Take Home Exercise - FSDS

1. Challenge: Exploring Gene Causality

2. Introduction to the challenge:

Phenotypes are traits or conditions that we can observe in an individual, which are influenced by genes. A causal gene is one that directly influences the development of a trait or disease. Variations in this gene can result in observable changes in the phenotype, making it a key target for genetic studies. Figuring out which genes are causal (for a particular phenotype) from many possible candidates is essential in fields like medicine, genetics, and biology. This knowledge helps in creating specific treatments and diagnostics by understanding the genetic foundations of different traits and conditions.

Now, what makes a gene causal?
Numerous researchers have pursued this question and published their findings in literature. Literature, which is accessible to us, has also been used to train Large Language Models - like OpenAI's and Meta's GPT and Llama respectively. It's plausible that these models, when asked about specific phenotypes or genes, generate embeddings that reflect the literature's consensus on their genetic associations.

3. Problem Statement to be solved:

You are given a dataset of phenotypes and their causal genes. You are provided with embeddings for the list of phenotypes and their corresponding causal genes. These embeddings were created by asking GPT3.5 to describe the gene or phenotype, and embedding them using OpenAI's text-embedding-3-large model. So, the embeddings for each phenotype and gene contain detailed information about their characteristics. **Your task is to conduct an exploratory analysis to determine *how these embeddings might indicate gene causality*.** In other words, does a pair of phenotype embedding and gene embedding give any signal for causality?

About the data sets to be used for this exercise

The dataset can be downloaded from here. Paths to the necessary files are relative to the `zenodo_directory` inside the zip file. Only use the files that are specified below for the analysis.

- **Phenotypes and genes:** A trait or condition, with associated genes. Out of these, only one gene is causal.

    - `zenodo_directory/data/benchmark_datasets/opentargets_step2.for_llm.tsv`
- **Ground Truth:** The gene considered causal for the phenotype. The ordering in this file is consistent with the above file.
    - `zenodo_directory/data/benchmark_datasets/opentargets_step2.labels`
- **Features:** 3072-dimensional embedding vector for all phenotypes and all genes. These are the features you need for the analysis.
    - `zenodo_directory/data/helper_datasets/gene_embeddings.csv`
    - `zenodo_directory/data/helper_datasets/phenotype_embeddings.csv`

Create your own data subsets-

After you downloaded the necessary files as mentioned above, we want you to create a unique dataset for yourself as described below.
[Imp : To ensure you work with a unique subset of the data, follow these steps for the `opentargets_step2.for_llm.tsv` file]:

- **Hash your name:** Convert your name to a hash value. Remove spaces and use the same case for all letters to maintain consistency. [E.g. if your name is Satish Kumar, use either satishkumar or SATISHKUMAR to create a hash value].
- **Use the hash as a seed:** This hash value will be used as a seed for random sampling. Sampling with a hash seed ensures that each participant consistently gets the same subset for their analyses.
- **Sample 500 phenotypes:** Using the seed, randomly sample 500 phenotypes from the dataset. This will be your unique dataset.
- **Document the hash in your submission:** Ensure you include the hash value in your submission to validate your unique dataset.

4. Task: Your task is to conduct an exploratory analysis to determine if these embeddings might indicate gene causality -

Here are some hints to help you with the analysis:

- **Mapping of Phenotypes to Genes:** Each phenotype is associated with multiple genes, one being causal and the rest, non-causal. You must devise a strategy to analyze these relationships.
  For example, you can create a dataset where one phenotype is mapped to its causal gene and also to several non-causal genes, labeling the pair as causal and non-causal respectively.
- **Dimensionality Reduction:** Apply suitable methods to reduce the dimensions of the embedding vectors to facilitate visualization of the relationships between phenotype and gene embeddings.

- **Vector Analysis:** Consider manipulating the embeddings to derive new vectors that may represent the relationships between genes and phenotypes. Analyze these derived vectors to explore potential insights into causality.
- **Clustering:** Experiment with different unsupervised methods to cluster the embeddings. Assess if these clusters can help identify distinct patterns or groups that suggest causal relationships between genes and phenotypes.

  Do you see any signal that separates causal gene-phenotype pairs from non causal ones from these experiments?

5. Evaluation Criteria:

- **Creativity and Innovation:** How you utilize the embeddings and the analytical methods to uncover insights.
- **Analytical Rigor:** The thoroughness and correctness of your exploratory analysis.
- **Presentation:** Clarity and effectiveness in communicating findings and insights by explaining why you did what you did.

6. Submission Requirements:

- **Jupyter Notebook:** A complete notebook (.ipynb file or Google colab link to your notebook) containing the code, detailed comments, and documentation.
- **Report:** A concise report summarizing your approach, methodology, results.
- **Video:** A video recording demonstrating your approach to solve the problem (you can walk us through the code and report here).

Make sure the links you share (Google Colab, YouTube, etc) are not private.

7. Additional Guidelines:

- Focus on the exploratory nature of the analysis; *there are no predefined correct answers*.
- You can use the following resources to aid your analysis:
  - Genotypes and Phenotypes
  - Principal Component Analysis (You are encouraged to explore more dimensionality reduction techniques)
  - What are embeddings?