Can We Trust LLMs for Medical Diagnosis? Evaluating the Robustness of Clinical Reasoning under Perturbation

Anonymous ACL submission

Abstract

Medical large language models (LLMs) have been widely proposed for medical diagnosis with relatively high accuracy. However, existing LLM evaluations often prioritize top-1 accuracy while ignoring the fragility of the reasoning process on realistic clinical notes that are plagued by noise and inconsistent formats. To explore the robustness of reasoning process, this study proposes an adversarial perturbation framework that consists of two strategies: semantic pruning of clinical notes to verify the attention limitation of LLMs, and noise injection to investigate the anti-interference capability of the reasoning process. The experiments are conducted on a realistic dataset, and the results verify the attention limitation and inadequate anti-interference capability of LLMs. These findings reveal the reasoning logic and provide a feasible solution for trustworthy LLMs.

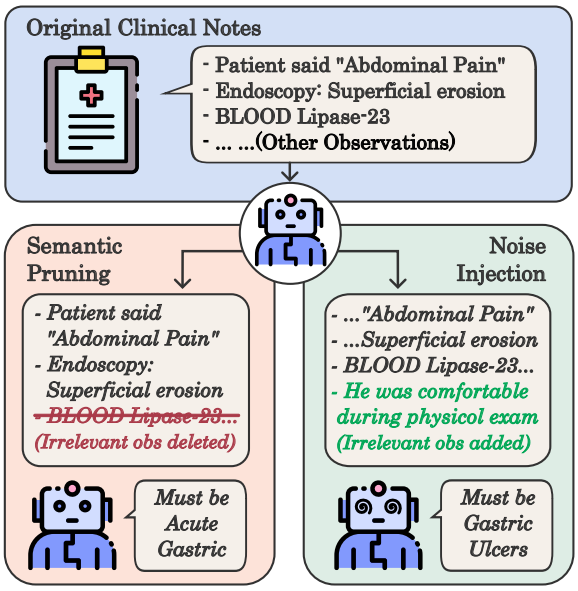

Figure 1: Perturbation strategies for clinical notes.

1 Introduction

Current large language models (LLMs) are advancing rapidly in the medical field, particularly performing impressively in question-answer (QA) tasks (Anil et al., 2023; Han et al., 2023; Jin et al., 2020). However, the question in QA tasks is relatively short and structured, while the real clinical scenarios for physicians are far more complex, involving long-context, messy clinical notes with significant noise (Gao et al., 2025; Hirschtick, 2006; Lehman et al., 2023). It has been shown that LLMs exhibit limited accuracy when processing long contexts and are sensitive to noise (Moradi and Samwald, 2022; Singh et al., 2024), because they are quite likely to overlook critical information (Liu et al., 2024; Finlayson et al., 2019).

Currently, predominant benchmarks typically prioritize Top-1 accuracy on standardized inputs, potentially ignoring the fragility of the reasoning process. Moreover, many evaluations of LLMs rely on medical QA (Pal et al., 2022; Li et al., 2023; Chen et al., 2025a) or natural language inference (NLI) (Jullien et al., 2023), while, in fact, a model might achieve the right diagnosis for the wrong reasons (Ye and Durrett, 2022; Liévin et al., 2024). For the medical field, it requires safe diagnosis with rigorous reasoning. (Li et al., 2025; Croskerry, 2009). Therefore, it is extremely significant to explore the robustness of reasoning process on realistic clinical notes.

To achieve this goal, we first utilize the DiReCT dataset (Wang et al., 2024), including real clinical notes that are long-context and with redundancy. Then we propose an adversarial perturbation (Jia and Liang, 2017) framework, comprising two strategies, that is, semantic pruning and noise injection, to perturb the clinical notes. The evaluation results show that models indeed generate more accurate and faithful responses after semantic pruning, while noise injection leads to unfaithful reasoning, which proves that LLMs have limited attention and anti-interference capability. Therefore,

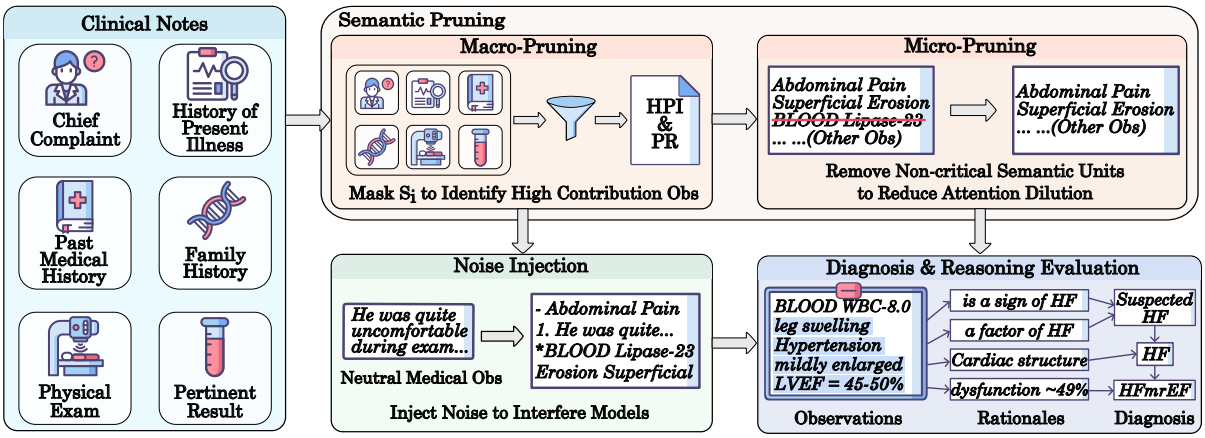

Figure 2: The pipeline of the adversarial perturbation framework.

enhancing signal-to-noise ratio of clinical notes is a feasible approach to improve the robustness of reasoning process.

2 Methodology

In this section, the adversarial perturbation framework will be introduced, involving semantic pruning and noise injection. Figure 2 illustrates the pipeline of the framework, of which four core components are detailed as follows.

2.1 Clinical Notes

To evaluate the realistic clinical scenarios, we utilize DiReCT dataset, proposed by Wang et al. (2024), where the real clinical notes $\mathcal{R}$ consist of 6 records as $\mathcal{R} = [r_1, r_2, ..., r_6]$, each of which corresponds clinical notes section demonstrating in Figure 2. The clinical diagnostic task (Wang et al., 2024) is formalized as a mapping function $f_\theta(R, \mathcal{K}) \rightarrow (d, \mathcal{E})$, where the model θ predicts a diagnosis d and rationale $\mathcal{E}$ based on clinical note $\mathcal{R}$ and knowledge graph $\mathcal{K}$. During the processes of manual annotation and model reasoning, observation o in R is extracted as the basis for their diagnosis.

2.2 Semantic Pruning

Semantic pruning is subdivided into macro pruning and micro pruning. These pruning manipulations can identify the key observations in the clinical note, thereby evaluating the attention stability of the models.

In **macro-pruning**, we iteratively apply a mask $\mathcal{T}_{mask}(R, i)$ to mask records r_i, generating an incomplete input $R'_i = R \setminus \{r_i\}$. By comparing the performance drop ΔAcc_i, we identify the high contribution records h from the clinical notes.

In **micro-pruning**, we select the h as a target record. Formally, a target record consists of a set of semantic units $U = \{u_1, u_2, ..., u_n\}$. A unit u_j is defined as either a self-contained sentence or a $\{examination, result\}$ pair. Instead of standard random dropout, which may compromise syntactic coherence, we employ an LLM to perform semantic unit deletion, selecting a subset $U_{del} \subset U (1 \leq |U_{del}| \leq 3)$, following the instruction as $P(d|R) \approx P(d|R \setminus U_{del})$, which achieves more concise clinical notes while remaining the diagnostic orientation invariant.

2.3 Noise Injection

Based on the result of macro-pruning, noise injection is conducted on the record h. We achieve noise injection via a controlled rewriting process by injecting syntactic noise and benign medical noise, which is operated by a LLM. Syntactic noise is created through list reordering and punctuation variation, while benign medical noise means several clinically plausible yet diagnostically neutral observations. These two kinds of noises, together defined as U_{noise}, are injected into the targets, following the instruction as $P(d|R) \approx P(d|R \cup U_{noise})$, which ensures that the noise injected maintains the same diagnostic orientation.

2.4 Evaluation Metrics

To rigorously quantify the resilience of clinical reasoning under perturbation, we adopt three evaluation metrics proposed in (Wang et al., 2024). Specifically, diagnostic accuracy indicates the correctness of the diagnosis, while observation completeness and reasoning faithfulness reveal the robustness of the reasoning process.

Model	Baseline	Micro-Pruning		Noise Injection	
	$\mathcal{A}^{diag}$ / ($\mathcal{F}^{all}$)	HPI	PR	HPI	PR
GPT-5-mini	76.52 (12.36)	75.15 (−1.37) (12.11 (−0.25))	75.93 (−0.59) (11.58 (−0.78))	77.30 (+0.78) (9.81 (−2.55))	74.17 (−2.35) (9.95 (−2.41))
Gemini-2.5-Flash	70.84 (14.01)	63.01 (−7.83) (11.20 (−2.81))	61.84 (−9.00) (11.18 (−2.83))	65.36 (−5.48) (9.61 (−4.40))	65.95 (−4.89) (10.67 (−3.34))
Llama-3.3-70B	71.04 (7.41)	69.28 (−1.76) (7.49 (+0.08))	71.43 (+0.39) (7.68 (+0.27))	70.06 (−0.98) (6.73 (−0.68))	70.45 (−0.59) (6.84 (−0.57))
Qwen3-235B	62.62 (14.41)	61.45 (−1.17) (13.72 (−0.69))	64.38 (+1.76) (15.05 (+0.64))	60.47 (−2.15) (10.84 (−3.57))	61.45 (−1.17) (11.22 (−3.19))
DeepSeek-V3	62.43 (13.09)	62.23 (−0.20) (13.14 (+0.05))	63.99 (+1.56) (13.40 (+0.31))	60.47 (−1.96) (10.68 (−2.41))	61.45 (−0.98) (11.05 (−2.04))

Table 1: Comprehensive robustness evaluation results. All values are averages and **red** indicates improvement.

Diagnostic Accuracy Assuming N samples in total, the accuracy for each sample is set to 1 if its diagnosis d equals to the ground truth d^*, while accuracy is 0, otherwise. The diagnostic accuracy $\mathcal{A}^{diag}$ can be represented as follows:

$$\mathcal{A}^{diag} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(d_i = d_i^*) \tag{1}$$

Observation Completeness Firstly, we define a semantic equivalence operator, denoted as $\equiv_{sem}$. The observations are supposed to be the critical information for diagnosis. Observation o (generated by LLMs) and o^* (the ground truth) are equal, i.e., $o \equiv_{sem} o^*$, if they refer to the same clinical entity or phenomenon. A sample usually consists of multiple observations. Thus, let $O = \{o_1, o_2, ..., o_i\}$ be the prediction and O^* be the ground truth for each sample. To assess the observation completeness, we calculate its precision O^{pre} and recall O^{rec}, as follows:

$$\mathcal{O}^{pre} = \frac{|\mathcal{O} \cap \mathcal{O}^*|}{|\mathcal{O}|}, \mathcal{O}^{rec} = \frac{|\mathcal{O} \cap \mathcal{O}^*|}{|\mathcal{O}^*|} \tag{2}$$

Reasoning Faithfulness We employ a stricter criterion that requires the rationales to provide an explicit linking of the observations to the diagnosis. Similar to observation completeness, a predicted rationale z equals to the ground truth z^*, i.e., $z \equiv_{sem} z^*$, only if it conveys the same reasoning logic throughout the diagnostic process. $\varepsilon = (o, z, d)$ is defined as a triplet for the observation, rationale, and diagnosis, while $\varepsilon^* = (o^*, z^*, d^*)$ is the corresponding ground truth. A prediction is deemed faithful only if the triplet is aligned perfectly with the ground truth. Let $m(\varepsilon, \varepsilon^*)$ be the count of correctly matched triplets, which can be represented as: $m(\varepsilon, \varepsilon^*) = \{o \equiv_{sem} o^*, z \equiv_{sem} z^*, d = d^*\}$. The reasoning faithfulness F^{all} is then computed as the coverage of ground truth:

$$\mathcal{F}^{all} = \frac{m(\varepsilon, \varepsilon^*)}{|\mathcal{O} \cup \mathcal{O}^*|} \tag{3}$$

which measures the density of valid reasoning chains relative to the total observations. It is used to penalize models when generating the correct diagnosis through incorrect evidence or flawed logic.

3 Experiment

3.1 Experimental Setup

Datasets DiReCT dataset (Wang et al., 2024) features a collection of human physician-annotated, fine-grained clinical reasoning chains. It comprises 511 annotated clinical notes across five medical domains, sampled from a publicly available database, MIMIC-IV (Johnson et al., 2023).

Implementation Details To ensure deterministic outputs for rigorous reasoning analysis, the generation temperature was set to 0 with a greedy decoding strategy. In contrast, the temperature for micro-pruning and noise injection is set to 0.7, aiming to obtain linguistic diversity in the adversarial examples. The evaluation pipeline processed a total of ~28,000 inferences across 511 clinical notes.

3.2 Experimental Results

We firstly conduct the macro-pruning process to find the most important records, then evaluate the metrics after micro-pruning and noise injection and compare the results with a baseline, the prior work of Wang et al. (2024). As shown in Table 1, all models for the baseline have a high diagnostic accuracy, but with low reasoning faithfulness, indicating

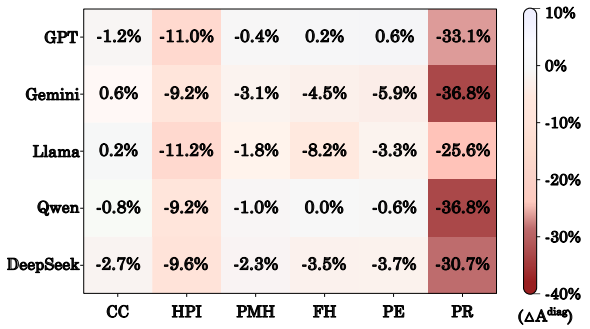

Figure 3: Heatmap of macro-pruning result. Values denote the percentage point change (ΔA^{diag}) in diagnostic accuracy relative to the Baseline.

Model	Baseline $\mathcal{O}^{pre}$ / ($\mathcal{O}^{rec}$)	HPI Deleted	PR Deleted
GPT	36.96 (69.74)	36.25 **(69.86)**	36.34 (69.15)
Gemini	39.91 (64.16)	**42.05** (54.58)	**42.44** (54.58)
Llama	29.97 (66.36)	**30.28** (65.80)	**30.51** **(67.32)**
Qwen	48.67 (47.93)	**49.04** **(49.03)**	**49.39** (47.93)
DeepSeek	45.73 (52.99)	45.49 (51.80)	45.08 (51.62)

Table 2: Observation completeness under micro-pruning. **Bold** are better performance than baseline.

Model	Baseline $\mathcal{O}^{pre}$ / ($\mathcal{O}^{rec}$)	HPI Modified	PR Modified
GPT	**36.96** (69.74)	36.49 (70.33)	35.31 (66.56)
Gemini	39.91 **(64.16)**	42.46 (56.58)	43.23 (57.67)
Llama	**29.97** (66.36)	29.82 (65.55)	29.78 (66.50)
Qwen	**48.67** (47.93)	47.83 (48.49)	48.21 (48.03)
DeepSeek	**45.73** **(52.99)**	43.47 (50.70)	44.96 (52.50)

Table 3: Observation completeness under noise injection. **Bold** indicates better performance in baseline.

that only focusing on diagnostic accuracy cannot obtain faithful diagnosis. In the following, we will detail the experimental results of the framework.

3.2.1 Results of Macro Pruning

In macro-pruning, each record is masked iteratively to evaluate models' attentional dependency. Figure 3 illustrates the deviation in diagnostic accuracy with and without pruning records. It shows that PR record shows the most severe impact with accuracy decreasing by at least 30%, while HPI record shows the second most significant impact, with a reduction of approximately 10%. Therefore, these two records contribute the most among the six records, which is the base for micro-pruning and noise injection.

3.2.2 Results of Micro Pruning

In Micro-pruning, only 1 to 3 units are deleted from HPI and PR record. Though ($\mathcal{A}^{diag}$) and ($\mathcal{F}^{all}$) remain marginal change for most scenarios, the results of pruning PI record exhibits consistent increase for both metrics on most models, which can prove that there is inherent attention limitations in LLMs when processing long contexts. We also evaluate the observation completeness metrics for micro-pruning. In Table 2, it can be seen that for most LLMs, both ($\mathcal{O}^{pre}$) and ($\mathcal{O}^{rec}$) consistently increase after pruning, which further validates the attention limitation problem.

3.2.3 Results of Noise Injection

The noise injection is also conducted for the record of HPI and PR. Table 1 shows the results of diagnosis accuracy and reasoning faithfulness. The performance of two metrics both decrease for almost all models. Notably, the decrease in reasoning faithfulness is even greater than in diagnostic accuracy. It indicates that the reasoning chain of models is easily interfered even only syntactic noise and benign medical noise are injected. The results of observation completeness, as shown in Table 3, provide more detailed evidence why reasoning fails. It can be seen that both the metrics on all models are worse than the baseline, demonstrating models' inadequate anti-interference capability.

4 Conclusion

It has been exhibited that the reasoning remains unfaithful even when the diagnostic results are correct. In this study, we explore the reasoning robustness of medical LLMs through an adversarial perturbation framework, in which semantic pruning removes redundancy while noise injection modifies the clinical notes. The experiments on semantic pruning verify the attention limitation of LLMs, and noise injection results indicate the fragile anti-interference capability in reasoning process. Therefore, our work identifies critical reasoning vulnerabilities in medical LLMs, providing a feasible way toward trustworthy LLMs in healthcare.

Limitations

While our work provides a comprehensive dissection of LLM reasoning behaviors under perturbation, several limitations remain, which are discussed below.

Scope of Evaluation Data Our analysis relies primarily on the DiReCT dataset. Although this benchmark spans five major medical specialties, the samples are still insufficient. Consequently, our findings only reflects the situation in certain specific fields, and may not fully extrapolate to the noisy, multimodal realistic hospital environments.

Proxy Nature of Perturbations Furthermore, the adversarial perturbations framework utilize the clinical noise that is provided by LLMs. The models are demanded to generate contents with limited syntactic coherence, which yet still remain substantially different from the realistic clinical notes that usually suffer from uncurated syntactic degradation. Therefore, the results of our evaluation may be upper bounds for the realistic scenarios.

Static Reasoning Limitation Our evaluation framework regards the diagnosis as a static, single-turn inference task. It can evaluate the reasoning capability on fixed inputs but ignoring the interactive nature of clinical practice. To be more practical, the models should be equipped with the ability of active information seeking or iterative hypothesis refinement. Multi-turn interactive diagnostic scenarios are the research direction for the next step.

Ethical considerations

Model Deployment Risks Our analysis reveals that even advanced LLMs have limited attention and anti-interference, thereby suffering performance collapse under too much noise. Thus, we urge practitioners to exercise extreme caution and maintain rigorous human oversight when deploying LLMs in diagnostic workflows. It is very risky to put too much trust in the current LLMs, especially in the medical domain.

Data Privacy We confirm that all data used is fully anonymized and we have adhered to the PhysioNet data use agreement.

Defensive Intent The adversarial framework is designed for stress-testing and theoretical analysis, aiming to facilitate the development of robust LLMs, rather than to malicious attack on medical AI systems.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Sijia Chen, Xiaomin Li, Mengxue Zhang, Eric Hanchen Jiang, Qingcheng Zeng, and Chen-Hsiang Yu. 2025b. Cares: Comprehensive evaluation of safety and adversarial robustness in medical llms. *arXiv preprint arXiv:2505.11413*.

Pat Croskerry. 2009. A universal model of diagnostic reasoning. *Academic medicine*, 84(8):1022–1028.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. 2019. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Robert E Hirschtick. 2006. Copy-and-paste. *Jama*, 295(20):2335–2336.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have. *A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv [cs. CL]*.

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

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226.

Shunsuke Kitada and Hitoshi Iyatomi. 2021. Attention meets perturbations: Robust and interpretable attention with adversarial training. *IEEE Access*, 9:92974–92985.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Conference on health, inference, and learning*, pages 578–597. PMLR.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Dongfang Li, Jindi Yu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023. Explaincpe: A free-text explanation benchmark of chinese pharmacist examination. *arXiv preprint arXiv:2305.12945*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, and 1 others. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

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

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

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*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Lujia Shen, Yuwen Pu, Shouling Ji, Changjiang Li, Xuhong Zhang, Chunpeng Ge, and Ting Wang. 2023. Improving the robustness of transformer-based large language models with dynamic attention. *arXiv preprint arXiv:2311.17400*.

Ayush Singh, Navpreet Singh, and Shubham Vatsal. 2024. Robustness of llms to perturbations in text. *arXiv preprint arXiv:2407.08989*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in neural information processing systems*, 37:74999–75011.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. 2023. Adversarial self-attention for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13727–13735.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Kun Zhang, Le Wu, Kui Yu, Guangyi Lv, and Dacao Zhang. 2025. Evaluating and improving robustness in large language models: A survey and future directions. *arXiv preprint arXiv:2506.11111*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and 1 others. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv e-prints*, pages arXiv–2306.

A Detailed Experimental Setup

A.1 Adversarial Perturbation Prompts

To ensure the reproducibility of our perturbation manipulations, we provide the full system prompts used by the GPT-4o agent. The prompt of micro-pruning and noise injection shows in Table 4 and Table 5

System Prompt for Micro Pruning
Suppose you are a professional clinical doctor. Your task is to **delete** 1 to 3 (can be adjusted by yourself according to the volume of content but must not exceed the limit) descriptions from the single input field named {target_input} of a patient's medical record to test model robustness. **Critical constraints**: - Modify content only in the field {target_input}. Do NOT mention or modify any other field (e.g., input1, input3, input5, etc.). - Perform ONLY 1 to 3 deletion operations in {target_input}. - If {target_input} has insufficient content to delete, return it unchanged and note "skipped" in the change log. - Delete self-contained units such as: a full sentence of description/observation, a clause containing a measurement indicator and its result, or a discrete item in a list (keep surrounding punctuation/format coherent). - Keep the remaining content complete, clinically coherent, and logically rigorous after deletion. - Do NOT delete the key descriptions that can lead to a change in sample's orientation of the disease. - Do NOT add new information or rewrite remaining content after deletion, so you should fulfill the requirements mentioned above before your deletion. Now, provide your response strictly in this format: Modified text: <the full rewritten version> Original input text: {content}

Table 4: The full prompt template used for the Micro Pruning.

System Prompt for Noise Injection
Suppose you are a professional clinical doctor. Your task is to **rephrase** only the following input field of a patient's medical record to test model robustness. The rewritten text must remain **clinically equivalent** to the original in meaning and diagnosis, while applying only the two types of modifications described below. 1. **Formatting modifications**: - Reorder list items or entries without changing their content meaning. - Adjust punctuation, indentation, or line breaks for stylistic variety. - Change enumeration or bullet formatting while keeping all information intact. 2. **Add content that is relevant to the context but does not affect the final diagnosis**: - Add minor, medically plausible but diagnostically irrelevant observations - These additions must be logically consistent and must never contradict, modify, or imply new diagnostic information. **Critical constraints**: - Modify content only the field named {target_input} in the provided JSON file. - Do not alter any other field, key, or structure. - Only 2 to 5 modification operations mentioned above can be performed in each input. - Preserve all medical facts, terminology, and diagnostic details. - Use professional clinical English. - Output only the modified text and a concise change log. Now, provide your response strictly in this format: Modified text: <the full rewritten version> Change log: <bullet-pointed list briefly describing what was changed and where> Original input text: {content}

Table 5: The full prompt template used for the Semantic Noise Injection.

A.2 Evaluation Protocols

Instead of designing new judgment criteria, we strictly adhered to the evaluation framework proposed by the DiReCT (Wang et al., 2024) benchmark to ensure comparability with prior work.

- Judge Model: Compared to the Llama-3-8B-Instruct model used in the original setup, we switched to the GPT-4.1-mini model to improve the accuracy of the evaluation.

- Observation Matching: To determine whether a predicted observation matches the ground truth, we utilize the zero-shot prompt from the observation matching function in the original setup. This prompt instructs the judge to decide if two text descriptions refer to the same clinical finding (Output: "Yes"/"No").

- For rationale evaluation, we use the few-shot prompt from the reason evaluation function in the original setup. This prompt includes five curated examples (samples of "Yes" and "No" pairs) to guide the judge in discriminating whether two reasoning paths explain a similar medical diagnosis premise.

Model	Diagnostic Accuracy		Observation Completeness			Reasoning Faith.	
	$\mathcal{A}^{cat}$	$\mathcal{A}^{diag}$	$\mathcal{O}^{pre}$	$\mathcal{O}^{rec}$	$\mathcal{O}^{cov}$	$\mathcal{F}^{obs}$	$\mathcal{F}^{all}$
GPT-5-mini	$\mathbf{0.875_{\pm 0.054}}$	$\mathbf{0.765_{\pm 0.038}}$	$0.370_{\pm 0.146}$	$\mathbf{0.697_{\pm 0.164}}$	$0.363_{\pm 0.151}$	$0.333_{\pm 0.243}$	$0.127_{\pm 0.097}$
Gemini-2.5-Flash	$0.847_{\pm 0.081}$	$0.708_{\pm 0.083}$	$0.399_{\pm 0.170}$	$0.642_{\pm 0.182}$	$0.375_{\pm 0.166}$	$\mathbf{0.340_{\pm 0.253}}$	$0.140_{\pm 0.130}$
Llama-3.3-70B	$0.806_{\pm 0.105}$	$0.710_{\pm 0.087}$	$0.300_{\pm 0.133}$	$0.664_{\pm 0.167}$	$0.292_{\pm 0.135}$	$0.220_{\pm 0.213}$	$0.074_{\pm 0.080}$
Qwen3-235B	$0.781_{\pm 0.115}$	$0.626_{\pm 0.084}$	$\mathbf{0.487_{\pm 0.213}}$	$0.479_{\pm 0.211}$	$\mathbf{0.391_{\pm 0.206}}$	$0.332_{\pm 0.290}$	$\mathbf{0.144_{\pm 0.144}}$
DeepSeek-V3	$0.798_{\pm 0.127}$	$0.624_{\pm 0.072}$	$0.457_{\pm 0.193}$	$0.530_{\pm 0.207}$	$0.387_{\pm 0.184}$	$0.309_{\pm 0.268}$	$0.131_{\pm 0.127}$

Table 6: Comprehensive baseline performance report. We present the mean scores and standard deviations ($\pm$SD) across all five functional layers of the evaluation pipeline using the original, unperturbed clinical notes. This table details the trade-offs between diagnostic accuracy ($\mathcal{A}^{diag}$), evidence extraction precision/recall ($\mathcal{O}^{pre}/\mathcal{O}^{rec}$), and reasoning faithfulness ($\mathcal{F}^{all}$), providing a holistic view of each model's baseline capabilities.

B Extended Performance Metrics

To provide a comprehensive view of model behavior, we report detailed performance breakdowns across medical domains and additional fine-grained metrics that were omitted from the main text for brevity.

B.1 Overview of Experimental Results

Table 6 provides a granular view of the models' baseline capabilities across all evaluation dimensions.

Although GPT-5-mini leads in top-level accuracy and demonstrates exceptional stability($A^{cat} \approx 87.5\%$, $A^{diag} \approx 76.5\%$), open-source weight models like Qwen have shown competitive performance on certain metrics and even surpassed it in some cases.

A critical observation is the pervasive Precision-Recall Gap. For instance, Llama-3.3 achieves a respectable recall of 66.4%, but a precision of only 30.0%. This indicates that current models, in their attempt to be comprehensive, fail to effectively filter out noise—extracting nearly 70% irrelevant information alongside valid evidence. This inefficiency in information filtering underscores the motivation for our micro-pruning.

B.2 Metrics Under Micro Pruning

As shown in Table 7, trimming content from both blocks resulted in improvements across nearly all model metrics, confirming that the performance gains stem from a higher signal-to-noise ratio in evidence retrieval.

While the absolute gains are subtle, the trend is significant: removing noise from objective data sections enables the model to capture more valid evidence, not less. This validates our hypothesis that dense, redundant clinical text actively suppresses the model's retrieval attention heads.

Model: Llama-3.3-70B			
Setting	Observation Completeness		
	$\mathcal{O}^{pre}$	$\mathcal{O}^{rec}$	$\mathcal{O}^{cov}$
Baseline	$0.300_{\pm 0.133}$	$0.664_{\pm 0.167}$	$0.292_{\pm 0.135}$
HPI Deleted	$0.303_{\pm 0.136}$	$0.658_{\pm 0.166}$	$0.294_{\pm 0.137}$
PR Deleted	$\mathbf{0.305_{\pm 0.132}}$	$\mathbf{0.673_{\pm 0.169}}$	$\mathbf{0.298_{\pm 0.133}}$

Table 7: The observation completeness of Llama in micro-pruning setting. Extraction quality metrics for *Llama-3.3-70B* under redundancy pruning. Boost in Precision when filtering the Pertinent Results.

B.3 Metrics Under Noise Injection

Compared to micro-pruning, which improve a model's diagnostic performance, noise injection introduce irrelevant content into the data text during the initial processing phase. This content does not aid reasoning or diagnosis and alters the original document format, making it more difficult for the model's attention to focus on key information. Ultimately, this leads to a decline in its diagnostic effectiveness.

Model: GPT-5-mini			
Setting	Observation Completeness		
	$\mathcal{O}^{pre}$	$\mathcal{O}^{rec}$	$\mathcal{O}^{cov}$
Baseline	$\mathbf{0.370_{\pm 0.146}}$	$0.697_{\pm 0.164}$	$\mathbf{0.363_{\pm 0.151}}$
HPI Modified	$0.365_{\pm 0.141}$	$\mathbf{0.703_{\pm 0.166}}$	$0.358_{\pm 0.145}$
PR Modified	$0.353_{\pm 0.156}$	$0.666_{\pm 0.209}$	$0.344_{\pm 0.158}$

Table 8: The Observation Completeness of GPT-5 in noise injection setting. Extraction quality metrics under semantic perturbation. Injecting noise to Pertinent Results distracts the model will cause a drop in Recall.

C Qualitative Case Studies

C.1 The Benefit of Redundancy Pruning

Table 9 provides a side-by-side comparison of Llama-3.3-70B's reasoning process before and after applying Micro-Pruning to remove non-contributory content from Pertinent Results.

Notably, to prevent the model from over-pruning during semantic pruning—which could result in the deletion of excessive content or critical indicators—we have implemented strict requirements for micro-pruning and restricted the number of pruning iterations to a very small range.

- Clinical Context: The patient presented with abdominal pain. The ground truth diagnosis is Acute Gastritis, supported by the endoscopic finding of "superficial erosion".
- Original Failure: In the baseline setting, the inputs contained a normal Lipase level (Lipase-23, normal range). The model hallucinated a more severe diagnosis (Gastric Ulcers).
- Correction via Pruning: By removing the single irrelevant Lipase entry, the model's extraction mechanism became sharpened. It focused on the "soft abdomen" and "erosion," and correctly predicted Acute Gastritis.

Ideally, pruning a lab result in Pertinent Results should not affect symptom extraction from other records. However, our case reveals a global attention dependency. The presence of the irrelevant Lipase entity appears to trigger a distraction effect, causing the model to allocate attention budget to processing this decoy. This dilutes its processing capability for other parts, resulting in a coarser, 'copy-paste' extraction strategy for the subjective complaints.

Once the decoy is removed, the model's attention map sharpens. It no longer over-extracts narrative noise from other records and successfully locks onto the definitive endoscopic evidence. This suggests that redundancy in one record can induce hallucinations or precision drops in globally distinct records, a critical insight for long-context clinical modeling.

C.2 The Fragility of Semantic Grounding

Table 10 demonstrates the instability of Gemini-2.5-Flash when distinguishing between heart failure subtypes.

- Clinical Context: The patient has heart failure features (high proBNP, edema) but a preserved ejection fraction (LVEF>55%), confirming a diagnosis of HFpEF.
- Baseline Success: The model successfully extracts the critical LVEF>55% value and correctly classifies the condition as HFpEF.
- Modification Failure: When benign narrative noise (e.g., "*The patient was comfortable*") is injected into the Pertinent Results records, the model fails to extract the LVEF value—the single most important metric for this classification. Instead, it over-interprets secondary signs like "Moderate cardiomegaly" and erroneously infers a reduced ejection fraction (HFrEF).

The semantic interpretation of a quantitative metric like LVEF > 55% should be robust to adjacent, clinically neutral narrative text. However, this failure case exposes a fragility in the model's information prioritization mechanism.

The injected noise acts as a semantic distractor, it alters the local context window of the finding section. We hypothesize that the model, when encountering this 'easier-to-process' natural language sentence, shifts its attention heads away from the dense, numerical LVEF data.

Consequently, without the explicit LVEF anchor, the model falls back on a heuristic association. This fallback logic is statistically common but clinically incorrect in this specific instance. This highlights that robustness in clinical LLMs is not just about resisting adversarial attacks, but about maintaining robustness against noise.

D Related Works

LLMs like Med-PaLM 2 (Singhal et al., 2023) and GPT-4 (Nori et al., 2023) have demonstrated expert-level performance on medical QA tasks. Recently, researchers have turned to adversarial mechanisms to evaluate complex clinical reasoning capabilities, involving the attentional stability (Wu et al., 2023; Kitada and Iyatomi, 2021; Shen et al., 2023) and the anti-interference capability(Chen et al., 2025b; Zhang et al., 2025).

D.1 Attention Limitation

LLMs is expected to handle not only QA tasks with short contexts but extensive patient histories effortlessly. However, "Lost in the Middle" (Liu et al.,

Baseline Input (w/ Redundancy)	Pruned Input
... ___ 01:00PM BLOOD Lipase-23 ← *Irrelevant Normal Lab* Endoscopy: Superficial erosion of the antrum	... *[Line Removed]* Endoscopy: Superficial erosion of the antrum
Extracted Observations	**Extracted Observations**
- "Abdominal Pain" - - "She feelike the LLQ is more sharp..." - - "Slight nausea with food..." - "Endoscopy: Superficial erosion"	- "Abdominal Pain" - "ABDOMEN - Soft, obese..." **- "Endoscopy: Superficial erosion"**
Diagnostic Prediction	**Diagnostic Prediction**
✗ Gastric Ulcers *(Implies deep tissue defect)*	**✓ Acute Gastritis** *(Matches superficial erosion)*

Table 9: The result demonstrated the impact of micro-pruning. Removing irrelevant lab data (Lipase) shifts the model's attention from subjective complaints back to objective endoscopic evidence.

Baseline Input	Noise Injected Input (w/ Redundancy)
... ___ CXR: ... (LVEF>55%). IMPRESSION: Mild interstitial edema.	... ___ CXR: ... (LVEF>55%). ... The patient was comfortable during the imaging procedure. IMPRESSION: Mild interstitial edema.
Extracted Observations	**Extracted Observations**
- "proBNP-7232*" - "Mild interstitial pulmonary edema" - "The left atrium is moderately dilated" **- "LVEF>55%"** ✓ *(Crucial Evidence)*	- "proBNP-7232*" - "Mild interstitial pulmonary edema" - "Moderate cardiomegaly" *[Missed LVEF>55%]*
Diagnostic Prediction	**Diagnostic Prediction**
✓ HFpEF (*Heart Failure w/ Preserved EF*)	**✗ HFrEF** (*Heart Failure w/ Reduced EF*)

Table 10: The result demonstrated the impact of noise injection. Injecting neutral sentences distracts the model from the critical LVEF value, leading to a misclassification of Heart Failure subtype.

2024) indicates that models' attention struggles to prioritize information in long sequences. This limitation is further substantiated by the LooGLE benchmark (Li et al., 2024), which reveals that current models fail to sustain complex reasoning capabilities as the input context expands. Similarly, Levy et al. (2024) found that LLMs often regard task-irrelevant text as valid signals. These findings are quantified by Hsieh et al. (2024), who demonstrate empirically a large gap between the claimed and the effective context length. Our finding in clinical diagnosis also aligns with these insights. We adopt multiple LLMs and conduct evaluation after semantic pruning. It can be seen that the performance for both diagnosis and reasoning would be improved after removing redundant informaition.

D.2 Anti-Interference Capability

Classic approaches generally utilize character-level noise or synonym substitution (Jia and Liang, 2017) to evaluate anti-interference capabilitie, but it is insufficient for complex scenarios, such as clinical diagnosis. For example, Zhu et al. (2023) proposed a framework to test the robustness of LLMs

against prompt variations, but only limited the text modification operations to simple character-level noise like structural perturbations. Recently, Wang et al. (2021) presents Adversarial GLUE, a new multi-task benchmark to quantitatively evaluate the vulnerabilities of LLMs under various types of adversarial attacks. A "human-craft" method is introduced to generate more complex and challenging noises. They demonstrated that the model's anti-interference capability is severely lacking when confronted with complex noise. However, it is significant to ensure the anti-interference of reasoning process in medical domain. In this work, thus, we propose a clinical noise injection strategy to simulate realistic clinician behaviors, aiming to explore models' reasoning robustness.