Evaluating Faithfulness and Consistency of LLMs in Medical Diagnoses

Hima Kammachi hkammachi@umass.edu
Khushei Meghana Meda kmeda@umass.edu
Rohit Bangalore Vijaya rohitbangalo@umass.edu

1 Problem statement

Large language models (LLMs) have shown strong capabilities in natural language understanding, yet their reliability in high-stakes domains such as medicine remains uncertain. Ensuring that model outputs are faithful to the input and self-consistent across variations is critical, as errors in medical reasoning can have serious consequences. In this work, we aim to evaluate the faithfulness (measures how well the model's output reflects the information actually relevant in the input — not what is plausible or assumed) and self-consistency (their ability to provide stable, semantically identical responses) of LLMs in medical question-answering tasks. We hypothesize that while LLMs may produce correct answers under standard conditions, their reasoning can degrade when inputs are modified, augmented with irrelevant information, or paraphrased. We further hypothesize that structured interventions, such as focused attention and chain-of-thought prompting, can improve model faithfulness by guiding attention to relevant evidence.

To test these hypotheses, we design experiments simulating challenging clinical reasoning scenarios, including irrelevant information overload, false premises, and input sensitivity. These tasks emulate the rigor of medical oral examinations, assessing whether models rely on medically relevant cues and maintain consistent reasoning under varied input conditions.

2 What you proposed vs. what you accomplished

- **Dataset Collection and Preprocessing:** We proposed to use the MedQA (English subset) and Complete Medical Symptom datasets, filtering for incomplete entries and formatting and adding scenarios to them for QA. – **COMPLETE**
- **Automated Faithfulness Experiments:** We proposed running specific adversarial tests including "Irrelevant Information Overload," "False Premises/Incorrect Facts," and "Leading Questions" to test model robustness. – **COMPLETE**
- **Consistency Testing (Paraphrasing):** We proposed paraphrasing to generate semantically equivalent questions to test if model answers remain consistent across input variations. – **COMPLETE**
- **Advanced Reasoning Tests (Causal & Attribution):** We proposed "Causal Direction Consistency" (symptoms $\rightarrow$ diagnosis vs. diagnosis $\rightarrow$ symptoms) and a manual "Input Attribution Test" (iteratively removing symptoms). – **COMPLETE**
- **Mitigation Strategies:** We proposed implementing "Focused Attention Chain-of-Thought (CoT)" prompting to improve faithfulness and reduce hallucinations. – **PARTIALLY COMPLETE** [We were able to test the fix for a few tests but could not complete the tests on others because the initial scenario generation and experiments took considerable time].
- **Multi-Model Evaluation:** We proposed evaluating across both closed-source (Gemini) and open-source (Gemma, Qwen) models using accuracy, precision, confidence intervals, and paired comparisons. – **COMPLETE**

3 Related Work

Faithfulness and robustness of large language models (LLMs) have received increasing attention, particularly in healthcare settings where unreliable reasoning can have high-stakes consequences. Prior work has shown that conventional

evaluation based on accuracy alone is insufficient to assess whether models rely on medically relevant evidence or produce stable and trustworthy decisions.

Several surveys and position papers highlight fundamental limitations in existing faithfulness evaluation practices. Xie et al. (2023) provide a systematic review of faithful AI in medicine, emphasizing that LLMs often exhibit overconfidence and that commonly used evaluation metrics fail to capture causal dependence on clinical evidence. More recently, Aljohani et al. (2025) survey trustworthiness concerns of LLMs in healthcare, identifying robustness, explainability, and faithfulness as key unresolved challenges.

A related line of work focuses on explanation faithfulness and attribution-based evaluation. Huang et al. (2023) study LLM-generated self-explanations and show that such explanations often diverge from attribution-based methods, raising concerns about their reliability. Parcalabescu and Frank (2023) propose self-consistency–based measures for evaluating explanations, demonstrating that alignment between predictions and explanations is not guaranteed. Earlier work by Jacovi and Goldberg (2020) formalizes the concept of faithfulness and argues that explanations should be evaluated based on causal influence rather than plausibility.

Several studies investigate self-consistency and prompting-based approaches as proxies for faithful reasoning. Madsen et al. (2024) evaluate whether self-explanations generated by LLMs are faithful and conclude that explanations can be misleading even when predictions are correct. In the biomedical domain, Fang et al. (2024) show that reasoning and faithfulness metrics often diverge from accuracy, reinforcing the need for intervention-based evaluation.

Mitigation strategies using prompting have also been explored. Fayyaz et al. (2024) compare prompting-based rationales with attribution-based explanations and find that structured prompting can influence model decisions but does not consistently guarantee faithful reasoning. Related work on multiple-choice evaluation further highlights structural biases introduced by prompt formats and answer constraints (Balepur et al., 2025). In contrast to prior work, our approach combines robustness testing through paraphrase consistency and adversarial perturbations with attribution-based redaction to jointly assess prediction stability, confidence modulation, and explanation faithfulness in a medical question-answering setting. By evaluating multiple model families on the MedQA dataset (Jin et al., 2021), our work provides a unified empirical analysis of when robustness and faithfulness improve and where they remain fundamentally limited.

4 Your dataset

To evaluate the faithfulness and consistency of Large Language Models (LLMs) in medical diagnostics, we utilize two primary datasets: MedQA and the Complete Medical Symptom Dataset. These datasets were selected for their depth in clinical scenarios and their ability to benchmark reasoning capabilities in high-stakes medical contexts.

Med QA We employ the MedQA dataset (Jin et al., 2021), a multilingual, open-domain medical question-resolving dataset. MedQA is designed to benchmark models on real-world clinical reasoning tasks, using questions sourced from medical board examinations in the United States, Mainland China, and Taiwan. The data set comprises over 60,000 multiple-choice questions in three languages: English (12,723), Traditional Chinese (34,251), and Simplified Chinese (14,123). Each question is paired with four answer options, and the questions cover a broad spectrum of medical specialties, including internal medicine, surgery, pediatrics, and pharmacology. We will be using the English dataset for our project.

Example data:


```
{"question": "A 23-year-old pregnant
↪  woman at 22 weeks gestation presents
↪  with burning upon urination....Which
↪  of the following is the best
↪  treatment for this patient?",
↪  "answer": "Nitrofurantoin",
↪  "options": {"A": "Ampicillin", "B":
↪  "Ceftriaxone", "C": "Ciprofloxacin",
↪  "D": "Doxycycline", "E":
↪  "Nitrofurantoin"}, "meta_info":
↪  "step2&3", "answer_idx": "E"}
```


Complete Medical Symptom Dataset We also use the Complete Medical Symptom Dataset, a large-scale clinical reasoning dataset available on Hugging Face [dataset]. The dataset contains approximately 1.3 million clinical cases, each describing patient symptoms along with associated medical information such as demographics and diagnoses. Its explicit symptom–diagnosis structure

makes it suitable for evaluating self-consistency and causal direction consistency in medical reasoning.

Example data:


```
{"symptoms":
[ "migraines headaches", " can't sleep",
↪  " whole body shaking shivering", "
↪  feel dizzy sometimes", "" ]
"label_text":[ "drug reaction" ]}
```


4.1 Data Preprocessing

To ensure the reliability of our faithfulness evaluation, we performed a series of cleaning and augmentation steps on the raw data.

Cleaning and Filtering: We processed the raw **MedQA** and **Complete Medical Symptom** datasets to remove entries with incomplete clinical histories or inconsistent metadata. This refinement was necessary to establish a high-quality baseline for our reasoning experiments.

Adversarial Augmentation: To facilitate our faithfulness stress tests, we expanded the dataset by appending new columns containing adversarial variations of the original questions. These augmentations included:

- **Irrelevant Information Injection:** We systematically added extraneous details (e.g., patient clothing, household activities) to the clinical vignettes to test for "information overload."
- **False Premise Injection:** We embedded factually incorrect physiological data (e.g., impossible blood pressure readings) to evaluate the model's ability to detect inconsistencies.
- **Paraphrasing:** We generated semantically equivalent versions of the questions, allowing us to test for self-consistency across different phrasings.

Causal Direction Augmentation: We augmented the dataset using the Complete Medical Symptom Dataset to generate reverse-formulation queries by mapping diagnoses to their associated symptom sets. This preprocessing step produced paired forward and reverse question instances for downstream causal direction consistency experiments.

Example data after adversarial augmentation:


```
{"question":"A 23-year-old pregnant
↪  woman at 22 weeks gestation presents
↪  with burning upon urination. ... The
↪  hospital cafeteria is serving
↪  shepherd's pie today, and the
↪  hallway is unusually noisy with the
↪  sounds of a visiting children's
↪  choir. She's been trying to learn
↪  how to knit, ... She's been
↪  meticulously tracking her calorie
↪  intake using a fitness app. ...
↪  She's been reading a lot about the
↪  benefits of meditation. Which of the
↪  following is the best treatment for
↪  this patient?","options": {"A":
↪  "Ampicillin", "B": "Ceftriaxone",
↪  "C": "Ciprofloxacin", "D":
↪  "Doxycycline", "E":
↪  "Nitrofurantoin"}, "meta_info":
↪  "step2&3", "answer_idx": "E"}
```


5 Baselines

To establish a performance benchmark for medical reasoning, we evaluated a diverse set of Large Language Models (LLMs), spanning both proprietary (closed-source) and open-weights families.

5.1 Selected Models

We utilized the following models as our primary baselines:

- **Gemini (Google):** We included Gemini to compare performance across different proprietary architectures and safety alignment strategies.
- **Gemma:** To assess the capabilities of open-weights models, we employed Gemma (specifically the instruction-tuned variant). This allows us to evaluate whether accessible, locally hostable models can compete with larger proprietary APIs in high-stakes medical contexts.
- **Qwen:** We included Qwen to evaluate a strong open-weights model trained at scale for instruction following and reasoning. This allows us to assess its faithfulness and consistency under medical false-premise stress tests alongside proprietary and other open models.

5.2 Baseline Implementation and Hyperparameters

We treat these models as zero-shot learners. In the baseline setting, models are presented with the clinical question without any "Chain-of-Thought" instructions or few-shot examples. This provides a raw measure of their native reasoning ability before our specific prompting interventions.

Hyperparameters: To ensure reproducibility and minimize randomness in our faithfulness evaluation, we set the `temperature` to **0.0** (or the lowest possible setting for each API) and used a fixed random seed where applicable. We set a `max_tokens` limit of 512 to allow sufficient space for the diagnosis while preventing excessive hallucination.

Data Splits: We utilized the official train split of the MedQA (USMLE) dataset for the faithfulness evaluations (5 of 6 experiments). No fine-tuning was performed on the training set; strictly pre-trained models were used to reflect a realistic "off-the-shelf" deployment scenario. For the self-consistency evaluation, we instead used the Complete Medical Symptom Dataset.

6 Your approach

Our approach centers on a systematic, "black-box" stress test of Large Language Models to quantify their faithfulness and self-consistency in medical settings. Unlike traditional benchmarks that solely measure accuracy, our framework evaluates *how* models arrive at their answers by perturbing inputs and observing shifts in reasoning.

6.1 Experimental Design

We developed a modular evaluation pipeline that ingests clinical questions from the MedQA and complete medical symptoms dataset and subjects them to a series of adversarial transformations. We maintained the original Multiple-Choice Question (MCQ) format, but assessed the models' robustness by manipulating the clinical vignettes provided in the prompt.

6.1.1 False Premise Injection

False Premise Construction: To evaluate faithfulness under incorrect factual assumptions, we augment MedQA questions with deliberately false clinical premises. These include physiologically impossible values (e.g., implausible vital signs), contradictory medical statements, and unsupported factual assertions that conflict with established clinical knowledge, while preserving the original diagnostic question structure. We then compare the generated question with the original question using conditional checks, to filter out bad scenario generations.

Behavioral Evaluation Protocol: Rather than measuring accuracy alone, we assess faithfulness based on how models respond to false premises. Model outputs are analyzed to determine whether the model (i) detects the impossibility of the premise, (ii) refuses to answer when appropriate, and (iii) expresses confidence levels consistent with epistemic uncertainty. Responses are categorized using a rule-based decision framework that distinguishes between refusal, uncertainty-aware reasoning, and confident propagation of false information.

Faithfulness Scoring: Each response is assigned a graded faithfulness score based on the model's behavior. Correct detection of impossibility followed by refusal is classified as *highly faithful*, while detection accompanied by low-confidence reasoning is considered *faithful*. Failure to detect false premises combined with high-confidence answers is labeled as *confabulating*. Intermediate behaviors are categorized as partially faithful or unfaithful. This graded scheme enables fine-grained analysis beyond binary correctness.

6.1.2 Causal Direction Consistency

Bidirectional Question Construction: To evaluate causal stability, we construct paired question instances using the Complete Medical Symptom Dataset. For each clinical case, we generate a forward formulation that maps symptoms to a diagnosis and a reverse formulation that maps a diagnosis to its associated symptom set. These paired instances share the same underlying clinical information but reverse the causal direction of reasoning.

Evaluation Protocol: Models are evaluated independently on both forward and reverse formulations. For the forward pass, standard diagnostic accuracy is measured by comparing predicted diagnoses against ground-truth labels. For the reverse pass, model outputs are analyzed to determine whether the predicted symptom sets are clinically consistent with the original forward formulation. To compare reverse-generated symptoms with the original symptom set, we normalize symptom strings and perform semantic matching using sentence-embedding cosine similarity.

Consistency Measurement: Causal direction consistency is assessed by measuring agreement between forward and reverse predictions for each paired instance. We label each example into one of four outcome categories based on (i) forward diag-

nostic correctness and (ii) reverse symptom consistency. Aggregate agreement rates are reported to quantify the stability of model reasoning under causal inversion, complementing forward-pass accuracy metrics.

6.1.3 Irrelevant Information Overload

Adversarial Sample Construction: To evaluate model robustness against distraction, we construct adversarial samples by injecting clinically irrelevant noise into questions from the MedQA dataset.We utilize an auxiliary LLM to augment clinical vignettes with extraneous details such as patient clothing, weather conditions, or unrelated daily activities, while strictly preserving the original medical facts and symptoms. This process tests the hypothesis that LLMs may rely on spurious correlations rather than causal medical evidence. We validate the semantic preservation of the clinical core by manually reviewing a subset of generated samples.

Evaluation Protocol and Consistency Measurement: Models are evaluated independently on both the original (Q_{orig}) and adversarial (Q_{adv}) formulations. We define *Faithfulness Consistency* as the agreement rate between the model's predictions on these paired inputs. A model is considered faithful for a given instance only if $Prediction(Q_{orig}) = Prediction(Q_{adv})$. We report the aggregate *Faithful Accuracy*, representing the percentage of cases where the model maintains its correct diagnosis despite the presence of irrelevant noise, along with 95% confidence intervals.

Mitigation Strategies: To counteract performance degradation, we implement a **Focused Attention Chain-of-Thought (CoT)** prompting strategy. Unlike standard zero-shot inference, this approach explicitly instructs the model to "think step-by-step" and "identify medically relevant evidence" before selecting an answer, aiming to mitigate the hallucination effects observed in the adversarial tests.

6.1.4 Leading Questions

Adversarial Sample Construction: To assess model susceptibility to confirmation bias and suggestion, we construct adversarial "leading" samples by injecting false causal assumptions or premature diagnostic conclusions into the question stem. As outlined in our methodology, we modify clinical vignettes to include misleading preambles—such as "Given that the patient likely has [Incorrect Diagnosis X]..."—or misstated facts that contradict the symptoms provided. This tests whether the model blindly accepts the suggested premise (bias) or faithfully grounds its reasoning in the reported clinical evidence (verification).

Evaluation Protocol and Consistency Measurement: Models are evaluated on both the neutral original questions (Q_{orig}) and the leading adversarial variants (Q_{lead}). We define *Suggestion Compliance* as the rate at which the model switches its prediction to the incorrect diagnosis explicitly suggested in the prompt. Conversely, *Faithful Robustness* is measured as the percentage of cases where the model rejects the misleading premise and retains the correct diagnosis derived from the original clinical facts (i.e., $Prediction(Q_{lead}) = Prediction(Q_{orig})$).

Mitigation Strategies: To counteract the effect of leading cues, we apply the **Focused Attention Chain-of-Thought (CoT)** strategy. By prompting the model to explicitly "verify the consistency of the suggested diagnosis with the clinical symptoms" before answering, we aim to reduce the rate of suggestion compliance and enforce evidence-based reasoning.

6.1.5 Paraphrase Consistency Evaluation

Paraphrase Construction: To evaluate robustness to surface-level linguistic variation, we construct paraphrased versions of the clinical question stem while preserving all underlying medical facts and answer options. Paraphrases are generated under explicit constraints that prohibit the introduction of new clinical information, resolution of ambiguities, or changes to diagnostic intent. Only the phrasing of the question text is modified, ensuring semantic equivalence with the original prompt.

Evaluation Protocol: Each question is evaluated in paired form: once using the original question and once using its paraphrased counterpart. Answer options remain unchanged across both queries. Models are prompted to select exactly one answer option (A–D), enabling direct comparison between original and paraphrased predictions.

Consistency Measurement: Paraphrase consistency is assessed by measuring prediction sta-

bility between the original and paraphrased inputs. We additionally track flip-direction outcomes, distinguishing between harmful transitions (correct→incorrect) and beneficial transitions (incorrect→correct). Aggregate stability and flip rates are reported to quantify the sensitivity of model predictions to non-substantive linguistic variation.

6.1.6 Input Attribution and Faithfulness Evaluation

Attribution Signal Extraction: To evaluate the faithfulness of model explanations, we identify words in the question stem that are influential to the model's decision. For open-weight models (Gemma and Qwen), importance signals are derived directly from next-token probability distributions over answer options. For proprietary models (Gemini), which do not expose internal logits, calibrated probability estimates are obtained via structured self-reporting prompts.

Redaction-Based Intervention: All words identified as influential are simultaneously redacted from the question stem and replaced with neutral placeholders, while preserving answer options. The model is then re-evaluated on the redacted input using the same inference settings. This intervention enables causal testing of whether the identified words materially influence model predictions or confidence.

Faithfulness Measurement: Attribution faithfulness is assessed by measuring prediction stability under redaction and changes in the probability assigned to the correct answer. We report stability rates, flip rates, and the change in confidence $\Delta P(correct)$, allowing us to distinguish between features that affect final decisions and those that primarily modulate model confidence.

6.2 Implementation Details

Models and Inference Settings We evaluate four large language models spanning both proprietary and open-weight families: Gemini (Flash family), accessed via API; Gemma-3-27B-IT, an open-weight instruction-tuned model; and Qwen, an open-weight instruction-tuned model. Open-weight models are executed locally, while proprietary models are queried via their respective APIs. For all experiments, models are prompted to select exactly one answer option (A–D) without generating explanations.

Experiments involving open-weight models are conducted on Google Colab Pro using NVIDIA T4 GPUs with high-RAM runtime configurations. We also incorporated timeouts into API executions to avoid getting per minute timeouts.

7 Results

7.1 Paraphrase Robustness

Table 1 summarizes paraphrase robustness results on MedQA across Gemini, Gemma, and Qwen. Gemini demonstrates strong robustness to semantically equivalent reformulations, achieving high prediction stability (86.0%) with only a small accuracy drop of 1.2%. This indicates that Gemini's predictions are largely invariant to surface-level linguistic variation.

In contrast, Gemma exhibits substantially lower stability (68.1%) and a larger accuracy drop (5.5%), suggesting greater sensitivity to paraphrasing. Qwen shows the lowest robustness among the evaluated models, with prediction stability of 50.5% and the largest accuracy drop (7.0%), indicating pronounced sensitivity to changes in question phrasing.

Across all models, correct→incorrect transitions occur more frequently than incorrect→correct transitions, indicating that paraphrasing more often degrades performance rather than correcting errors. These results highlight significant differences in self-consistency across model families, with Gemini exhibiting the strongest paraphrase robustness and Qwen the weakest.

7.2 Input Attribution and Faithfulness

Table 2 reports results for attribution-based redaction experiments. We evaluate whether words identified as influential for a model's prediction are causally important by measuring prediction stability and changes in confidence after redaction. Gemini and Gemma both retain the same predicted answer in over 80% of cases after redaction, achieving prediction stabilities of 82.7% and 83.8%, respectively. Qwen exhibits lower stability (73.5%), indicating a greater tendency for prediction changes when influential words are removed. Despite high prediction stability for Gemini and Gemma, redaction frequently reduces the probability assigned to the correct answer. Positive mean values of $\Delta P(correct)$ across all models indicate that identified influential words contribute to model confidence even when the final prediction remains unchanged. Qwen shows the highest

Table 1: Paraphrase Robustness on MedQA (accuracy, stability, and flip breakdown). Stability confidence intervals are 95% Wilson binomial CIs.

Model	Orig Acc	Pert Acc	Stability	Flip Rate	Stab 95% CI	C→I	I→C
Gemini	71.80%	70.60%	86.00%	14.00%	[0.827, 0.888]	6.20%	5.00%
Gemma	43.70%	38.20%	68.10%	31.90%	[0.651, 0.709]	13.50%	8.00%
Qwen	42.50%	35.50%	50.50%	49.50%	[0.436, 0.574]	18.50%	11.50%

Table 2: Attribution Faithfulness via Redaction (stability and confidence change). $\Delta P(correct) = P_{base}(correct) - P_{redacted}(correct)$.

Model	Stability	Flip Rate	Stab 95% CI	Mean ΔP	% $\Delta P > 0$
Gemini	82.67%	17.33%	[0.785, 0.862]	0.0411	41.07%
Gemma	83.77%	16.23%	[0.828, 0.847]	0.0090	45.62%
Qwen	73.51%	26.49%	[0.667, 0.793]	0.0326	51.35%

proportion of cases with $\Delta P(correct) > 0$, suggesting stronger reliance on the redacted tokens for confidence calibration.

Overall, these findings suggest that word-level attribution captures confidence modulation more reliably than discrete prediction changes, highlighting limitations of redaction-based attribution as a proxy for faithful causal explanation.

7.3 Leading Questions

Table 3 shows that Gemma is highly susceptible to leading questions under baseline prompting, achieving a faithful accuracy of only 46.0%. Applying the Focused Attention CoT strategy yields a substantial improvement to 62.0%, with non-overlapping confidence intervals, indicating that structured reasoning prompts mitigate prompt-induced bias. In contrast to Gemma, Gemini exhibits stronger baseline robustness to leading questions, achieving a faithful accuracy of 66.7% (95% CI $[55.9, 76.0]$). However, it still succumbs to suggestion in roughly one-third of cases, indicating that even capable models can be biased by false premises. The application of the Focused Attention CoT strategy produces a dramatic improvement, boosting faithful accuracy to 95.1% (95% CI $[88.0, 98.1]$). The non-overlapping confidence intervals confirm that this performance gain is statistically significant, suggesting that explicitly prompting Gemini to verify clinical consistency effectively neutralizes its tendency to comply with misleading user cues.

7.4 Information Overload

Table 4 shows that Gemini is largely robust to irrelevant information overload even without mitigation. CoT prompting yields only a modest improvement, with overlapping confidence intervals, suggesting limited additional benefit when baseline robustness is already high. Gemma exhibits a relatively strong baseline faithful accuracy of 75.1% (95% CI $[59.8, 85.8]$), indicating that it is naturally more resistant to distracting noise than to leading suggestions (where it scored only 46.0%). The application of Focused Attention CoT yields a marginal improvement to 77.5% (95% CI $[60.2, 86.9]$). Overall, these results indicate that Chain-of-Thought prompting provides the largest gains in adversarial settings involving biased or misleading cues, while offering smaller improvements when models already demonstrate strong baseline robustness.

Table 3: Leading Question Robustness. Faithful accuracy before and after applying Focused Attention Chain-of-Thought (CoT) prompting.

Setting	Faithfulness Score	95% CI
Gemma	46.0%	[36.6, 55.7]
Gemma with CoT Fix	62.0%	[52.2, 70.9]
Gemini	66.7%	[55.9, 76.0]
Gemini with CoT Fix	95.1%	[88.0, 98.1]

Table 4: Irrelevant Information Overload. Faithful accuracy before and after applying Focused Attention Chain-of-Thought (CoT) prompting.

Setting	Faithfulness Score	95% CI
Gemma	75.1%	[59.8, 85.8]
Gemma with CoT Fix	77.5%	[60.2, 86.9]
Gemini	86.1%	[76.8, 92.0]
Gemini with CoT Fix	88.6%	[79.7, 93.9]

7.5 False Premise Injection.

Table 5 reports the core quantitative metrics for false premise evaluation, while Table 6 summarizes the distribution of faithfulness behaviors. Although both models exhibit comparable baseline diagnostic accuracy on unmodified inputs, their behavior diverges substantially when exposed to incorrect premises. Qwen achieves a markedly higher average faithfulness score (0.367) than Gemma (0.199), indicating a greater ability to detect, mitigate, or appropriately respond to false clinical assumptions.

Behavioral analysis further highlights distinct failure modes. As shown in Table 6, Gemma confabulates in the majority of cases (67.4%), frequently propagating false premises with high confidence. In contrast, Qwen exhibits a more balanced distribution, with a substantially lower confabulation rate (12.6%) and a higher proportion of partially faithful and faithful responses. Supporting metrics in Table 5 show that Qwen flags false

Model	N	Avg Faithfulness	Flag Rate	Same Prediction Rate
Qwen-2.5-3B	707	0.367 [0.355, 0.379]	0.110 [0.089, 0.136]	0.444 [0.408, 0.481]
Gemma	939	0.199 [0.190, 0.208]	0.012 [0.007, 0.021]	0.741 [0.712, 0.768]

Table 5: False premise evaluation metrics with 95% confidence intervals. Avg Faithfulness is computed using bootstrap resampling (10,000 iterations); Flag Rate and Same Prediction Rate use Wilson confidence intervals.

Model	Highly Faithful	Faithful	Partially Faithful	Unfaithful	Confabulating
Qwen-2.5-3B	–	9.6%	44.1%	33.7%	12.6%
Gemma	0.1%	0.4%	29.6%	2.4%	67.4%

Table 6: Distribution of faithfulness behavior categories under false premise injection. Percentages are computed over all evaluated instances per model.

premises more often and is less likely to repeat its original prediction under perturbed inputs, suggesting greater sensitivity to contradictory information. Together, these results demonstrate that robustness to false premises is not captured by accuracy alone and varies significantly across models.

7.6 Causal Direction Consistency.

Table 7 summarizes forward diagnostic accuracy and causal consistency scores for both models. Forward accuracy is comparable between Gemma and Qwen, indicating similar performance on the standard symptom-to-diagnosis task. In contrast, causal consistency scores are substantially lower for both models, demonstrating that correct forward predictions do not reliably translate into stable reasoning when the direction of inference is reversed.

Despite comparable forward accuracy, both models exhibit weak bidirectional agreement, with causal consistency scores below 0.10 on average. The overlapping confidence intervals suggest no statistically significant difference between the models on this metric. These results indicate that causal direction consistency captures a stricter and distinct property of reasoning stability that is not reflected by forward diagnostic accuracy alone.

8 Error Analysis

8.1 Paraphrase-Induced Errors

Paraphrase-induced errors are most prevalent in questions containing long clinical descriptions or subtle distinctions between answer options. In such cases, rewording the question can alter the relative salience of clinical cues, leading to prediction flips even when the underlying medical content is preserved.

This effect is more pronounced for Gemma and

Model	N	Forward Accuracy	Causal Consistency Score
Gemma	122	0.189 [0.123, 0.269]	0.097 [0.076, 0.119]
Qwen-2.5-3B	99	0.202 [0.128, 0.295]	0.088 [0.061, 0.120]

Table 7: Causal direction consistency results with 95% confidence intervals. Forward accuracy measures standard diagnostic performance (symptoms $\rightarrow$ diagnosis), while the causal consistency score measures agreement between forward and reverse (diagnosis $\rightarrow$ symptoms) reasoning.

Qwen than for Gemini, consistent with their lower paraphrase stability scores. These models appear more sensitive to surface-level linguistic variation, suggesting weaker invariance to syntactic and lexical reformulations. In contrast, Gemini demonstrates greater robustness, with fewer harmful correct$\rightarrow$incorrect transitions under paraphrasing.

8.2 Attribution Limitations and Distributed Reasoning

Attribution-based redaction experiments reveal that removing words identified as influential frequently does not change the predicted answer, although it often reduces the probability assigned to the correct option. This indicates that such words contribute more strongly to confidence modulation than to discrete decision changes.

One plausible explanation is distributed reasoning, where model predictions are supported by multiple overlapping contextual cues rather than a small set of critical tokens. Alternatively, these results highlight limitations of word-level redaction as a proxy for causal explanation, as simple removal-based interventions fail to capture token interactions, redundancy, and higher-order dependencies. Taken together, the observed error patterns suggest that robustness failures under paraphrasing and limitations of attribution-based explanations arise from a combination of model sensitivity to linguistic variation and constraints of the evaluation methodology. These findings motivate the use of robustness- and faithfulness-aware evaluation metrics alongside traditional accuracy-based benchmarks, particularly in high-stakes domains such as medical decision-making.

8.3 False premise

The baseline is the accuracy of the models on the original questions from the dataset. Both the models do much better on prognosis questions, maybe because they are pattern based and somewhat direct, and they fail on mechanism-to-diagnosis questions requiring multi-step reason-

ing.

8.4 Causal direction consistency

The baseline for this experiment is the forward diagnostic accuracy of each model on the original symptom lists. Analysis of failure cases reveals two recurring patterns. First, models frequently struggle with short, decontextualized symptom lists; because these inputs omit patient-specific attributes such as age, sex, and clinical history, they differ substantially from the contextualized inputs commonly encountered during training. Second, performance degrades for cases involving multiple disconnected or atypical symptoms, in contrast to classic diagnostic presentations with well-defined symptom clusters, which are handled well.

8.5 Leading Questions

Analysis of errors in the Leading Questions experiment reveals a phenomenon known as *sycophancy*, where the model prioritizes agreement with the user's stated premise over faithful interpretation of the clinical evidence. When a prompt includes a false causal assumption (e.g., "Given the patient likely has [Incorrect Diagnosis]..."), models frequently hallucinate supporting symptoms or suppress contradictory facts to align with the suggested conclusion.

For instance, in cases where the clinical presentation clearly indicated a viral infection, leading prompts suggesting a bacterial etiology caused models to disproportionately weight minor symptoms (like low-grade fever) that fit the bacterial narrative while ignoring clear viral indicators (like clear lungs). This indicates that the models' attention mechanisms are heavily biased toward the immediate instruction context, effectively overriding the latent medical knowledge retrieved from the symptom description. The success of the CoT mitigation (Section 7.3) further supports this: when forced to explicitly verify the premise before answering, the model acts as a critic rather than a complier, significantly reducing error rates.

8.6 Irrelevant Information Overload

Errors in the Irrelevant Information Overload task primarily stem from *associative retrieval failures*, where the model latches onto strong semantic keywords in the noise rather than the causal medical chain. A prime example observed in our qualitative analysis (and highlighted in our proposal) involved a vignette about Sudden Infant Death Syndrome (SIDS). When the text was augmented with irrelevant details about the mother visiting a market in a "tropical country," the model shifted its prediction from "supine sleep position" (correct) to "avoiding cassava" (incorrect).

This error demonstrates that the token "tropical" triggered a strong retrieval association with tropical dietary toxins, which overpowered the clinically relevant signal of the infant's death. Unlike leading questions, where the model tries to satisfy the user, these errors are driven by *salience bias*,the irrelevant information is linguistically distinct or rare (e.g., specific clothing colors, exotic locations), causing the model to attend to it disproportionately. This suggests that while models are robust to general paraphrasing, they remain vulnerable to "distractor attacks" that introduce high-entropy concepts into the context window.

9 Contributions of group members

All the three team members performed multiple tests

- Khushei: False premises and causal direction consistency
- Hima: Paraphrase Consistency and Input attribution Test
- Rohit: Leading questions and irrelevant information overloading

10 Conclusion

In this work, we conducted a comprehensive evaluation of Large Language Models in medical diagnostics, moving beyond standard accuracy to assess the critical dimensions of faithfulness and self-consistency. Our experiments reveal a significant dichotomy between "getting the answer right" and "reasoning correctly." While proprietary models like Gemini demonstrate impressive robustness to surface-level linguistic variation (paraphrasing) and irrelevant noise, they remain susceptible to cognitive biases, such as leading questions, where they can be "tricked" into incorrect diagnoses 33% of the time.

A surprising finding was the divergence in failure modes among open-weights models. While Gemma showed higher susceptibility to confirmation bias, Qwen demonstrated superior detection of false premises, suggesting that model architecture and training data significantly influence spe-

cific reasoning behaviors beyond general capability. Furthermore, our "Causal Direction" experiment highlighted a universal weakness: all models struggled to maintain consistency when reasoning from diagnosis to symptoms, despite high forward accuracy.

One of the most challenging aspects of this project was developing rigorous metrics for "faithfulness." Our attribution experiments revealed that standard redaction-based methods are often insufficient, as models distribute reasoning across many tokens; removing "influential" words often modulated confidence rather than flipping predictions. Additionally, generating high-quality adversarial scenarios for the mitigation phase proved computationally expensive, limiting the scope of our CoT testing.

Future work should focus on exploring mechanistic interpretability techniques to better understand where medical knowledge is stored versus where reasoning errors occur. Ultimately, this study underscores that for LLMs to be safely deployed in healthcare, they must be optimized for robust, evidence-based reasoning, not just leaderboard accuracy.

11 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes. We used ChatGpt

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - We used AI tools for high-level editing support, including grammar correction, clarity improvements, and restructuring of already-written text in sections such as the methodology description and results discussion.
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - AI assistance was primarily used for minor rewriting, formatting suggestions, and improving readability of text that was originally written by the authors. All technical decisions, experimental design, implementation, data analysis, and interpretation of results were performed by the authors. AI-generated suggestions were reviewed and edited for correctness, and no content was included without verification.

References

Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. 2025. A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *Preprint*, arXiv:2502.14127.

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911.

Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. Evaluating human alignment and model faithfulness of llm rationale. *arXiv preprint arXiv:2407.00219*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

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.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? *Preprint*, arXiv:2401.07927.

Letitia Parcalabescu and Anette Frank. 2023. On measuring faithfulness or self-consistency of natural language explanations. *arXiv preprint arXiv:2311.07466*.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: a systematic review with large language models and beyond. *MedRxiv*.