Panacea Final Report

Improving Evaluation Metrics for Clinical Trial Design and Matching Tasks

Paper Overview

The paper titled **"Panacea: A foundation model for clinical trial search, summarization, design, and recruitment"** introduces a large language model (LLM)-based system specifically tailored for clinical development workflows. Panacea is presented as a versatile tool that supports several high-impact tasks, including identifying relevant trials, summarizing eligibility criteria, proposing novel trial designs, and matching patients to suitable trials. The system leverages the capabilities of frontier foundation models like GPT-4 and Gemini Pro, augmented with prompt engineering and medical grounding techniques to align with domain-specific demands.

A central contribution of the Panacea paper is the development of datasets and evaluation strategies for benchmarking LLMs across diverse clinical trial tasks. The authors propose modular tasks such as trial design and patient-trial matching and provide initial scoring pipelines using automatic metrics and expert annotation. Notably, the authors emphasize the growing role of generative models in automating complex, previously manual processes in clinical trial design and recruitment.

Panacea demonstrates strong performance on both information retrieval and generative subtasks, with results suggesting the feasibility of deploying LLMs to support early-stage trial development. Furthermore, the framework offers reproducible methods for measuring LLM effectiveness and highlights promising directions for integrating human oversight and iterative improvement. Despite these strengths, the paper acknowledges gaps in evaluating more nuanced or context-sensitive tasks where traditional metrics may not capture model efficacy fully.

Limitation Addressed and Project Goals

One key limitation discussed in the Panacea paper is the **need for better evaluation metrics** that capture the full complexity of clinical trial tasks. Current automatic metrics (such as BLEU and ROUGE) often fall short when used in isolation, especially in high-stakes applications like trial design or patient-trial matching. These limitations make it difficult to assess whether the model outputs are genuinely useful to domain experts or aligned with clinical goals.

To address this gap, **I conducted a targeted evaluation** of two of Panacea's core tasks: (1) clinical trial design and (2) patient-trial matching. My goal was to compare different evaluation metrics—both automated and model-based—and determine which more effectively captures

quality and utility. By focusing on a user-centered lens, I aimed to identify whether Panacea's outputs aligned with expert reasoning, and to examine if automatic scores alone are sufficient.

Methods

Task Setup and Dataset

I evaluated two generative subtasks from the Panacea framework:

- **Clinical Trial Design** – the model proposes a trial protocol based on a medical condition or disease name.

- **Patient-Trial Matching** – the model predicts which clinical trials best match a synthetic patient profile and provides a justification for each selection.

For the design task, outputs included the trial phase, intervention, and eligibility criteria. For the matching task, the output consisted of a **list of matched trials** and **rationales**—not a binary yes/no judgment for each trial. Instead, matches were evaluated based on how well the set of predicted trials aligned with the gold-standard matches.

I selected five examples per task from Panacea's evaluation data. This subset maintained a manageable annotation load while capturing variation in clinical contexts.

Evaluation Pipeline: I compared three evaluation strategies:

- **Automatic Metrics**:
    - BLEU and ROUGE-L, computed against provided gold-standard references (used only for the design task).

- **Model-Based Scoring**:
    - A second LLM rated each output on a 1–5 scale using a rubric focused on fluency, factual accuracy, relevance, and completeness.

- **Baseline vs. Enhanced Pipeline**:
    - The baseline relied only on BLEU and ROUGE-L.
    - The enhanced pipeline combined automatic metrics with rubric-based LLM scoring for a more nuanced evaluation.

For the matching task, an automatic **trial set overlap score** (0–5) was used in place of BLEU/ROUGE, since matching is a selection task rather than text generation.

Annotation and Scoring Process

For each example, I prompted an LLM to apply a consistent rubric. I also calculated BLEU and ROUGE-L scores to analyze how well surface-level metrics aligned with semantic quality in the design task. While automatic metrics emphasized textual overlap, the LLM scoring captured domain-sensitive reasoning and structure.

For the **matching task**, an automatic score was computed based on the **degree of overlap** between the model's predicted trials and the gold-standard set. This **graded score (0–5)** reflects the quality of trial selection, not binary correctness.

Summary of Scores: The following averages reflect the metrics used:

Task	BLEU	ROUGE-L	Auto Score	LLM Score(Rubric-Based)
Clinical Trial Design	0.123	0.608	3.8	4.14
Patient-Trial Matching	N/A	N/A	5.0	3.74

These scores provided the basis for my comparative analysis across the two tasks.

Results

Clinical Trial Design

The LLM-based rubric gave an average score of 4.14, slightly above the auto score of 3.8, and significantly more informative than BLEU alone. In some examples, BLEU was as low as 0.095 despite outputs being clinically valid and well-structured. ROUGE-L fared slightly better at 0.608, capturing more relevance than BLEU.

For example, in Design Example 4, the auto score was 2.5, but the LLM awarded a 4.7, citing a strong logical structure and appropriate eligibility criteria. This example illustrates the disconnect between n-gram overlap metrics and clinical utility.

Overall, LLM-based scoring better reflected semantic appropriateness, logical consistency, and adherence to task structure—areas where BLEU and ROUGE often fell short due to paraphrasing or stylistic variation.

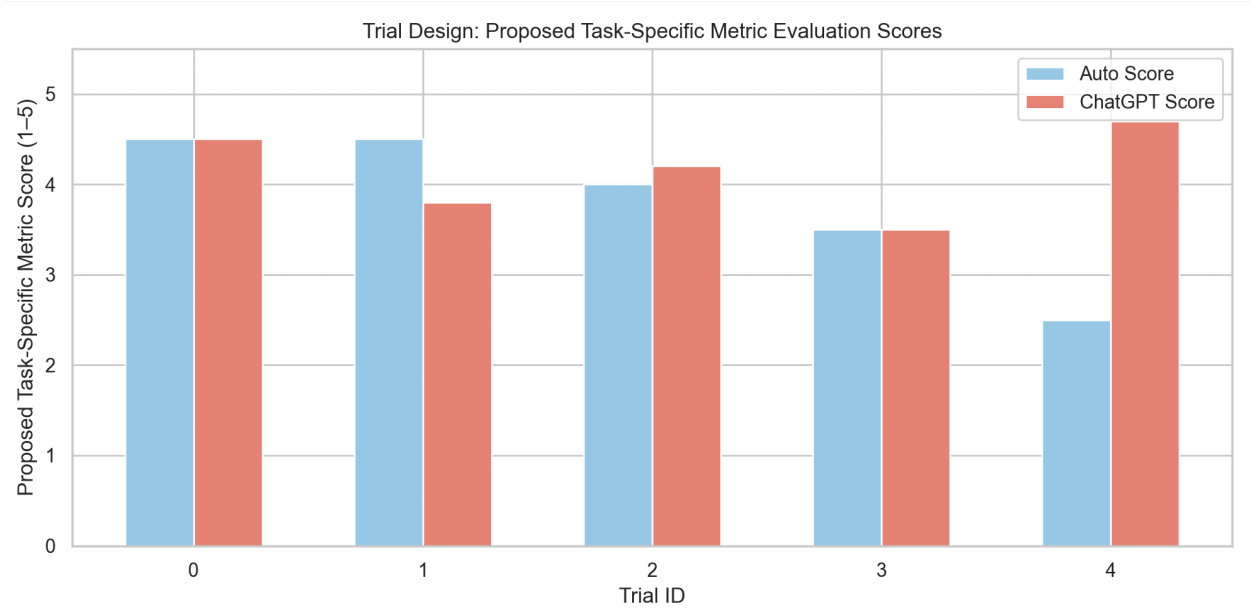
Patient-Trial Matching

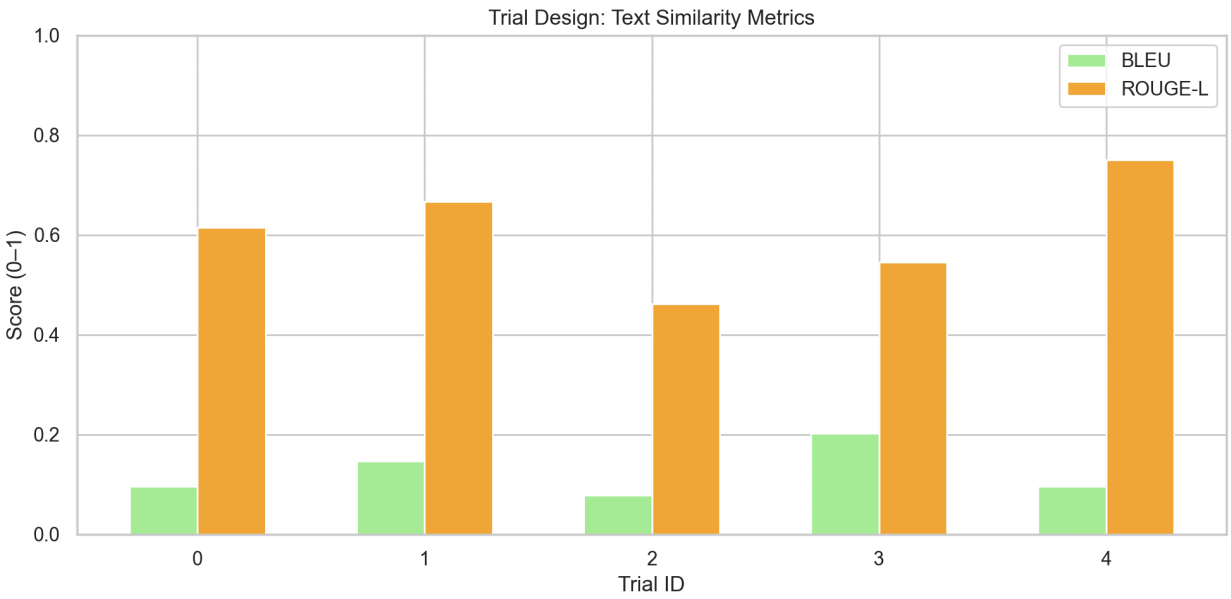
The automatic score was based on the **overlap between predicted and reference trials**, with each example receiving a score between 0 and 5. In this subset, every example received a **score of 5.0**, indicating perfect or near-perfect overlap with the gold-standard set. However, these scores did **not account for the quality of justifications or clinical reasoning.**

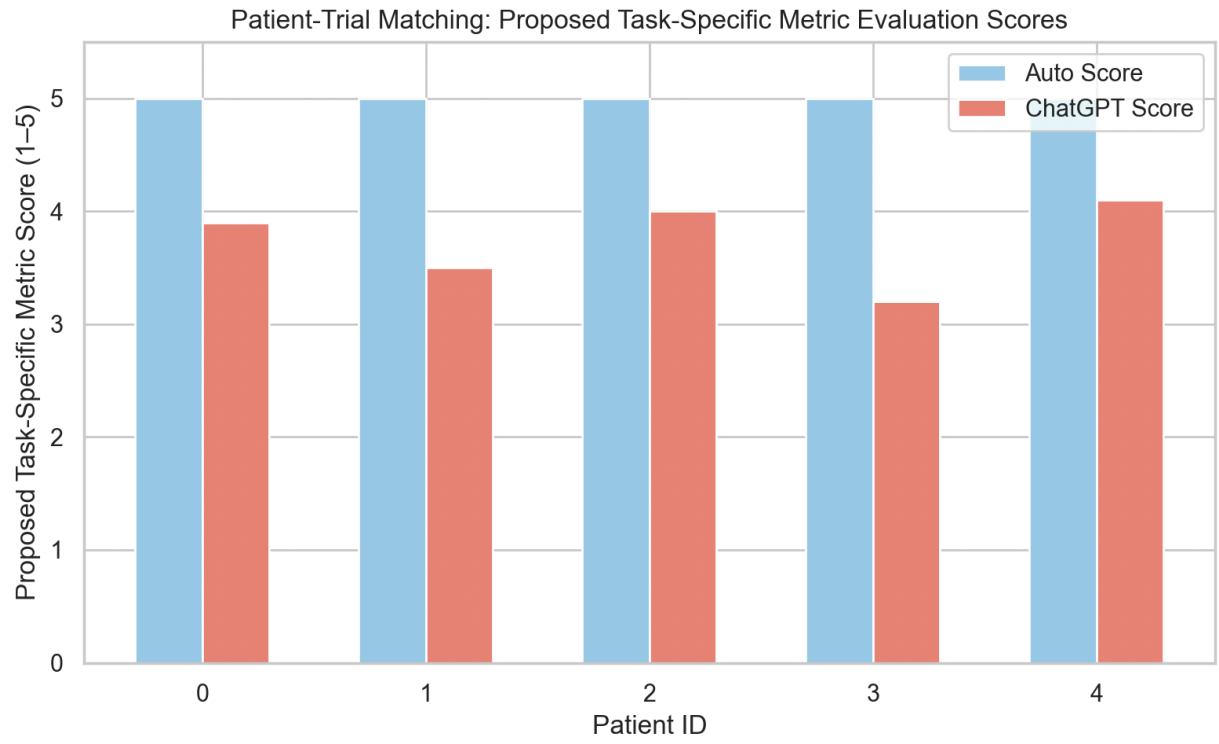
In contrast, LLM scores for matching outputs ranged from **3.2 to 4.1**, showing greater sensitivity to **clinical logic, justification clarity**, and **consideration of contraindications**.

For instance, Matching Example 3 was rated a 5 by the overlap-based auto score but only 3.2 by the LLM, which cited vague justification and overlooked clinical contraindications. This discrepancy highlights how automatic matching scores—while useful for coverage—can miss **important nuances in medical reasoning**.

The enhanced evaluation method (LLM-based rubric) revealed **finer-grained distinctions** in clarity, justification strength, and domain alignment that automatic overlap scores could not capture.




Trial Design: Text Similarity Metrics
BLEU
ROUGE-L
Score (0–1)
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Trial ID


Patient-Trial Matching: Proposed Task-Specific Metric Evaluation Scores
Auto Score
ChatGPT Score
Proposed Task-Specific Metric Score (1–5)
5
4
3
2
1
0
0
1
2
3
4
Patient ID

Limitations

While the enhanced evaluation method offered richer insights, several limitations remain:

- LLM-based scoring introduces potential bias, depending on prompt design and rubric clarity.
- The sample size of five examples per task limits generalizability.
- No domain experts were involved in the rating process, meaning clinical alignment was inferred from the LLM's judgment.
- Automatic metrics like BLEU and ROUGE are still useful in specific contexts but often fail to capture semantic fidelity when outputs differ in phrasing.
- The rubric could be improved with task-specific calibration, especially for structured outputs like trial protocols.
- All evaluated examples were synthetic or curated, which may not fully represent real-world complexity or ambiguity in clinical data.

Conclusion

This evaluation supports the Panacea paper's claim that **conventional automatic metrics are not sufficient** for evaluating complex clinical tasks. BLEU and ROUGE offered limited insight into semantic quality or task adherence—especially in low-overlap, paraphrased responses.

For both tasks, **LLM-based scoring** more reliably captured **structural soundness, clinical relevance, and overall quality**. In the Patient-Trial Matching task specifically, the evaluation revealed that **trial set overlap scores alone can be misleading** if they ignore the justification and medical logic behind selections.

This rubric-based approach proved more capable of identifying useful outputs and detecting subtle errors, making it a valuable supplement to standard metrics in clinical model evaluation.

Additionally, the enhanced pipeline offers a **practical middle ground** between automated evaluation and costly expert review. It enables scalable, semi-structured assessments that can guide model development, deployment, and fine-tuning in sensitive medical contexts.

Future Work

- **Expand the dataset** to include dozens or hundreds of examples for statistically meaningful comparisons across model families.

- **Conduct a comparative study** where domain experts and ChatGPT score the same outputs, measuring inter-rater reliability and alignment.

- **Develop better clinical metrics**, possibly using ontology-based tools (e.g., UMLS concept overlap) or structural completeness checklists.

- Integrate the pipeline into a **user-facing evaluation dashboard**, where model developers can submit outputs and receive real-time, rubric-based scoring.

- Explore fine-tuning or instruction tuning models on benchmark-aligned data, using the enhanced evaluation system as a feedback loop.