Project Instructions: Clinical Language Model Evaluation

# **Overview of the Task**

We have prepared 50 anonymized patient cases, each including:

* **Clinical Information** about each patient
* **Treatment Plans** using this clinical information, we generated personalized treatment plans for each patient with three different AI large language models (LLMs), each designed to provide recommendations based on established clinical guidelines.

The models used are:

|  |  |
| --- | --- |
| **In-house RAG Model** | Developed in our lab, based on NIH guidelines, and PubMed references |
| **Claude Sonnet 3.5** | An advanced AI, guided by prompts for NIH-compliant treatment plans |
| **Llama 3** | Another advanced AI, also guided by prompts for NIH-compliant plans |

The goal of this evaluation is to assess each treatment plan’s **quality**, **accuracy**, **reasoning**, and **personalization**. Additionally, we aim to compare the performance of these three models to better understand their strengths, weaknesses, and areas for improvement. Your expert evaluations will play a crucial role in achieving these objectives and advancing the development of AI-driven treatment planning.

# **Documents provided for the Evaluation**

## Word Document: Patient Cases and AI-Generated Treatment Plans

Each patient case includes:

1. **Patient Description**: A detailed clinical case summary.
2. **Ground Truth Treatment Plan**: Provided for reference only. Note that this is not evaluated and may not always be accurate or complete.
3. **Treatment Plans from four LLMs**: Model 1 (NIH), Model 2 (NIH), Model 3 (NIH), Model 3 (PubMed)

## Excel Document: Evaluation Template

The evaluation template where you will record assessments. Further Information in “Evaluation Criteria”.

# **Evaluation Criteria**

## Clinical and Scientific Consensus

Assess how well the treatment plan aligns with established clinical guidelines and scientific evidence.

|  |  |
| --- | --- |
| **Aligned** | Consistent with current guidelines and supported by evidence |
| **Partially Aligned** | Partially consistent with current guidelines and supported by evidence |
| **Opposed** | Contradicts established guidelines or evidence |
| **No** **Consensus** | Addresses areas where no clear clinical or scientific agreement exists |

## Reasoning and Chain of Thought

Evaluate the logical progression and clarity of reasoning in the treatment plan.

|  |  |
| --- | --- |
| **Disjointed Chain of Thought** | Lacks a clear and logical structure; reasoning is fragmented and difficult to follow |
| **Partially Coherent Chain of Thought** | Demonstrates some logical progression but has gaps or inconsistencies; alternative approaches are not well addressed |
| **Comprehensive and Insightful Chain of Thought** | Clear, logical, and well-constructed reasoning, incorporating evidence and considering alternatives effectively |

## Content Accuracy and Completeness

Determine the factual accuracy and thoroughness of the plan.

|  |  |
| --- | --- |
| **Incorrect with harm** | Contains errors that could harm the patient |
| **Incorrect but no harm** | Contains errors that do not negatively impact patient care |
| **Incomplete** | Missing major components of a comprehensive treatment plan |
| **Partially complete** | Covers most essential elements but lacks some important details |
| **Content is Good and Complete** | Accurate and addresses all necessary aspects of treatment |

## Personalization of Treatment Plan

Assess how well the plan is tailored to the specific patient case.

|  |  |
| --- | --- |
| **Not personalized** | Generic, does not consider patient-specific factors |
| **Partially personalized** | Incorporates some relevant patient-specific factors |
| **Highly personalized** | Fully tailored to the patient’s unique characteristics and needs |

In the Excel Document:

Columns B-E for Model 1 (NIH)

Columns F-I for Model 2 (NIH)

Columns J-M for Model 3 (NIH)

Columns N-Q for Model 3 (PubMed)

## Comments Section (Column R)

* **Optional:** Provide any additional observations, concerns, or suggestions regarding the LLM-generated treatment plan or comments of your review **only if** needed