REPORT PROJECT

PROJECT:	NAME:	DUE DATE
AI for Medical Transciptions	Bach Nhat	5/12/2025

SUMMARY

A medical dataset that supports research and model development without exposing sensitive or personally identifiable patient information, in compliance with HIPAA (Health Insurance Portability and Accountability Act) privacy regulations.

PROBLEMS

- The dataset still contains a substantial amount of data.
- A confusion matrix is used to examine how the model misclassifies one class as another.
- Some instances are labeled as cardiology despite containing gastrointestinal-related content, indicating potential labeling noise.
- The dataset also exhibits class imbalance.

PROCESS

1. Utilize AI to identify topics for the project.
2. Search for available datasets on Kaggle.
=>https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions
3. Develop an NLP model using Python.
4. Review dataset evaluations on Kaggle to facilitate data comparison.

ENHANCE

- Use BioClinicalBERT / PubMedBERT models instead of the standard BERT.
- Apply semi-supervised relabeling by leveraging the trained model to detect “mislabelled” samples.
- Normalize and reduce noise in the data.