Wolters Kluwer: Health - Language model

AI-powered differential diagnosis support

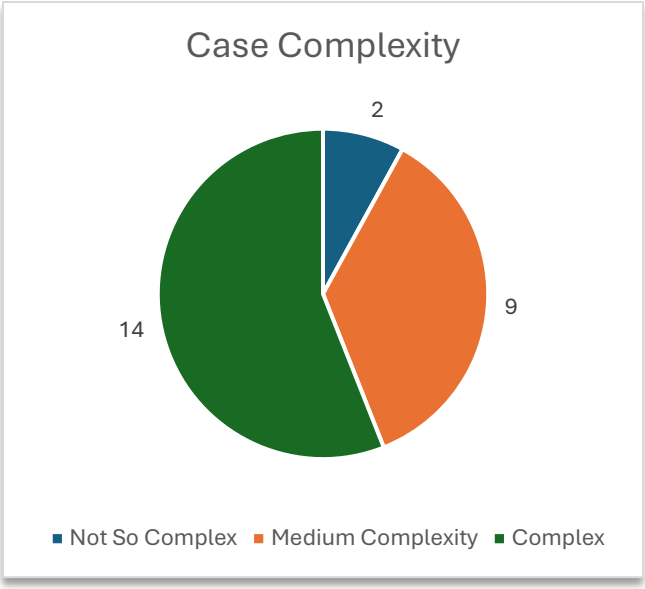
Project Abstract:

In partnership with Wolters Kluwer Health, this project evaluates the potential of state-of-the-art large language models to generate accurate differential diagnoses and actionable next-step recommendations. Leveraging 25 real NEJM case studies, we compare out-of-the-box LLM knowledge against findings mentioned in the Clinicopathological Conferences (and informed by the latest pre-print findings). Performance is measured via ranking metrics (MRR, DCG) and a bespoke 5-point BOND scale for next-step accuracy. Our working prototype built using Python/Streamlit delivers ranked differentials, life-threat alerts, and structured care plans.

Data Sources:

The New England Journal of Medicine's Clinicopathological Conference (CPC) series is a long-standing educational forum that guides readers through real-world diagnostic puzzles. Each installment begins with a detailed clinical vignette—history, exam findings, laboratory and imaging data—followed by a step-by-step expert "think-aloud" differential that highlights key decision points. The case concludes with the definitive pathological or diagnostic reveal, offering a master class in integrating clinical clues with definitive evidence.

We had access to 25 NEJM CPC case records ranging from complex multi-system or atypical presentations needing broad, multidisciplinary work-ups to not so complex straightforward, and narrow differentials resolved with routine evaluation and basic labs/imaging.



Deliverable Breakdown:

An objective-driven breakdown of the project's main goal and deliverables are mentioned below:

1. **Case-Study Dataset Assembly:** Curate 25 real NEJM Clinicopathological Conference reports as structured test cases for homogeneity.
2. **Model Selection and Input Engineering:** Design and iterate on different kinds of prompts as well as account for model hallucinations.
3. **Quantitative Evaluation:** Calculate evaluative metrics such as Mean Reciprocal Rank (MRR) and DCG (Discounted Cumulative Gain) to evaluate the out-of-box performance of LLMs.
4. **Prototype Development:** Build a clickable prototype which envisions the use case of the project.
5. **Innovative Product Design:** Emulate patient-clinician dialogs to test LLMs' ability to manage dynamic history-taking and follow-up questioning.

Case-Study Dataset Assembly:

All 25 NEJM CPC case records were broken down into sections that allowed creating structured inputs (JSON) for inputting to the LLMs and for easy replication of the experiments.

The sectional breakdown can be summarized as:

- Case Number
- Case Summary

- History of Patient Illness
- Differential Diagnosis
- Reason for Differentials

The case record breakdown can be accessed here:

NEJM Case Record Breakdown.xlsx

Model Selection and Input Engineering:

Choice of Large Language Model:

The following models were selected for analyzing the history of patient illness:

- **OpenAI ChatGPT o4-mini-high**

A compact reasoning model from OpenAI, o4-mini-high delivers fast, cost-efficient performance on technical tasks, especially in math, coding, and scientific domains, while retaining multimodal (vision and tool) capabilities previously limited to larger models.

- **Google Gemini 2.5 Pro**

Google's flagship reasoning model on Vertex AI, Gemini 2.5 Pro handles text, code, images, audio, and video with token limits up to 1 million, making it ideal for complex, multimodal problems.

- **xAI Grok 3**

Developed by xAI and integrated with X, Grok 3 advances multi-step reasoning through large-scale reinforcement learning, allowing it to deliberate for seconds or minutes, self-correct errors, and explore alternative solutions.

- **Perplexity Research**

Perplexity's Deep Research model automates expert-level investigations by running dozens of searches, ingesting hundreds of sources, and synthesizing comprehensive reports.

API vs Web Interface:

We used Gemini 2.5 Pro Preview 05-06 APIs in our Google Colab notebook to generate differential diagnoses.

For the remaining 3 LLMs, we used their respective paid web interfaces.

Prompting Techniques:

We designed 2 prompts that were used for generating differential diagnoses via the LLMs:

- A generic prompt that is quick to write and easy for humans to tweak but it trades off strict structure and programmatic rigor.
- An “Atom of Thought” prompt which gave bullet-proof consistency, better chain-of-thought auditing, and seamless integration into pipelines.

Both these prompts can be accessed via the attached documents:

[1] Generic Differential Diagnosis

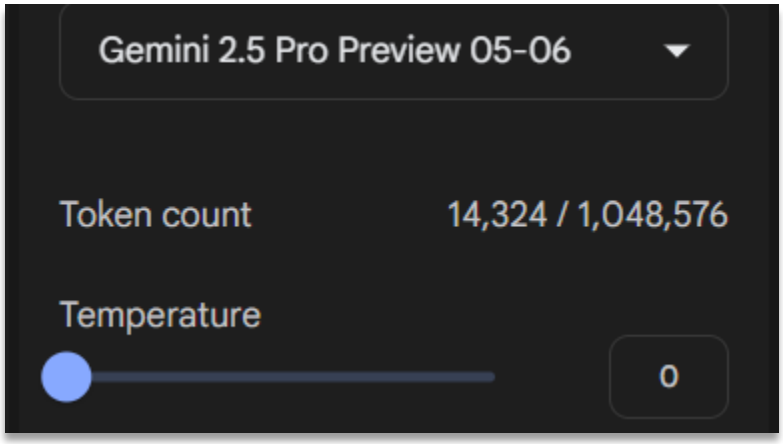
[2] Atom of Thought Differential Diagnosis

Parameter Tuning:

In the Google Colab notebooks, temperature was programmatically set to 0 to reduce hallucinations and creativity of solutions:

```
generate_content_config = types.GenerateContentConfig(
    temperature=0,
    response_mime_type="application/json",
)
```

For the web interfaces, we turned the temperature to 0 wherever there was provision:



Quantitative Evaluation:

Process of Evaluation:

In the absence of domain knowledge experts, we extracted the published NEJM "ground-truth" diagnoses and then ran each LLM's outputs against these labels using two programmatic methods:

- **Semantic Similarity Using Embeddings:**

We leverage pretrained sentence-transformer models to convert both the ground-truth diagnosis and each model's predicted diagnoses into high-dimensional vectors ("embeddings"). By computing the cosine similarity between the correct-diagnosis embedding and each prediction's embedding, we quantify how semantically close they are (even if the wording differs).

Key steps:

1. Embedding Generation – Load a SentenceTransformer (e.g. all-MiniLM-L6-v2) and encode texts into tensors.
2. Cosine Similarity – Compute util.cos_sim() between the ground-truth vector and each predicted-diagnosis vector.
3. Threshold Matching – Declare a prediction "correct" if similarity ≥ 0.85 (configurable).
4. Metric Computation – Use the first matching rank to calculate:
    a. MRR: Reciprocal of that rank (0 if no match).
    b. NDCG@k: Discounted gain at cut-off k, normalizing by the ideal score.
    c. Top-1 Accuracy: Fraction of cases where the top prediction passes the threshold.

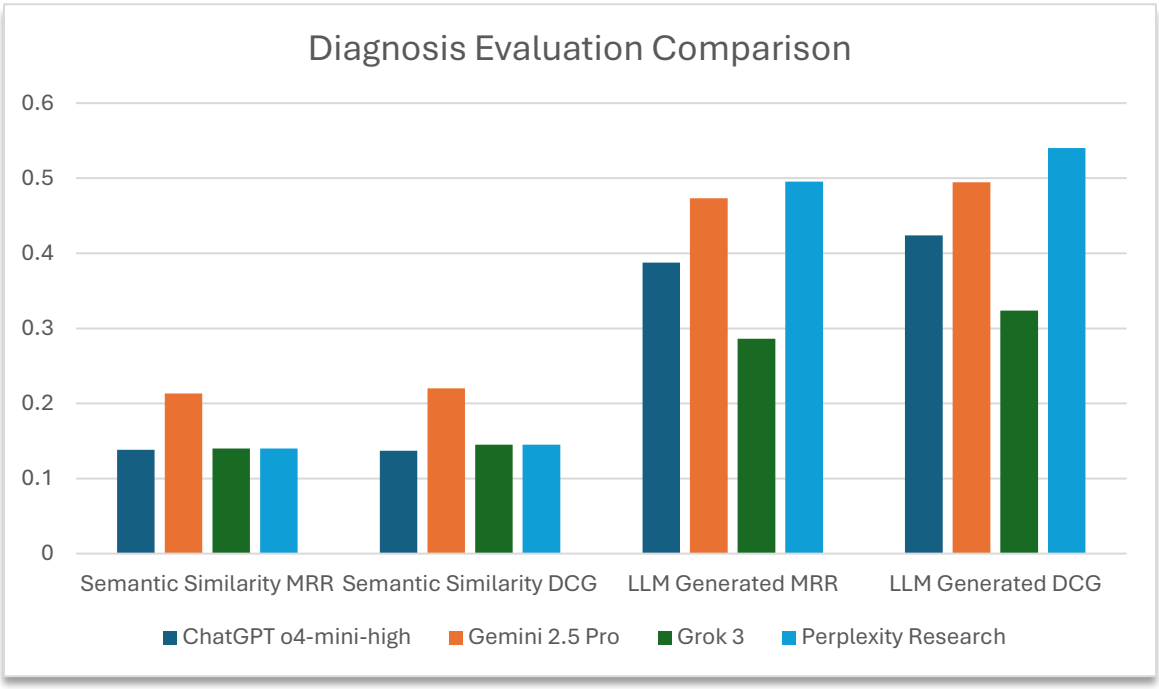
- **Cross-LLM Consensus**

We boost confidence by requiring agreement across multiple models before accepting a diagnosis. After loading each LLM's ranked outputs and the NEJM "ground-truth" labels, we programmatically query a reference LLM (e.g. Gemini 2.5 Pro) with a zero-temperature prompt to ask "YES/NO" if each prediction semantically matches the gold standard. We record the first rank at which any two models agree and then compute MRR, NDCG@k, and Top-1 accuracy on those consensus hits. This approach filters out idiosyncratic errors and highlights diagnoses jointly endorsed by independent LLMs.

Key steps:

1. Load Data – Import the NEJM ground-truth diagnoses and each LLM's ranked predictions.
2. Match Predictions – For each model and case, ask a reference LLM (e.g. Gemini 2.5 Pro) "YES/NO" if a predicted diagnosis semantically matches the gold-standard, and note the first rank where it says YES.
3. Form Consensus – For each case, keep only those diagnoses that at least two different LLMs agree match (i.e. both returned YES).
4. Calculate Metrics – Using the consensus ranks, compute MRR, NDCG@k, and Top-1 accuracy treating unmatched or single-model hits as zero.
5. Review Results – Report overall consensus performance and per-case details showing which models agreed and at what rank.

Based on the two methods, we find that **Cross-LLM Consensus is the better evaluation method:**



Differential Generating LLM	Semantic Similarity MRR	Semantic Similarity DCG	LLM Generated MRR	LLM Generated DCG
Gemini 2.5 Pro	0.2133	0.22	0.4733	0.4945
Grok 3	0.14	0.1452	0.286	0.3239
ChatGPT o4-mini-high	0.1384	0.1372	0.3878	0.4239
Perplexity Research	0.14	0.1452	0.4955	0.5402

Statistical Metrics:

The metrics we used for evaluating the out-of-box performance of LLMs are described below:

- **Mean Reciprocal Rank (MRR)**

MRR is a simple metric for evaluating ranked lists: for each query, take the reciprocal of the position of the first relevant result (1 for a top-rank hit, ½ for second place, ⅓ for third, etc.), then average those reciprocals over all queries. It tells you how far down users must look before they find something useful—higher MRR means relevant items appear earlier on average.

- **Discounted Cumulative Gain (DCG)**

DCG measures the total "gain" (relevance) of items in a ranked list, but with diminishing returns the deeper you go. You sum each item's relevance score, divided by $\log_2(\text{position} + 1)$, so that higher-ranked items contribute much more than lower ones. It captures both how many relevant items you retrieve and how well they're ordered.

- **Top-1 Accuracy**

Top-1 accuracy is the percentage of cases in which the model's highest-ranked prediction (rank 1) is exactly the correct or relevant item.

Prototype Development:

Top Differential Diagnosis:

Diagnosis 1: Congestive Heart Failure (Confidence level: 70%)

The patient's symptoms of bilateral ankle swelling, syncope, exertional dyspnea, and a history of coronary stent placement suggest a cardiac origin. Congestive heart failure can cause fluid to accumulate in the body, leading to edema, especially in the lower extremities. Syncope can occur due to decreased cardiac output.

Diagnosis 2: Infective Endocarditis (Confidence level: 60%)

The patient's fever, fatigue, new scattered ecchymoses, and arthralgias, along with a history of coronary stent placement, could suggest infective endocarditis. This is a serious infection of the heart valves or endocardium that can lead to septic emboli, causing various symptoms including those presented.

Diagnosis 3: Autoimmune Disease (e.g., Rheumatoid Arthritis, Systemic Lupus Erythematosus) (Confidence level: 50%)

The patient's symptoms of diffuse myalgia, arthralgias, anorexia, and fatigue could suggest an autoimmune disease. Rheumatoid arthritis can cause joint pain and swelling, while systemic lupus erythematosus can cause a variety of symptoms including fatigue, fever, and myalgia.

Next Steps

- Perform a comprehensive cardiac evaluation, including EKG, echocardiogram, and possibly a stress test to assess for congestive heart failure.
- Obtain blood cultures and possibly an echocardiogram to evaluate for infective endocarditis.
- Conduct autoimmune serology tests, including rheumatoid factor, ANA, and anti-dsDNA, to evaluate for autoimmune diseases.
- Monitor the patient closely for any changes in symptoms or new symptoms, and adjust the differential diagnosis and treatment plan accordingly.