Datasets:

ArguGPT: <https://www.kaggle.com/datasets/alejopaullier/argugpt>

DAIGT Proper Train Dataset: <https://www.kaggle.com/datasets/thedrcat/daigt-proper-train-dataset/>

Notebooks:

Train your own Tokenizer: <https://www.kaggle.com/code/datafan07/train-your-own-tokenizer>

KerasNLP: <https://www.kaggle.com/code/alexia/kerasnlp-starter-notebook-llm-detect-ai-generate>

12/8/23 Meeting

* Features: Split excerpts into different features
  + Metadata
    - Average sentence length
    - Run on sentences
    - Count of punctuation
    - Typos/grammar
    - Words frequency distribution
    - Sentence structure (passive vs active voice)
    - Num sentences per paragraph
  + Content based
    - String comparison between prompt/source and essay
    - Semantics
    - Diversity of vocabulary (NLP)
      * Repetitiveness within text
      * Same phrases as source/prompt
    - Sentiment
    - Predictability of language
  + Adding other models