SLM Evaluation Task

**Background:**

In this task, you will evaluate how effectively a small language model (SLM) predicts Emergency Department (ED) disposition decisions — such as whether to admit or discharge a patient — for individuals with traumatic brain injury (TBI). The focus is on evaluating the predicted ED disposition and reasoning for 30 patients, whose summary is provided at the top of each page.

The language model is given detailed patient information and asked to predict a disposition along with a clinical justification. Importantly, two different prompting strategies are used to generate two different responses: Option 1 and Option 2. These options are randomly ordered and flipped for each case to minimize presentation bias.

**Role**:

For each patient case (p1 to p30), you will:

1. Read the patient summary at the top – Every patient case is referred to as Patient A.
2. Review both LLM-generated responses (Option 1 and Option 2) including their predicted ED disposition and clinical reasoning.
3. Evaluate each response independently using the following three metrics:
   1. Clinical Relevance of the reasoning
   2. Clinical Reasoning Quality
   3. Overall Usefulness.

The patient summary is provided at the top and Option 1 and Option 2 are provided below explaining the reasoning and its predicted ED disposition for Patient A (Shown below)

A screenshot of a computer

AI-generated content may be incorrect.

Each metric is rated using a slider located directly below each option.

A screenshot of a computer

AI-generated content may be incorrect.

**Instructions for Evaluation:**

1. Use the sliders below Option 1 to rate Option 1, and the sliders below Option 2 to rate Option 2.
2. Once you have rated both options, click the "Next" button to proceed to the next patient.
3. After each evaluation, your scores are automatically saved to a file named: evaluation\_log.csv.
4. If you made a mistake and want to go back and change it. Open the evaluation\_log.csv file on excel and remove the row that you want to redo.

### ****Progress and Completion:****

### You will evaluate ****30 patients in total****.

### You may ****pause and resume**** your evaluation — the app will continue from where you left off.

### Once all 30 patients are completed, please ****send the**** evaluation\_log.csv ****file**** to: ****pranav.manjunath@duke.edu****

**Evaluation Metrics:**

1. **Clinical Relevance of Reasoning (1-5)**

Assesses whether the LLM’s reasoning focuses on clinically pertinent features (e.g., vital signs, imaging, guidelines) rather than tangential or irrelevant details.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Entirely off-topic (no clinical relevance) | Mostly irrelevant (mentions some clinical terms but misses key factors) | Partially relevant (identifies some key factors but includes irrelevant details) | Mostly relevant (focuses on most key clinical factors, minor irrelevant tangents) | Completely relevant (all reasoning centers on appropriate clinical features without extraneous information).” |

1. **Reasoning Quality (1–5)**

Assesses the logical coherence and depth of the LLM’s explanation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| No identifiable reasoning (jumps to a conclusion without explanation) | Minimal/fragmented reasoning (only mentions one or two facts, no clear logic) | Basic reasoning (identifies some relevant factors but lacks clear logical flow) | Good reasoning (identifies multiple factors, logical flow is mostly clear) | Excellent reasoning (comprehensive, stepwise logic; each step follows coherently). |

1. **Practical Usefulness (1–5)**

Assesses how actionable and helpful the LLM’s response would be for actual clinical decision-making.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Not useful (no actionable guidance; would not aid my decision)) | Slightly useful (minimal guidance; clinician would need to find most information elsewhere) | Moderately useful (some actionable points, but key details are missing) | Very useful (clear guidance on disposition, with most relevant details) | Extremely useful (comprehensive guidance that could directly inform ED disposition) |

### ****Note on LLM Response Sections****

Each LLM response is typically structured into **four sections**. However, for the purposes of this evaluation, we have **removed Section 2**. Below is an overview of the remaining sections you will see:

* **Section 1:** A recap of the most important clinical features relevant to predicting the ED disposition for Patient A.
* **Section 3:** The model’s clinical reasoning explaining why it chose the specific disposition.
* **Section 4:** The final ED disposition prediction.

**Why Section 2 is excluded:**  
Section 2 usually contains a comparison between Patient A and similar past patients. In our setup, we have provided the model with those similar patients, so any comparison made in the response can be assumed to be based on accurate information. To streamline your review, we’ve removed this section.

Please focus your evaluation on **Sections 1, 3, and 4**.

How to Run the AI Evaluation App

This guide will walk you through the steps to run a Streamlit-based AI evaluation app, even if you've never installed Python before.

# Step 1: Install Python

1. Go to: https://www.python.org/downloads  
2. Click the big yellow “Download Python 3.x.x” button.  
3. Run the installer.  
4. ✅ IMPORTANT: During install, check the box that says “Add Python to PATH”.

# Step 2: Install Streamlit

1. Open the Terminal (Mac) or Command Prompt (Windows).  
2. Type the following command and press Enter:  
 pip install streamlit pandas

# Step 3: Download the App Files

You should receive a folder (e.g., `app1/`) containing:  
- app.py (the main app)  
- patient\_set.json (the data)  
  
Make sure all the files are in one folder.

# Step 4: Open the App

1. Open Terminal or Command Prompt.  
2. Navigate to the folder where the files are stored. For example:  
 cd Desktop/app1  
3. Run the app with this command:  
 streamlit run app.py

# Step 5: Use the App

The app will open automatically in your web browser. You can now start evaluating responses. All your results will be stored in a csv called evaluations\_log.csv. You can cancel out of it at any time and restart the app and it will continue from where its left off!

# Next Time

To run the app again, just repeat Step 4. You do not need to reinstall anything.