Evaluation Summary Table with Final Diagnoses Provided by Models

Case Identifier	Model Used	Final Diagnosis from Record	Model's Final Diagnosis	Correct Diagnosis in List	Reciprocal Rank (MRR)	DCG Score	Performance Interpretation
NEJMcpc2402493	ChatGPT 4	Leptospirosis	Leptospirosis	Yes	1.0	1.585	Strong
NEJMcpc2402493	Claude 3.7 Sonnet	Leptospirosis	Leptospirosis	Yes	1.0	3.32193	Strong
NEJMcpc2402496 (GPT)	ChatGPT 4	Brain abscess due to Listeria monocytogenes	Primary CNS Lymphoma	No	0 (Ranked Lowest)	Not Applicable	Weak
NEJMcpc2402496 (Claude)	Claude 3.7 Sonnet	Brain abscess due to Listeria monocytogenes	Cerebral Abscess	Yes (3rd in List)	0.333	1.585	Moderate
NEJMcpc2402498	ChatGPT 4	Postpartum nephrotic syndrome	Nephrotic Syndrome	Yes	1.0	3.32193	Strong
NEJMcpc2402498	Claude 3.7 Sonnet	Postpartum nephrotic syndrome	Nephrotic Syndrome	Yes	1.0	3.32193	Strong

Comparative Analysis of ChatGPT 4 and Claude 3.7 Sonnet in Medical Diagnostics

In an effort to assess the effectiveness of AI models in medical diagnostics, I conducted a structured evaluation of ChatGPT 4 and Claude 3.7 Sonnet. This analysis focused on their ability to generate differential diagnoses for specific medical cases drawn from the New England Journal of Medicine (NEJM). The models were provided with detailed patient histories and were tasked with generating a list of possible diagnoses, which were then compared against the actual final diagnoses recorded in the NEJM case files.

Methodology:

- **Models Tested:** ChatGPT 4 and Claude 3.7 Sonnet
- **Data Source:** NEJM medical case records
- **Evaluation Metrics:** Mean Reciprocal Rank (MRR) and Discounted Cumulative Gain (DCG)

- **Procedure:**
    - Provided each model with a patient case.
    - Retrieved the top 5 differential diagnoses from each model.
    - Assessed whether the actual diagnosis appeared in the model's differential list.
    - Calculated MRR and DCG scores to quantitatively evaluate the model's ranking quality.

Cases Evaluated:

1. **NEJMcpc2402493**
2. **NEJMcpc2402496**
3. **NEJMcpc2402498**

Key Findings:

- **ChatGPT 4** showed a strong ability to correctly identify and rank the actual diagnoses in most cases, particularly excelling in the cases of NEJMcpc2402493 (Leptospirosis) and NEJMcpc2402498 (Postpartum nephrotic syndrome).
- **Claude 3.7 Sonnet** demonstrated moderate performance, accurately identifying the correct diagnosis but sometimes ranking it lower in the differential list. This was notably observed in NEJMcpc2402496, where it recognized 'Brain Abscess due to Listeria monocytogenes' but only placed it third.
- Both models performed exceptionally well when the input data was clear and detailed, particularly for cases with distinctive clinical presentations and confirmatory lab results.

Performance Overview:

- **ChatGPT 4:**
    - Strong in NEJMcpc2402493 and NEJMcpc2402498 with top ranks and high DCG scores.
    - Weaker in NEJMcpc2402496, missing the actual diagnosis in its differentials.
- **Claude 3.7 Sonnet:**
    - Consistently moderate, correctly identifying the actual diagnosis but with lower rankings.
    - Showed strength in capturing a wider range of possible conditions, reflecting its broader analytical approach.

Conclusion:

The analysis highlights the potential of AI models like ChatGPT 4 and Claude 3.7 Sonnet to assist in medical diagnostics. While both models demonstrated promising capabilities, there remains room for improvement in terms of ranking accuracy and model sensitivity to less common diagnoses. These findings suggest that while AI can significantly aid the diagnostic process, it should complement rather than replace traditional diagnostic methods.

Recommendations for Clinical Integration:

- AI models should be used as adjunct tools in diagnostic processes, especially for complex cases where multiple differential diagnoses are considered.
- Continuous training on diverse medical cases will enhance the models' accuracy and reliability.
- Clinicians should critically review AI-generated diagnostics, particularly in cases where AI model confidence diverges from clinical expectations.

Medical AI Assitant Prototype:

Prototype Overview:

I've built a prototype of the "Differential Diagnosis Assistant," a web-based application aimed at transforming how medical professionals generate differential diagnoses. This prototype is designed to demonstrate the core functionalities that will be enhanced once we gain access to the necessary API keys for integrating advanced AI models.

Key Features of the Differential Diagnosis Assistant:

1. **Patient Information Input:**
    - Users can manually input comprehensive patient data including age, gender, chief complaint, and symptom duration. This flexibility ensures that the application can be used in various clinical scenarios.
2. **Symptom Detailing:**
    - The application allows for detailed entry of symptoms, providing a structured format for data input which is crucial for accurate AI analysis.
3. **Case Report Upload:**
    - There is functionality to upload case reports directly in PDF or JSON format, facilitating easy data entry for complex cases and ensuring that the application can handle a wide range of medical documents.
4. **Generate Differential Diagnosis:**
    - With the preliminary setup, users can simulate generating differential diagnoses based on the input data. This feature will be fully powered by AI once API integration is complete, offering real-time, intelligent diagnostic suggestions.
5. **Clinical Timeline:**

- A visual timeline feature is included to help visualize the progression of a patient's symptoms and interventions, aiding clinicians in their decision-making process.

6. **Quick Actions:**

   - Facilities for saving cases, sharing reports, printing, and adding notes are integrated, enhancing the utility and collaborative potential of the application in real clinical environments.

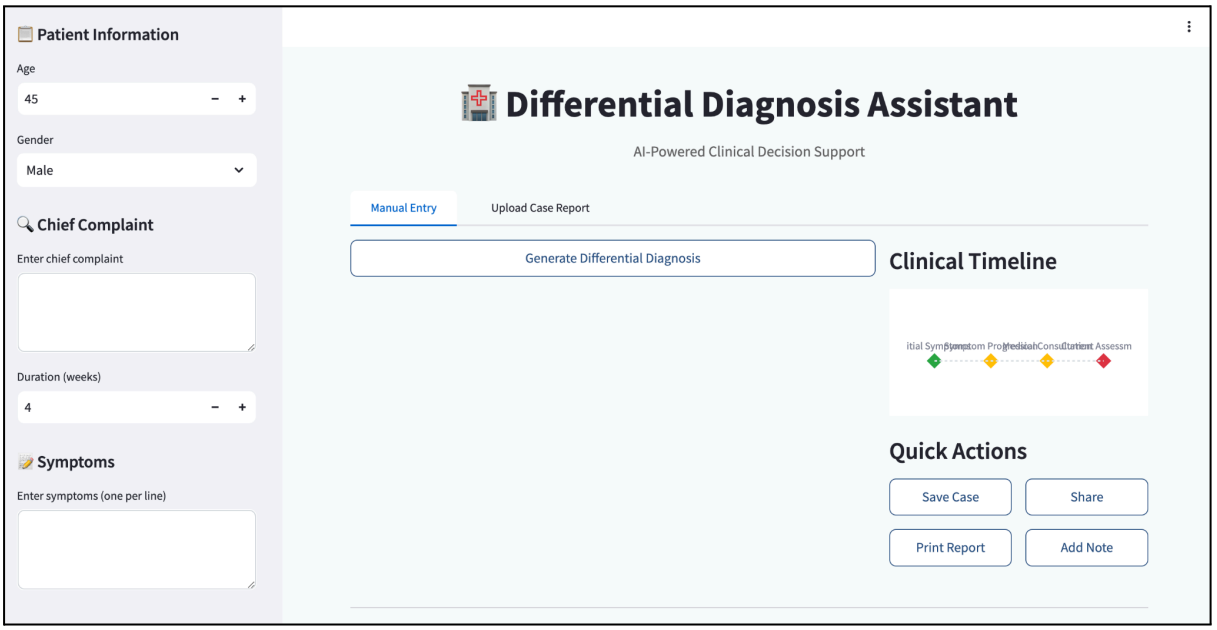
Future Enhancements:

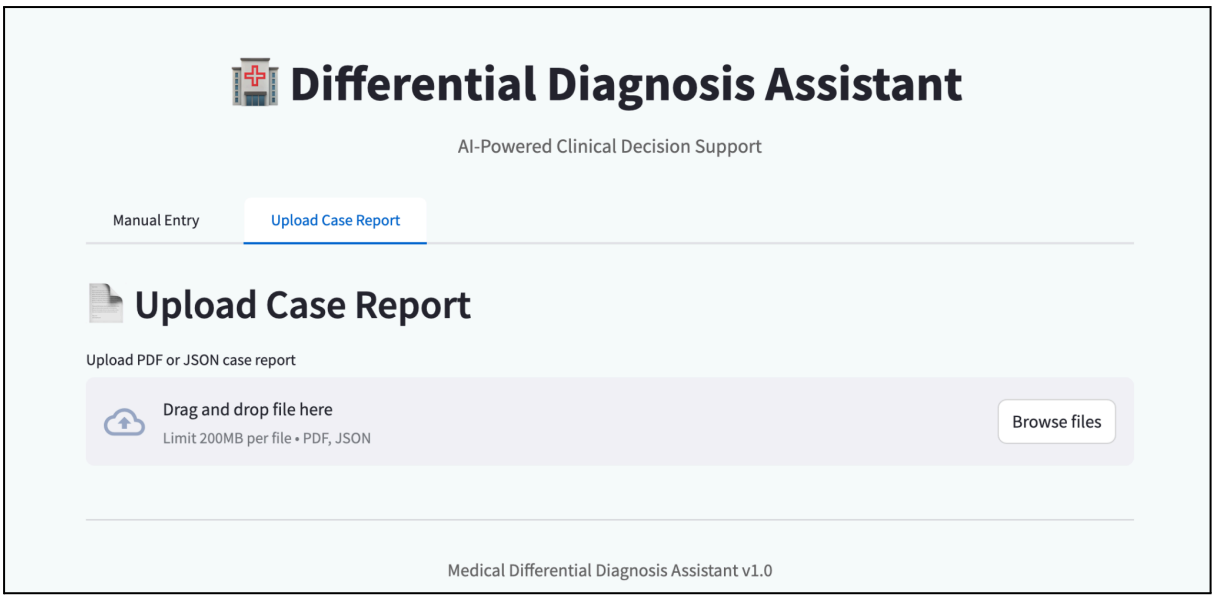
- **AI Integration:** Upon acquiring access to API keys, the application will integrate sophisticated AI models to enhance the accuracy and relevance of differential diagnoses generated. This will leverage cutting-edge technology to provide support based on the latest medical data and diagnostic algorithms.

- **User Interface Improvements:** Continuous improvements to the user interface will ensure that the application remains user-friendly and efficient, adapting to the needs of its clinical users.

Conclusion:

The "Differential Diagnosis Assistant" prototype is a stepping stone towards a fully integrated AI-powered tool that will significantly aid in clinical decision-making. It is designed to evolve, incorporating advanced AI functionalities to provide precise and timely medical diagnoses that support healthcare professionals in delivering superior patient care. This initial phase is crucial for demonstrating feasibility and gathering user feedback, which will guide the final product's development.




Differential Diagnosis Assistant
AI-Powered Clinical Decision Support
Manual Entry
Upload Case Report
Upload Case Report
Upload PDF or JSON case report
Drag and drop file here
Limit 200MB per file • PDF, JSON
Browse files
Medical Differential Diagnosis Assistant v1.0