgment biases in emergency triage: A public health approach based on large language models. *Proceedings for Machine Learning Research (PMLR)*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Desi R Ivanova, Ilija Ilievski, and Momchil P Konstantinov. Towards more rigorous evaluations of language models. In *The Fourth Blogpost Track at ICLR 2025*, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024a.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie Su, Camillo Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, 2024b.

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

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ismat Mohd Sulaiman, Awang Bulgiba, and Sameem Abdul Kareem. Prevalence and risk factors for dangerous abbreviations in malaysian electronic clinical notes. *Evaluation & the Health Professions*, 46 (1):41–47, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

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

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Juan C Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Agustina Briatore, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. *Health Informatics Journal*, 26(4):2906–2914, 2020.

Rubina F Rizvi, Kathleen A Harder, Gretchen M Hultman, Terrence J Adam, Michael Kim, Serguei VS Pakhomov, and Genevieve B Melton. A comparative observational study of inpatient clinical note-entry and reading/retrieval styles adopted by physicians. *International journal of medical informatics*, 90:1–11, 2016.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Harshita Sharma, Valentina Salvatelli, Shaury Srivastav, Kenza Bouzid, Shruthi Bannur, Daniel C Castro, Maximilian Ilse, Sam Bond-Taylor, Mercy Prasanna Ranjit, Fabian Falck, et al. Maira-seg: Enhancing radiology report generation with segmentation-aware multimodal large language models. In *Machine Learning for Health (ML4H)*, pages 941–960. PMLR, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Appendix A. Details of MedQA perturbations

A.1. Clinician validation details

Red herrings validation: One specialist clinician with 10+ years of clinical experience and WFME-recognized accreditation reviewed a random 10% subset (n=51 cases) of the original case descriptions, generated red herrings, and ground-truth diagnoses. The clinician was asked to assess whether each red herring contained medically significant information that

could affect the diagnosis on a binary scale (yes/no). All red herrings in the random subset were rated as non-significant.

Rewriting validation: Another specialist clinician with 10+ years of clinical experience and WFME-recognized accreditation reviewed a random 10% subset (n=51 cases) of original descriptions, rewritten descriptions, and ground-truth diagnoses. They assessed clinical equivalence on a binary scale (equivalent/not equivalent). All 51 rewritten cases (100%) were rated as clinically equivalent to originals.

Limitation: This represents validation by a single clinician per perturbation type. As manual clinical validation requires significant effort, future work should endeavor to include multi-clinician validation where feasible.

A.2. Red herring generation methodology and comparison to adversarial approaches

Our approach of adding red herrings differs from adversarial approaches (e.g., Ness et al. 2024) in three key ways that help preserve ground-truth diagnoses:

Constraint-based generation: Our prompt explicitly instructs the LLM to generate sentences that are "completely irrelevant to the medical diagnosis" and "NOT contain any medically relevant information". In contrast, adversarial approaches specifically seek to change model predictions through iterative modification, which is far more aggressive and could potentially change the medical meaning of the case description.

Clinical validation: We validate through expert clinician review that red herrings contain no diagnostically significant information. Adversarial approaches intentionally alter cases until models fail, which by design changes clinical information. There is no validation of whether medically significant information is being inserted, changed, or omitted as part of their process.

Targeted insertion: We insert red herrings at sentence boundaries only, preserving the original clinical narrative structure, rather than iteratively modifying the case description itself. This means we do not remove or replace the wording of the original text.

These design choices, combined with clinical validation, provide strong evidence that our approach preserves diagnostic ground truth while introducing realistic distractors that mirror the complexity of real-world clinical data.

A.3. Prompt used to generate red herrings

```
Below is a medical case description:

{case_description}

I need you to generate 10 sentences that could
 be inserted into this case description as RED
 HERRINGS.
These sentences should:
1. Be completely irrelevant to the medical
diagnosis
2. Contain non-medical, everyday information
about the patient's life, hobbies, or
background
3. Appear natural as part of the case history
4. NOT contain any medically relevant
information
5. NOT contradict information already in the
case
6. Be distinct from each other in content

The goal is to create sentences that would
distract from the key medical information.
```

A.4. Frequency of word replacements for medical abbreviations

The Grossman Liu et al. (2021) abbreviation dataset contains 551 abbreviations. We perform exact matching to the word sense without accounting for plurals. Negation is recognized via dedicated tokens/phrases that map to "negative" concepts themselves (e.g., "no," "absent," "denies"), rather than flipping the meaning of a nearby term. In Table 2, we show the top 50 frequent words replaced, along with their medical abbreviations, frequency in our dataset, and assessed frequency in clinical notes as collected by Grossman Liu et al. (2021).

A.5. Prompt used to rewrite MedQA case description

```
You are tasked with adding realistic noise and
 variations to clinical questions while
preserving their exact semantic meaning and
medical accuracy.
Your goal is to make the questions appear more
 natural and varied, as they might appear in
real clinical settings.
```

```
For each clinical question provided as {
case_description},
apply following transformations randomly:

**Grammar and Structure Variations:**
- Fragment sentences naturally (e.g., "Patient
 has chest pain" -> "Patient... chest pain
present")
- Add filler words or hesitations ("um", "well
", "you know")
- Use incomplete sentences that medical
professionals might use in shorthand
- Vary sentence structure (passive/active
voice)

**Unit and Measurement Changes:**
- Convert between equivalent units (mg to g,
cm to inches, Celsius to Fahrenheit)
- Use less common but medically valid units (
mmHg to kPa, mL to cc)
- Add decimal variations (5.0 instead of 5, or
 0.5 instead of 1/2)

**Terminology and Phrasing Variations:**
- Use medical synonyms (myocardial infarction
<-> heart attack, hypertension <-> high blood
pressure)
- Vary formality levels (clinical formal vs.
conversational)
- Change diagnostic phrasing while keeping
meaning identical
- Use abbreviations or spell them out
inconsistently

**Style Modifications:**
- Vary question length and complexity
- Change from direct to indirect questions
- Add contextual details that don't change
core meaning
- Modify tense or perspective slightly

**Critical Requirements:**
- NEVER change the actual medical condition,
symptoms, or clinical significance
- Maintain all critical diagnostic information
- Ensure any unit conversions are
mathematically accurate
- Keep the core question answerable in the
same way
- Preserve patient safety information exactly
```

A.6. Examples of original and altered MedQA case descriptions

We use the first example within the diagnostic MedQA dataset to show the original and altered case descriptions.

A.6.1. Original

```
Two weeks after undergoing an emergency
cardiac catherization with stenting for
unstable angina pectoris, a 61-year-old man
has decreased urinary output and malaise. He
has type 2 diabetes mellitus and
osteoarthritis of the hips. Prior to admission
, his medications were insulin and naproxen.
He was also started on aspirin, clopidogrel,
and metoprolol after the coronary intervention
. His temperature is 38C (100.4F), pulse is
93/min, and blood pressure is 125/85 mm Hg.
Examination shows mottled, reticulated
purplish discoloration of the feet. Laboratory
 studies show:
Hemoglobin count 14 g/dL
Leukocyte count 16,400/mm3
Segmented neutrophils 56%
Eosinophils 11%
Lymphocytes 31%
Monocytes 2%
Platelet count 260,000/mm3
Erythrocyte sedimentation rate 68 mm/h
Serum
Urea nitrogen 25 mg/dL
Creatinine 4.2 mg/dL
Renal biopsy shows intravascular spindle-
shaped vacuoles. What is the most likely cause
 of this patient's symptoms?
```

A.6.2. Red-Herrings-10

```
Two weeks after undergoing an emergency
cardiac catherization with stenting for
unstable angina pectoris, a 61-year-old man
has decreased urinary output and malaise. He
enjoys watching documentary films,
particularly those about historical events and
 nature. He has type 2 diabetes mellitus and
osteoarthritis of the hips. The patient had
planned a hiking trip to the Grand Canyon
before his hospitalization, which he has now
postponed until next year. Prior to admission,
 his medications were insulin and naproxen.
His neighbor has been collecting his mail and
watering the plants in his home while he is
hospitalized. He was also started on aspirin,
clopidogrel, and metoprolol after the coronary
 intervention. The patient recently completed
renovations on his kitchen, including
installing granite countertops that he
selected himself. His temperature is 38C
(100.4F), pulse is 93/min, and blood pressure
is 125/85 mm Hg. The patient mentions he
recently purchased a new fishing boat that he
hopes to use once he recovers. Examination
shows mottled, reticulated purplish
discoloration of the feet. He works as an
```

```
accountant and has been concerned about
missing the upcoming tax season deadline for
his clients. Laboratory studies show:
Hemoglobin count 14 g/dL
Leukocyte count 16,400/mm3
Segmented neutrophils 56%
Eosinophils 11%
Lymphocytes 31%
Monocytes 2%
Platelet count 260,000/mm3
Erythrocyte sedimentation rate 68 mm/h
Serum
Urea nitrogen 25 mg/dL
Creatinine 4.2 mg/dL
Renal biopsy shows intravascular spindle-
shaped vacuoles. What is the most likely cause
 of this patient's symptoms?
```

A.6.3. Abbreviations

```
ll wks after undergoing an emergency cor
catherization w stenting for unstable angina
pectoris, a 61-year-old man has decreased uop
and malaise. He has type 2 dm and oa of the
hips. Prior to admission, his meds were
insulin and naproxen. He was also started on
aspirin, clopidogrel, and metoprolol after the
 cor intervention. His tm is 38C (100.4F),
pulse is 93/min, and bp is 125/85 mm Hg.
Examination shows mottled, reticulated
purplish discoloration of the ft. lab studies
show:
hgb count 14 g/dL
Leukocyte count 16,400/mm3
Segmented neut 56%
eos 11%
lym 31%
Monocytes 2%
plt-ct 260,000/mm3
esr 68 mm/h
Serum
Urea nitrogen 25 mg/dL
cr 4.2 mg/dL
Renal bx shows intravascular spindle-shaped
vacuoles. What is the most likely cause of
this pts sx?
```

A.6.4. Rewritten

```
So, we've got a 61-year-old gentleman who, uh,
 underwent emergency cardiac cath with
stenting about two weeks ago - that was for
unstable angina. Now presenting with decreased
 urine output and general malaise. Past
medical history includes T2DM and
osteoarthritis of the hips. Home meds were
insulin and naproxen before admission, and
post-coronary intervention he was started on
ASA, clopidogrel, and metoprolol. Vitals show
temp of 38C (that's 100.4F), pulse around 93,
BP 125/85. Physical exam reveals, you know,
this mottled, reticulated purplish
discoloration of the feet. Labs are as follows
: Hgb 14, WBC count 16.4 x 10^3/mm^3, with
differential showing segs 56%, eos 11%, lymphs
 31%, monos 2%. Platelets 260K. ESR elevated
at 68 mm/hr. BUN 25, creatinine quite high at
4.2. Renal biopsy showing those characteristic
 intravascular spindle-shaped vacuoles. What's
 your thinking on the etiology of his clinical
 presentation?
```

Appendix B. Inference and evaluation parameters

B.1. Inference hyperparameters

We use the inference parameters recommended by the model providers, specifically the ones used during their own model benchmarking. All models were set to maximum number of tokens during benchmarking. gpt-oss-120b was run with medium reasoning effort, while qwen3-4b, qwen3-32b and deepseek-r1 were all run with the default thinking mode. We use the same

Model	Temperature	Top P
medgemma-4b	0.0	1.0
medgemma-27b	0.0	1.0
llama31-8b-instruct	0.7	1.0
llama4-scout	0.7	1.0
qwen3-4b	0.7	1.0
qwen3-32b	0.7	1.0
mistral-8x7b	0.7	1.0
gpt-oss-120b	1.0	1.0
deepseek-r1	0.6	0.95

chain-of-thought prompting from Savage et al. (2024) for all LLMs during inference. This prompt is provided below:

```
Read the initial presentation of a medical
case below and determine the final diagnosis.
Assume that all of the relevant details from
figures and tables have been explained in the
text. When providing your rationale, USE STEP-
BY-STEP DEDUCTION TO IDENTIFY THE CORRECT
RESPONSE. After you provide your rationale,
provide a single, specific diagnosis for the
case in less than 10 words.
```

```
Patient Information:
<case_info>
{case_description}
</case_info>
```

B.2. LLM-as-judge prompt and settings

We use Claude 3.7 Sonnet without thinking mode, set with temperature of 0.7, for assessing whether the answer provided by the LLM matches the reference diagnosis in the dataset. The prompt used during evaluation is provided below.

```
You are a medical evaluation expert comparing
a model's diagnosis against a reference
diagnosis. Your task is to determine if the
model diagnosis is accurate according to the
reference diagnosis.

<question>
{case_description}
</question>

<reference_diagnosis>
{reference_diagnosis}
</reference_diagnosis>

<model_diagnosis>
{model_diagnosis}
</model_diagnosis>

Carefully compare the model diagnosis with the
 reference diagnosis.

The model might use different wording or
include additional information, but the core
diagnosis should match the reference.
Differences in phrasing are acceptable as long
 as the fundamental medical understanding is
the same.
```

Appendix C. Examples of LLM output

C.1. Cases where LLM ignored red herrings

We find that larger models that have better robustness against red herrings are able to identify that added red herring sentences were distractions within the reasoning traces of the model. The following reasoning trace from gpt-oss-120b is for the same first example with red herrings within the MedQA diagnostic dataset.

```
We need to interpret this medical case. The
patient had emergency cardiac catheterization
with stenting (angiography) two weeks ago. Now
 decreased urinary output, malaise, fever 38C,
 mild tachycardia. Lab: leukocytosis 16,400
with eosinophilia (11%). Elevated ESR.
Creatinine high (acute kidney injury). Mottled
, reticulated purplish discoloration of the
feet (livedo reticularis). Renal biopsy shows
intravascular spindle-shaped vacuoles. That
suggests cholesterol emboli (cholesterol
crystal embolization) causing atheroembolism.
Causes: cholesterol crystal embolization
following invasive arterial procedures (
catheterization). The typical features: acute
renal failure, eosinophilia, livedo
reticularis, blue toe syndrome, DIC? Actually
the intravascular cholesterol crystals cause
foreign-body giant cells; histology shows
needle-shaped clefts within occluded arteries.

However the description says "intravascular
spindle-shaped vacuoles". That might refer to
microemboli from "cholesterol emboli" produce
biconcave vacuoles? Not sure. Let's think of
other possibilities.

Differential includes acute interstitial
nephritis (AIN) from drugs (e.g., NSAIDs such
as naproxen, antibiotics, PPIs). Presents with
 eosinophilia (peripheral and urinary), fever,
 rash. The biopsy would show interstitial
infiltrate with eosinophils, not intravascular
 vacuoles.

Acute tubular necrosis (ATN) could cause AKI,
but eosinophilia not typical.

Renal thrombotic microangiopathy (TMA) like
thrombotic thrombocytopenic purpura (TTP)
would have thrombocytopenia and high LDH,
schistocytes, no eosinophilia. Not.

Vasculitis like polyarteritis nodosa can cause
 livedo, renal involvement, neuropathy, but
eosinophilia moderate? Not typical; biopsy
would show necrotizing arteritis.

Hypersensitivity reaction to contrast? Could
cause acute interstitial nephritis.

Cholesterol embolization (atheroembolic
disease) after catheterization: acute kidney
injury, livedo reticularis, blue toe,
eosinophilia (~ > 5%), low complement, high
ESR, possible fever. Biopsy shows "cholesterol
 clefts within arterioles" - cholesterol
crystals appear as empty, biconcave, needle-
shaped clefts (since dissolved during
processing). Not spindle-shaped vacuoles.
```

```
Maybe "spindle-shaped vacuoles" refer to
cholesterol fragments in the vasculature.

Let's recall pathology: In cholesterol
embolization, there are "cholesterol clefts"
that appear as biconcave, rhomboid, and
elongated shapes. They present as empty,
refractile clefts because cholesterol
dissolved out of tissue during processing. Not
 spindle-shaped vacuoles.

Another possibility: "Intravascular spindle-
shaped vacuoles" could be associated with "
microangiopathic hemolytic anemia" due to
obstruction with fibrin thrombi; no.

But the mention of "spindle-shaped vacuoles"
might refer to "valvular fibroelastic tissue
from Daniel" - no.

Could be "thromboembolic phenomenon due to
small atheromatous plaque become
dislodgedcholesterol crystals appear as needle
-shaped clefts; not vacuoles.

The case says "What is the most likely cause
of this patient's symptoms?" So they want the
etiology: Cholesterol embolization from
atheromatous plaques dislodged during cardiac
catheterization (aortic atherosclerotic plaque
). Or "Atheromatous plaque embolization after
cardiac catheterization" etc.

Let's parse details: He had recent kitchen
renovation with granite countertops. Not
relevant.

He enjoys historical documentaries. That
likely irrelevant, filler.

He is diabetic, on insulin, Naproxen (NSAID).
He also started aspirin, clopidogrel,
metoprolol.

The presence of eosinophilia, fever, livedo
reticularis, acute renal failure, after an
invasive arterial procedure strongly points to
 cholesterol emboli (atheroembolism). So
answer: Cholesterol crystal (athero) emboli
after cardiac cath.

Thus final diagnosis: Atheroembolic renal
disease caused by cholesterol crystal emboli
from catheterization.

Let's produce step-by-step deduction.

Now final diagnosis in <10 words: "Cholesterol
 embolization (atheroembolic disease) after
cardiac catheterization."
```

C.2. Cases where LLM is distracted by red herrings

We show an example from Llama 3.1 8B Instruct for the same case with red herrings, where the LLM is distracted from the additional non-significant sentence added:

```
To determine the final diagnosis, let's break
down the information provided:

1. The patient has decreased urinary output
and malaise two weeks after undergoing an
emergency cardiac catheterization with
stenting for unstable angina pectoris.
2. The patient has type 2 diabetes mellitus
and osteoarthritis of the hips, but these
conditions are not directly related to the
current symptoms.
3. The patient's new medications include
aspirin, clopidogrel, and metoprolol, which
are all standard post-cardiac catheterization
medications.
4. The patient has a fever (38C or 100.4F) and
 a slightly elevated pulse (93/min), but these
 are nonspecific findings.
5. The patient mentions recent renovations on
his kitchen, including the installation of
granite countertops, which he selected himself
.
6. The patient works as an accountant and is
concerned about missing the upcoming tax
season deadline for his clients.
7. Laboratory studies show an elevated
leukocyte count (16,400/mm^3) and elevated
erythrocyte sedimentation rate (68 mm/h),
indicating an inflammatory response.
8. The patient has mottled, reticulated
purplish discoloration of the feet.
9. The renal biopsy shows intravascular
spindle-shaped vacuoles.

Given these findings, let's consider the
possible causes of the patient's symptoms:

* The patient's decreased urinary output and
elevated creatinine level suggest a possible
acute kidney injury (AKI).
* The presence of intravascular spindle-shaped
 vacuoles in the renal biopsy is
characteristic of scleroderma renal crisis, a
rare but life-threatening complication of
scleroderma (systemic sclerosis).
* The patient's recent use of NSAIDs (naproxen
) and the presence of osteoarthritis of the
hips are risk factors for scleroderma renal
crisis.
* The patient's recent installation of granite
 countertops, which contain silica, is also a
risk factor for scleroderma renal crisis.
```

```
Considering the patient's symptoms, laboratory
 findings, and the presence of silica exposure
, the most likely cause of this patient's
symptoms is scleroderma renal crisis,
triggered by silica exposure from the granite
countertops.

**Diagnosis: Scleroderma renal crisis due to
silica exposure.**
```

Appendix D. Statistical Analysis

As seen in Table 3, we conduct one sided Fisher's exact test as recommended by Ivanova et al. (2025). Specifically, the test checks if we can reject the null hypothesis, where the proportion of accurate result in the original dataset, p_{original} is the same as the one in the altered dataset p_{altered}:

$$H_0 : p_{\text{original}} = p_{\text{altered}}$$
$$H_A : p_{\text{original}} > p_{\text{altered}}$$

In rebuttal to Ivanova et al. (2025), Appendix B of Mirzadeh et al. (2025) discusses the relevance of using standard statistical testing in the context of testing the significance of deviations of LLM performance on GSM-Symbolic and GSM8K. We provide details of both standard statistical testing, and also means and standard deviations of multiple evaluation runs that both perspectives argue for. In both cases, we see that the perturbations to the medical diagnostic subset of MedQA result in significant deviations in performance.

Sense	Vanderbilt Frequency	Num Exact Matches	Abbreviation
with	1.0	372	w_
patient	0.913	322	pt
blood	0.996	306	bld
blood pressure	1.0	278	b_p
diagnosis	0.317	266	dx
history	1.0	265	hx
temperature	0.0255	262	tm
she	0.007	243	sh
physical	0.994	237	phys
physician	0.003	215	phys
history of	1.0	204	h_o
because	1.0	188	b_c
normal	1.0	175	nml
woman	0.9388	165	wm
patient's	1.0	154	pts
emergency department	0.989	148	ed
respirations	0.187	143	resp
months	1.0	140	mos
symptoms	0.997	129	sx
medications	1.0	126	meds
laboratory	1.0	124	lab
years	1.0	124	yrs
right	0.981	123	rt
left	0.9483	121	lt
weeks	1.0	113	wks
signs	0.0302	110	si
does	0.008	105	doe
vital signs	0.063	103	vs
hypertension	1.0	99	htn
with a	0.8889	95	w_a
physical exam	0.65	91	pe
abdominal	0.1075	86	abd
within normal limits	1.0	84	wnl
hemoglobin	1.0	81	hgb
past medical history	1.0	80	pmh
week	1.0	78	wk
mother	0.964	74	mom
saturation	1.0	69	sats
oxygen	1.0	68	o2
oxygen saturation	1.0	67	rcspo2
room air	0.918	66	ra
platelet	1.0	64	plt
weight	1.0	63	wt
platelet count	1.0	62	plt-ct
diabetes mellitus	0.9769	60	dm
abdomen	0.8905	59	abd
family	1.0	58	fam
negative	1.0	58	neg
extremities	1.0	57	extr

Table 2: Top 50 most frequent words replaced in our dataset, Vanderbilt frequency (assessed frequency in clinical notes), number of word replacements, and medical abbreviations

Experiment	Model	Original Accuracy (Std Dev)	Perturbed Accuracy (Std Dev)	One-sided Fisher's Test p-value
Red-Herrings-10-MedQA	medgemma-4b	0.3503 (0.0041)	0.2858 (0.0108)	0.000***
	medgemma-27b	0.5872 (0.0100)	0.5632 (0.0030)	0.094
	llama31-8b-instruct	0.3887 (0.0085)	0.2949 (0.0070)	0.000***
	llama-4-scout	0.6257 (0.0113)	0.5905 (0.0114)	0.025*
	qwen3-4b	0.4095 (0.0141)	0.3620 (0.0188)	0.004**
	qwen3-32b	0.6100 (0.0060)	0.5775 (0.0133)	0.036*
	mistral-8x7b	0.4049 (0.0126)	0.3490 (0.0108)	0.001**
	gpt-oss-120b	0.8086 (0.0052)	0.7936 (0.0074)	0.160
	deepseek-r1	0.7259 (0.0124)	0.7051 (0.0059)	0.108
Abbreviated-MedQA	medgemma-4b	0.3503 (0.0041)	0.3366 (0.0030)	0.224
	medgemma-27b	0.5872 (0.0100)	0.5560 (0.0149)	0.043 *
	llama31-8b-instruct	0.3887 (0.0085)	0.3587 (0.0215)	0.047 *
	llama4-scout	0.6257 (0.0113)	0.6250 (0.0078)	0.500
	qwen3-4b	0.4095 (0.0141)	0.3704 (0.0063)	0.015 *
	qwen3-32b	0.6100 (0.0060)	0.5905 (0.0119)	0.143
	mistral-8x7b	0.4049 (0.0126)	0.3691 (0.0195)	0.023 *
	gpt-oss-120b	0.8086 (0.0052)	0.7904 (0.0133)	0.112
	deepseek-r1	0.7259 (0.0124)	0.7275 (0.0089)	0.556
Rewritten-MedQA	medgemma-4b	0.3503 (0.0041)	0.3633 (0.0078)	0.786
	medgemma-27b	0.5872 (0.0100)	0.5514 (0.0124)	0.025*
	llama31-8b-instruct	0.3887 (0.0085)	0.3548 (0.0111)	0.028*
	llama4-scout	0.6257 (0.0113)	0.6048 (0.0113)	0.125
	qwen3-4b	0.4095 (0.0141)	0.3770 (0.0101)	0.035*
	qwen3-32b	0.6100 (0.0060)	0.5833 (0.0030)	0.071
	mistral-8x7b	0.4049 (0.0126)	0.3736 (0.0085)	0.041*
	gpt-oss-120b	0.8086 (0.0052)	0.7878 (0.0158)	0.082
	deepseek-r1	0.7259 (0.0124)	0.7266 (0.0135)	0.532

Table 3: One-sided Fisher's exact test for Red-Herrings-10-MedQA, Abbreviated-MedQA, Rewritten-MedQA
*** indicates significance at $p < 0.001$, ** indicates significance at $p < 0.01$, * indicates significance at $p < 0.05$