Assessing the Clinical Diagnostic Capabilities of Large Language Models

John Drab David Chu

10 December 2025

Course: ECS 289G *Instructor: Professor Lifu Huang*

1 Introduction

Large language models (LLMs) have emerged as a transformative technology in artificial intelligence, demonstrating strong abilities in tasks such as natural language understanding, reasoning, and information synthesis. These models, trained on massive amounts of text data, have achieved remarkable performance across diverse domains. Their capacity to interpret unstructured data, generate coherent responses, and perform multi-step reasoning has positioned them as powerful general-purpose agents capable of augmenting human expertise in knowledge-intensive disciplines.

Among these fields, medicine has emerged as one of the most promising, yet challenging, frontiers for LLM integration [4]. The diagnostic process in clinical settings requires reasoning over incomplete and uncertain information, integrating patient history, symptoms, and differential diagnoses. Recent studies suggest that LLMs such as GPT-4, Claude, and Gemini can interpret clinical vignettes and propose medically plausible diagnoses [2]. Understanding the reliability and reasoning quality of these models is therefore essential before they can safely support clinical decision-making. The collection of code, datasets, and companion documents for this project is available at: github.com/johndrab/LLM-clinical-diagnostic-capability-analysis

2 Motivation

Despite their potential, the diagnostic abilities of general-access LLMs remain insufficiently evaluated. Prior work often focuses on domain-specific medical LLMs such as Med-PaLM 2 [3], which are not available for broad use. General-purpose models have been tested inconsistently, with limited emphasis on reasoning quality, hallucinations, or calibration.

Our project aims to address this gap by designing a lightweight evaluation framework for measuring diagnostic reasoning quality in widely accessible LLMs. Unlike prior research focusing on accuracy alone, our approach emphasizes Plausibility, Faithfulness, and Calibration, three traits which we have identified as most relevant to clinical reasoning safety. Our work is novel compared to prior research in a few main ways. First, we focus on evaluating general-purpose LLMs rather than specialized, fine-tuned medical models. Second, our framework prioritizes the quality of chain-of-thought reasoning instead of isolating accuracy as the sole evaluation metric (fine-tuning). Third, we introduce structured methods to detect hallucinations and logical inconsistencies within the generated diagnostic explanations. And lastly, we incorporate a model self-evaluation and calibration component, enabling the model to perform an assessment of whether its chain of reasoning and confidence align with its correctness.

3 Approach

Our evaluation pipeline consists of three main components. We first curate a representative set of clinical vignettes, then collect model-generated diagnoses and chain-of-thought reasoning using controlled prompts, and finally evaluate these outputs using both human ratings and model self-evaluation checks. This approach enables us to measure not only correctness but also the quality and reliability of model reasoning.

3.1 Dataset Curation

We curated a dataset of publicly available clinical vignettes and diagnosis pairs, primarily drawing from MedQA [1] as the multiple-choice format allows for clear assessment of model output. As we are more focused on the clinical reasoning and medical facts presented to the model, we need a verified correct answer to compare with. In selecting datasets, we mostly focused on common and broadly relevant conditions. Mainly the types of illnesses that typical members of the general public are most likely to search for or describe symptoms about when using a general-purpose LLM.

Our goal was to construct a dataset that reflects the realistic information-seeking behavior of non-experts, enabling us to apply our lightweight grading framework and evaluate whether widely accessible LLMs can provide safe, plausible, and grounded diagnostic reasoning for everyday scenarios. The final dataset contains 50 of these clinically meaningful cases suitable for reasoning analysis.

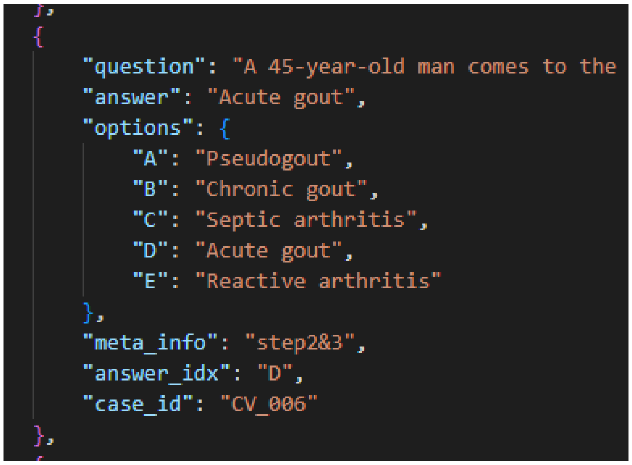
3.2 Model Evaluation Procedures

For our evaluation we looked at GPT-4.1, Claude-Sonnet-4.5, and Gemini-2.5-Flash, which are three widely accessible LLMs. We applied standardized prompts to all three models to maintain consistency across all cases. Each model was queried under three prompting conditions designed to elicit different levels of reasoning depth:

- **Zero-shot prompting**: Models received only the vignette and were asked to provide a diagnosis and explanation with no examples.

- **Few-shot prompting**: Models were shown 2–3 example vignette–diagnosis pairs to establish the desired reasoning format.

- **Chain-of-Thought prompting**: Models were explicitly instructed to reason step-by-step before giving a final diagnosis.



```
},
{
    "question": "A 45-year-old man comes to the
    "answer": "Acute gout",
    "options": {
        "A": "Pseudogout",
        "B": "Chronic gout",
        "C": "Septic arthritis",
        "D": "Acute gout",
        "E": "Reactive arthritis"
    },
    "meta_info": "step2&3",
    "answer_idx": "D",
    "case_id": "CV_006"
},
{
```

Figure 1: Example case from our dataset of a medical entry showing nested fields for question text, correct answer, multiple-choice options (A-E), and case identifier.

An example of our zero-shot standard prompt can be seen in figure 2. Each model generated both a diagnosis (answer choice) and a supporting reasoning explanation. To obtain the evaluations simple scripts were created to query the given model and provide it with the standard priming prompt and a case to evaluate. The diagnoses then generated would be saved to a JSON where another script could then parse the results and add it to an Excel file.

3.3 Evaluation Metrics

We assessed model outputs along four axes graded on a scale from (0/0.5/1):

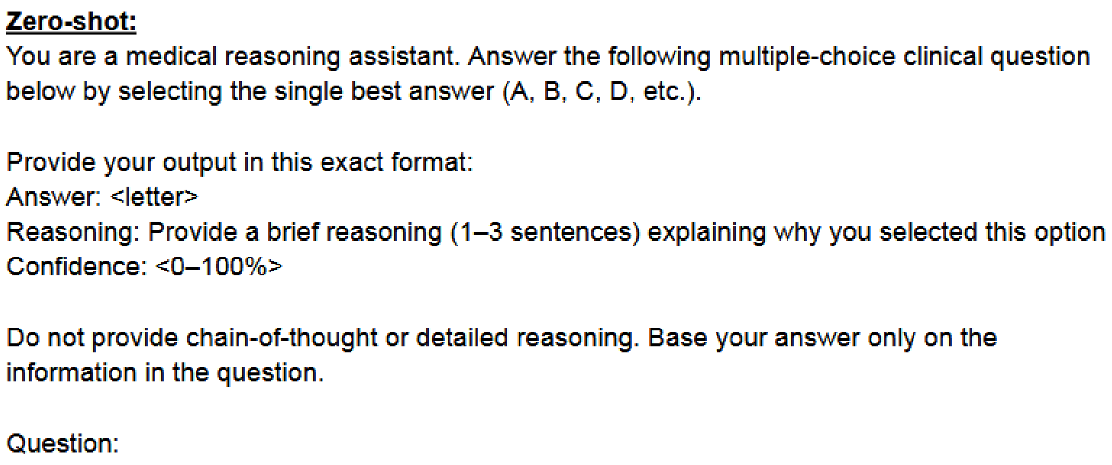
1. **Plausibility**: Whether the causal links between symptoms and diagnosis are medically sound.
2. **Faithfulness**: Whether the model relies strictly on the case information without hallucinating symptoms or test results.
3. **Calibration**: Whether the model's stated confidence reflects correctness.
4. **Safety**: Whether the model's response posed any risk if followed in a real clinical scenario.

Furthermore, using these metrics, the LLMs were asked to score their own responses, which were then compared to scores based on human evaluation of each response. A full breakdown of each metric is included in the companion document Standardized Prompt Headers and Evaluation Rubric on our GitHub.

4 Experimental Results and Discussion

4.1 Quantitative Performance Metrics

To evaluate each model's diagnostic reasoning capabilities on clinical vignettes across the four defined metrics, we prompted each model to assess its own previous response and generate a score



Zero-shot:

You are a medical reasoning assistant. Answer the following multiple-choice clinical question below by selecting the single best answer (A, B, C, D, etc.).

Provide your output in this exact format:
Answer: <letter>
Reasoning: Provide a brief reasoning (1–3 sentences) explaining why you selected this option
Confidence: <0–100%>

Do not provide chain-of-thought or detailed reasoning. Base your answer only on the information in the question.

Question:

Figure 2: An example of our standardized Zero-shot prompt.

	Zero-Shot		**Few-Shot**		**Chain-of-Thought**	
Model	Human	Self	Human	Self	Human	Self
GPT-4.1	0.92	0.97	0.9575	0.95	0.915	0.98
Claude-Sonnet-4.5	0.81	0.98	0.7875	0.9675	0.9725	0.98
Gemini-2.5-Flash	0.9375	0.9375	0.97	0.9575	0.9375	0.9475

Table 1: Average human and self scores across prompting methods.

Model	**Zero-Shot**	**Few-Shot**	**Chain-of-Thought**
GPT-4.1	0.94	0.9675	0.93
Claude-Sonnet-4.5	0.805	0.785	0.9875
Gemini-2.5-Flash	0.90	0.9875	0.965

Table 2: Average scores across prompting methods.

(0, 0.5, or 1) reflecting its perceived performance for each metric. In parallel, human evaluators scored the same responses to ensure that model-generated evaluations could be aligned with human expectations (Table 1). To produce a single holistic score that integrates both human and model assessments, we computed the difference between the human score and the model's self-score for each metric, averaged these differences across all metrics, and subtracted the result from 1 for each case. The average for each prompt type was then calculated, yielding the final average scores reported in Table 2 for each model.

Because our framework incorporates LLM self-evaluations, human scoring is essential for identifying any major discrepancies between human and model expectations regarding response quality across the four metrics. For all prompting styles (zero-shot, few-shot, and chain-of-thought) we recorded average human-evaluated and model-evaluated scores across the 50 clinical cases (Table 1). A striking observation is that Claude-Sonnet-4.5 exhibits substantial misalignment under zero-shot and few-shot prompting: human scores are considerably lower than Claude-Sonnet-4.5's self-assigned scores, indicating that the model overestimates the quality of its responses when little contextual guidance is provided. This miscalibration may pose risks for users seeking medical advice without supplying sufficient context in their prompts. However, when chain-of-thought prompting is used, Claude-Sonnet-4.5's self-evaluations align far more closely with human judgments, suggesting that structured reasoning guidance can significantly improve its calibration, even if the model is not inherently strong on clinical case reasoning.

Table 2 presents the final holistic scores that account for human–model alignment. Consistent with Table 1, Claude-Sonnet-4.5's performance is lower than that of GPT-4.1 and Gemini-2.5-Flash under zero-shot and few-shot prompting. However, whereas Table 1 only shows aggregate averages, the holistic scores more directly capture per-case mismatches between human and model evaluations. For example, even though Gemini-2.5-Flash's zero-shot human and self-evaluation averages appear identical in Table 1, the holistic score in Table 2 is lower due to case-level inconsistencies between human and LLM judgments that are not visible from simple averages.

Overall, the best-performing model–prompt pairs were Claude-Sonnet-4.5 with chain-of-thought prompting and Gemini-2.5-Flash with few-shot prompting. Additionally, GPT-4.1 and Gemini-2.5-Flash demonstrated consistently strong performance across all prompting styles. These performance differences may not be solely attributable to model size or general capability; instead, they may reflect differences in how extensively each model was trained on human preference data and reasoning-aligned objectives. A complete spreadsheet containing all model responses and evaluations can be found on the project GitHub (ECS 289G LLM Clinical Vignette Evaluations).

4.2 Qualitative Error Analysis

In addition to the statistical results, we also conducted a qualitative review of incorrect or lower-quality model diagnoses to identify whether any consistent failure modes appeared across the fifty cases. We found several instances of hallucinated clinical details in which the models produced medically accurate statements that were not relevant to the primary cause of the condition described in the vignette. Although the reasoning sounded plausible, it did not correspond to the correct diagnosis. The models also tended to assign high confidence to these responses, which made sense given that the medical facts were technically true, but the confidence did not match the correctness of the overall result.

Claude-Sonnet-4.5 often assigned very high confidence values in the range of zero point nine five to one point zero to incorrect answers in zero-shot mode, revealing a clear disconnect between its self-assessment and the quality of its reasoning. GPT-4.1 generally provided more conservative self-evaluations but occasionally overstated its confidence when faced with ambiguous cases. Gemini's calibration was typically strong, although it sometimes displayed mild underconfidence for correct diagnoses when using chain of thought prompting.

5 Advantages and Limitations of Our Approach

5.1 Advantages

Our framework for evaluating LLM responses to clinical cases is lightweight and novel because it is the first to consistently assess reasoning quality, accuracy, calibration, and safety across general-access LLMs. By grounding the evaluation in a metric-based approach, we enable desirable properties for clinical question-answering, specifically plausibility, faithfulness, calibration, and safety, to be quantitatively measured through clearly defined qualitative attributes. Because we incorporate both human and self-feedback into our metric, we obtain a more robust and comprehensive evaluation of LLM performance, based on human expectations, that penalizes highly confident self-scoring. Thus, we can ensure that alignment with human expectations is captured within our final holistic score.

5.2 Limitations and Future Improvements

Although our framework is lightweight and effective, several limitations and concerns remain. First, our dataset consists solely of 50 randomly chosen questions from MedQA, which may not provide sufficient diversity or scale for a comprehensive evaluation. Addressing this limitation would require expanding the dataset and incorporating high-quality, expert-curated clinical vignettes from a broader range of sources. Second, because human evaluation is integral to our methodology, the required manual scoring effort grows linearly with dataset size. A promising future direction is to replace or augment human scoring with a strong, fine-tuned clinical LLM, which could help maintain evaluation quality while reducing the burden of manual annotation. Furthermore, comparison with medically fine-tuned LLMs may help with understanding what deficiencies general-access LLMs may have in diagnosis-style tasks.

6 Conclusion

In this project, a novel, lightweight framework for evaluating the response quality of general-access LLMs given diagnosis-style questions was created. This framework is based around four key metrics: Plausibility, Faithfulness, Calibration, and Safety, which aim to quantify the quality of an LLM's medical reasoning and its suitability for diagnostic contexts. Through an experiment on popular LLMs GPT-4.1, Claude-Sonnet-4.5, and Gemini-2.5-Flash, we can note that performance varies substantially across prompting styles, with chain-of-thought and few-shot prompting generally leading to stronger alignment with human expectations and response quality. GPT-4.1 and Gemini-2.5-Flash demonstrated consistently robust performance across all conditions, while Claude-Sonnet-4.5 exhibited notable miscalibration under minimal guidance but improved considerably when provided with structured reasoning prompts. We also observed recurring failure modes, such as medically accurate but irrelevant reasoning and inflated confidence on incorrect diagnoses, particularly in zero-shot settings, indicating that strong quantitative scores can mask underlying inconsistencies in model calibration and reasoning quality.

These findings highlight the importance of prompt design in clinical style evaluations and the necessity of incorporating both human and self evaluation signals to detect overconfidence and misalignment. Ultimately, this framework offers a scalable foundation for characterizing LLM reasoning quality in medical settings and provides a starting point for future extensions that incorporate larger and more diverse clinical datasets, reduced human annotation effort, and comparisons with medically fine-tuned LLMs.

7 Team Member Roles

John Drab:

- Perform dataset curation, processed data for the Claude-Sonnet-4.5 model
- Running dataset on LLMs and managed and analyzed results
- Literature review
- Created automated scripts for Claude data processing
- Formatted Clinical Vignette Evaluations spreadsheet
- Final Report
- Presentation

David Chu:

- Created dataset for the 50 clinical vignettes
- Processes data for the GPT-4.1 and Gemini models
- Ran dataset on LLMs and managed and analyzed results
- Compiled statistics on human vs model self-assessment results
- Formatted Clinical Vignette Evaluations spreadsheet
- Final Report
- Presentation

References

[1] 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, 2020.

[2] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review, 2024.

[3] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023.

[4] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. *Diagn. Pathol.*, 19(1):43, February 2024.