Evaluating Large Language Models on Rare Disease Diagnosis: A Case Study using House M.D.

Arsh Gupta [1], **Ajay Narayanan Sridhar** [1], **Bonam Mingole** [1], **Amulya Yadav** [1]

[1]The Pennsylvania State University
abg6210@psu.edu, afs6372@psu.edu, bjm6940@psu.edu, amulya@psu.edu

Abstract

Large language models (LLMs) have demonstrated capabilities across diverse domains, yet their performance on rare disease diagnosis from narrative medical cases remains underexplored. We introduce a novel dataset of 176 symptom-diagnosis pairs extracted from *House M.D.*, a medical television series validated for teaching rare disease recognition in medical education. We evaluate four state-of-the-art LLMs such as GPT 4o mini, GPT 5 mini, Gemini 2.5 Flash, and Gemini 2.5 Pro on narrative-based diagnostic reasoning tasks. Results show significant variation in performance, ranging from 16.48% to 38.64% accuracy, with newer model generations demonstrating a 2.3× improvement. While all models face substantial challenges with rare disease diagnosis, the observed improvement across architectures suggests promising directions for future development. Our educationally validated benchmark establishes baseline performance metrics for narrative medical reasoning and provides a publicly accessible evaluation framework for advancing AI-assisted diagnosis research.

Introduction

The rapid advancement of large language models (LLMs) has opened new opportunities for applying natural language understanding to complex domains requiring reasoning under uncertainty (Jerrentrup et al. 2015; Mechler et al. 2017; Sanges et al. 2018). In healthcare, accurate diagnosis from patient-reported symptoms remains a critical challenge where LLMs could provide value for clinical decision support and medical education. However, evaluating LLM performance in medical diagnosis faces significant barriers due to privacy constraints, limited dataset availability, and the formalized nature of existing medical datasets.

Medical television dramas, particularly *House M.D.*, offer a unique solution by providing rich, case-based scenarios where symptoms and diagnostic outcomes are clearly described in narrative form (Jerrentrup et al. 2015; Mechler et al. 2017; Cambra-Badii et al. 2020). The educational value of *House M.D.* has been well-established, with successful use in medical education to teach about rare diseases and clinical reasoning (Jerrentrup et al. 2015; Mechler et al. 2017). Research shows that 49.6% of health sciences students watch medical dramas regularly, and these shows effectively convey medical knowledge (Cambra-Badii et al. 2020).

In this work, we evaluate LLM performance on symptom-to-diagnosis prediction using a novel dataset of 176 cases constructed from *House M.D.* episodes. Each case pairs symptom descriptions with corresponding disease diagnoses, demanding that models identify medical clues within narrative descriptions, closely mirroring clinical practice. Our dataset is publicly available on Kaggle[1]

Our work addresses three challenges: (i) general-purpose LLMs lack specialization for narrative medical reasoning, (ii) available datasets rarely capture narrative-driven diagnostic structure, and (iii) limited dataset sizes raise generalizability concerns. We evaluate multiple state-of-the-art LLMs like GPT 5 Mini (OpenAI 2025), GPT 4o Mini (OpenAI 2024), Gemini 2.5 Flash (DeepMind 2025a), and Gemini 2.5 Pro (DeepMind 2025b) to assess diagnostic reasoning capabilities across different architectures. Our experiments reveal modest performance on rare disease diagnostic tasks, highlighting domain-specific challenges in medical reasoning from narrative cases.

Our contributions are: (1) a novel dataset of symptom-disease mappings from publicly available narrative cases reflecting clinical reasoning patterns, and (2) comprehensive baseline evaluation across multiple LLM architectures demonstrating current limitations in narrative medical reasoning.

Results establish clear performance baselines across different model families, indicating that narrative-based rare disease diagnosis remains a significant challenge for current LLMs. These findings suggest that educationally validated medical narratives can serve as valuable benchmarks for measuring LLM diagnostic capabilities, establishing a foundation for future research in AI-assisted medical diagnosis and highlighting the need for domain-specific adaptations.

Related Work

Medical television dramas have gained recognition as effective educational tools with structured narrative formats suitable for both human learning and machine learning applica-

[1]https://bit.ly/4p9ltW8

tions. Jerrentrup et al. (Jerrentrup et al. 2015) demonstrated that *House M.D.* can be successfully integrated into medical curricula for teaching rare diseases and complex diagnostic scenarios. Cambra-Badii et al. (Cambra-Badii et al. 2020) found that 49.6% of health sciences students regularly watch medical dramas, with *House M.D.* among the most popular, and that these shows effectively teach bioethical and professional practice issues.

Sanges et al. (Sanges et al. 2018) and Sarrafpour et al. (Sarrafpour et al. 2019) validated case-based learning for rare disease education, showing significant improvements in students' diagnostic recognition abilities. These findings suggest that structured symptom-diagnosis relationships in narrative medical content may provide valuable evaluation data for AI systems reasoning through clinical presentations.

Beyond educational validation, studies have highlighted the unique coverage of rare diseases in medical television. Mechler et al. (Mechler et al. 2017) analyzed orphan diseases featured in *House M.D.*, showing the series frequently presents rare diseases seldom encountered in training, making it valuable for both medical students and AI diagnostic evaluation.

Schaefer and von Hirschhausen (Schaefer and von Hirschhausen 2016) demonstrated that medical television helps viewers recognize symptoms and seek appropriate consultation. Furman and Clayton (Furman and Clayton 2015) analyzed genetic concepts in medical shows, highlighting that narrative case presentations closely parallel diagnostic reasoning challenges, suggesting clear utility for evaluating automated diagnostic systems.

Recent advances emphasize simulation and case-based learning approaches mirroring structured narratives in medical television. Rattani et al. (Rattani et al. 2019) identified narrative methods as particularly effective for complex scenarios, while Rasul et al. (Rasul et al. 2021) validated real-scenario approaches in medical education. These studies demonstrate that well-structured narrative cases effectively bridge theoretical knowledge and practical application, indicating their potential for evaluating LLMs on diagnostic reasoning patterns.

While existing literature supports the educational value of medical television for human learners, limited research explores its application to machine learning systems. Our work addresses this gap by systematically extracting symptom-diagnosis pairs from *House M.D.* episodes and evaluating LLM diagnostic capabilities across multiple architectures, extending established educational value of medical narratives to AI evaluation.

Dataset Creation

We constructed our dataset from the publicly available *House M.D.* wiki (https://house.fandom.com), extracting narrative content from all 176 episodes across eight seasons. The data extraction involved three stages: (1) web scraping using `BeautifulSoup` (Richardson 2024), which allows for robust parsing of HTML content, (2) structured prompt generation where each evaluated model (GPT-4o mini (OpenAI 2024), GPT-5 Mini (OpenAI 2025), Gemini 2.5 Flash (DeepMind 2025a), Gemini 2.5 Pro (DeepMind 2025b) independently transforms raw narrative episodes into standardized medical case formats, and (3) quality filtering to ensure clinical detail, diagnostic clarity, and alignment with real-world medical reasoning (Jerrentrup et al. 2015; Cambra-Badii et al. 2020; Sanges et al. 2018).

Our final dataset consists of 176 symptom-diagnosis pairs spanning diverse medical specialties, with emphasis on complex diagnostic scenarios requiring multi-step reasoning. The complete dataset and model evaluation results are publicly available on Kaggle and GitHub.[2]

Educational Validation: Jerrentrup et al. (Jerrentrup et al. 2015) demonstrated successful integration of *House M.D.* into medical curricula for teaching rare disease recognition.

Rare Disease Coverage: Mechler et al. (Mechler et al. 2017) found the show frequently features orphan diseases underrepresented in traditional medical datasets, addressing a gap in medical AI evaluation benchmarks.

Clinical Realism: Despite dramatic elements, the show employs medical consultants to ensure clinical accuracy and follows a consistent diagnostic framework mirroring practice.

Accessibility: Unlike proprietary medical datasets, *House M.D.* content is publicly available, enabling reproducible research while avoiding ethical constraints of real patient data.

Narrative Context: The narrative format preserves temporal symptom progression, demographics, and clinical decision-making that structured datasets often lose (Cambra-Badii et al. 2020; Schaefer and von Hirschhausen 2016).

While our dataset reflects limitations of fictional content, including dramatic exaggeration and complex case focus, these characteristics may benefit evaluation by providing challenging edge cases that test model robustness. The educational validation of *House M.D.* by medical professionals provides confidence that extracted scenarios contain clinically meaningful information suitable for AI evaluation (Cambra-Badii et al. 2020; Schaefer and von Hirschhausen 2016).

Methodology

We designed a straightforward evaluation pipeline to assess LLM performance on symptom-to-diagnosis prediction tasks across multiple model architectures, consisting of prompt construction, model inference, and evaluation metrics.

Stage	**Process**	**Example**
Exact Match	Prediction in ground truth	*Lupus* in *Systemic Lupus*
Fuzzy Match	Token similarity (thresh. = 0.8)	*Sarcoidosis* $\approx$ *Sarcoid* (0.89)
Classif.	Binary outcome	Correct / Incorrect

Table 1: Fuzzy matching workflow

[2]Code: https://bit.ly/481OSuD, https://bit.ly/4oCxqnj;
Dataset: https://bit.ly/4p9ltW8;
Results: https://bit.ly/3JYGEeG, https://bit.ly/47FHXs1

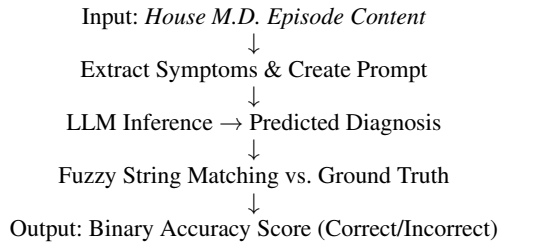

Figure 1: Evaluation pipeline for LLM-based diagnosis

Model Selection and Configuration

We evaluated four state-of-the-art LLMs: GPT-4o Mini, GPT-5 Mini, Gemini 2.5 Flash, and Gemini 2.5 Pro. This selection spans different model families (OpenAI and Google) and capability levels, enabling assessment of diagnostic reasoning across various architectures and training approaches.

The model was configured with standard parameters: temperature set to 0.0 to ensure deterministic outputs, maximum token length of 1500 to accommodate detailed diagnostic reasoning, and no additional system prompts beyond the diagnostic instruction to avoid introducing bias toward specific medical frameworks.

Prompt Design and Inference

Our prompts follow a structured medical case presentation format designed to simulate realistic clinical scenarios. Each prompt contains patient demographic information, symptom descriptions with temporal progression, relevant medical history, and initial diagnostic workup results. The prompts explicitly request a single primary diagnosis while encouraging the model to provide supporting reasoning.

For each case, models generate diagnostic responses in a single-pass approach without iterative refinement. Model responses were collected systematically across all 176 cases with consistent experimental conditions.

Prompt	Ground Truth Disease
"A 27-year-old male presents to the ER after an episode of acute aphonia and syncope during his wedding ceremony. Initially, malingering was suspected, but he subsequently developed a productive cough, cyanosis, and was found to have a pleural effusion. His workup revealed a positive mononucleosis test, which is atypical for his age....."	Arnold-Chiari malformation

Table 2: Example evaluation case from Gemini 2.5 Pro showing only a portion of the Prompt and Ground Truth Disease.

Evaluation Metrics

We evaluated predictions using fuzzy string matching against ground truth diagnoses, addressing the challenge that medical conditions have multiple valid names. Our algorithm employs Python's `SequenceMatcher` with a 0.8 similarity threshold, performing exact substring matching first, then token-wise fuzzy comparison. Final accuracy is computed as the proportion of correctly classified cases, providing clear performance benchmarks while accommodating medical terminology ambiguity.

Limitations

Our methodology has acknowledged limitations. Fuzzy matching may miss semantically equivalent diagnoses using substantially different terminology, and the binary metric may not capture partial credit for related diagnoses. However, our approach provides a systematic and reproducible framework for evaluating LLM diagnostic performance across multiple architectures.

Results

Overall Performance

We evaluated four state-of-the-art LLMs on our *House M.D.* diagnostic dataset using fuzzy string matching. Table 3 summarizes the accuracy across all models.

Model	Correct	Accuracy (%)
GPT-4o-mini	29/176	16.48
Gemini 2.5 Flash	58/176	32.95
GPT-5-mini	65/176	36.93
Gemini 2.5 Pro	68/176	38.64

Table 3: Diagnostic accuracy across LLM models

Performance varied significantly across model architectures, with Gemini 2.5 Pro achieving the highest accuracy at 38.64%, followed by GPT-5 Mini at 36.93%, Gemini 2.5 Flash at 32.95%, and GPT-4o Mini at 16.48%. Despite these differences, all models demonstrated substantial challenges with rare disease diagnostic reasoning.

Performance Analysis

Performance varied not only across models but also across seasons, as shown in Table 4. Season 1 achieved the highest accuracy at 56.52%, while Season 5 showed the lowest at 20.83%. This variation suggests that diagnostic complexity varies throughout the series, with later seasons potentially featuring more challenging rare disease cases. However, the relatively strong performance in Season 8 (52.38%) indicates that temporal progression alone does not fully explain accuracy differences rather case-specific diagnostic complexity appears to be the primary driver.

Across all models, performance was better on common conditions with distinctive symptom presentations (meningitis, myocardial infarction, pulmonary embolism). All models struggled with rare diseases (neurocysticercosis,

Season	Episodes	Correct	Accuracy (%)
Season 1	23	13	56.52
Season 2	24	7	29.17
Season 3	24	8	33.33
Season 4	16	7	43.75
Season 5	24	5	20.83
Season 6	21	8	38.10
Season 7	23	9	39.13
Season 8	21	11	52.38

Table 4: Per-season diagnostic accuracy for Gemini 2.5 Pro

Erdheim-Chester disease), multi-system autoimmune disorders (systemic lupus erythematosus, sarcoidosis), and toxicological cases requiring integration of exposure history with clinical presentation.

The performance gap between models suggests that architectural differences and training approaches significantly impact diagnostic reasoning capabilities. GPT-5-mini and Gemini 2.5 Pro's superior performance indicates that newer model generations with enhanced reasoning capabilities show meaningful improvements over earlier versions, though substantial limitations remain.

Implications

These results establish important baseline performance metrics for narrative-based rare disease diagnosis and demonstrate that current LLMs show promising capabilities in medical reasoning tasks. The improvement from GPT-4o Mini (16.48%) to Gemini 2.5 Pro (38.64%) indicates that the field is making meaningful progress toward clinically useful AI diagnostic systems. While absolute accuracy levels indicate room for improvement, it is important to contextualize these results: our benchmark exclusively features diagnostically challenging cases that often puzzle expert physicians, representing a substantially harder evaluation task than typical medical AI benchmarks. The ability to correctly diagnose nearly 40% of these exceptionally difficult cases demonstrates meaningful medical reasoning capabilities and establishes a solid foundation for future improvements through domain-specific fine-tuning, integration with medical knowledge bases, or hybrid reasoning approaches.

Discussion

Our results demonstrate significant variation in LLM diagnostic reasoning capabilities, with performance ranging from 16.48% (GPT-4o Mini) to 38.64% (Gemini 2.5 Pro). This 2.3× improvement highlights rapid progress in medical reasoning, though rare disease diagnosis remains challenging for current general-purpose LLMs (Schaefer and von Hirschhausen 2016; Mechler et al. 2017).

Our *House M.D.* dataset addresses a gap in medical AI evaluation by providing narrative-based diagnostic scenarios testing reasoning rather than factual recall which offers a meaningful benchmark for evaluating AI diagnostic capabilities validated by the show's documented use in medical education (Jerrentrup et al. 2015).

Limitations include potential bias from fictional narratives, lack of expert medical validation, and a binary accuracy metric that does not capture clinical significance of errors (Cambra-Badii et al. 2020). Models frequently provided confident but incorrect explanations, raising concerns for clinical deployment without specialized training and validation.

Despite these limitations, the substantial improvement across model generations is encouraging. Our benchmark establishes baseline metrics for narrative-based rare disease diagnosis and provides a foundation for future improvements through domain-specific fine-tuning or hybrid approaches combining LLMs with medical knowledge bases.

Future Directions and Research Opportunities

Several promising directions emerge from our findings. First, expanding the dataset to include additional medical television series (*Grey's Anatomy*, *The Good Doctor*, *ER*) would provide broader diagnostic scenario coverage and reduce bias toward *House M.D.*'s rare disease focus. Second, expert medical validation of extracted symptom-diagnosis pairs would strengthen dataset reliability and enable comparison with existing clinical benchmarks.

The 2.3× performance improvement from GPT-4o Mini (16.48%) to Gemini 2.5 Pro (38.64%) suggests substantial room for further gains through domain-specific fine-tuning. Integrating medical knowledge bases (UMLS, SNOMED-CT) with LLM reasoning may address rare disease recognition limitations. Finally, hybrid approaches combining narrative understanding with structured medical knowledge and uncertainty quantification mechanisms could mitigate the confident misdiagnosis problem observed across all models, moving toward clinically useful diagnostic support systems.

Conclusion

We introduce a novel dataset of 176 diagnostically challenging cases derived from *House M.D.* and establish baseline performance across four state-of-the-art LLMs. Results show significant variation (16.48% to 38.64%), with the 2.3× improvement across model generations demonstrating rapid progress in medical reasoning capabilities. However, even the best-performing models indicate that rare disease diagnosis remains challenging for current LLMs.

This work contributes an educationally validated benchmark for narrative-based medical reasoning and establishes clear performance baselines for future research. Our findings highlight both the promise of LLMs in medical AI and the need for continued development through domain-specific training, expert validation, and hybrid approaches to achieve clinically useful diagnostic systems.

References

Cambra-Badii, L.; et al. 2020. Medical drama viewing habits and educational impact on health sciences students. *BMC Medical Education*, 20: 200–210.

DeepMind, G. 2025a. Gemini 2.5 Flash: our cost-efficient thinking model. https://developers.googleblog.com/en/start-building-with-gemini-2-5-flash/.

DeepMind, G. 2025b. Gemini 2.5 Pro: our most advanced reasoning model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

Furman, R.; and Clayton, S. 2015. Teaching genetics concepts through medical television shows. *Genetics Education Review*, 8: 50–60.

Jerrentrup, A.; et al. 2015. Using House M.D. for teaching rare diseases in medical curricula. *Medical Education Journal*, 49: 123–130.

Mechler, H.; et al. 2017. Orphan diseases and medical education in House M.D. *Journal of Medical Media Studies*, 5: 23–34.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. 2025. GPT-5 Mini: A streamlined version of GPT-5. https://platform.openai.com/docs/models/gpt-5-mini.

Rasul, F.; et al. 2021. Teaching professionalism using real-world scenarios in undergraduate medical education. *Medical Education Journal*, 55: 321–332.

Rattani, A.; et al. 2019. Narrative methods and simulation in medical ethics and professionalism education. *BMC Medical Ethics*, 20: 78–88.

Richardson, L. 2024. Beautiful Soup 4: Pythonic HTML and XML parsing. https://pypi.org/project/beautifulsoup4/. Accessed: 2025-10-19.

Sanges, S.; et al. 2018. Role-play simulation and case-based learning for rare disease education. *Advances in Medical Education*, 12: 45–55.

Sarrafpour, M.; et al. 2019. Simulation and career-computer learning for awareness of rare diseases. *Medical Teacher*, 41: 1100–1110.

Schaefer, K.; and von Hirschhausen, E. 2016. Entertainment media as a tool for rare disease awareness. *Medical Education Online*, 21: 1–9.