GenomicGPT: AI-Powered Framework for Gene-Disease Relationship

Harvard Rare Disease Hackathon 2025

Team Pumpkin Seeds

Praveena Jothsna Pendyala, Keerthana Goka, Kunal Malhan

1. Introduction

The discovery of gene-disease relationships plays a crucial role in rare disease diagnosis and research. However, a major challenge faced by clinicians and researchers is the fragmented nature of relevant literature and genetic datasets. Research papers, reviews, and datasets are dispersed across multiple sources such as PubMed, ClinGen, and GeneReview, making it difficult to access comprehensive information quickly. Additionally, there is no standardized confidence score to assess the strength of these relationships, leaving medical professionals to manually interpret large volumes of data. To address these issues, we propose GenomicGPT, a Generative AI-powered system designed to:

- Extract relevant literature from multiple sources using Agentic AI, LangChain, and Retrieval-Augmented Generation (RAG) systems.
- Generate structured summaries of research papers to improve accessibility.
- Assign a confidence score to gene-disease associations based on extracted evidence.
- Enhance scoring further by applying Machine Learning (ML) models to available genetic datasets.

By integrating Natural Language Processing (NLP), AI-driven literature mining, and ML-based evidence validation, our solution streamlines research workflows and provides actionable insights for clinicians, researchers, and rare disease patients.

2. Problem Statement

Development of a quantitative framework to determine the mechanism of disease for gene-disease relationships

3. Challenges

1. Fragmented Literature & Genetic Data

- Research articles on gene-disease pairs are scattered across multiple repositories.
- Genetic datasets are stored in separate databases, making retrieval difficult.
- Clinicians struggle to access relevant information quickly for diagnosis.

2. Lack of a Standardized Confidence Score

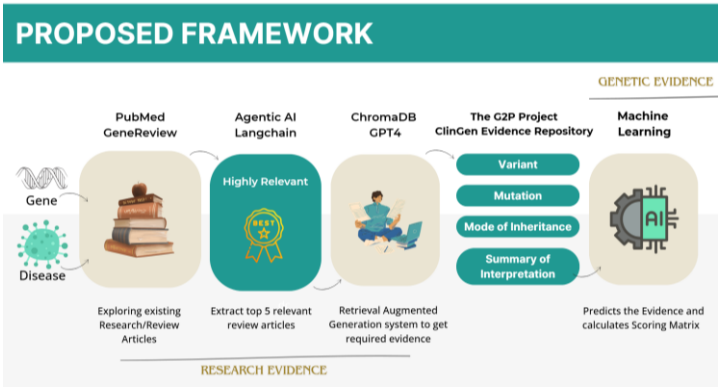
- Current research does not provide a quantitative measure of confidence for gene-disease associations.
- Clinicians must manually interpret literature, leading to time-consuming and subjective decision-making.

3. Inaccessibility of Research for Non-Experts

- Families of rare disease patients lack a centralized platform to access authentic and technical literature.

4. Proposed Solution: GenomicGPT

GenomicGPT is a three-layered AI-driven framework that integrates literature mining, NLP, and ML-based scoring to streamline gene-disease discovery.



4.1 AI-Driven Literature Extraction & Summarization

- Agentic AI & LangChain are used to retrieve top-ranked review papers from sources like PubMed and ClinGen.
- RAG (Retrieval-Augmented Generation) is applied using ChromaDB and GPT-4 to summarize key insights.

- Clinicians and researchers can access all relevant literature in a single interface, reducing the time spent searching for data.

4.2 Confidence Scoring from Literature Evidence

- Extracted literature is analyzed using AI models to assign confidence scores based on strength and frequency of gene-disease associations.
- The scoring mechanism provides quantitative evidence, helping researchers prioritize findings.

4.3 ML-Based Refinement Using Genetic Datasets

- Beyond literature scoring, ML models are applied to genetic datasets to further refine confidence levels.
- This approach combines textual evidence with real-world genetic data, making the results more robust and scientifically validated.

4. Target Audience & Use Cases

GenomicGPT is designed to serve multiple stakeholders in the rare disease research and diagnostic ecosystem.

1. Geneticists & Rare Disease Centers

- Use Case: Clinicians diagnosing rare diseases can retrieve highly relevant literature and confidence scores in one place, expediting the diagnostic process.

2. Researchers Investigating Disease Mechanisms

- Use Case: Scientists studying disease mechanisms can leverage AI-generated confidence scores as a starting point for further validation and research.

3. Families of Rare Disease Patients

- Use Case: Non-experts can access a curated, authentic repository of research papers to better understand rare conditions affecting their families.

5. User Interface & Platform Design

We developed an intuitive web-based platform that allows users to:

- Search for a gene-disease pair and retrieve relevant literature.
- View confidence scores for associations based on AI-generated insights.
- Explore genetic datasets to validate findings further.

6. Conclusion & Future Scope

GenomicGPT presents a novel AI-driven approach to solving critical gaps in gene-disease research by integrating literature mining, NLP-driven summarization, and ML-based confidence scoring.

Key Benefits:

✓ Accelerates rare disease diagnosis by providing instant access to curated research.

✓ Reduces clinician workload by summarizing vast biomedical literature.

✓ Improves research workflows by offering data-driven confidence scores.

Future Enhancements:

- Fine-tuning models with domain-specific datasets (e.g., ClinVar, Orphanet).
- Integrating real-time research updates using AI-powered search agents for the general public to access.