# Project Report: Exploring Gene Causality Using Phenotype-Gene Embeddings

## Summary

This project investigated the potential of using GPT-3.5 generated embeddings to identify gene-phenotype causality. It utilized embedding-based methods, dimensionality reduction, clustering analysis, and causality prediction to explore relationships between genes and phenotypes.

## Methodology

1. **Embedding Generation:** GPT-3.5 generated embeddings for genes and phenotypes.
2. **Dimensionality Reduction:** PCA and UMAP were applied to reduce the dimensionality of embeddings.
3. **Clustering Analysis:** KMeans and Agglomerative clustering were used to identify clusters of genes and phenotypes. Silhouette scores evaluated clustering quality.
4. **Causality Prediction:** Cosine similarity and Euclidean distance were used to predict causal relationships between genes and phenotypes.
5. **Accuracy Evaluation:** Accuracy of clustering and causality predictions were calculated against ground truth data.

## Results

* **Clustering Analysis:** KMeans and Agglomerative clustering revealed potential relationships between genes and phenotypes within the reduced embedding space. Silhouette scores provided an assessment of cluster quality.
* **Causality Prediction:** Cosine similarity and Euclidean distance, alone and combined, were used to predict causality between genes and phenotypes.
* **Accuracy Results:** KMeans and Agglomerative cluster accuracies, Euclidean distance accuracy, and combined accuracy were calculated and analyzed. Higher accuracy indicated better alignment with ground truth causality.

## Conclusion

The project successfully demonstrated the feasibility of using embedding-based methods to explore gene-phenotype causality. The combined use of cosine similarity and Euclidean distance enhanced causality prediction accuracy.

## Future Steps

Future work will focus on:

* Expanding the dataset.
* Improving embedding generation.
* Exploring advanced clustering algorithms.
* Implementing network analysis and causality inference methods.
* Validating results through biological experiments.
* Enhancing interpretability and explainability.
* Incorporating dynamic modeling.
* Investigating clinical applications.

By addressing these future steps, the project can be further developed to enhance its accuracy, robustness, and applicability for advancing medical and genetic research.