Medgenee: Leveraging ChatGPT for Gene Discovery in Research Papers

A Novel Approach to Biomedical Informatics

Amy Hardy
| ah64343 |
| AI 395T | AI in Healthcare | Spring 2024 |

Abstract

This paper presents a novel approach to extracting information about genes from scientific research papers using ChatGPT, a large language model (LLM). Medgenee proposes a methodology to identify genes mentioned in abstracts related to specific health topics, such as Alzheimer's disease, and analyze their relevance to the disease and associated comorbidities. By integrating natural language processing (NLP) techniques with domain-specific knowledge about gene research, our method aims to provide insights into both well-known and novel genes associated with the disease.

Introduction

Gene Discovery

The exploration of genetic foundations within biological systems is vital to biomedical research. The identification and study of genes, particularly those associated with diseases, are paramount for advancing the understanding of complex biological mechanisms and for the development of targeted therapies. Gene discovery has a profound impact on medical diagnostics, treatment strategies, and the broader field of personalized medicine. As we go forward in the genomic era, the ability to efficiently identify and characterize genes related to specific health conditions becomes increasingly important. This has propelled the adoption of advanced computational tools to handle the vast data derived from genomic studies, enhancing both the speed and accuracy of genetic discoveries.

LLMs

Among these computational tools, LLMs such as ChatGPT have emerged as powerful resources. These models, trained on large corpuses of text, have the ability to understand and generate human-like text, making them important in processing and interpreting vast amounts of unstructured data. ChatGPT has demonstrated remarkable capabilities in natural language understanding, making it a promising tool for biomedical informatics. Its potential to assist in gene discovery lies in its ability to parse and extract relevant information from scientific literature, a task that is both time-consuming and complex for human researchers.

By integrating advanced NLP capabilities with domain-specific inquiries, researchers can quickly identify mentions of genes and their contextual relevance to various diseases. This approach not only accelerates the pace of gene discovery but also opens new avenues for exploring genetic links that are less understood or previously unrecognized in the scientific community.

Background

The integration of AI, particularly NLP, into gene discovery is an area of growing interest within the scientific community. Several pioneering studies

have demonstrated the efficacy of using AI to extract and analyze or summarize data from biomedical texts [1]. Another study found that ChatGPT was capable of assisting doctors in diagnosis [2]. These developments display the transformative potential of AI in enhancing medical research, establishing a strong foundation for further exploration into the use of LLMs in this field.

Prompt Engineering for LLMs

Prompt engineering involves the strategic formulation of input queries to elicit the most informative and accurate responses from a language model. This technique is crucial when utilizing models like ChatGPT for specialized tasks. Effective prompt engineering can improve the model's performance by guiding it to focus on specific elements within the text, thus ensuring that the outputs are relevant and useful for the task at hand. For instance, prompts can be designed to ask the model to identify and summarize mentions of genetic markers associated with specific diseases, or to extract discussions of novel gene-disease correlations from newly published research articles. The art and science of prompt engineering plays a pivotal role in leveraging the full capabilities of LLMs in a targeted and efficient manner.

Recent Gene Discoveries

The field of gene research continuously unveils new insights into the molecular mechanisms underlying various diseases. Recent advancements have highlighted not only the identification of disease-associated genes but also their roles in pathogenesis and potential as therapeutic targets. For example, the discovery of genes related to Alzheimer’s disease has opened new avenues for treatment and prevention strategies, emphasizing the importance of these findings in improving patient outcomes [3]. The continuous influx of genomic data from global research initiatives underscores the need for advanced tools like LLMs to manage and interpret these vast datasets. Understanding these recent discoveries and their implications helps to appreciate the importance of developing new methods for gene discovery, which can accelerate and refine our understanding of complex diseases.

Methodology

Gathering Scientific Research Articles for Data

The initial step in our methodology involves collecting a substantial dataset of scientific abstracts, which serve as the primary source for gene discovery. This process is executed by accessing the PubMed database, a comprehensive repository of biomedical literature. Using a specifically tailored search query, we extract abstracts that mention or relate to specific health conditions of interest. The query is designed to maximize the relevance of the retrieved documents, focusing on recent publications and those with a high citation index to ensure the inclusion of impactful research. The automated scraping of these abstracts is facilitated by using Beautiful Soup, a Python package, which allows us to systematically access PubMed's resources, providing a streamlined and efficient method for gathering data.

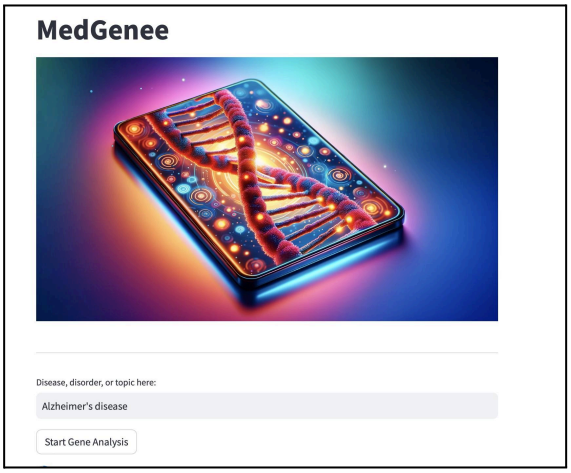
Integrating Streamlit for User Interaction

Following the collection of data, our methodology incorporates the use of the Streamlit library to create an interactive dashboard. This tool enables researchers and healthcare professionals to engage directly with the findings of our project, allowing them to input specific topics, diseases, or terms of interest and receive customized results.

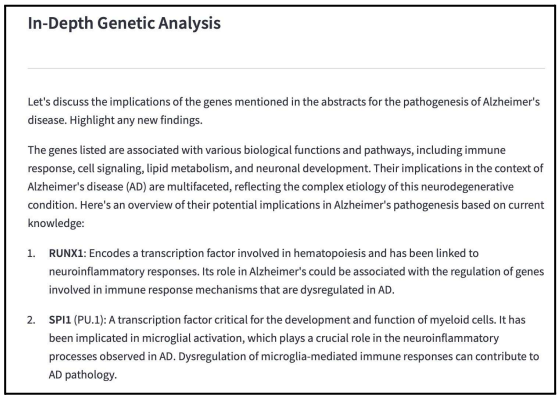
Streamlit is a Python library designed to facilitate the quick creation of web apps for machine learning and data science projects. It is highly

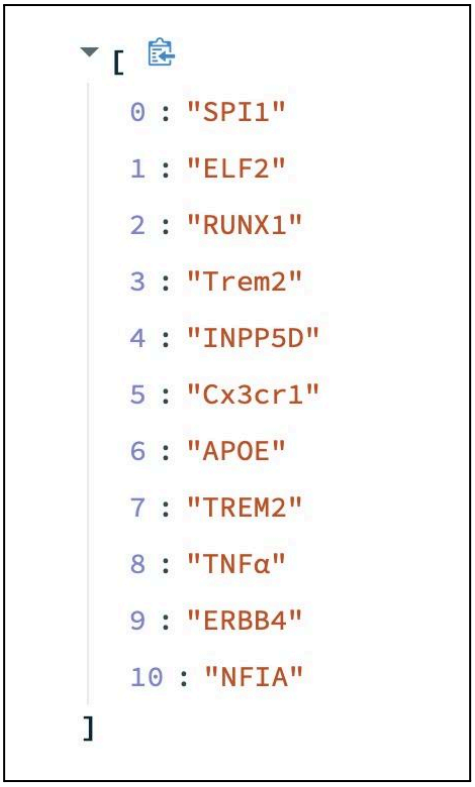
favored for its ease of use and ability to transform complex data scripts into shareable web apps. For our project, Streamlit will serve as the interface through which users can explore the gene discovery outputs generated by ChatGPT. The implementation of the dashboard begins with setting up a simple interface where users can enter a search query related to a particular biomedical topic or disease. This input is then processed by the backend, which retrieves the relevant abstracts from PubMed.

Features of the Streamlit Dashboard



- Search and Query Input: Users can type in a disease or topic to generate a focused query for gene-related information.





- Results Display: After processing the user's query through ChatGPT with the appropriate prompts, the dashboard displays a list of identified genes. This section is dynamically updated based on the query results. It will also display the analysis performed by ChatGPT on the genes and the disease.

By leveraging Streamlit, we not only facilitate a more interactive and user-friendly way for individuals to engage with the project's outputs but also ensure that our tool can be easily accessed and utilized across different devices and platforms. This approach significantly enhances the practical utility of our gene discovery methodology, making it a valuable resource for ongoing research and clinical applications.

Prompting ChatGPT for Gene Discovery

The process of developing effective prompts involves understanding both the capabilities of the language model and the specific requirements of the gene discovery task. The prompts must be

clear, direct, and structured in a way that maximizes the extraction of relevant information. Below are examples of the types of prompts we employed:

Direct Query Prompt:

```
Identify all gene names mentioned in
this abstract.
```

This prompt directs ChatGPT to scan the text for gene names.

Contextual Inquiry Prompt:

```
Discuss the implications of the genes
mentioned in this abstract for the
pathogenesis of {input_topic}.
Highlight any new findings.
```

This prompt encourages the model to analyze the abstract not only for gene names but also for any new research findings or hypotheses regarding these genes, thereby providing insights into cutting-edge research.

Comparative Analysis Prompt:

```
Compare the genes listed in this
abstract with those known to be
associated with {input_topic} and note
any novel genes.
```

This prompt tasks ChatGPT with identifying overlaps and discrepancies between the genes mentioned in the abstract and those already known to be associated with the disease, spotlighting potentially novel genes.

Results

The application was tested focusing on Alzheimer's disease as a case study. Abstracts relevant to Alzheimer's were collected, with a particular emphasis on genes implicated in its pathogenesis. Following analysis, the application provided a comprehensive summary of the genes discussed in these abstracts.

The results of the testing were promising. The application successfully identified a range of genes associated with AD, including both well-established factors like APOE and emerging candidates like TREM2 and TNFα. This indicates the application's capability to uncover both established and novel gene-disease correlations.

Overall, the results suggest that the application holds significant potential as a valuable tool for gene discovery in biomedical research. By efficiently extracting relevant genetic information from scientific literature, it can aid researchers in uncovering new insights into complex diseases, paving the way for enhanced understanding and targeted therapeutic interventions.

Discussion

The Argument for LLMs in Gene Discovery

Leveraging LLMs for gene discovery presents a compelling argument for revolutionizing biomedical informatics. By harnessing the NLP capabilities of these models, researchers can streamline the process of identifying genes associated with specific health conditions, thereby accelerating the pace of genetic research. The ability of LLMs to sift through vast amounts of scientific literature and extract relevant information holds immense potential for uncovering both well-established and novel gene-disease correlations.

Limitations

While LLMs offer promising capabilities for gene discovery, it is essential to acknowledge their limitations. One primary concern is the potential for bias in the training data, which may influence the model's outputs and interpretation of scientific literature. Additionally, LLMs may struggle with

ambiguity or nuanced language in research abstracts, leading to errors or misinterpretations in gene extraction. Addressing these challenges requires ongoing efforts to refine model training methodologies and enhance the robustness of natural language algorithms.

Also, the reliance on abstracts as the primary source of data may limit the depth of information available for gene discovery. Abstracts often provide concise summaries of research findings, omitting crucial details that may impact the interpretation of gene-disease associations. Future endeavors should explore ways to integrate full-text articles and supplementary data sources to enrich the gene discovery process and mitigate potential biases introduced by abstracts alone.

Future Research

Future research should focus on leveraging LLMs not only for gene extraction but also for advanced analysis and hypothesis generation. Expanding the capabilities of LLMs to interpret complex biological data and generate novel insights will enhance their utility in driving forward genetic research. By exploring innovative approaches that harness the computational power of LLMs for hypothesis generation, researchers can uncover hidden patterns and connections within genomic data, leading to breakthrough discoveries in disease mechanisms and therapeutic targets. Embracing interdisciplinary collaboration and developing transparent frameworks for model interpretation are essential for realizing the full potential of LLMs in advancing biomedical research.

In future versions of MedGenee specifically, expanding the number of prompts to encompass a wider range of information and incorporating additional data sources beyond abstracts, such as full-text articles and supplementary datasets, could further enhance its efficacy in gene discovery. Additionally, integrating feedback mechanisms to continuously refine the model's performance and collaborating with domain experts to validate findings would augment its utility and reliability in advancing genetic research.

References

[1] Hake, J., Crowley, M., Coy, A., Shanks, D., Eoff, A., Kirmer-Voss, K., Dhanda, G., & Parente, D. J. (2024). Quality, Accuracy, and Bias in ChatGPT-Based Summarization of Medical Abstracts. The Annals of Family Medicine, 22(2), 113-120. https://doi.org/10.1370/afm.3075

[2] Mehnen, L., Gruarin, S., Vasileva, M., & Knapp, B. (2023). ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. medRxiv. https://doi.org/10.1101/2023.04.20.23288859

[3] Ataei, B., Hokmabadi, M., Asadi, S., Asadifard, E., Aghaei Zarch, S. M., Najafi, S., & Bagheri-Mohammadi, S. (2024). A review of the advances, insights, and prospects of gene therapy for Alzheimer's disease: A novel target for therapeutic medicine. Gene, 912, 148368. https://doi.org/10.1016/j.gene.2024.148368

Accessing MedGenee

To access and use the MedGenee demo dashboard: visit https://medgenee.streamlit.app. Please be aware that this dashboard is connected to a personal OpenAI payment account, so while testing is welcome and appreciated, please keep your queries to a minimum. The dashboard will no longer be publicly accessible two weeks after the end of the course.