Executive Summary

This report evaluates the diagnostic accuracy of an AI model deployed via the Perplexity interface using 25 clinical case summaries inspired by New England Journal of Medicine (NEJM) Clinicopathological Conferences (CPCs). Each case was presented in structured format, with the model tasked to return a ranked list of the top 8 differential diagnoses. We then assessed its ranking quality using common information retrieval metrics including Top-K hit rates, Discounted Cumulative Gain (DCG), and Mean Reciprocal Rank (MRR).

While the model demonstrated reasonable coverage, its reliability for top-ranked predictions remains limited: it correctly identified the diagnosis as the top prediction in 11 of 25 cases (44%) and within the top 3 predictions in 16 cases (64%). The correct answer was within the top 8 predictions in 19 cases (76%). However, the average DCG (0.5402) and MRR (0.4955) indicate that relevant diagnoses were not consistently ranked near the top.

Methodology

a. Dataset

We curated 25 clinical cases simulating NEJM CPC formats, encompassing a wide range of specialties including infectious disease, cardiology, psychiatry, neurology, and hematology. For each case, the model was prompted with structured patient information including symptoms, laboratory findings, imaging results, and relevant history. The final diagnosis was excluded from the prompt to avoid information leakage.

b. Prompting and Output

PROMPT_1 = """

Prompt: NEJM Medical Case Analysis with Atom-of-Thought Reasoning (JSON Output, using percentage confidence)

Goal:

Analyze a provided NEJM medical case record and generate a differential diagnosis (top 8) ranked by likelihood as a percentage, along with the final diagnosis. Justify each ranking using atom-of-thought reasoning and suggest next diagnostic steps/tests a physician would perform. The response must be formatted as structured JSON.

Context Dump:

You are a highly advanced medical AI trained in clinical reasoning, differential diagnosis, and diagnostic testing...

```
Return Format (JSON Structure):

{
  "case_id": "<unique_case_id_or_filename>",
  "case_summary": "<brief summary>",
  "differential_diagnosis": [
    { "diagnosis": "...", "reasoning": "...", "confidence_percent": "...%" },
    ...
  ],
  "final_diagnosis": {
    "diagnosis": "...",
    "justification": "..."
  },
  "next_steps_recommended_tests": ["...", "..."]
}
"""
```

We removed the PDFs containing the correct diagnosis from each case, reducing some specific details and retaining only patient summaries and relevant medical history. The remaining information was then input into the LLM. Cases were submitted to Perplexity using its Advanced Research mode via the web interface. The model was instructed to return only a ranked list of the top 8 differential diagnoses with explanation.

c. **Evaluation Criteria**

For each model output, we checked whether the correct final diagnosis appeared within the top-1, top-3, or top-8 predictions. If found, we recorded its rank. From this, we computed:

- **Top-1 / Top-3 / Top-8 accuracy**: Hit rate at each rank threshold.
- **Discounted Cumulative Gain (DCG)**: if the correct diagnosis is present.
- **Mean Reciprocal Rank (MRR)**: if the correct diagnosis is present, else 0.

Results Summary

- **Top-1 Accuracy**: 44% (11/25)

- **Top-3 Accuracy**: 64% (16/25)
- **Top-8 Accuracy**: 76% (19/25)
- **Average DCG**: 0.5402
- **Average MRR**: 0.4955

Metrics were averaged over all 25 cases.

Interpretation & Limitations

While the model successfully displayed the correct diagnosis in the top 8 in most cases, its top-1 accuracy was still below clinical standards for high-risk applications. Diagnoses often appeared lower in the list, reducing their practical utility in decision-making settings. In addition, DCG/MMR values were missing or unranked for many cases because the correct diagnosis appeared outside the top 8.

In addition, we observed that in cases where the correct diagnosis was particularly long or complex—such as “disseminated Mycobacterium kansasii and Mycobacterium abscessus infection due to Mendelian susceptibility to mycobacterial disease”—the AI was often unable to accurately reproduce the full diagnostic phrase. However, it still frequently identified the core components (e.g., “disseminated mycobacterial infection”), suggesting that semantic understanding was partially effective despite the vocabulary mismatch. This nuance is not fully captured by the exact match ranking metric and warrants further investigation.

Per-Case Evaluation Table

Case Number	Final Diagnosis	Top-1 Hit	Top-3 Hit	Rank	DCG	MMR
NEJMcpc2100279	Infective endocarditis due to Haemophilus parainfluenzae.	TRUE	TRUE	1	1	1
NEJMcpc2300900	AL amyloidosis.	TRUE	TRUE	1	1	1
NEJMcpc2309383	Common variable immunodeficiency.	FALSE	FALSE	N/A	0	0
NEJMcpc2309500	Sweet’s syndrome.	FALSE	FALSE	N/A	0	0

NEJMcpc2309726	Nutritional optic neuropathy...	TRUE	TRUE	1	1	1
NEJMcpc2312734	Myasthenia gravis.	FALSE	FALSE	8	0.3155	0.125
NEJMcpc2312735	Postpartum obsessive–compulsive disorder.	TRUE	TRUE	1	1	1
NEJMcpc2402483	Overlap syndrome with lupus nephritis + amyloidosis.	TRUE	TRUE	1	1	1
NEJMcpc2402485	24-Hydroxylase deficiency due to CYP24A1 variant.	FALSE	FALSE	N/A	0	0
NEJMcpc2402486	Rosai–Dorfman–Destombes disease.	FALSE	FALSE	6	0.3562	0.1667
NEJMcpc2402487	Intralobar bronchopulmonary sequestration.	FALSE	FALSE	N/A	0	0
NEJMcpc2402488	Paraneoplastic encephalomyelitis with CAA.	TRUE	TRUE	1	1	1
NEJMcpc2402489	Metastatic adenoid cystic carcinoma of the breast.	TRUE	TRUE	1	1	1
NEJMcpc2402490	Disseminated Mycobacterium kansasii & abscessus infections.	TRUE	TRUE	1	1	1

NEJMcpc2402491	Bronchopneumonia likely due to influenza.	FALSE	TRUE	2	0.6309	0.5
NEJMcpc2402492	Legionella infection with rhabdomyolysis.	TRUE	TRUE	1	1	1
NEJMcpc2402493	Icteric leptospirosis.	FALSE	TRUE	3	0.5	0.3333
NEJMcpc2402496	Brain abscess due to Listeria monocytogenes.	FALSE	FALSE	4	0.4307	0.25
NEJMcpc2402498	Postpartum nephrotic syndrome due to FSGS.	FALSE	TRUE	2	0.6309	0.5
NEJMcpc2402499	Trichobezoar.	FALSE	FALSE	N/A	0	0
NEJMcpc2402500	Psychotic disorder due to general medical condition (postictal psychosis).	TRUE	TRUE	1	1	1
NEJMcpc2402504	Cryptococcal meningoencephalitis.	FALSE	FALSE	N/A	0	0
NEJMcpc2402505	Post-MI reentrant ventricular tachycardia.	FALSE	TRUE	2	0.6309	0.5
NEJMcpc2412511	Posterior reversible encephalopathy syndrome (PRES) in sickle cell disease.	FALSE	TRUE	2	0.6309	0.5

NEJMcpc2412513	Sarcoidosis (Löfgren's syndrome).	TRUE	TRUE	1	1	1

Conclusion

Our analysis suggests that while Perplexity's AI model demonstrates partial diagnostic reasoning ability, it currently lacks the consistency and precision needed for clinical deployment without expert oversight. It may offer value as an exploratory tool or to assist with differential diagnosis generation—but not as a stand-alone decision-maker. Future improvements should prioritize rank precision, clinical terminology mapping, and robustness across specialties.