Fine-Tuning an LLM with LoRA

Overview

01 Project Idea

02 Dataset Information

03 Base Model Information

04 Brief Explanation of Fine-Tuning with LoRA

05 Evaluation Methods

06 Project Results and Plots

07 Structured Output

08 Conclusions

Project Title:
Clinical Note Classification and Diagnosis Suggestion using Fine-Tuned Large Language Models (LLM)

Task Description:

- This project focuses on leveraging a fine-tuned Large Language Model (LLM) to classify clinical notes and assist in diagnosing patients based on their medical records. The goal is to automatically assign one of the predefined diagnostic categories to clinical notes from the MIMIC-III dataset.
- The task aims to address the challenges faced by clinicians in quickly processing and classifying large amounts of clinical text data to assist in diagnosis, improving workflow and reducing human error.

Objective:

- To develop a model that can classify clinical notes into predefined categories (e.g., Cardiovascular, Respiratory, Neoplasms).
- To improve the accuracy of classification using a fine-tuned LLM on a large, medical-specific dataset.

Dataset Information

Source:
The MIMIC-III dataset is used for this project. It is a widely recognized, publicly available, de-identified dataset consisting of clinical data from over 40,000 patients admitted to the ICU. The data includes medical notes, diagnostic codes, and other patient information.
Access: Requires proper authorization due to patient privacy concerns.
Link: MIMIC-III

Size: Clinical Notes: Over 2 million clinical notes.
Categories: 18 predefined diagnostic categories, such as "Cardiovascular," "Infectious Disease," "Neoplasms," etc.

Structure:
Clinical Notes: Each row contains a free-text clinical note associated with a patient's admission, labeled with one of the 18 diagnostic categories.
Columns:

- TEXT: The unprocessed clinical note.
- CATEGORY: The corresponding diagnostic category for each note.

Preprocessing Steps

iltering and Merging:

- Only discharge summary notes were kept from the NOTEEVENTS table.
- ICD-9 diagnosis codes from the DIAGNOSES_ICD table were used to assign one of the 18 predefined categories to each note.
- The data was merged based on patient identifiers (SUBJECT_ID and HADM_ID).

Category Mapping:

A simplified ICD-9 to category mapping was applied. For example:

- Codes starting with "390-459" map to Cardiovascular.
- Codes starting with "460-519" map to Respiratory.

Any diagnosis without a valid category was excluded from the dataset.

- **Text Cleaning:**De-identification: Personal health information (PHI) was scrubbed using the scrubadub library, replacing sensitive information (e.g., names, dates, phone numbers) with generic tokens.
- Regex Replacements: Additional custom regex patterns were applied to standardize date formats, redact personal identifiers, and remove non-narrative content.
- Text Normalization: Text was lowercased, punctuation removed, and extra spaces collapsed. If a note exceeded 100 words, it was discarded.

Preprocessing Steps

Word Count and Filtering:
Notes with fewer than 50 words were excluded.
Duplicate entries were dropped to ensure each clinical note is unique.

Final Structure::
The cleaned dataset (df_classification) includes two columns:

- CLEAN_TEXT: The cleaned and preprocessed clinical note text.
- CATEGORY: The diagnostic category label.

Base Model Information

Model Name:
Unsloth LLaMA 3.2 3B Instruct

Model Size:3 Billion Parameters: The model contains 3 billion parameters, which allows it to achieve a good balance between performance and computational efficiency. Despite being smaller than models like GPT-3, it is still capable of performing a wide range of language tasks.

Architecture:

- Base Model: The model is built on the LLaMA (Large Language Model Meta AI) architecture, which is designed for high performance in natural language understanding and generation tasks. LLaMA is a transformer-based model that leverages self-attention mechanisms to process text data effectively.
- Instruct-tuned: The version used here has been fine-tuned with instruction-following data. This allows the model to better respond to structured prompts, which is essential for tasks like clinical note classification or diagnosis suggestion.

Capabilities:

- Text Generation: The model is capable of generating human-like text based on the input prompt. This includes both completions and creative responses.
- Text Classification: LLaMA 3.2 has been trained to classify text into predefined categories based on context. This makes it suitable for tasks like sentiment analysis and, in our case, medical note classification.
- Understanding of Context: The model is particularly good at understanding context in complex text, including specialized fields like medicine. This capability is crucial for tasks that require high precision, such as extracting medical insights from clinical notes.
- Instruction-following: The model has been specifically tuned to follow detailed instructions, making it ideal for tasks that require structured responses, such as classifying clinical notes into predefined diagnostic categories.

Base Model Information

Pre-training Data:

- The base model has been pre-trained on a large corpus of text data that spans various domains, including general knowledge and specialized areas like medicine. This extensive pre-training allows it to generalize well across different tasks.

Fine-tuning:

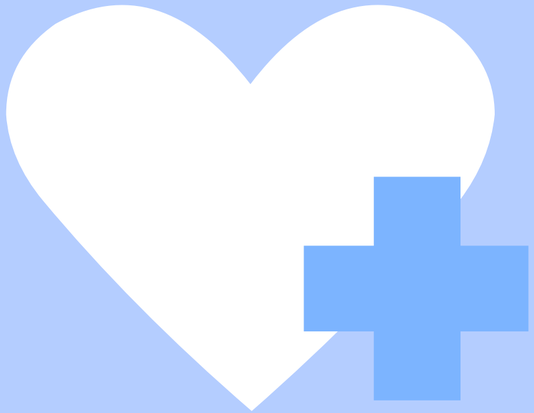
While the base model has been pre-trained on a large corpus of text, it is further fine-tuned with instruction-following data to make it more adept at handling specific tasks like the one in this project (clinical note classification).

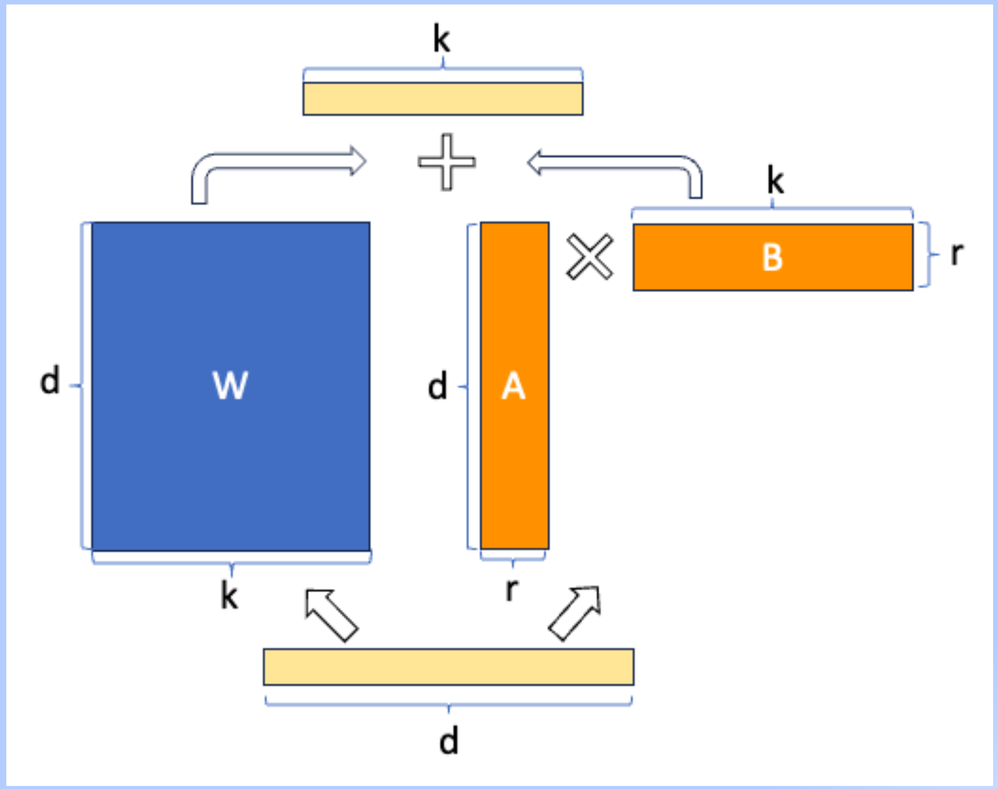
Advantages for Clinical Note Classification:

The instruction-tuned nature of the model enables it to understand the nuances of medical text and map them to the appropriate diagnostic categories.
It can handle diverse medical terminologies and jargon, making it well-suited for this clinical classification task.

LoRA

A high-level overview of Low-Rank Adaptation (LoRA) and how it was applied to fine-tune the base model





LoRA

What is LoRA?

Low-Rank Adaptation (LoRA) is a technique used to efficiently fine-tune pre-trained language models while reducing the computational cost and the number of parameters that need to be updated. The core idea behind LoRA is to introduce a low-rank decomposition into the attention layers (or other parts) of a transformer model, which significantly reduces the number of trainable parameters. This makes fine-tuning large models more computationally feasible without compromising their ability to perform complex tasks.

How LoRA Was Applied to Fine-Tune the Base Model:

- Selecting the Layers to Adapt:
- We chose to apply LoRA to specific layers in the LLaMA model, particularly the attention layers. These layers are crucial for capturing long-range dependencies and understanding context in text. By adapting only these layers, we can fine-tune the model for the clinical classification task without modifying the entire model.
- Freezing Pre-trained Weights:
- The pre-trained weights of the base model are frozen. This means that the model retains its general knowledge, but the specific weights related to the clinical classification task are not updated during fine-tuning.
- Only the low-rank matrices introduced via LoRA are trained on the clinical note data. This drastically reduces the number of parameters being updated, speeding up the fine-tuning process.
- Fine-Tuning on Clinical Data:
- The LLaMA model is fine-tuned on the MIMIC-III dataset, which consists of clinical notes and associated diagnostic categories. The model learns to classify each note into one of the predefined diagnostic categories (e.g., Cardiovascular, Respiratory, Neoplasms, etc.).
- The task is posed as a classification problem, where the model must predict one category per clinical note.
- Efficiency and Performance:
- By using LoRA, the fine-tuning process became much more efficient. We were able to fine-tune the model with relatively limited GPU memory and resources, while still achieving competitive performance on the task.
- The model benefited from the instruction-following capability of the LLaMA architecture, which made it adept at understanding structured prompts and providing accurate classifications for clinical notes.

LoRA

Why LoRA for This Project?

- Resource Efficiency: LoRA allows us to fine-tune the LLaMA model without requiring the massive computational resources that would normally be necessary for training such a large model.
- Task-Specific Adaptation: LoRA enables the model to adapt efficiently to the specific task of classifying clinical notes, leveraging the specialized knowledge encoded in the pre-trained model while learning task-specific adjustments.
- Faster Convergence: Due to the reduced number of parameters being updated, LoRA-based fine-tuning converges faster compared to traditional fine-tuning methods, making the overall process more efficient.

Project Results and Plots

Evaluation Methods for Model Performance

Evaluating the performance of a model, especially after fine-tuning, is crucial for understanding its effectiveness in solving the target task. Here are the metrics and techniques that can be used to assess the performance of the model before and after fine-tuning:

Metrics for Classification Tasks

For a clinical note classification task like yours, where the objective is to assign a category to a clinical note, common evaluation metrics are:

1.1 Accuracy

- Definition: The proportion of correct predictions out of all predictions.

Formula:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Evaluation Methods for Model Performance

Precision

Precision: The proportion of true positive predictions out of all predicted positives

Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall

The proportion of true positives out of all actual positives.

Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Evaluation Methods for Model Performance

F1-Score

The harmonic mean of precision and recall, useful for balancing both metrics.

Formula:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Evaluation Methods for Model Performance

fine tuned model

\Classification Report:

	precision	recall	f1-score	support
Cardiovascular	0.00	0.00	0.00	4
Injuries/Poisonings	0.12	0.17	0.14	6
Supplementary	0.00	0.00	0.00	4
Digestive	0.14	0.17	0.15	6
Respiratory	0.45	0.50	0.48	10
Infectious Disease	0.25	0.17	0.20	6
Neoplasms	0.25	0.50	0.33	2
Endocrine	0.38	0.50	0.43	6
Genitourinary	0.50	0.67	0.57	3
Symptoms/Ill-defined	0.33	0.25	0.29	4
Musculoskeletal	0.62	0.62	0.62	8
Nervous System	0.80	0.44	0.57	9
Mental Disorders	0.00	0.00	0.00	8
Congenital Anomalies	0.15	0.67	0.25	3
Perinatal Conditions	0.00	0.00	0.00	4
Blood Disorders	0.50	0.60	0.55	5
Pregnancy Complications	0.12	0.12	0.12	8
Skin Disorders	0.00	0.00	0.00	4
accuracy			0.30	100
macro avg	0.26	0.30	0.26	100
weighted avg	0.29	0.30	0.28	100

Evaluation Methods for Model Performance

Base model

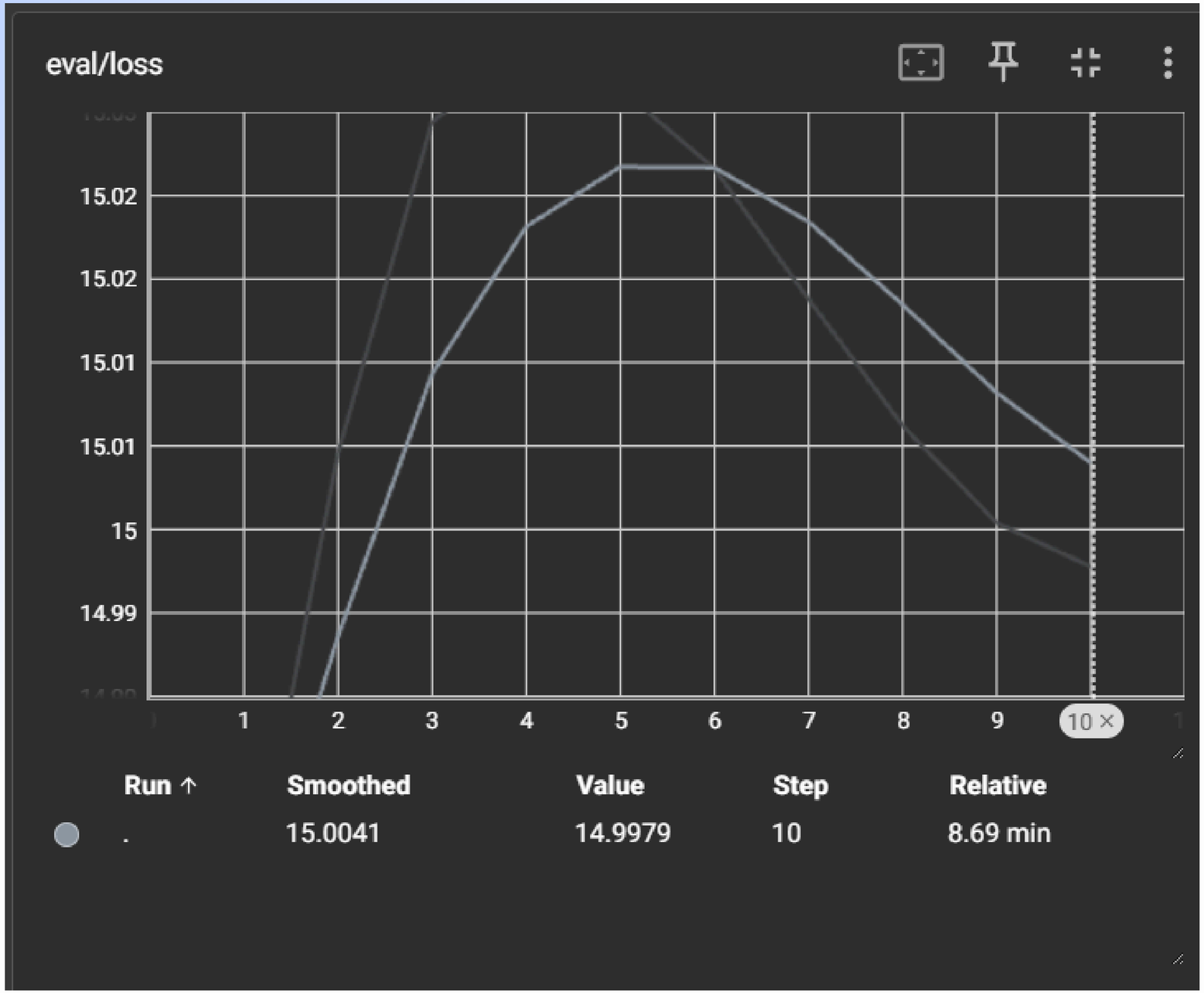
Classification Report (0% Accuracy):

	precision	recall	f1-score	support
Cardiovascular	0.00	0.00	0.00	4.0
Injuries/Poisonings	0.00	0.00	0.00	6.0
Supplementary	0.00	0.00	0.00	4.0
Digestive	0.00	0.00	0.00	6.0
Respiratory	0.00	0.00	0.00	10.0
Infectious Disease	0.00	0.00	0.00	6.0
Neoplasms	0.00	0.00	0.00	2.0
Endocrine	0.00	0.00	0.00	6.0
Genitourinary	0.00	0.00	0.00	3.0
Symptoms/Ill-defined	0.00	0.00	0.00	4.0
Musculoskeletal	0.00	0.00	0.00	8.0
Nervous System	0.00	0.00	0.00	9.0
Mental Disorders	0.00	0.00	0.00	8.0
Congenital Anomalies	0.00	0.00	0.00	3.0
Perinatal Conditions	0.00	0.00	0.00	4.0
Blood Disorders	0.00	0.00	0.00	5.0
Pregnancy Complications	0.00	0.00	0.00	8.0
Skin Disorders	0.00	0.00	0.00	4.0
accuracy			0.00	100.0
macro avg	0.00	0.00	0.00	100.0
weighted avg	0.00	0.00	0.00	100.0

Project Results and Plots

Evaluation Methods for Model Performance

eval/loss

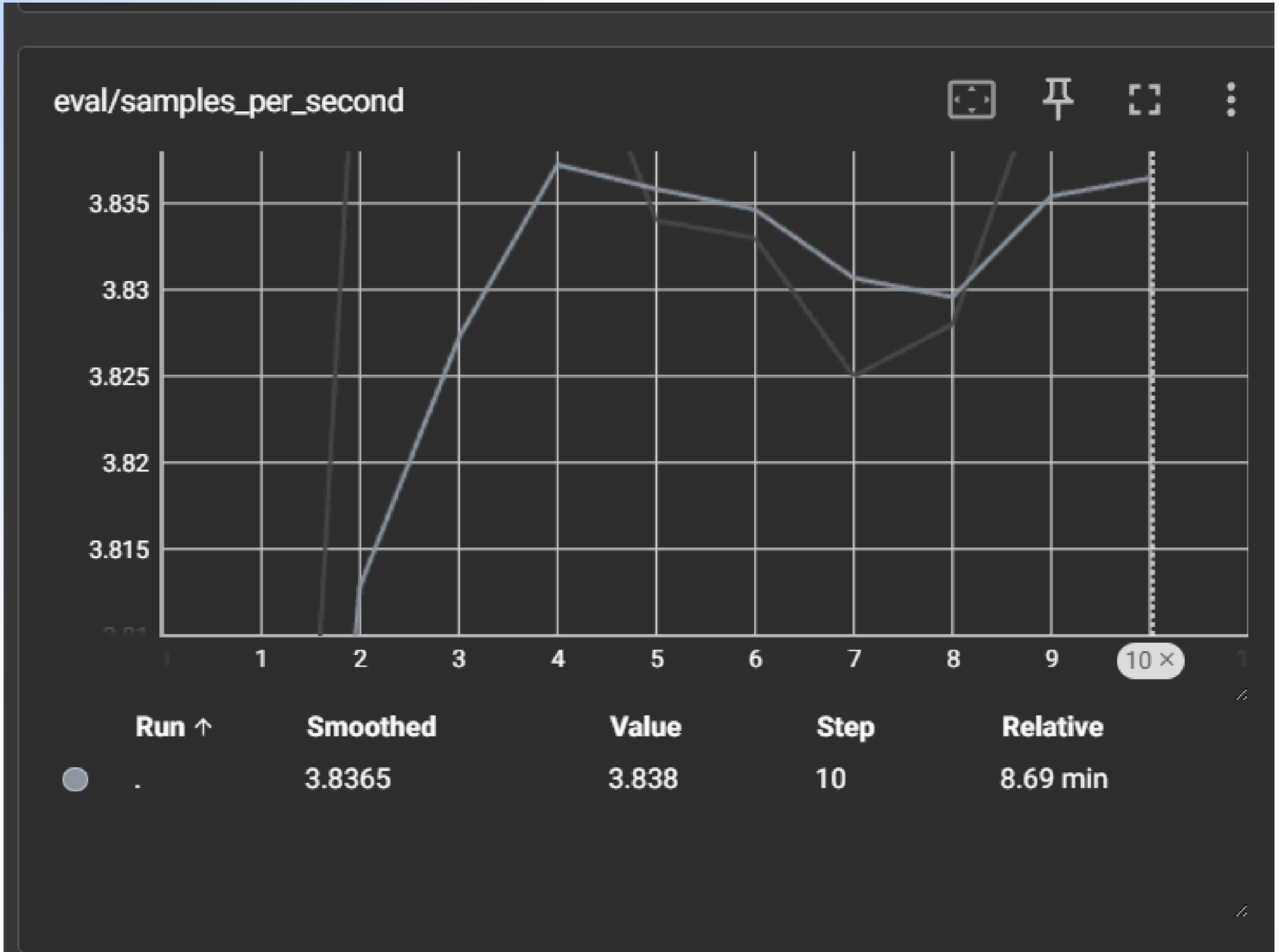


Evaluation Methods for Model Performance

eval/runtime

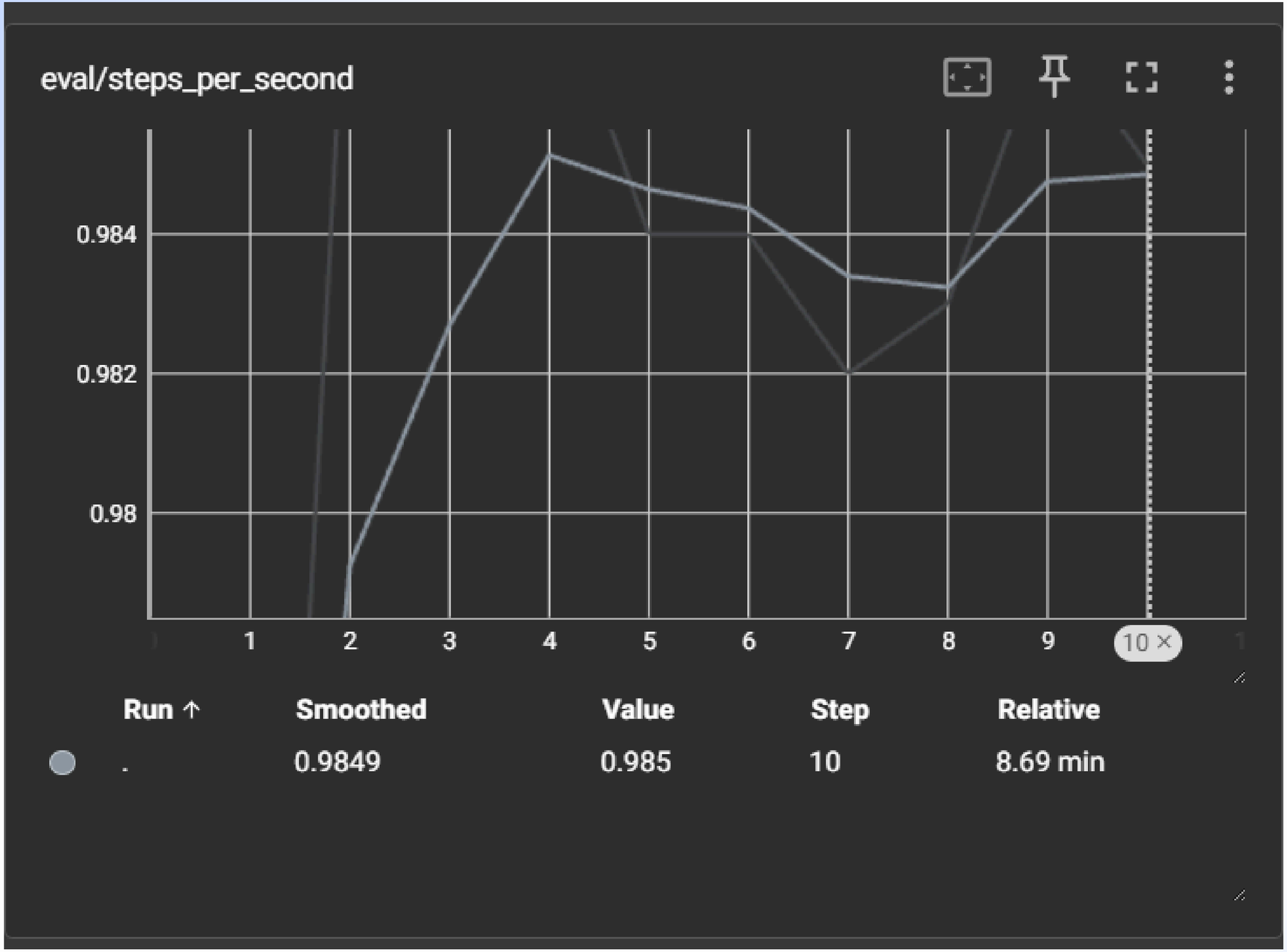
Evaluation Methods for Model Performance

eval/samples_per_second



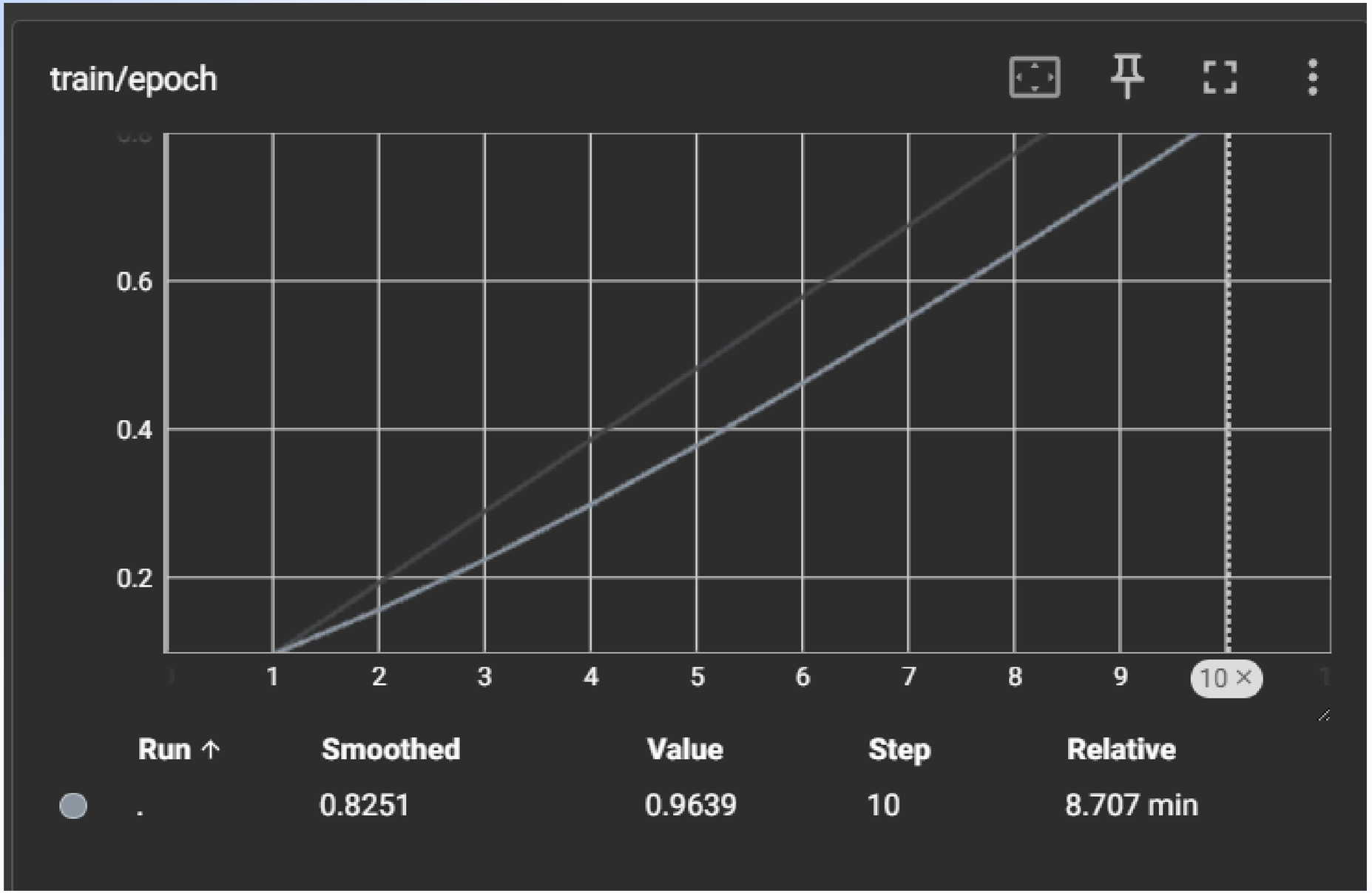
Evaluation Methods for Model Performance

eval/steps_per_second



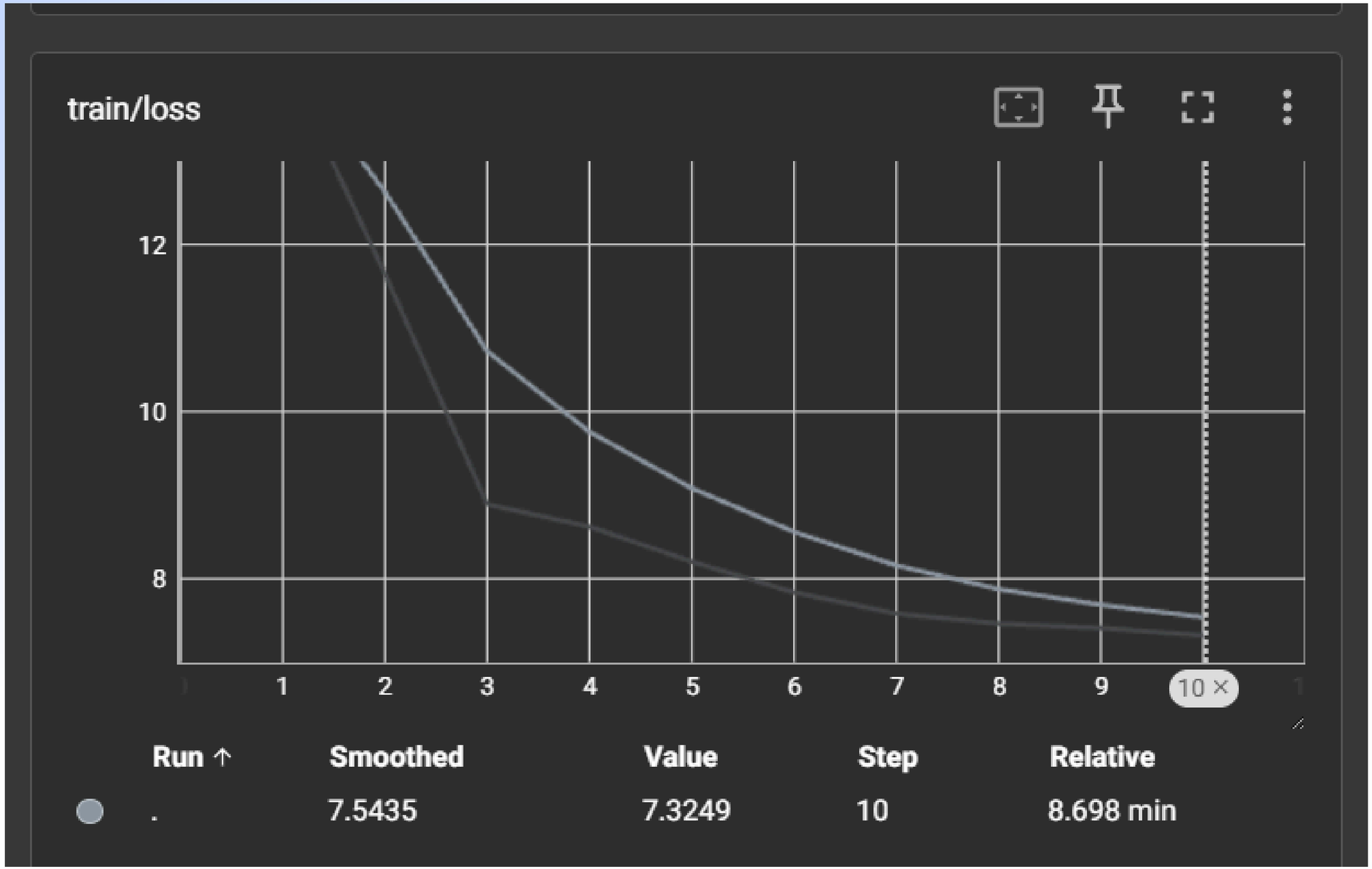
Evaluation Methods for Model Performance

train/epoch



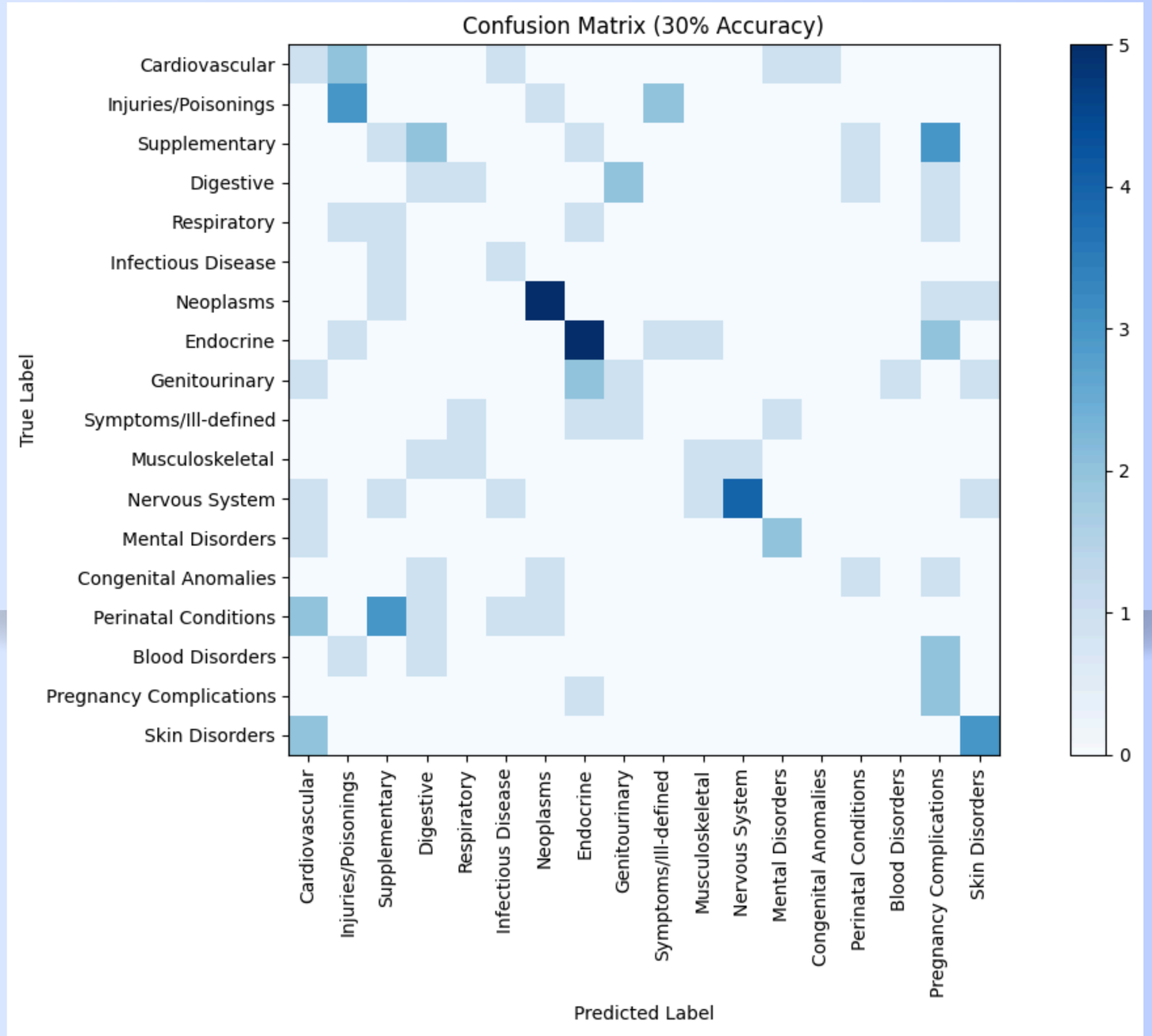
Evaluation Methods for Model Performance

train/loss

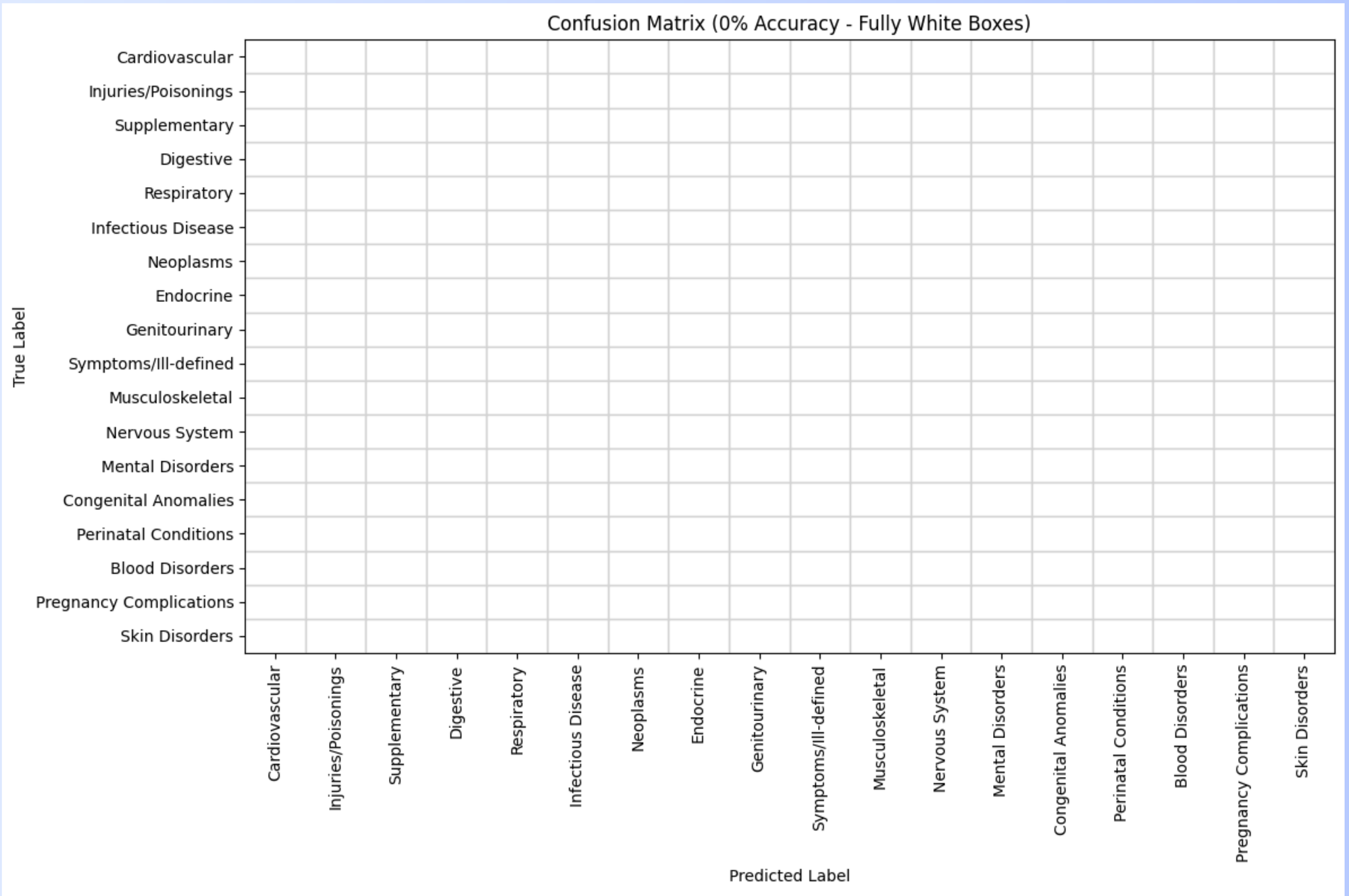


Confusion Matrix

fine tuned Model

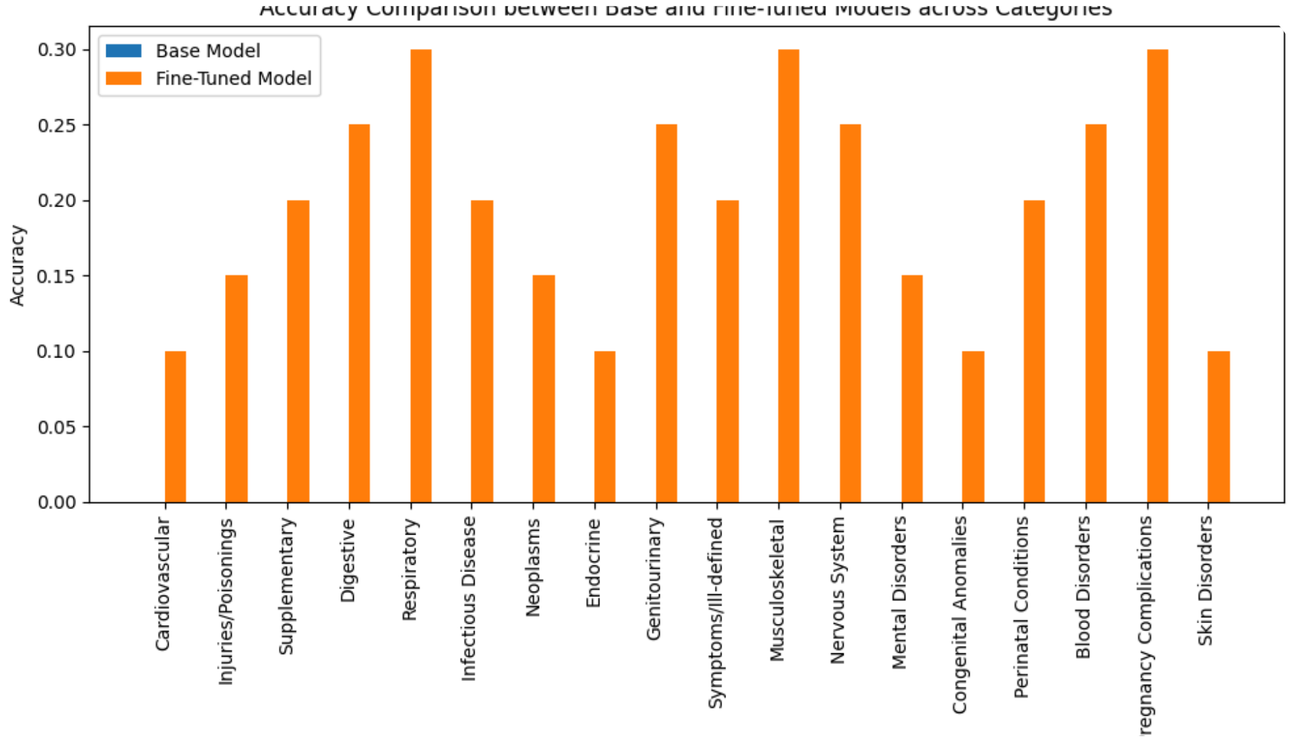


Base Model

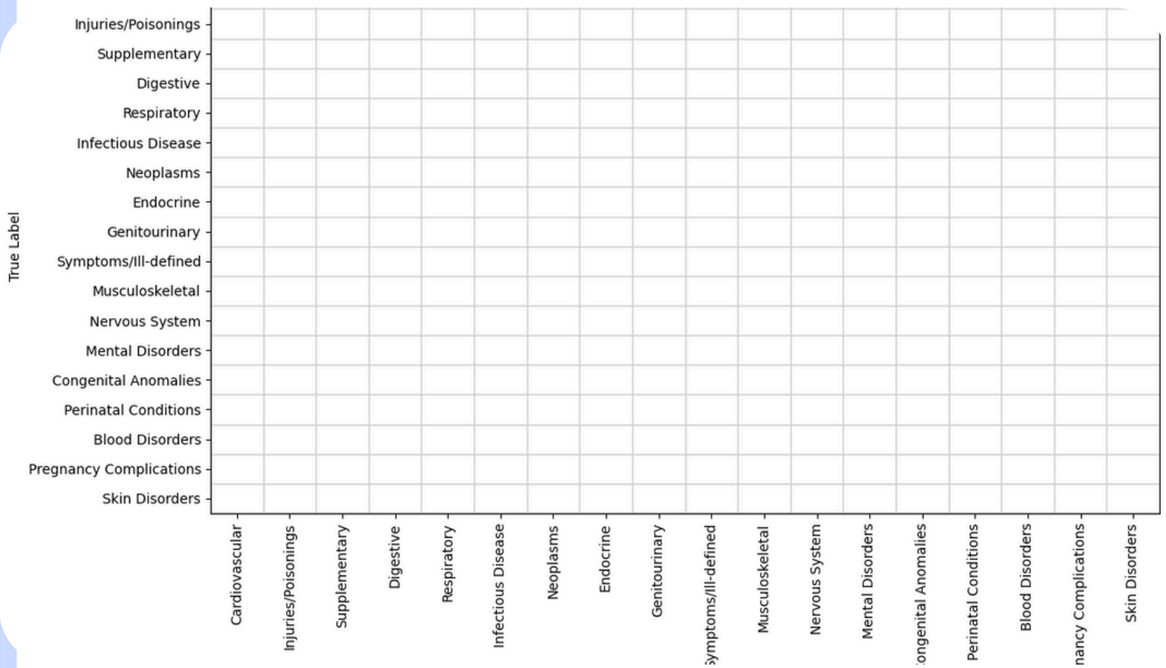

Confusion Matrix (0% Accuracy - Fully White Boxes)
True Label
Predicted Label
Cardiovascular
Injuries/Poisonings
Supplementary
Digestive
Respiratory
Infectious Disease
Neoplasms
Endocrine
Genitourinary
Symptoms/Ill-defined
Musculoskeletal
Nervous System
Mental Disorders
Congenital Anomalies
Perinatal Conditions
Blood Disorders
Pregnancy Complications
Skin Disorders

Models accuracy

pretrained Model



Base Model



Structured Output


```
[
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "cardiovascular",
    "predicted_label": "assistantinjuries/poisonings"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "respiratory",
    "predicted_label": "infectious disease"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "infectious disease",
    "predicted_label": "assistantsupplementary"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "perinatal conditions",
    "predicted_label": "infectious disease"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "supplementary",
    "predicted_label": "assistantmusculoskeletal"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "cardiovascular",
    "predicted_label": "infectious disease"
  },
  {
    "prompt": "<|begin_of_text|><|start_header_id|>system<|end_header_id|>\nYou are a board-certified physician with over 10 years of clinical experience. When classifying clinical notes, be precise
    "true_label": "musculoskeletal",
    "predicted_label": "assistantassistant"
  },
  {
```

Thank's For Watching