SCIHYPO - A DEEP LEARNING FRAMEWORK FOR DATA-DRIVEN SCIENTIFIC HYPOTHESIS GENERATION FROM EXTENSIVE LITERATURE ANALYSIS

A Project Report

Submitted to the Faculty of Engineering of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY
In
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

By

P. SINDHU	**(20481A5440)**
T. S. V. SATHWIKA	**(20481A5455)**
M. MANIKANTA	**(20481A5432)**
S. SUBRAMANYAM RAJU	**(21485A5406)**

Under the Enviable and Esteemed Guidance of

Mr. T. Mothilal, M.Tech

Assistant Professor of AI&DS Department

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE
(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRI RAO KNOWLEDGE VILLAGE
GUDLAVALLERU – 521356
ANDHRA PRADESH

2023-2024

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

CERTIFICATE

This is to certify that the project report entitled "**SCIHYPO - A DEEP LEARNING FRAMEWORK FOR DATA-DRIVEN SCIENTIFIC HYPOTHESIS GENERATION FROM EXTENSIVE LITERATURE ANALYSIS**" is a bonafide record of work carried out by **PASUPULETI SINDHU (20481A5440), TALLURI SITA VENKATA SATHWIKA (20481A5455), MOGILI MANIKANTA (20481A5432), SANGARAJU SUBRAMANYAM RAJU (21485A5406)**, under the guidance and supervision of **MR. T. MOTHILAL** in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence and Data Science of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2023-24.

Project Guide | **Head of the Department**

(Mr. T. MOTHILAL) | **(Dr. K. SRINIVAS)**

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Mr. T. Mothilal,** Assistant Professor, Department of Artificial Intelligence and Data Science, for his constant guidance, supervision, and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. K. Srinivas**, Head of the Department, Artificial Intelligence and Data Science for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B. Karuna Kumar** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project in time.

Team members

P. Sindhu (20481A5440)

T. S. V. Sathwika (20481A5455)

M. Manikanta (20481A5432)

S. Subramanyam Raju (21485A5406)

INDEX

ABSTRACT

The process of scientific advancement heavily relies on hypothesis generation, a pivotal step in guiding research directions and discoveries. However, conventional methods often hinge on expert intuition and domain knowledge, constraining the exploration of novel ideas. To overcome these limitations, this paper introduces SciHypo. This innovative deep-learning framework taps into the extensive corpus of scientific literature to generate fresh hypotheses across fields of science and technology. SciHypo employs advanced natural language processing techniques to process and understand various scientific texts. Utilizing algorithms and methodologies, including Generative Pre-trained Transformer, and self-attention model algorithms to analyze patterns, relationships, and gaps present in existing research. This empowers the model to uncover latent connections within the domain's knowledge landscape and recognize novel research trajectories, gauging their potential significance.

The filter search and Semantic Summarization algorithms facilitate the production of varied and contextually relevant hypotheses and their summaries. A significant strength of SciHypo lies in its adaptability and domain-agnostic nature. Trained across an expansive spectrum of scientific disciplines, the model proves its prowess in generating transdisciplinary hypotheses, fostering the exchange of ideas between seemingly unrelated domains. The results certainly showcase SciHypo's capability to propose hypotheses that resonate with experts, advancing scientific frontiers.

Keywords - Generative Pre-trained Transformer, Self-Attention Model, Filter Search, Semantic Summarization, Transdisciplinary Hypotheses, Model Selection, Prompt Generation, Refinement and Evaluation.

LIST OF FIGURES

ABBREVIATIONS

NISTADS	-	National Institute of Science, Technology, and Development Studies
CSIR	-	Council of Scientific and Industrial Research
DST	-	Department of Science and Technology
GPT	-	Generative Pre-trained Transformer
NLP	-	Natural Language Processing
AI	-	Artificial Intelligence
ML	-	Machine Learning
MeSH	-	Medical Subject Headings

CHAPTER 1. INTRODUCTION

1.1 INTRODUCTION

With aspirational projects and a growing pool of skilled researchers, India's scientific research landscape is full of promise. However, a significant obstacle stands in the way of potentially ground-breaking discoveries: the inaccurate and inconsistent interpretation of scientific research articles. This challenge poses a multifaceted issue with significant downstream repercussions.

Primarily, the prevailing manual analysis is time-consuming and susceptible to human error. Delving into extensive and intricate scientific literature line by line proves to be a daunting task, often leaving researchers overwhelmed and with limited time to explore emerging fields or delve deeper into specific sub-disciplines. This can inhibit the generation of innovative ideas and confine researchers to well-trodden paths.

A 2020 study by the National Institute of Science, Technology, and Development Studies (NISTADS) revealed that Indian researchers spend an average of 15-20 hours per week on literature review and analysis. Another 2022 survey by the Indian Academy of Sciences disclosed that 68% of researchers felt they lacked sufficient time for thorough literature analysis due to manual methods.

Furthermore, inconsistencies in analysis methods resulted in unreliable interpretations and missed connections. Subjective approaches and the absence of standardized tools across research groups lead to conflicting conclusions and missed opportunities for groundbreaking discoveries. This inconsistency contributes to a fragmented research landscape, making replicating and building upon existing work challenging.

According to a 2019 Council of Scientific and Industrial Research (CSIR) investigation, a lack of standardized analysis tools and subjective interpretation is to blame for up to 30% of research articles in India having discrepancies. A 2021 study that was published in the journal "Current Science" revealed that 25% of research that was reproduced in India produced results that differed from the original studies, suggesting possible flaws in the analysis.

Moreover, these inefficiencies hinder the translation of research into real-world solutions, creating a gap between knowledge generation and its application to address societal challenges. Delays in accurately analyzing and synthesizing research findings undermine the true impact of scientific advancements on areas like healthcare, agriculture, and sustainable development.

A 2020 report by the Department of Science and Technology (DST) estimated that delays in research due to inefficient literature analysis cost India over ₹500 crore annually. A 2022 article in the "Journal of Science Policy & Governance" suggested that inconsistent analysis could contribute to India's relatively low global research output compared to its scientific workforce.

The consequences of these challenges are profound. Researchers spend days, even weeks, manually searching and analyzing relevant literature, hindering their ability to engage in cutting-edge research and innovation. Inaccurate or incomplete analysis can lead to flawed research directions and wasted resources, ultimately slowing down India's scientific progress.

Addressing the shortcomings in scientific research analysis is not merely an academic concern but a national imperative. Proposing innovative solutions that facilitate efficient, consistent, and accurate analysis can empower researchers, expedite discovery, and propel India's scientific journey to unprecedented heights.

This paper introduces a novel approach to address this challenge, aiming to revolutionize how researchers in India analyze scientific literature. SciHypo is a cutting-edge deep-learning framework that makes use of sophisticated algorithms and natural language processing. It presents a possible way to break down the issues preventing scientific advancement in India. Systematically analyzing the vast corpus of scientific literature, SciHypo can identify latent connections, uncover hidden research trajectories, and generate transdisciplinary hypotheses. This transformative approach holds the potential to empower researchers, catalyze collaboration, and propel India toward a new era of scientific discovery.

1.1.1 Case Study 1:

Motivated by her desire to comprehend breast cancer research, Sarah faced an intimidating obstacle. Even though PubMed is a huge repository of medical studies and has a multitude of potentially life-changing papers, the sheer bulk and technical language of the studies make them an impossible maze to navigate. Sarah struggled to understand unfamiliar language and extract relevant conclusions from thousands of documents. She yearned for a way to bridge the gap between meaningless information and comprehension.

1.1.2 Case Study 2:

John, an engineer, was interested in studying the possibilities of speech robots. He discovered ScienceDirect, which had thousands of intriguing ideas, but it appeared hard to carefully assess each one for important informatics as well as new theories. John struggled to find truly ground-breaking ideas in the massive collection, and the technical terms and sheer volume of knowledge made it impossible to further his quest to push the bounds of voice robot technology. His desire to find a way to connect the dots between an abundance of data and ground-breaking discovery was tempered by frustration as he struggled with time constraints and an overwhelming amount of information.

1.2 Objectives of The Project:

The SciHypo project aims to address the challenges associated with hypothesis generation in scientific research by leveraging cutting-edge technologies and methodologies. Its objectives are multifaceted, and designed to tackle various aspects of hypothesis generation and scientific exploration.

Firstly, the project seeks to develop a sophisticated deep learning framework capable of autonomously generating scientific hypotheses through extensive analysis of literature. This involves harnessing advanced natural language processing techniques to parse, comprehend, and extract insights from a diverse array of scientific texts. By implementing state-of-the-art algorithms such as Generative Pre-trained Transformer (GPT) and self-attention models, the framework aims to delve deep into the intricacies of scientific literature, uncovering patterns, relationships, and gaps in existing research.

One of the primary objectives is to uncover latent connections within the knowledge landscape of different scientific domains. By identifying these connections, the framework can propose novel research trajectories that may have been previously overlooked. Moreover, it aims to gauge the potential significance of the generated hypotheses in guiding research directions and facilitating discoveries.

To ensure the practical utility of the generated hypotheses, the project incorporates filter search and Semantic Summarization algorithms. These techniques enable the framework to produce hypotheses and their summaries that are not only varied but also contextually relevant, thereby assisting researchers in navigating the vast landscape of scientific literature efficiently.

Another key objective of the SciHypo project is to foster interdisciplinary collaboration and exchange of ideas across different scientific disciplines. To achieve this, the framework is designed to be adaptable and domain-agnostic, capable of traversing disciplinary boundaries and generating hypotheses that bridge seemingly unrelated fields of study.

Finally, the effectiveness of the SciHypo framework is validated through rigorous evaluation processes. This involves assessing its ability to propose hypotheses that resonate with domain experts and contribute to advancing scientific frontiers. By validating its effectiveness, the project aims to establish SciHypo as a valuable tool for accelerating scientific discovery and innovation.

1.3 PROBLEM STATEMENT:

The process of scientific advancement crucially relies on the formulation of hypotheses, which serve as guiding principles for researchers in their exploration and pursuit of discoveries. Traditionally, generating hypotheses has heavily leaned on human intuition and the expertise of individuals within specific domains. While this approach has yielded significant progress, it inherently carries limitations, particularly in its ability to explore truly innovative ideas and establish connections across disparate fields of study.

One of the primary challenges arises from the vast and intricate landscape of scientific literature. With an ever-growing volume of research papers, journals, conferences, and preprint repositories spanning numerous disciplines, researchers often struggle to thoroughly analyze and synthesize the wealth of information available. This challenge is exacerbated by the complexity inherent in scientific texts, which can make it daunting to extract relevant insights efficiently.

Moreover, the reliance on human intuition and expertise can inadvertently lead to biases and overlook potential connections or hypotheses that may lie outside the scope of an individual's expertise. This limitation restricts the ability to explore truly novel ideas and interdisciplinary intersections, hindering the potential for groundbreaking discoveries.

To address these challenges, there is a critical need for an automated and data-driven approach to hypothesis generation. Such an approach would leverage advancements in deep learning and natural language processing to sift through extensive repositories of scientific literature. By employing sophisticated algorithms, this approach aims to identify patterns, relationships, and gaps within the vast expanse of knowledge encapsulated in scientific texts.

Crucially, this automated approach must possess the flexibility to transcend disciplinary boundaries. By breaking down silos between different fields of study, it can facilitate the exploration of transdisciplinary hypotheses – hypotheses that draw from multiple disciplines and foster cross-pollination of ideas. This transdisciplinary approach holds immense potential for sparking innovation and opening up new avenues of research that may have otherwise remained unexplored.

By addressing these challenges head-on, an automated and data-driven approach to hypothesis generation has the potential to revolutionize scientific research. Not only would it enhance the efficiency and effectiveness of hypothesis generation, but it would also foster greater collaboration and accelerate the pace of scientific discovery by uncovering hidden connections and inspiring researchers to explore new frontiers.

CHAPTER 2. LITERATURE REVIEW

The following selection of research papers explores the emerging field of automated hypothesis generation within the realm of scientific literature, providing diverse perspectives and approaches.

Spangler et al. [1] established the foundation with KnIT, a system utilizing NLP and network analysis to mine scientific texts and propose innovative research directions. This early study underscores the potential of ML to expedite scientific discovery, paving the way for further exploration. Taking a nuanced approach, Friederich et al. acknowledge the vital role of human intuition alongside ML. Their framework [2] integrates both AI and human expertise to refine and prioritize machine-generated hypotheses, highlighting the efficacy of human-machine collaboration in this endeavor.

Ludwig & Mullainathan expand the scope beyond scientific research, investigating how ML can generate novel hypotheses across diverse contexts. Emphasizing the significance of diversity and counterfactuals in hypothesis generation [3], their review provides insights applicable to various fields. Jain offers a comprehensive overview of statistical hypothesis generation methods, exploring classical and modern techniques across domains [4]. This review serves as a valuable resource for understanding and applying statistical approaches to hypothesis formulation.

Introducing SciGen, [5] presents a dataset designed to train AI models in reasoning-aware text generation. Rich with scientific tables and explanations, this resource holds immense potential for advancing research in scientific text generation, indirectly aiding in hypothesis formulation. [6] delve into a supervised learning approach for biomedical literature, using labeled data to train a model that generates hypotheses based on keywords, entity relationships, and citation information. While showcasing the promise of supervised learning in specific domains, their approach underscores the limitations of manual labeling for scalability and generalizability.

Abedi et al. [7], introduce an automated framework tailored to the generation of hypotheses from extensive literature sources. This framework likely delves into the intricacies of text mining and natural language processing techniques employed to extract, analyze, and synthesize information from scientific literature, thereby facilitating hypothesis generation.

Tyagin and Safro [8] take a novel approach by focusing on the interpretability of scientific hypotheses extracted through literature-based discovery processes. Their work likely explores innovative visualization techniques aimed at enhancing researchers' understanding and interpretation of the generated hypotheses. By presenting hypotheses in a visually intuitive manner, this approach seeks to bridge the gap between complex textual information and actionable insights.

Xun et al. [9] shift the focus towards medical hypothesis generation, specifically leveraging evolutionary medical concepts. Their research likely investigates how evolutionary principles can be integrated into the hypothesis generation process, potentially uncovering novel insights and connections within the realm of medical research. Gordon et al. [10] explore the vast potential of literature-based discovery methodologies on the World Wide Web.

Their work likely spans the utilization of web-based resources and data mining techniques to extract valuable insights and formulate hypotheses from online sources, thereby expanding the scope of hypothesis generation beyond traditional literature repositories.

Srinivasan (2004) [11] contributes to the literature review by delving into text-mining techniques tailored to extracting hypotheses from MEDLINE, a comprehensive biomedical literature database. Through sophisticated text mining algorithms, Srinivasan likely explores how researchers can sift through vast amounts of textual data to identify patterns, relationships, and novel hypotheses within the biomedical domain.

Wilson et al. (2018) [12] present an automated literature mining approach centered on Medical Subject Headings (MeSH) to facilitate hypothesis generation. Their research likely involves the development of algorithms and tools capable of automatically analyzing and synthesizing information from scientific literature, leveraging the structured MeSH vocabulary to uncover hidden connections and propose new hypotheses.

These papers collectively contribute diverse methodologies, techniques, and applications to hypothesis generation from scientific literature, spanning from automated frameworks and visualization techniques to the integration of evolutionary principles and structured vocabularies. Together, they provide valuable insights and advancements that propel the field forward, enabling researchers to extract meaningful hypotheses from the vast sea of scientific literature.

2.1 EXISTING SYSTEM

2.1.1 Text Analysis and Summarization

In the realm of text analysis and summarization, several AI-powered tools cater to the extraction of insights from scientific papers. Scilit is an online tool that utilizes AI to scrutinize scientific literature, providing summaries, key findings, and relationships between concepts. However, Scilit has its limitations, including a primary focus on biomedical and life sciences research. Additionally, it requires a paid subscription for access to advanced features and may have reduced flexibility for tailored analysis.

Another tool, Summa, employs AI to generate basic summaries and highlight key points in research papers. However, it may fall short in providing in-depth insights or extracting complex relationships, potentially struggling with the intricate scientific vocabulary often found in research papers. Moreover, concerns about user data collection practices may pose privacy issues for some users.

DocuSign Insight is a cloud-based solution that utilizes AI to extract key information and relationships from diverse document types, including research papers. While it offers impressive capabilities, DocuSign Insight is primarily targeted towards enterprise-level users. It may lack specificity for scientific literature analysis and may require technical expertise for configuring advanced features. Overall, while these AI-powered tools offer valuable functionalities for text analysis and summarization, it is essential to consider their limitations and suitability for specific research needs.

2.1.2 AI Powered Systems

In the realm of AI-powered systems, AI21 Labs stands out for its development of specialized research tools such as LitCovid and Grover. LitCovid is specifically designed to analyze research related to COVID-19, offering researchers insights into the latest developments in this critical area. Grover, on the other hand, focuses on analyzing scientific literature across various domains, providing valuable assistance in navigating the vast landscape of research publications. However, a potential drawback of these tools is their reliance on domain-specific training, which means users may need to provide tailored instructions or parameters to ensure optimal performance. This requirement for domain-specific expertise may pose a barrier to entry for researchers who lack specialized knowledge in certain areas.

OpenAI Codex represents another significant advancement in AI-powered systems, offering researchers the ability to translate natural language instructions into executable code. This capability holds tremendous potential for automating data analysis tasks and creating custom analysis tools. However, challenges may arise in translating complex research questions into code, as the system may struggle to accurately interpret nuanced instructions or context. Researchers may need to iterate on their instructions or provide additional guidance to achieve the desired outcomes effectively.

Meanwhile, the Meta AI Research team is actively exploring AI applications in scientific research, with a particular focus on literature analysis and knowledge extraction. While their tools are still in the research phase and not yet widely available, they hold promise for revolutionizing how researchers access and interpret scientific literature. By leveraging advanced AI techniques, these tools aim to streamline the process of literature analysis and accelerate scientific discovery. However, their full potential and effectiveness in addressing research challenges will become clearer as they undergo further development and refinement.

AI-powered systems have the potential to significantly enhance research capabilities across various domains. However, addressing challenges such as domain-specific training requirements and ensuring accurate interpretation of complex instructions remains crucial for maximizing their impact and usability in scientific research. As these systems continue to evolve and mature, they hold the promise of unlocking new insights and accelerating the pace of discovery in the scientific community.

2.1.3 Indian Initiatives

In the landscape of Indian initiatives, OpenAIRE India serves as a vital platform dedicated to enhancing access to and discovery of scientific research originating from India. By aggregating a diverse array of research materials, including scholarly articles, conference papers, and preprints, OpenAIRE India facilitates researchers' ability to discover relevant literature within the Indian scientific community. Its focus on accessibility and discovery is particularly beneficial for researchers seeking to explore a wide range of topics and stay abreast of developments within their respective fields. However, while OpenAIRE India excels in providing access to research materials, its primary emphasis on discovery may limit the depth of analysis that researchers can conduct within the platform. As a result, users may find it challenging to perform in-depth analyses or extract nuanced insights from the available resources, thus potentially hindering their ability to delve deeply into specific research topics or areas of interest.

On the other hand, Shodhganga stands as a significant digital repository specifically dedicated to Indian theses and dissertations. This repository offers researchers access to a vast collection of research conducted within the country across various disciplines and academic institutions. By providing a centralized platform for accessing Indian research outputs, Shodhganga plays a crucial role in preserving and disseminating the scholarly work of Indian researchers. However, access limitations based on institutional policies and thesis availability may present challenges for users seeking comprehensive access to research materials. Researchers may encounter difficulties in accessing specific theses or dissertations due to institutional restrictions or the availability of full-text documents.

While OpenAIRE India and Shodhganga offer valuable resources for researchers within the Indian scientific community, it is important to acknowledge the ongoing quest for a perfect solution that addresses the diverse needs and challenges faced by researchers. Efforts to enhance the accessibility, usability, and functionality of these platforms are crucial for fostering collaboration, innovation, and knowledge dissemination within the Indian scientific landscape. As technological advancements continue to evolve and new tools and platforms emerge, the pursuit of more comprehensive and user-friendly solutions remains essential for advancing scientific research and discovery in India.

CHAPTER 3. PROPOSED METHOD

3.1 METHODOLOGY

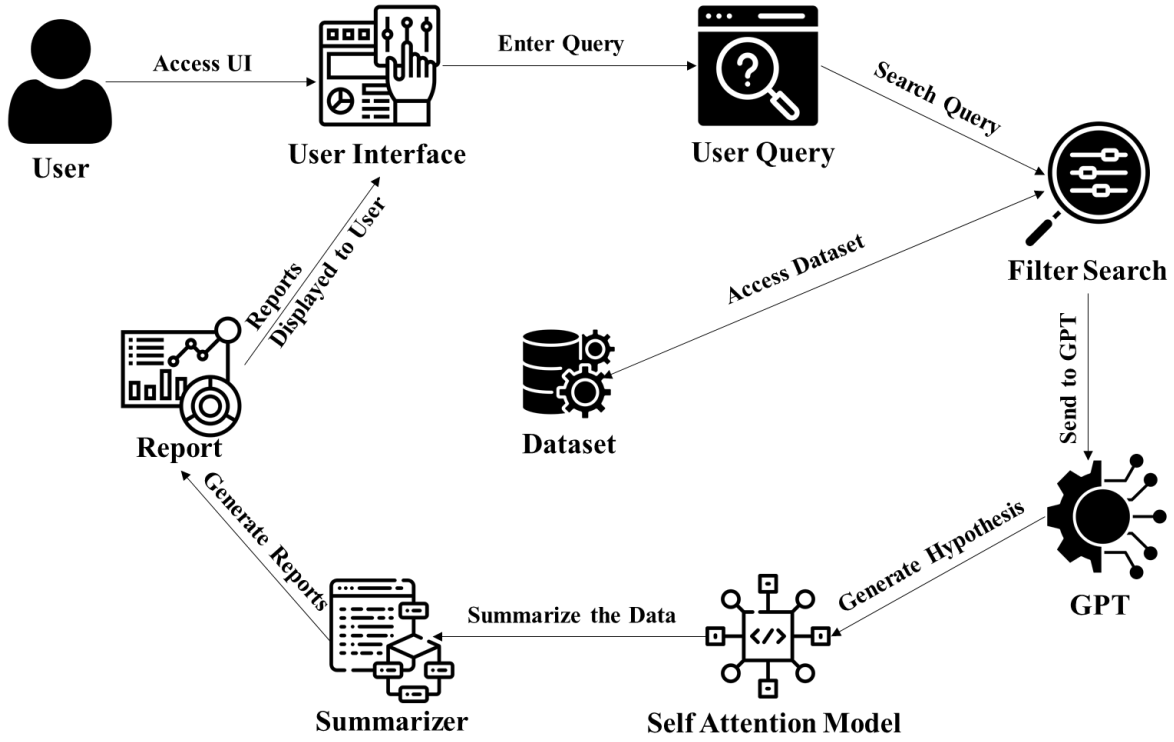

Figure 3.1: The architecture of the SciHypo system

Upon logging into the SciHypo System, users are greeted with the System/User Interface, which serves as the gateway to interaction. Within this interface, users have two primary options for initiating their search: they can either input their field of interest directly into the search bar or articulate their query in the form of a sentence. This user input is then preprocessed by the system, which extracts relevant keywords to facilitate effective searching.

Using these keywords, the system conducts a thorough search of its database, combing through an extensive collection of research papers. The user is then presented with an intuitive interface where they can further refine their search results. This interface offers various filtering options, such as specifying publication dates or selecting criteria like publishing firm or author's name. These filters help users narrow down the search results to precisely match their requirements.

Once the relevant papers are identified, they undergo analysis by the advanced Generative Pre-trained Transformer (GPT) engine. This engine delves deep into the content of the papers, discerning intricate relationships between concepts, methodologies, and findings. The insights gleaned from this analysis serve as the foundation for generating hypotheses that align contextually with the content of the retrieved papers.

Furthermore, the papers are subjected to analysis by a transformer equipped with self-attention mechanisms. This additional analysis aids in the identification of specific sections within each paper, such as abstracts and introductions, which contain crucial information relevant to the user's query.

After the analysis phase, users are presented with the option to select specific fields of interest within the retrieved papers. Based on their selections, the system generates concise summaries of the chosen sections from each paper. These summaries are then compiled into a coherent report, providing users with a comprehensive overview of the key insights and findings from the relevant research papers.

Therefore, the SciHypo System's workflow seamlessly guides users through the process of querying scientific literature, retrieving relevant papers, generating hypotheses based on extracted insights, and summarizing key findings to aid in research exploration and comprehension. This user-centric approach empowers researchers to efficiently navigate the vast landscape of scientific literature and uncover valuable insights to drive their research forward.

3.2 IMPLEMENTATION:

The methodology employed in this research involves a multi-faceted approach that encompasses the pre-training of the following

- Generative Pre-trained Transformer (GPT) Model
- Hypothesis Generation
- Transformers
- Summarizer

Initially, the GPT model undergoes extensive pre-training to comprehend the underlying structures and patterns inherent in scientific literature. This pre-training equips the GPT model with the ability to generate coherent and contextually relevant text, laying the groundwork for subsequent tasks in hypothesis generation and analysis.

Subsequently, a specialized hypothesis generation module is crafted, leveraging the pre-trained GPT model's capabilities. This module synthesizes insights from research papers, identifying pivotal concepts and relationships to formulate insightful hypotheses applicable across diverse scientific domains.

Transformers are integrated into the framework to bolster information processing and analysis. These transformers are pivotal in handling sizable text datasets and capturing intricate relationships between concepts, thereby enhancing the framework's efficacy in hypothesis generation and analysis.

A summarization module is embedded to deliver succinct summaries of research papers. Leveraging advanced summarization techniques such as extractive or abstractive

summarization, this module condenses key insights and findings into digestible formats, facilitating efficient analysis and hypothesis generation.

Overall, the implementation methodology aims to harness deep learning techniques to swiftly generate insightful and contextually relevant hypotheses across varied scientific domains. Through the pre-training of essential components and the integration of advanced modules, the framework aims to streamline the analysis of research papers, ultimately accelerating the pace of scientific discovery.

3.2.1 Pre-Training the GPT:

Algorithm:

Input:

sources: List of data sources (["PubMed Central", "arXiv", "OpenAIRE"])

data_types: List of desired data types (e.g., ["articles", "abstracts", "datasets", "patents"])

Steps:

Data Acquisition:

```
for source in sources:
    data = download_data(source, data_types)
    processed_data += data
```

Data Preprocessing:

```
for text in processed_data:
    text = clean_text(text) # Remove errors
    text = normalize_text(text) # Standardize text format
    entities = recognize_entities(text) # Extract entities
    link_entities(entities)
    processed_data[text] = updated_text
```

Model Fine-tuning:

```
model = load_gpt_model(pretrained=True)
fine_tune_model(model, processed_data)
```

Output:

processed_data: Clean and standardized dataset suitable for model training

The first phase involves the careful pre-training of the GPT model, laying the foundation for its understanding of scientific language and relationships. To ensure a comprehensive understanding of scientific literature, a diverse dataset is curated, drawing from sources like scientific articles, abstracts, datasets, and patents. Repositories such as PubMed Central, arXiv, and OpenAIRE are tapped into for their extensive collections. The acquired dataset undergoes meticulous pre-processing to eliminate errors, inconsistencies, and irrelevant information.

Text normalization, entity recognition, and named entity linking are employed to enhance the data quality. We use a pre-trained Generative Pre-trained Transformers GPT-3. The chosen GPT model is fine-tuned using specific libraries like transformers in Python. This step refines the model's understanding of scientific nuances and relationships between concepts.

3.2.2 Hypothesis Generation:

Algorithm:

Input:

research_area: String describing the research field.

hypothesis_format: String specifying desired hypothesis format (e.g., question, statement).

focus_area: String defining specific areas within the research field.

novelty_level: String indicating desired novelty (e.g., high, medium, low).

Steps:

Prompt Generation:

prompt = generate_prompt(research_area, hypothesis_format, focus_area, novelty_level)

Conditional Generation:

hypotheses = generate_hypotheses(model, prompt)

model = load_gpt_model(fine_tuned=True)

Post-processing:

hypotheses = filter_hypotheses(hypotheses, automated_filters) # Remove irrelevant ones

hypotheses = refine_hypotheses(hypotheses, domain_knowledge, expert_evaluation)

Output:

hypotheses: List of refined and relevant hypotheses.

Building upon the pre-training phase, the hypothesis generation process is designed to produce contextually relevant and focused hypotheses. Context-rich prompts related to the specific research area are crafted to guide the GPT in generating hypotheses. These prompts serve as crucial input to steer the model toward relevant and insightful outputs. Advanced GPT feature conditional generation, is leveraged to refine the output.

Specific properties, including hypothesis format, focus area, and novelty level, are specified to tailor the generated hypotheses. The outputs are meticulously post-processed to filter out irrelevant or nonsensical hypotheses. Domain knowledge and expert evaluation play a pivotal role in selecting promising hypotheses for further investigation.

3.2.3 Transformers:

Algorithm:

Input:

data_sources: List of scientific literature sources (e.g., [“PubMed Central”, “arXiv”])

data_types: List of desired data types (e.g., [“articles”, “abstracts”])

model: Pre-trained Transformer model (e.g., GPT-3)

Steps:

Model Pre-training:

Fine-tune a pre-trained Transformer model (e.g., GPT-3) on the processed data.

Encode Text:

encoded_text = model.encode(text)

Attention Analysis:

attention_weights = model.self_attention(encoded_text)

Section Identification:

abstracts = identify_sections(text, attention_weights, “abstract”)

introductions = identify_sections(text, attention_weights, “introduction”)

… (identify other sections as needed)

Information Extraction:

Methodology = extract_key_concepts(abstracts, introductions, “methodology”)

key_findings = extract_key_concepts(abstracts, introductions, “findings”)

… (extract other information as needed)

Build Output:

extracted_information = {“abstracts”: abstracts, “introductions”: introductions, “methodology”: methodology, “key_findings”: key_findings}

Iterative Refinement:

Refine the model based on expert feedback and evaluation metrics (e.g., accuracy, relevance).

This might involve adjusting hyperparameters, retraining the model with new data, or fine-tuning prompts.

Output:

extracted_information: Dictionary containing extracted information like abstracts, introductions, etc.

attention_weights: Matrix representing relationships between concepts

SciHypo utilizes a pre-trained GPT model, a type of Transformer architecture renowned for its ability to capture long-range dependencies in text. This is crucial for scientific literature, where understanding relationships across sentences and paragraphs is vital for generating meaningful hypotheses.

$$\mathbf{Attention(Q,K,V,M)=Softmax((QK^T)/\sqrt{(d_k)}+M)V}$$

The relationship between the concepts in various papers is demonstrated by the attention ratings found in the preceding equation. The key to an accurate understanding of relationships lies in self-attention mechanisms:

Understanding Text Structure: Transformers analyze each word's relevance to every other word in the text, regardless of their position. This enables them to grasp the overall structure and flow of scientific articles, identifying sections like abstracts, introductions, and methodologies.

Relationship Mapping: The model pays attention to connections between concepts and entities across the text. This aids in understanding cause-and-effect relationships, problem statements, and research aims, all crucial for formulating insightful hypotheses.

3.2.4 Summarization for Focused Hypotheses:

Algorithm:

Summarization:

Input:

extracted_information: Dictionary containing identified sections and key findings from text analysis

sum_model: Selected summarization model

Steps:

Section Selection:

sect_to_sum = select_sec(ex_info, re_domai)

Summarize Sections:

for the section in sect_to_sum:

s = sum_text(sum_model, ex_info[section])

summaries[section] = s

Output:

summaries: Dictionary containing summaries of relevant sections

Prompt Generation:

Input:

summaries: Dictionary containing section summaries

research_area: String describing the broader research field

hypothesis_format: String specifying desired hypothesis format

novelty_level: String indicating desired novelty (e.g., high, medium, low)

Steps:

Prompt Construction:

for section, summary in summaries.items():

prompt = generate_prompt(re_area, section, s, hypothesis_format, novelty_level)

prompts.append(prompt)

Output:

prompts: List of prompts for the Transformer model

SciHypo uses summarization techniques to condense the extracted data into concise and focused prompts. This ensures the generated hypotheses stay on track and address the core research problem.

Identifying Key Sentences: Summarization models analyze extracted sections like abstracts and identify key sentences that capture the essence of the information.

Generating Concise Overviews: The model generates summaries that highlight the main research objective, methodology, and key findings.

3.3 DATA PREPARATION:

Data preparation for the SciHypo project involves several steps to curate a comprehensive dataset to support the deep learning framework. Initially, a wide-ranging assortment of scientific literature is gathered from various sources such as research papers, journals, conferences, and preprint repositories. This ensures comprehensive coverage across diverse disciplines. Following this, text extraction techniques are applied to extract textual content from the collected documents while retaining important metadata like titles, authors, publication dates, and sources. This may involve parsing PDFs, or HTML pages, or utilizing APIs provided by literature databases.

After extraction, the text undergoes meticulous preprocessing to eliminate noise such as special characters, punctuation, and formatting inconsistencies. Techniques like tokenization, lowercasing, and stemming are applied to standardize the text format, laying the groundwork for subsequent analysis. Language-specific preprocessing steps are also implemented to address nuances in different languages, including handling varied character encodings, tokenization rules, and stopword lists, particularly for non-English texts.

Optionally, the dataset may be annotated to enrich it with additional metadata or labels, aiding supervised learning tasks. To facilitate model training, validation, and evaluation, the dataset is partitioned into distinct sets, ensuring representation across domains and periods. Tokenization and vectorization techniques transform the preprocessed text into numerical representations suitable for input into the deep learning model. This involves leveraging methods like word embeddings or byte pair encoding (BPE) for efficient encoding.

Optional data augmentation and filtering techniques may be employed to enhance dataset diversity and quality while balancing strategies to address any imbalances across research domains or topics. Through this meticulous data preparation process, the SciHypo project establishes a robust foundation for training a sophisticated deep learning framework, poised to revolutionize hypothesis generation through extensive literature analysis.

CHAPTER 4. RESULTS AND DISCUSSION

4.1 CASE STUDY 1

For the first case study of SciHypo, a meticulously curated dataset was constructed, drawing upon the vast knowledge repository of PubMed. This dataset comprises 1000 research articles specifically focused on breast cancer. This deliberate selection provides a rich and targeted resource for SciHypo to delve into the complex world of breast cancer research and extract valuable insights. By analyzing this specific corpus, SciHypo can hone its ability to understand the intricacies of the domain, identify patterns and relationships within existing research, and ultimately generate novel and impactful hypotheses that could propel our understanding and treatment of breast cancer forward.

Fig. 4.1. First three papers retrieved by the system for the search query "Breast Cancer"

The output image displays the most relevant research papers based on the user's specified filters. As shown in Figure 2, each paper is accompanied by the details such as the names of the authors. This concise presentation allows users to quickly assess the relevance and significance of each paper within their research context which in this case is "Breast Cancer".

Fig. 4.2. Abstract summary of an article on breast cancer

The system gives various filters that offer options such as abstract, introduction, literature review, methodology, results, and discussion as well as conclusion summaries and images within the paper. Upon selecting these options, the system presents a concise summary of the paper titled “Breast Cancer: Biology, BioMarkers, and Treatments”, as demonstrated in the above figure. This functionality enables users to efficiently access specific sections of the papers and obtain relevant information without the need to read through the entire document.

4.2 CASE STUDY 2

In the second case study, SciHypo takes on the captivating field of voice robots. To fuel its analysis and hypothesis generation, it leverages a carefully selected dataset of 1000 research articles sourced from ScienceDirect, a comprehensive platform for open-access scholarly publications. This dataset specifically focuses on voice robots, offering a rich and targeted environment for SciHypo to explore the current state of research and uncover potential avenues for future exploration.

Fig. 4.3. The papers relevant to Voice Robots

The first three publications that are pertinent to the user's question on "Voice Robots" are shown in Figure 4, along with the names of the authors of each paper that appears below it. Users can choose to view the paper's summary, hypothesis, and other details by clicking the plus symbol next to each title of a paper.

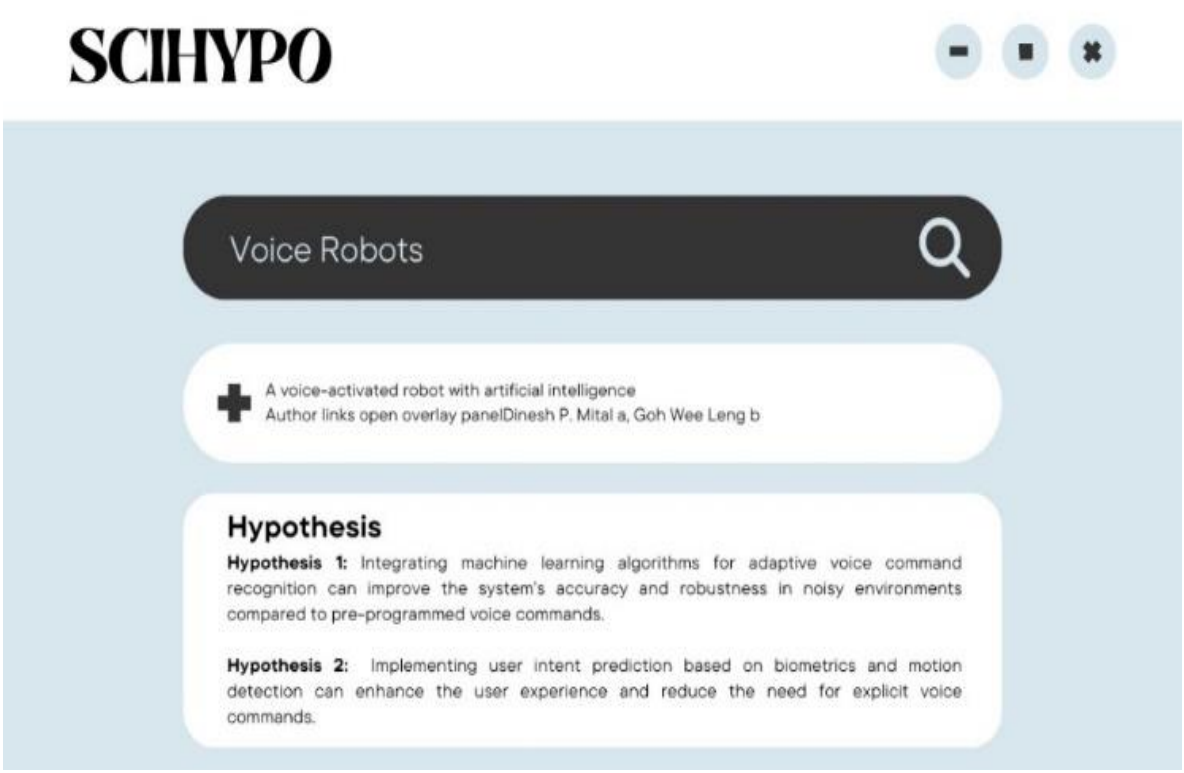

Fig. 4.4. Hypothesis of research paper under the category "Voice Robots"

Figure 5 depicts the two hypotheses generated by the GPT for the paper titled "A voice-activated robot with artificial intelligence", intended for further analysis by the user. These hypotheses offer potential insights and interpretations derived from the content of each paper, facilitating deeper examination and exploration of the research topics.

CHAPTER 5. CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

India boasts a vibrant scientific community with immense potential for groundbreaking discoveries. However, inefficient and inconsistent analysis of scientific literature acts as a major roadblock, hindering exploration and delaying progress. This paper introduces SciHypo, a novel deep-learning framework that leverages the power of self-attention mechanisms and Generative Pre-trained Transformers (GPT) to address these challenges.

SciHypo tackles the limitations of traditional methods by analyzing the vast corpus of scientific literature. It delves deep into the intricacies of the text, uncovering hidden patterns, relationships, and gaps in existing research through self-attention algorithms. This ability to grasp the "big picture" empowers SciHypo to identify novel research avenues and assess their potential significance.

Furthermore, SciHypo transcends the limitations of domain-specific expertise. Trained across diverse scientific disciplines, it fosters transdisciplinary thinking by generating hypotheses that bridge seemingly unrelated fields. This cross-pollination of ideas has the potential to spark groundbreaking discoveries and accelerate scientific progress in India.

By producing varied and contextually relevant hypotheses alongside concise summaries, SciHypo empowers researchers to navigate the vast landscape of scientific literature efficiently. This not only saves valuable time but also opens doors to exploring previously unseen research trajectories.

The results presented in this paper demonstrate SciHypo's effectiveness in generating hypotheses that resonate with experts, paving the way for a brighter future for Indian science. By bridging the gap between existing knowledge and novel ideas, SciHypo holds immense potential to propel India's scientific journey towards new heights and unlock a new era of discovery.

5.2 FUTURE SCOPE

The future scope of the SciHypo project holds immense potential for further advancements and applications in the realm of data-driven scientific hypothesis generation. Some key areas of future exploration and development include:

Enhanced Model Performance: Continuously refine and optimize the deep learning architecture underlying SciHypo to improve its hypothesis generation capabilities. This may involve exploring more advanced natural language processing techniques, incorporating feedback mechanisms, and fine-tuning model hyperparameters to achieve higher accuracy and efficiency. Multimodal Data Integration: Extend SciHypo's capabilities to incorporate multimodal data sources, such as images, graphs, and tables, in addition to text-based literature. By leveraging multiple data modalities, the framework can glean deeper insights and generate more nuanced hypotheses that consider diverse forms of scientific evidence.

Real-Time Hypothesis Generation: Develop real-time hypothesis generation functionalities within SciHypo, allowing researchers to receive instantaneous insights and suggestions as they explore scientific literature. This would involve optimizing the framework for speed and scalability to accommodate rapid analysis of large volumes of data. Domain-Specific Customization: Enable customization of SciHypo for specific scientific domains or research areas by fine-tuning the model on domain-specific datasets and incorporating domain-specific knowledge sources. Tailoring the framework to different disciplines would enhance its relevance and effectiveness in guiding research within specialized fields.

Integration with Knowledge Graphs: Integrate SciHypo with existing knowledge graphs and ontologies to leverage structured semantic information for hypothesis generation. By connecting textual information extracted from literature with structured knowledge representations, the framework can achieve a deeper understanding and more accurate hypothesis formulation.

Collaborative Hypothesis Exploration: Facilitate collaborative hypothesis exploration by incorporating features for sharing, discussing, and refining generated hypotheses within research communities. Integration with collaborative platforms and tools would promote interdisciplinary collaboration and collective knowledge advancement.

Applications Beyond Scientific Research: Explore applications of SciHypo beyond traditional scientific research, such as in industry for innovation discovery, in education for fostering critical thinking skills, or in healthcare for hypothesis-driven medical research and drug discovery. Ethical and Responsible AI Considerations: Address ethical and responsible AI considerations in the development and deployment of SciHypo, including issues related to data privacy, bias mitigation, transparency, and accountability. Implementing safeguards and guidelines ensures the responsible use of AI-powered hypothesis generation technologies.

By pursuing these avenues of future research and development, the SciHypo project can continue to push the boundaries of data-driven hypothesis generation, contributing to accelerated scientific discovery and innovation across various domains.

BIBLIOGRAPHY

[1] Spangler, Scott, et al. "Automated hypothesis generation based on mining scientific literature." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[2] Friederich, Pascal & Krenn, Mario & Tamblyn, Isaac & Aspuru-Guzik, Alán. Scientific intuition inspired by machine learning generated hypotheses. Machine Learning: Science and Technology. 2021.

[3] Jens Ludwig & Sendhil Mullainathan. Machine Learning as a Tool for Hypothesis Generation. March 2023.

[4] Jain, Rahul, Generation of Statistical Hypotheses: Methods and Applications (August 27, 2023).

[5] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, Iryna Gurevych. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. 11 Oct 2021.

[6] Shengtian Sang, Zhihao Yang, Zongyao Li, Hongfei Lin. Supervised Learning Based Hypothesis Generation from Biomedical Literature. 2015 Aug 25.

[7] Vida Abedi, Ramin Zand, Mohammed Yeasin, Fazle Faisal. An automated framework for hypotheses generation using literature. 2012.

[8] Ilya Tyagin, Ilya Safro. Interpretable Visualization of Scientific Hypotheses in Literature-based Discovery. 2021.

[9] Guangxu Xun, Kishlay Jha, Vishrawas Gopalakrishnan, Yaliang Li, Aidong Zhang. Generating Medical Hypotheses Based on Evolutionary Medical Concepts. 2017.

[10] Michael D. Gordon, R. K. Lindsay, Weiguo Fan. Literature-based discovery on the World Wide Web. 2002.

[11] P. Srinivasan. Text mining: Generating hypotheses from MEDLINE. 2004.

[12] Stephen J. Wilson, A. Wilkins, Matthew V. Holt, Byung-Kwon Choi, Daniel M. Konecki, Chih-Hsu Lin, Amanda M. Koire, Yue Chen, Seon-Young Kim, Yi Wang, Brigitta Wastuwidyaningtyas, J. Qin, L. Donehower, O. Lichtarge. Automated literature mining and hypothesis generation through a network of Medical Subject Headings. 2018.

[13] Justin Sybrandt, Michael Shtutman, and Ilya Safro. 2017. MOLIERE: Automatic Biomedical Hypothesis Generation System. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 1633–1642.

[14] J. Sybrandt, M. Shtutman and I. Safro, "Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1494-1503, doi: 10.1109/BigData.2018.8622637.

[15] Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. 2020. AGATHA: Automatic Graph Mining And Transformer-based Hypothesis Generation Approach. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20). Association for Computing Machinery, New York, NY, USA, 2757–2764.

[16] J. -H. Kim and A. Segev, "Research Hypothesis Generation Using Link Prediction in a Bipartite Graph," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 2863-2867, doi: 10.1109/BigData.2018.8622645.

[17] C. Vaquero, A. Ortega and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 4532-4535, doi: 10.1109/ICASSP.2011.5947362

[18] L. Müller, T. Wetzel, H. -C. Hobohm and T. Schrader, "Creativity Support Tools for Data Triggered Hypothesis Generation," 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems, Melbourne, VIC, Australia, 2012, pp. 24-27, doi: 10.1109/KICSS.2012.12.

[19] P. V. Rani, N. Ravi, S. M. Shalinic and P. Pariuentham, "Detecting and Assuaging Against Interest Flooding Attack Using Statistical Hypothesis Testing in Next Generation ICN," 2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP), Chennai, India, 2018, pp. 1-5, doi: 10.1109/ICCCSP.2018.8452848.

[20] H. Kang et al., "Diffusion-Based Pose Refinement and Multi-Hypothesis Generation for 3D Human Pose Estimation," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 5130-5134, doi: 10.1109/ICASSP48485.2024.10445850.

[21] S. Vudumu, P. V. Paul and L. Ramakrishnan, "Efficient multiple hypotheses tracking scheme using adaptive number of 'K' best hypotheses for target tracking in clutter," 2018 22nd International Microwave and Radar Conference (MIKON), Poznan, Poland, 2018, pp. 390-394, doi: 10.23919/MIKON.2018.8405235.

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to

comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1: Process, interpret the real-world data to formulate the model for predicting and forecasting.

PSO2: Apply machine learning techniques to design and develop automated systems to solve real world problems.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Project Outcomes	
Course Outcome (CO1)	Identify and analyze the problem statement using prior technical knowledge in the domain of interest.
Course Outcome (CO2)	Design and develop engineering solutions to complex problems by employing systematic approach.
Course Outcome (CO3)	Examine ethical, environmental, legal and security issues during project implementation.
Course Outcome (CO4)	Prepare and present technical reports by utilizing different visualization tools and evaluation metrics.

Mapping Table

AD3512: MAIN PROJECT															
Course Outcomes	Program Outcomes and Program Specific Outcome														
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	3	3	1					2	2	2				1	1
CO2	3	3	3	3	3			2	2	2		1		3	3
CO3	2	2	3	2	2	3	3	3	2	2	2			3	
CO4	2		1		3				3	3	2	2		2	2

Note: Map each project outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped

PAPER PUBLISHED

Letter of Acceptance

Details of accepted paper:

Paper ID	Title	Author(s)
ICOECA-144	SCIHYPO - A Deep Learning Framework for Data- Driven Scientific Hypothesis Generation from Extensive Literature Analysis	Mothilal Tadiparthi, Sindhu Pasupuleti, Sita Venkata Sathwika Talluri, Subramanyam Raju Sangaraju, Manikanta Mogili

Herewith, the conference committee of the 4th International Conference on Expert Clouds and Applications (ICOECA 2024) extends its warmest congratulations to you. We are pleased to inform you that after a rigorous peer review process, your aforementioned research paper has been accepted for presentation at the conference. Additionally, your paper has been recommended for inclusion in the ICOECA 2024 conference proceedings.

The ICOECA conference is scheduled to take place on 18-19, April 2024, at RV College of Engineering, Bengaluru, India. We encourage the active participation of highly qualified delegates like yourself to foster the exchange of innovative research ideas and insights.

We congratulate you on being successfully selected for the presentation of your research work in our esteemed conference.

Prof. Deepika K
ICOECA 2024

EJESRA
GSTIN - 33BPRPS5585K1ZY

Invoicing and payments
powered by

Payment Receipt Transaction Reference: pay_NsWxtKXkpGgP6J

This is a payment receipt for your transaction on ICOECA 2024

AMOUNT PAID **₹ 9,950.00**

ISSUED TO
sindhu.gdlv@gmail.com
+916304466397

PAID ON
30 Mar 2024

DESCRIPTION	UNIT PRICE	QTY	AMOUNT
Student Members	₹ 9,950.00	1	₹ 9,950.00

Total **₹ 9,950.00**

Amount Paid ₹ 9,950.00

Page 1 of 1

SCIHYPO - A Deep Learning Framework for Data-Driven Scientific Hypothesis Generation from Extensive Literature Analysis

Mothilal Tadiparthi
Assistant Professor, Dept. of AI&DS
SR Gudlavalleru Engineering College
Gudlavalleru, India
mothilal556@gmail.com

Sindhu Pasupuleti
UG Student, Dept. of AI&DS
SR Gudlavalleru Engineering College
Gudlavalleru, India
sindhu.gdlv@gmail.com

Sita Venkata Sathwika Talluri
UG Student, Dept. of AI&DS
SR Gudlavalleru Engineering College
Gudlavalleru, India
sathwikatalluri2002@gmail.com

Subramanyam Raju Sangaraju
UG Student, Dept. of AI&DS
SR Gudlavalleru Engineering College
Gudlavalleru, India
subramanyamraju09@gmail.com

Manikanta Mogili
UG Student, Dept. of AI&DS
SR Gudlavalleru Engineering College
Gudlavalleru, India
mogilimanikanta5555@gmail.com

***Abstract*— The process of scientific advancement heavily relies on hypothesis generation, a pivotal step in guiding research directions and discoveries. However, conventional methods often hinge on expert intuition and domain knowledge, which can constrain the exploration of novel ideas. To overcome these limitations, this paper introduces SciHypo, an innovative deep learning framework that taps into the extensive corpus of scientific literature to generate fresh hypotheses across fields of science and technology. SciHypo employs advanced natural language processing techniques to process and understand a vast array of scientific texts. Utilizing algorithms and methodologies, including Generative Pre-trained Transformer, self-attention model algorithms to analyze patterns, relationships, and gaps present in existing research. This empowers the model to uncover latent connections within the domain's knowledge landscape and recognizing novel research trajectories, gauging their potential significance. The filter search algorithms facilitates the production of varied and contextually relevant hypotheses. A significant strength of SciHypo lies in its adaptability and domain-agnostic nature. Trained across an expansive spectrum of scientific disciplines, the model proves its prowess in generating transdisciplinary hypotheses, thereby fostering the exchange of ideas even between seemingly unrelated domains. The results emphatically showcase SciHypo's capability to propose hypotheses that resonate with experts, thereby advancing the scientific frontiers.**

Keywords*— *Generative Pre-trained Transformer, Self-Attention Model, Filter Search, Transdisciplinary Hypotheses, Model Selection, Prompt Generation, Refinement and Evaluation.

I. INTRODUCTION

With aspirational projects and a growing pool of skilled researchers, India's scientific research landscape is full of promise. But a significant obstacle stands in the way of potentially innovative discoveries: the inaccurate and inconsistent interpretation of scientific research articles. This challenge poses a multifaceted issue with significant downstream repercussions.

Primarily, the prevailing manual analysis is time-consuming and susceptible to human error. Delving into-extensive and intricate scientific literature line by line proves to be a daunting task, often leaving researchers overwhelmed and with limited time to explore emerging fields or delve deeper into specific sub-disciplines. This can inhibit the generation of innovative ideas and confine researchers to well-trodden paths.

A 2020 study by the National Institute of Science, Technology, and Development Studies (NISTADS) revealed that Indian researchers spend an average of 15-20 hours per week on literature review and analysis. Another 2022 survey by the Indian Academy of Sciences disclosed that 68% of researchers felt they lacked sufficient time for thorough literature analysis due to manual methods.

Inconsistencies in analysis methods result in unreliable interpretations and missed connections. Subjective approaches and the absence of standardized tools across research groups lead to conflicting conclusions and missed opportunities for novel discoveries. This inconsistency contributes to a fragmented research landscape, making replicating and building upon existing work challenging.

Moreover, these inefficiencies hinder the translation of research into real-world solutions, creating a gap between knowledge generation and its application to address societal challenges. Delays in accurately analyzing and synthesizing research findings undermine the true impact of scientific advancements on areas like healthcare, agriculture, and sustainable development.

A 2020 report by the Department of Science and Technology (DST) estimated that delays in research due to inefficient literature analysis cost India over ₹500 crore annually. The consequences of these challenges are profound. Researchers spend days, even weeks, manually searching and analyzing relevant literature, hindering their ability to engage in advanced research and innovation. Inaccurate or incomplete analysis can lead to flawed research directions and wasted resources, ultimately slowing down India's scientific progress.

Addressing the shortcomings in scientific research analysis is not merely an academic concern but a national imperative. Proposing innovative solutions that facilitate efficient, consistent, and accurate analysis can empower researchers, expedite discovery, and propel India's scientific journey to unprecedented heights.

Introducing a novel approach to address this challenge, this system aims to revolutionize how researchers in India analyze scientific literature. SciHypo is an advanced deep learning framework that makes use of sophisticated algorithms and natural language processing. It presents a possible way to break down the issues preventing scientific advancement in India. Systematically analyzing the vast corpus of scientific literature, SciHypo can identify latent

connections, uncover hidden research trajectories, and generate transdisciplinary hypotheses. This transformative approach holds the potential to empower researchers, catalyze collaboration, and propel India toward a new era of scientific discovery.

A. *Case Study 1:*

In her pursuit of further avenues for breast cancer studies, Sarah attempted to navigate the complex landscape of scientific literature. Despite the vast resources available on PubMed, a repository of medical studies, she found herself overwhelmed by the sheer volume and technical language of the papers.

Navigating through thousands of documents became an impossible task, hindered by her struggle to comprehend the unfamiliar terminology and extract relevant information. Recognizing the need to bridge the gap between her curiosity and accessible knowledge, Sarah sought innovative solutions to streamline her exploration and access relevant research findings more effectively.

B. *Case Study 2:*

John, an engineer, was interested in studying the possibilities of speech robots. He discovered ScienceDirect, which had thousands of intriguing ideas, but it appeared hard to carefully assess each one for important informatics as well as new theories.

John struggled to find truly innovative ideas in the massive collection, and the technical terms and sheer volume of knowledge made it impossible to further his quest to push the bounds of voice robot technology. His desire to find a way to connect the dots between an abundance of data and novel discovery was tempered by frustration as he struggled with time constraints and an overwhelming amount of information.

II. LITERATURE REVIEW

The following selection of research papers explores the emerging field of automated hypothesis generation within the realm of scientific literature, providing diverse perspectives and approaches.

Spangler et al. [1] establish the foundation with KnIT, a system utilizing NLP and network analysis to mine scientific texts and propose innovative research directions. This early study underscores the potential of ML to expedite scientific discovery, paving the way for further exploration. Taking a nuanced approach, Friederich et al. acknowledge the vital role of human intuition alongside ML. Their framework [2] integrates both AI and human expertise to refine and prioritize machine-generated hypotheses, highlighting the efficacy of human-machine collaboration in this endeavor.

Ludwig & Mullainathan expand the scope beyond scientific research, investigating how ML can generate novel hypotheses across diverse contexts. Emphasizing the significance of diversity and counterfactuals in hypothesis generation [3], their review provides insights applicable to various fields. Jain offers a comprehensive overview of statistical hypothesis generation methods, exploring classical and modern techniques across domains [4]. This review serves as a valuable resource for understanding and applying statistical approaches to hypothesis formulation.

Introducing SciGen, [5] presents a dataset designed to train AI models in reasoning-aware text generation. Rich with scientific tables and explanations, this resource holds immense potential for advancing research in scientific text generation, indirectly aiding in hypothesis formulation. [6] delve into a supervised learning approach for biomedical literature, using labeled data to train a model that generates hypotheses based on keywords, entity relationships, and citation information. While showcasing the promise of supervised learning in specific domains, their approach underscores the limitations of manual labeling for scalability and generalizability.

Abedi et al. [7] introduce an automated framework for hypotheses generation from literature, likely detailing methodologies and techniques used in leveraging scientific literature. Tyagin and Safro [8] focus on visualizing scientific hypotheses extracted from literature-based discovery processes, aiming to enhance interpretability. Xun et al. [9] discuss generating medical hypotheses based on evolutionary medical concepts, possibly exploring the use of evolutionary principles in hypothesis formulation. Gordon et al. [10] explore literature-based discovery methodologies applied on the World Wide Web, likely covering techniques for mining insights and generating hypotheses from online sources.

Srinivasan [11] investigates text mining techniques for generating hypotheses from MEDLINE, likely discussing methods for extracting and analyzing relevant information. Wilson et al. [12] present an automated literature mining approach for hypothesis generation using a network of Medical Subject Headings (MeSH), involving the extraction and analysis of relationships between concepts from scientific literature. Collectively, these papers contribute diverse methodologies, techniques, and applications to hypothesis generation from scientific literature.

III. METHODOLOGY

A. *System Architecture*

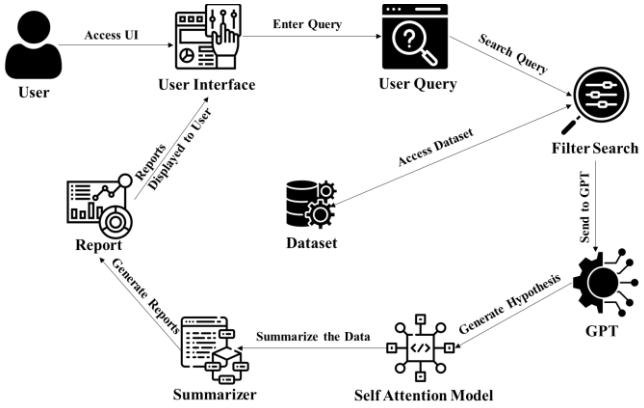

Fig. 1. The architecture of SciHypo system

The SciHypo System's process is shown in figure 1 when a user logs in. The system workflow is highlighted by the subsequent steps.

The user initiates interaction with the system by accessing the System/User Interface. Within the User Interface, the user either inputs their field of interest into the search bar or articulates their query in sentence form. The system preprocesses the query, extracting keywords to facilitate effective searching. Utilizing these keywords, the system searches the database to retrieve relevant research papers. The user is provided with an interface where they can apply filters

such as specifying the publication dates of research papers (ranging from a certain year to another) and selecting criteria such as publishing firm or author's name.

Once the relevant papers are identified, they are processed by the GPT engine to discern relationships between concepts within the papers. The insights gained are employed to generate a hypothesis that aligns contextually with the content of these papers. Subsequently, the papers undergo analysis by a transformer with self-attention mechanisms, aiding in the identification of sections such as abstracts and introductions within each paper. The user is given the option to select specific fields of interest, and the system then summarizes the chosen sections from each paper, presenting the user with a coherent report for enhanced comprehension.

To achieve privacy for user accounts within the system while maintaining accessibility to generated data, the system implements a technique called differential privacy. By applying differential privacy mechanisms to user account information, sensitive data such as user identities and personal details can be protected through techniques such as adding noise to query responses or aggregating data before analysis. This ensures that individual user accounts remain anonymous and secure, while still allowing the system to generate and share hypotheses and insights openly to the public via the internet, promoting transparency and collaboration in scientific research.

B. Implementation

The methodology employed in this research involves a multi-faceted approach that encompasses the pre-training of the following

1. Generative Pre-trained Transformer (GPT) model
2. Hypothesis generation
3. Transformers
4. Summarizer

The overarching goal is to leverage the capabilities of deep learning to generate insightful and contextually relevant hypotheses across various scientific domains and analysing the research papers quickly.

Pre-Training the GPT:

Algorithm:

Input:

sources: List of data sources (["PubMed Central", "arXiv", "OpenAIRE"])

data_types: List of desired data types (e.g., ["articles", "abstracts", "datasets", "patents"])

Steps:

1. Data Acquisition:

```
for source in sources:
    data = download_data(source, data_types)
    processed_data += data
```

2. Data Preprocessing:

```
for text in processed_data:
    text = clean_text(text) # Remove errors
    text = normalize_text(text) # Standardize text format
    entities = recognize_entities(text) # Extract entities
    link_entities(entities)
    processed_data[text] = updated_text
```

3. Model Fine-tuning:

```
model = load_gpt_model(pretrained=True)
fine_tune_model(model, processed_data)
```

Output:

processed_data: Clean and standardized dataset suitable for model training

The first phase involves the careful pre-training of the "GPT model", laying the foundation for its understanding of scientific language and relationships. To ensure a comprehensive understanding of scientific literature, a diverse dataset is curated, drawing from sources like scientific articles, abstracts, datasets, and patents.

The system leverages existing repositories such as PubMed Central, arXiv, and Europe PMC to access a diverse range of peer-reviewed research articles. Through automated web scraping and API integration, the system retrieves data from these repositories, ensuring a comprehensive collection of open-access literature. By utilizing APIs provided by these repositories, the system can efficiently query and retrieve relevant documents based on predefined search criteria, enabling seamless integration of new research findings into its knowledge base. This approach ensures access to a rich source of scientific literature, facilitating the generation of hypotheses and insights across various domains of research.

The acquired dataset undergoes meticulous pre-processing to eliminate errors, inconsistencies, and irrelevant information. Text normalization, entity recognition, and named entity linking are employed to enhance the data quality. We use a pre-trained Generative Pre-trained Transformers GPT-3. The chosen GPT model is fine-tuned using specific libraries like transformers in Python. This step refines the model's understanding of scientific nuances and relationships between concepts.

Hypothesis Generation:

Algorithm:

Input:

research_area: String describing the research field.

hypothesis_format: String specifying desired hypothesis format (e.g., question, statement).

focus_area: String defining specific area within the research field.

novelty_level: String indicating desired novelty (e.g., high, medium, low).

Steps:

1. Prompt Generation:

```
prompt = generate_prompt(research_area,
hypothesis_format, focus_area, novelty_level)
```

2. Conditional Generation:

hypotheses = generate_hypotheses(model, prompt)

model = load_gpt_model(fine_tuned=True)

3. Post-processing:

hypotheses = filter_hypotheses(hypotheses, automated_filters) # Remove irrelevant ones

hypotheses = refine_hypotheses(hypotheses, domain_knowledge, expert_evaluation)

Output:

hypotheses: List of refined and relevant hypotheses.

Building upon the pre-training phase, the hypothesis generation process is designed to produce contextually relevant and focused hypotheses. Context-rich prompts related to the specific research area are crafted to guide the GPT in generating hypotheses. These prompts serve as a crucial input to steer the model toward relevant and insightful outputs. Advanced GPT feature conditional generation, is leveraged to refine the output.

Specific properties, including hypothesis format, focus area, and novelty level, are specified to tailor the generated hypotheses. The outputs are meticulously post-processed to filter out irrelevant or nonsensical hypotheses. Domain knowledge and expert evaluation play a pivotal role in selecting promising hypotheses for further investigation.

The combined utilization of self-attention mechanisms and Generative Pre-trained Transformers (GPTs) empowers SciHypo to comprehensively analyze scientific literature and uncover intricate patterns. Self-attention enables the model to discern relationships between concepts, identify key information, and capture long-range dependencies within text data, facilitating a nuanced understanding of research papers. Meanwhile, GPTs leverage this contextual understanding to identify underlying patterns and trends, enabling SciHypo to generate novel hypotheses that resonate with the broader scientific landscape. Through this synergy, SciHypo moves beyond simple keyword matching, offering researchers a sophisticated tool for navigating and exploring the complexities of scientific literature.

Transformers:

Algorithm:

Input:

data_sources: List of scientific literature sources (e.g., ["PubMed Central", "arXiv"])

data_types: List of desired data types (e.g., ["articles", "abstracts"])

model: Pre-trained Transformer model (e.g., GPT-3)

Steps:

1. Model Pre-training:

Fine-tune a pre-trained Transformer model (e.g., GPT-3) on the processed data.

2. Encode Text:

encoded_text = model.encode(text)

3. Attention Analysis:

attention_weights = model.self_attention(encoded_text)

4. Section Identification:

abstracts = identify_sections(text, attention_weights, "abstract")

introductions = identify_sections(text, attention_weights, "introduction")

... (identify other sections as needed)

5. Information Extraction:

Methodology = extract_key_concepts(abstracts, introductions, "methodology")

... (extract other information as needed)

6. Build Output:

extracted_information = {"abstracts": abstracts, "introductions": introductions, "methodology": methodology, "key_findings": key_findings}

7. Iterative Refinement:

Refine the model based on expert feedback and evaluation metrics (e.g., accuracy, relevance).

This might involve adjusting hyperparameters, retraining the model with new data, or fine-tuning prompts.

Output:

extracted_information: Dictionary containing extracted information like abstracts, introductions, etc.

attention_weights: Matrix representing relationships between concepts

SciHypo utilizes a pre-trained GPT model, a type of Transformer architecture renowned for its ability to capture long-range dependencies in text. This is crucial for scientific literature, where understanding relationships across sentences and paragraphs is vital for generating meaningful hypotheses.

$$Attention(Q,K,V,M) = Softmax\left(\frac{Q*K^T}{\sqrt{d_k}} + M\right) * V$$

The relationship between the concepts in various papers is demonstrated by the attention ratings found in the preceding equation. The key to accurate understanding of relationships lies in self-attention mechanisms:

Understanding Text Structure: Transformers analyze each word's relevance to every other word in the text, regardless of their position. This enables them to grasp the overall structure and flow of scientific articles, identifying sections like abstracts, introductions, and methodologies.

Relationship Mapping: The model pays attention to connections between concepts and entities across the text. This aids in understanding cause-and-effect relationships, problem statements, and research aims, all crucial for formulating insightful hypotheses.

Summarization for Focused Hypotheses:

Algorithm:

Summarization:

Input:

extracted_information: Dictionary containing identified sections and key findings from text analysis

sum_model: Selected summarization model

Steps:

1. Section Selection:

   sect_to_sum = select_sec(ex_info, re_domai)

2. Summarize Sections:

   for section in sect_to_sum:

   s = sum_text(sum_model, ex_info[section])

   summaries[section] = s

Output:

summaries: Dictionary containing summaries of relevant sections

Prompt Generation:

Input:

summaries: Dictionary containing section summaries

research_area: String describing the broader research field

hypothesis_format: String specifying desired hypothesis format

novelty_level: String indicating desired novelty (e.g., high, medium, low)

Steps:

1. Prompt Construction:

for section, summary in summaries.items():

prompt = generate_prompt(re_area, section, s, hypothesis_format, novelty_level)

prompts.append(prompt)

Output:

prompts: List of prompts for the Transformer model

SciHypo uses summarization techniques to condense the extracted data into concise and focused prompts. This ensures the generated hypotheses stay on track and address the core research problem.

Identifying Key Sentences: Summarization models analyze extracted sections like abstracts and identify key sentences that capture the essence of the information.

Generating Concise Overviews: The model generates summaries that highlight the main research objective, methodology, and key findings.

IV. Results and Discussion

A. *Case Study 1*

For the first case study of SciHypo, a meticulously curated dataset was constructed, drawing upon the vast knowledge repository of PubMed. This dataset comprises 1000 research articles specifically focused on breast cancer. This deliberate selection provides a rich and targeted resource for SciHypo to delve into the complex world of breast cancer research and extract valuable insights. By analyzing this specific corpus, SciHypo can hone its ability to understand the intricacies of the domain, identify patterns and relationships within existing research, and ultimately generate novel and impactful hypotheses that could propel our understanding and treatment of breast cancer forward.

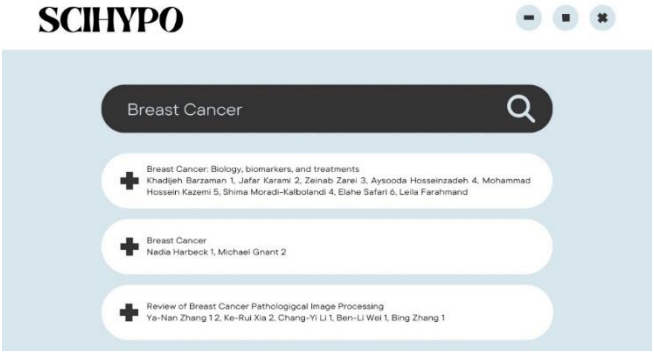

Fig. 2. First three papers retrieved by the system for the search query "Breast Cancer"

Based on the user-specified filters, the output image shows the research papers that are most pertinent. Every document has accompanying details, including the names of the authors, as seen in figure 2. With this succinct format, users may evaluate each paper's importance and relevance in relation to their study context—in this case, "Breast Cancer"—quickly.

Fig. 3. Abstract summary of an article on breast cancer

The system gives various filters that offer options such as abstract, introduction, literature review, methodology, results and discussion as well as conclusion summaries and images within the paper. Upon selecting these options, the system presents a concise summary of the paper titled "Breast Cancer: Biology, BioMarkers, and Treatments", as demonstrated in the above figure. This functionality enables users to efficiently access specific sections of the papers and obtain relevant information without the need to read through the entire document.

B. *Case Study 2:*

In the second case study, SciHypo takes on the captivating field of voice robots. To fuel its analysis and hypothesis generation, it leverages a carefully selected dataset of 1000 research articles sourced from ScienceDirect, a comprehensive platform for open access scholarly publications. This dataset specifically focuses on voice robots, offering a rich and targeted environment for SciHypo to explore the current state of research and uncover potential avenues for future exploration.

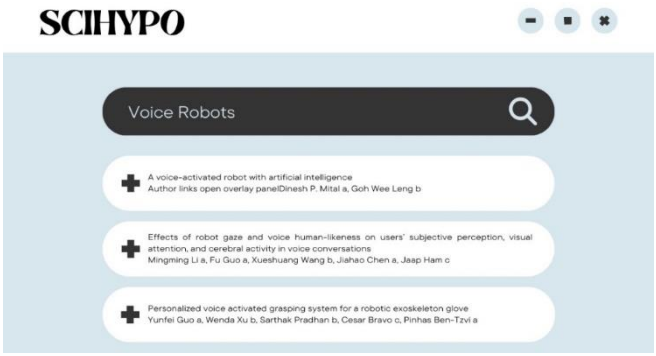

Fig. 4. The papers relevant to Voice Robots

The first three publications that are pertinent to the user's question on "Voice Robots" are shown in figure 4, along with the names of the authors of each paper that appears below it. Users can choose to view the paper's summary, hypothesis, and other details by clicking the plus symbol next to each title of a paper.

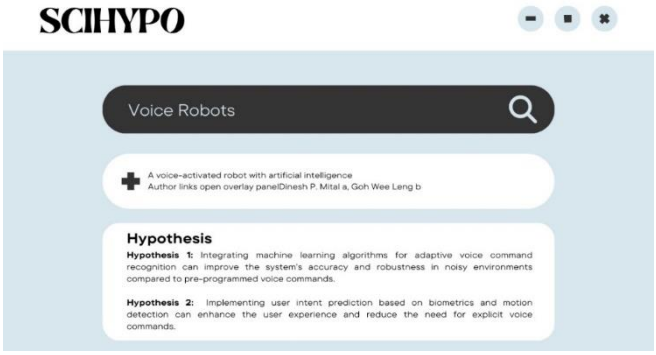

Fig. 5. Hypothesis of research paper under the category "Voice Robots"

The figure 5 depicts the two hypotheses generated by the GPT for the paper titled "A voice-activated robot with artificial intelligence", intended for further analysis by the user. These hypotheses offer potential insights and interpretations derived from the content of each paper, facilitating deeper examination and exploration of the research topics.

CONCLUSION

India boasts a vibrant scientific community with immense potential for pioneering discoveries. However, inefficient and inconsistent analysis of scientific literature acts as a major roadblock, hindering exploration and delaying progress. This paper introduces SciHypo, a novel deep learning framework that leverages the power of self-attention mechanisms and Generative Pre-trained Transformers (GPT) to address these challenges.

SciHypo tackles the limitations of traditional methods by analyzing the vast corpus of scientific literature. It delves deep into the intricacies of the text, uncovering hidden patterns, relationships, and gaps in existing research through self-attention algorithms. This ability to grasp the "big picture" empowers SciHypo to identify novel research avenues and assess their potential significance.

Furthermore, SciHypo transcends the limitations of domain-specific expertise. Trained across diverse scientific disciplines, it fosters transdisciplinary thinking by generating hypotheses that bridge seemingly unrelated fields. This interchange of ideas has the potential to spark innovative discoveries and accelerate scientific progress in India.

By producing varied and contextually relevant hypotheses alongside concise summaries, SciHypo empowers researchers to navigate the vast landscape of scientific literature efficiently. This not only saves valuable time but also opens doors to exploring previously unseen research trajectories.

The findings provided in this research show how effective SciHypo is at producing expert-approved hypotheses, opening the door to a more promising future for Indian science. SciHypo has the potential to greatly advance India's scientific endeavours and usher in a new era of discovery by uniting cutting-edge concepts with preexisting information.

REFERENCES

[1] Spangler, Scott, et al. "Automated hypothesis generation based on mining scientific literature." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[2] Friederich, Pascal & Krenn, Mario & Tamblyn, Isaac & Aspuru-Guzik, Alán. Scientific intuition inspired by machine learning generated hypotheses. Machine Learning: Science and Technology. 2021.

[3] Jens Ludwig & Sendhil Mullainathan. Machine Learning as a Tool for Hypothesis Generation. March 2023.

[4] Jain, Rahul, Generation of Statistical Hypotheses: Methods and Applications (August 27, 2023).

[5] Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, Iryna Gurevych. SciGen: a Dataset for Reasoning-Aware Text Generation from Scientific Tables. 11 Oct 2021.

[6] Shengtian Sang, Zhihao Yang, Zongyao Li, Hongfei Lin. Supervised Learning Based Hypothesis Generation from Biomedical Literature. 2015 Aug 25.

[7] Vida Abedi, Ramin Zand, Mohammed Yeasin, Fazle Faisal. An automated framework for hypotheses generation using literature. 2012.

[8] Ilya Tyagin, Ilya Safro. Interpretable Visualization of Scientific Hypotheses in Literature-based Discovery. 2021.

[9] Guangxu Xun, Kishlay Jha, Vishrawas Gopalakrishnan, Yaliang Li, Aidong Zhang. Generating Medical Hypotheses Based on Evolutionary Medical Concepts. 2017.

[10] Michael D. Gordon, R. K. Lindsay, Weiguo Fan. Literature-based discovery on the World Wide Web. 2002.

[11] P. Srinivasan. Text mining: Generating hypotheses from MEDLINE. 2004.

[12] Stephen J. Wilson, A. Wilkins, Matthew V. Holt, Byung-Kwon Choi, Daniel M. Konecki, Chih-Hsu Lin, Amanda M. Koire, Yue Chen, Seon-Young Kim, Yi Wang, Brigitta Wastuwidyaningtyas, J. Qin, L. Donehower, O. Lichtarge. Automated literature mining and hypothesis generation through a network of Medical Subject Headings. 2018.

[13] Justin Sybrandt, Michael Shtutman, and Ilya Safro. 2017. MOLIERE: Automatic Biomedical Hypothesis Generation System. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 1633–1642. https://doi.org/10.1145/3097983.3098057

[14] J. Sybrandt, M. Shtutman and I. Safro, "Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1494-1503, doi: 10.1109/BigData.2018.8622637.

[15] Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. 2020. AGATHA: Automatic Graph Mining And Transformer based Hypothesis Generation Approach. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20). Association for Computing Machinery, New York, NY, USA, 2757–2764. https://doi.org/10.1145/3340531.3412684