Instruction

You are a medical data analyst evaluating the accuracy of ChatGPT's differential diagnosis capabilities using Mean Reciprocal Rank (MRR) and Discounted Cumulative Gain (DCG). Your task is to assess the ranking quality of ChatGPT's differentials compared to the actual final diagnosis recorded in medical case files.

Context

I have extracted patient histories from four medical case records and used ChatGPT to generate differential diagnoses along with reasoning. Now, I want to quantitatively evaluate ChatGPT's performance in predicting the final diagnosis mentioned in the case records.

Input Format

You will be provided with the following information for each case:

1. **Case Name** (Identifier for the case)
2. **ChatGPT's Differential Diagnoses** (A ranked list of possible diagnoses)
3. **Final Diagnosis from the Case Record** (The correct diagnosis for the patient)

Task Breakdown

1. **Interpret and Compare Diagnoses**
    - Assess whether the final diagnosis appears in ChatGPT's differential list.
    - If present, determine its rank position.
    - If absent, consider it ranked at the lowest possible position.
2. **Calculate **MRR (Mean Reciprocal Rank)**:
    - Compute the reciprocal rank (1/rank) for each case where the correct diagnosis appears.
    - Take the mean over all cases to get the final MRR score.
3. **Calculate **DCG (Discounted Cumulative Gain)**:
    - Assign relevance scores (higher for earlier ranks).
    - Use the logarithmic discounting formula to compute DCG for ChatGPT's rankings.
    - Normalize with Ideal DCG (IDCG) if necessary.

4. **Aggregate Performance Assessment**
    - Compute the average MRR and DCG across all cases to measure ChatGPT's overall effectiveness.
    - Provide a brief interpretation of the scores.

Output Format

- For each case, return:
    - The **reciprocal rank** of the correct diagnosis.
    - The **DCG score** for the differential ranking.
    - A short **interpretation** of whether ChatGPT's performance was strong, moderate, or weak.
- Finally, return **aggregate MRR and DCG scores**, along with a conclusion on ChatGPT's overall diagnostic accuracy based on patient history alone.

Additional Notes

- If a differential list contains multiple plausible diagnoses, consider medical reasoning to rank their relevance.
- Use **step-by-step reasoning** when computing the metrics to maintain transparency.
- Clearly state any assumptions made during ranking evaluation.