Environment Setup

* Set up a Python 9.x environment in a Unix based OS (tested on Ubuntu 20.04)
* Install following packages:
  + openai 0.28.0
  + numpy 1.24.2
  + pandas 1.5.3
  + tiktoken 0.4.0
  + scikit-learn 1.0.2
  + nltk 3.8.1
    - install stopwords for nltk
  + xgboost 1.7.4

Download Third-Party Data

* Download *MIMIC-SBDH.csv* from (<https://github.com/hibaahsan/MIMIC-SBDH>)
* Download *ADMISSIONS.csv* and *NOTEEVENTS.csv* from (<https://physionet.org/content/mimiciii/1.4/>)
* Place these files in the *AnnotateGPT-Codebase* folder

Running AnnotateGPT – SHARECOPY.ipynb

* Open the file using a Jupyter IDE
* Run every cell in order, carefully reading and completing the set up steps
* *“XX-XX-gpt-train.pkl”* can be moved to the appropriate subfolder inside the SBDH-Annotated-Sets folder if one wishes to use them to train XGBoost models
* *“XX-XX-cohen\_calc\_set.pkl”* should be left in the main folder if one wishes to calculate Cohen’s kappa results from them. Cohen’s kappa calculations are performed in the Results section of *AnnotateGPT – SHARECOPY.ipynb*
* Results may vary slightly due to the nature of GPT annotation

Running XGBoost Training – SHARECOPY.ipynb

* Open the file using a Jupyter IDE
* Run every cell in order, carefully reading and completing the set up steps
* You can run the demo for Community XGBoost model training using AnnotateGPT training sets without annotation by simply running all cells in order (third party data required)
* Results may vary slightly due to the nature of GPT annotation and XGBoost training