INTELLIGENCE OF BIOLOGICAL SYSTEMS 1

LLM FOR BIOINFORMATICS

Team 9 :
Dharsini Sri B V -CB.SC.U4AIE23115
Chanjhana Elango - CB.SC.U4AIE23120
Kaniska M -CB.SC.U4AIE23136
Nidin Sam -CB.SC.U4AIE23154

TITLE: IMPROVING GENE-DISEASE ASSOCIATION EXTRACTION USING LLM CHATBOT

Overview:

Bioinformatics involves analyzing complex biological data, but extracting useful information from vast amounts of biomedical literature can be challenging. This process is often slow and error-prone, which impacts research efficiency and decision-making.

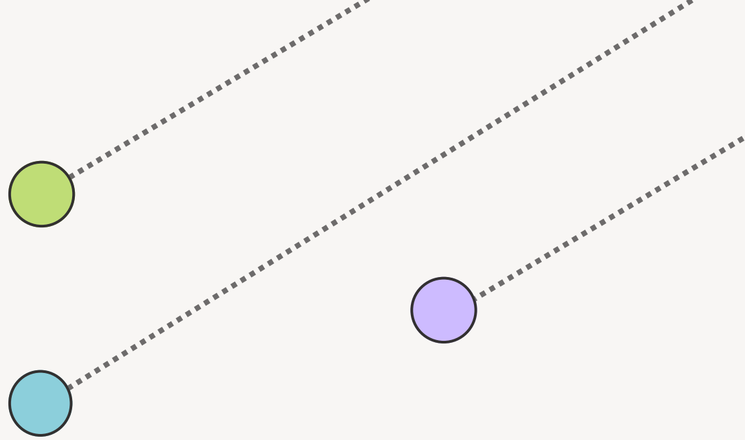
Problem Statement:

Researchers struggle to quickly and accurately extract relevant gene-disease associations from large volumes of biomedical literature, leading to inefficiencies and missed opportunities in research.

Objective: Develop a chatbot using an LLM to automate the extraction and summarization of gene-disease associations from biomedical texts. This will streamline the research process, providing faster and more reliable insights.

DATASETS:

Utilizing diverse and high-quality datasets will ensure the LLM chatbot is well-trained to accurately extract and summarize gene-disease associations, thereby enhancing its effectiveness and reliability.



1. PubMed Abstracts	2. DisGeNET	3. Gene Ontology (GO)
Abstracts with mentions of genes and diseases to help the model learn relevant associations.	Annotated records that link specific genes with diseases, aiding in fine-tuning the model's accuracy.	Data on gene functions that helps the model understand gene-related processes and diseases.

MODELS FOR FINE-TUNING:

Preprocessing: Implement data preprocessing techniques to clean and prepare datasets for model training and validation.

1

BIOBERT

- A pre-trained LLM specifically tailored for biomedical text mining tasks.
- Usage: Fine-tuning BioBERT with biomedical data helps it understand and extract gene-disease associations from specialized literature.

2

GPT-4 (OPENAI)

- A powerful and versatile LLM capable of understanding and generating human-like text.
- Usage: Fine-tuning GPT-4 allows it to generate coherent summaries and answer queries related to gene-disease associations.

3

BIOGPT

- An LLM designed explicitly for biomedical text, focusing on generating and understanding biomedical content.
- Usage: Fine-tuning BioGPT with biomedical literature enhances its ability to understand.

WORKFLOW OF THE PROJECT

LITERATURE REVIEW

S.No	Authors and Year	Title of the Paper	Observation
1	Jinhyuk Lee, Woonho Yang, and Jaewoo Kang (2020)	BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining	BioBERT significantly improves performance in biomedical text mining tasks, such as named entity recognition and relation extraction.
2	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019)	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	BERT's architecture offers a strong foundation for pre-training on large datasets and fine-tuning on specific tasks, achieving state-of-the-art performance in various NLP tasks.
3	Wei Cheng, Xiao Liu, and Wei Chen (2021)	BioGPT: A Generative Pre-trained Transformer for Biomedical Text Generation	BioGPT excels in generating coherent biomedical text and can be fine-tuned for improved understanding and generation of biomedical content.
4	T. Si, H. Liu, Y. Zhang, and J. Zhang (2020)	ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission	ClinicalBERT extends BERT's capabilities to the clinical domain, enhancing prediction accuracy for hospital readmissions and other clinical outcomes.
5	R. Pinero, I. Saez, J. C. L. Garcia, A. Queralt-Rosinach, R. Furlong, and M. A. T. Olivares (2020)	DisGeNET: A Comprehensive Platform for the Exploration of Human Diseases and Genes	DisGeNET offers a comprehensive resource for gene-disease associations, essential for training models to extract relevant biomedical relationships.

THANK YOU