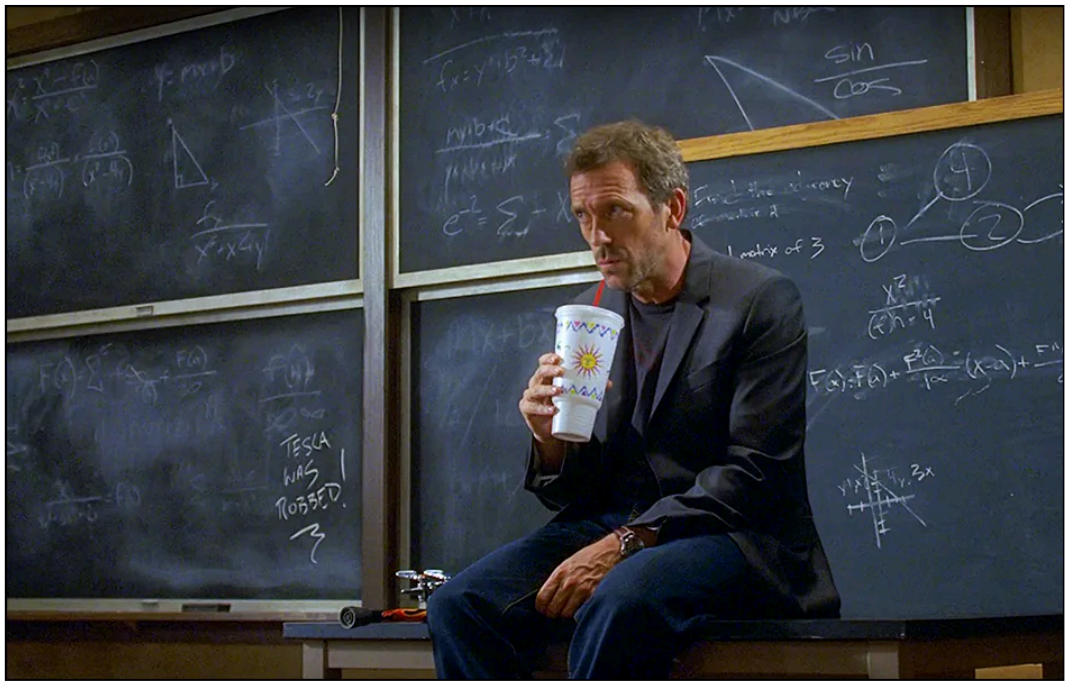
Using the ‘House’ TV Show To Develop AI’s Diagnostic Capabilities

Originally published at https://www.unite.ai/using-the-house-tv-show-to-develop-ais-diagnostic-capabilities/



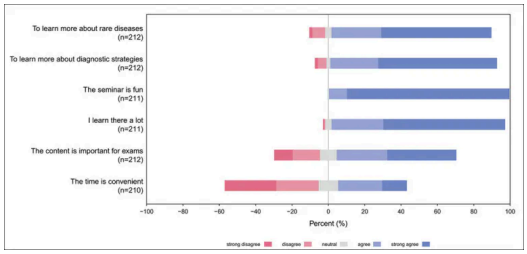
Though rare disease diagnosis is a particularly hard challenge for AI (as it is for humans), popular language models ChatGPT and Gemini show promising performance when trained on diagnostic cases from the popular ‘House’ medical drama.

Nearly half of all health sciences students regularly watch medical dramas such as *House*, *Grey’s Anatomy*, and *Scrubs*. Though this kind of material can only be used for didactic purposes with a lot of filtering and framing, due to the risk of spreading dangerous misinformation, the standard of research for dramas featuring medical conditions tends to be quite high (though accuracy varies across productions).

Unsurprisingly, doctors often originate, advise on and/or write TV medical dramas. In such cases, extensive medical domain knowledge is advantageous not only to accurately convey medical issues, but also to ideate suggestions for new and interesting storylines.

One of the most studiously-researched medical shows of the recent ‘golden age’ of TV is *House* (aka *House MD*), wherein the eccentricities of the lead character and the huge fluctuations in the supporting cast, entertaining as these were, took second place to the ‘disease of the week’.

In fact, out of 177 episodes aired across its eight season run, *House* provided an assiduous 176 diagnostic case studies. Though the show ended in 2012, by 2015 it was already in use as a teaching tool, with a special *Dr. House* seminar offering improved results in comparison to standard seminar fare, even though attending offered no student credits:



From a 2015 study, diverse reasons why medical students wanted to attend a diagnostic seminar that leveraged information from the ‘House’ TV show. The seminars were scheduled at a deliberately challenging time, and conveyed no study credits; in spite of these factors, the initiative was a hit. Source

House and AI

Though the use of *House* and other diverse TV shows has been proven in multiple studies to be an effective adjunct aide to learning, for medical students, little of this approach has been attempted so far in a machine learning context.

Now, a new paper from Pennsylvania State University has made an initial foray in this direction, by developing a dataset featuring all usable 176 *House* case studies, formulated into a narrative-driven diagnostic structure, subsequently evaluated on popular LLMs from OpenAI and Google.

Despite the difficulty of this challenge (which characterizes one of the most difficult fields in biological sciences), the researchers found that more recent versions of ChatGPT and Gemini showed improvement over older versions, indicating that the evolutionary trend of model development is likely to lean effectively into diagnostic processes over time.

The paper states:

> *'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.'*

Besides establishing performance baselines against which future efforts can be evaluated, the authors note that the new dataset – which they are making publicly available – solves the lack of narrative process inside existing medical datasets, and is easily available, in contrast with the gate-kept culture of standard medical datasets.

The new work is titled *Evaluating Large Language Models on Rare Disease Diagnosis: A Case Study using House M.D*, and comes from four researchers at Penn State*.

Data

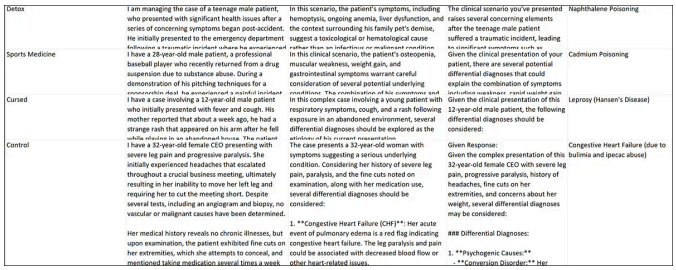
To populate their dataset, the authors used publicly available material from the long-established House Wiki fandom site. Narrative content was extracted and distilled using the popular Beautiful Soup framework, which can extract structural data from the HTML source of web pages.

After the base narratives were harvested in this way, four LLMs were used to transform the output into standardized case formatting. The models used were GPT-4o mini; GPT-5 Mini; Gemini 2.5 Flash; and Gemini 2.5 Pro. Finally, quality filtering was applied, to ensure that the dataset had appropriate clinical detail, and alignment with the current state of the art in medical reasoning.

The authors observe that 'orphan' diseases (a.k.a., rare diseases) are underrepresented in standard medical databases; in certain instances, their coverage in the *House* show may represent an unusual percentage of their total existing coverage.

The authors concede that the utility of a data source of this type has to be tempered with caution in regard to the artistic license that may get prioritized at times in the development of medical drama:

> *'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].'*



Detox	I am managing the case of a teenage male patient, who presented with significant health issues after a series of concerning symptoms began post-accident. He initially presented to the emergency department following a traumatic incident where he experienced	In this scenario, the patient's symptoms, including hemoptysis, ongoing anemia, liver dysfunction, and the context surrounding his family pet's demise, suggest a toxicological or hematological cause rather than an infectious or malignant condition	The clinical scenario you've presented raises several concerning elements after the teenage male patient suffered a traumatic incident, leading to significant symptoms such as	Naphthalene Poisoning
Sports Medicine	I have a 28-year-old male patient, a professional baseball player who recently returned from a drug suspension due to substance abuse. During a demonstration of his pitching techniques for a	In this clinical scenario, the patient's osteopenia, muscular weakness, weight gain, and gastrointestinal symptoms warrant careful consideration of several potential underlying conditions. The combination of his symptoms and	Given the clinical presentation of your patient, there are several potential differential diagnoses that could explain the combination of symptoms including weakness, rapid weight gain,	Cadmium Poisoning
Cursed	I have a case involving a 12-year-old male patient who initially presented with fever and cough. His mother reported that about a week ago, he had a strange rash that appeared on his arm after he fell while playing in an abandoned house. The patient	In this complex case involving a young patient with respiratory symptoms, cough, and a rash following exposure in an abandoned environment, several differential diagnoses should be explored as the etiology of his current presentation	Given the clinical presentation of this 12-year-old male patient, the following differential diagnoses should be considered:	Leprosy (Hansen's Disease)
Control	I have a 32-year-old female CEO presenting with severe leg pain and progressive paralysis. She initially experienced headaches that escalated throughout a crucial business meeting, ultimately resulting in her inability to move her left leg and requiring her to cut the meeting short. Despite several tests, including an angiogram and biopsy, no vascular or malignant causes have been determined. Her medical history reveals no chronic illnesses, but upon examination, the patient exhibited fine cuts on her extremities, which she attempts to conceal, and mentioned taking medication several times a week	The case presents a 32-year-old woman with symptoms suggesting a serious underlying condition. Considering her history of severe leg pain, paralysis, and the fine cuts noted on examination, along with her medication use, several differential diagnoses should be considered: 1. **Congestive Heart Failure (CHF)**: Her acute event of pulmonary edema is a red flag indicating congestive heart failure. The leg paralysis and pain could be associated with decreased blood flow or other heart-related issues.	Given Response: Given the complex presentation of this 32-year-old female CEO with severe leg pain, progressive paralysis, history of headaches, fine cuts on her extremities, and concerns about her weight, several differential diagnoses may be considered: ### Differential Diagnoses: 1. **Psychogenic Causes:** - **Conversion Disorder:** Her	Congestive Heart Failure (due to bulimia and ipecac abuse)

Examples from the dataset generated for the project. Source

Tests

To evaluate model accuracy on narrative diagnostic tasks, the authors designed a simple pipeline combining prompt generation, model inference, and scoring.

The four aforementioned LLMs were tested, with each model configured with temperature set to zero (ensuring deterministic rather than 'creative' output), and with a maximum token length of 1,500 – an allowance designed to accommodate complex diagnostic reasoning. No additional system prompts were used to frame the queries further.

The prompts themselves adhered to a standard structured medical case presentation format – the kind viewers will be most familiar with from medical dramas when a new patient/disease is introduced, and a doctor summarizes an overview for the benefit of other doctors present (effectively, though, for the viewers' benefit).

Each prompt presented a clinical narrative comprising demographic details; a timeline of symptoms; relevant medical history; and early diagnostic findings. The model is briefed to identify a single primary diagnosis, and to justify its conclusion with reasoning.

Each model generated its diagnostic response in a single pass, without any iterative refinement; and responses were collected under consistent conditions across all 176 cases:

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

An illustrative example showing a narrative clinical prompt and its corresponding ground truth diagnosis, as used for testing Gemini 2.5 Pro.
Source

For metrics, predictions were evaluated using a 'fuzzy' string-matching procedure designed to account for ambiguity in medical terminology. The approach used Python's SequenceMatcher library, with a similarity threshold of 0.8, beginning with exact substring matching and falling back to token-wise comparison when necessary. Accuracy was calculated as the proportion of cases classified correctly under these conditions:

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

The 'fuzzy matching' workflow used by the researchers.

The authors note that fuzzy matching could mean that semantically identical diagnoses that use different terminology could be missed, but present their approach as the most reproducible that could meet all the project's constraints.

Results

Diagnostic accuracy varied widely across models, with Gemini 2.5 Pro performing best 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 struggled with the demands of diagnostic reasoning for rare diseases:

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

Results for diagnostic accuracy across the four models trialed.

The authors note also that performance varied across seasons of the show:

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

Varying accuracy across the diverse seasons of House, but not with any obvious curve or clear reason.

The paper states:

> *'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.'*

Models performed more reliably when diagnosing common conditions with recognizable symptoms, such as meningitis, myocardial infarction, and pulmonary embolism – but consistently struggled with rare diseases such as neurocysticercosis and Erdheim-Chester disease, as well as complex autoimmune disorders like systemic lupus erythematosus and sarcoidosis. Performance also dropped on toxicological cases that required linking exposure history to clinical signs.

The authors suggest that the variation in accuracy between models points to meaningful differences in architecture and training strategy, with the stronger performance of GPT-5 Mini and Gemini 2.5 Pro indicating that newer generations of LLM benefit from improved reasoning capabilities – although their results still reveal clear limitations in handling complex diagnostic tasks.

The results, they contend, provide baseline metrics for narrative-based rare disease diagnosis, strongly indicating that current language models are beginning to show useful medical reasoning abilities.

The jump in performance from GPT-4o Mini at 16.48% to Gemini 2.5 Pro at 38.64%, the paper concludes, signals steady progress toward clinically applicable AI support tools.

While the researchers concede that accuracy levels remain modest, the benchmark focuses exclusively on highly complex cases that routinely challenge even trained physicians, and the ability to correctly identify the diagnosis in nearly 40% of these difficult examples points to genuine reasoning capacity, laying the groundwork for future improvements through targeted fine-tuning, structured medical knowledge integration, or hybrid reasoning strategies.

Conclusion

There are some obvious dangers in re-purposing TV show narratives into real-world medical datasets – even in cases, as with *House*, where the source material has a high level of qualified medical contributions and/or oversight.

It's interesting to note that a typical episode of *House* effectively acts as a summarization machine for a series of medical entries that may not be directly accessible on the internet for the average person, or for data sources that present the information in a far more fragmented and non-linear way.

Having a doctor actually write the screenplay for an episode, as frequently happened with *House*, could be used by researchers as some form of 'sign-off' on the content; but this ignores the fact that artistic considerations may have influenced the presentation of the disease in the episode.

This leaves the data in the condition of so many other potentially useful data sources for training: in need of a fresh layer of expensive, qualified human oversight.

* *Please note that this very short paper does not follow the customary template, and I have adapted coverage to accommodate this.*

First published Monday, November 17, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai