AI-Powered Academic Research Assistant

Submitted in partial fulfillment of the requirements
for the degree of

B.E. Computer Engineering

By

Daniel Ferreira	**Roll No:54**	**212255**
Prem Tatkari	**Roll No:56**	**212262**
Divyesh Mistry	**Roll No:57**	**212258**
Kyran Almeida	**Roll No:58**	**212252**

Guide
Prof. Pradnya Sawant
Assistant Professor

Department of Computer Engineering
St. Francis Institute of Technology
(Engineering College)

University of Mumbai

2023-2024

CERTIFICATE

This is to certify that the project entitled **"AI-Powered Academic Research Assistant"** is a bonafide work of **"Daniel Ferreira(Roll no.54), Prem Tatkari(Roll no.56), Divyesh Mistry(Roll no.57) and Kyran Almeida(Roll no.58)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

(Prof. Pradnya Sawant)

Guide

(Dr. Kavita Sonawane)	**(Dr. Sincy George)**
Head Of Department	**Principal**

Project Approval Report for B.E.

This project report entitled ***"AI-Powered Academic Research Assistant"*** by ***Daniel Ferreira(Roll no.:54), Prem Tatkari(Roll no.:56), Divyesh Mistry(Roll no:57) and Kyran Almeida(Roll no:58)*** is approved for the degree of ***B.E. in Computer Engineering***.

Examiners

1. ____________________

2. ____________________

Date: 31/10/2023

Place: Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included; we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in this submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Daniel Ferreira (54)

Prem Tatkari (56)

Divyesh Mistry (57)

Kyran Almeida (58)

Date: 31/10/2023

Abstract

In the ever-expanding landscape of academic research, the demand for efficient and intelligent tools to assist researchers in navigating, comprehending, and synthesising vast amounts of information has never been more pressing. This project introduces an innovative AI-powered Academic Research Assistant—a dynamic and sophisticated system designed to revolutionise the research process. Harnessing the capabilities of cutting-edge natural language processing and machine learning techniques, the AI-powered assistant seamlessly integrates into the researcher's workflow. As a result, the AI-powered Academic Research Assistant stands poised to reshape the landscape of academic research. It offers a tangible solution to the information overload conundrum, amplifying research efficiency, knowledge dissemination, and cross-disciplinary collaboration. This project not only exemplifies technological innovation but also underscores the potential for AI to augment human intelligence in tackling complex, knowledge-intensive challenges.

Keywords: *Machine Learning, AI, Academic Research Assistant, Natural Language Processing*

Contents

List of Figures

Chapter 1

Introduction

An Academic Research Assistant is a crucial role in the field of academic research, providing support to faculty members and research teams. They assist in conducting literature reviews, gathering and analysing data, and preparing reports or manuscripts for publication. Research assistants often work closely with professors or researchers, helping them design experiments, develop research methodologies, and formulate research questions. Our project aims to provide researchers with an AI-powered academic research assistant. It will ease the process of information gathering for researchers.

1.1 Description

AI-Powered Academic Research Assistant is a one-step solution to all your research needs. We plan to develop an application that is easy-to-operate and solves many problems for researchers. Our platform lets you find research papers related to your area of interest along with their summaries and key takeaways.

1.2 Problem Formulation

Researchers often spend significant time and effort searching for relevant research papers, extracting useful information, and organising their findings. This project aims to automate these tasks and provide researchers with a streamlined and personalised research experience. The solution is an AI-powered research assistant that uses web scraping, machine learning, and natural language processing to improve the research process.

To address these issues, there is a need for a more streamlined and automated job placement system that facilitates better communication, reduces inefficiencies, and provides data-driven insights for continuous improvement. Such a system would benefit both students and employers by expanding opportunities and ensuring a more inclusive and efficient job placement process.

1.3 Motivation

Embarking on the journey to create an AI-powered Academic Research Assistant holds the promise of transforming the way knowledge is discovered and shared. Imagine a world where researchers can delve into the depths of information with unprecedented ease, where connections between seemingly unrelated studies are unveiled effortlessly, and where time-consuming tasks are automated, freeing up intellectual energy for true innovation. This initiative is about providing researchers, educators, and learners with a tool that can catalyse breakthroughs, expedite learning, and foster worldwide collaboration, not just technology. By pursuing this endeavour, we are not only embracing the cutting edge of AI, but also being a catalyst for academic development. The difficulties we'll face and the solutions we'll devise have the potential to reshape research methodology, leaving an enduring mark on how we advance human knowledge. Our commitment to developing an AI-powered Academic Research Assistant is an investment not just in the future of education and research, but also in our own development as a creative visionary.

1.4 Proposed Solution

The proposed solution is an AI-powered academic research assistant that utilises web scraping, machine learning, and natural language processing techniques to enhance the research process. The system will scrape relevant data from academic websites, journals ensuring comprehensive coverage of research sources. The project will culminate in the development of a user-friendly web interface where users can input queries, browse search results, and explore papers.

1.5 Scope of the project

The scope of this project encompasses the development of a comprehensive AI-powered Academic Research Assistant, designed to enhance every facet of the research process. It will involve implementing natural language processing techniques to efficiently parse and comprehend diverse academic literature. The assistant will be capable of extracting key insights, and generating concise summaries. The project will require building an intuitive user interface, ensuring seamless integration into researchers' workflows. Ethical considerations, data privacy, and responsible AI usage will be integral components of the development. Overall, the scope envisions a transformative tool that streamlines information synthesis, encourages knowledge exchange, and catalyses academic innovation.

Chapter 2

Review of Literature

LongT5 [1], a new text-to-text transformer for long sequences. LongT5 is based on the T5 transformer architecture, but it makes several modifications to improve its performance on long sequences. These modifications include: Using a new attention mechanism that is more efficient for long sequences. Pre-training LongT5 on a dataset of long sequences. Scaling up the size of LongT5. On the evaluation of LongT5 on several summarization and question-answering tasks. They show that LongT5 outperforms the original T5 model on these tasks, and it achieves state-of-the-art results on some of them. However, LongT5 is not yet as widely used as other language models, so there is less research available on its strengths and weaknesses.

PEGASUS [2], a new pre-training method for abstractive summarization. PEGASUS is based on the Transformer language model, but it makes several modifications to improve its performance on abstractive summarization. These modifications include: Using a new objective function that encourages the model to generate summaries that are both informative and fluent. Pre-training PEGASUS on a dataset of human-written summaries and their corresponding source documents. On the evaluation of PEGASUS on several summarization tasks. They show that PEGASUS outperforms the state-of-the-art on these tasks. But the authors of the paper do not compare their model to other models that have been pre-trained on similar datasets.

On Extractive and Abstractive Neural Document Summarization with Transformer Language Models [3] presents a study on extractive and abstractive neural document summarization with transformer language models. They compare the performance of several transformer-based models on a variety of summarization tasks. They show that transformer-based models can achieve state-of-the-art results on both extractive and abstractive summarization tasks. They trained the transformer models from scratch. But such training is very expensive. LongT5, PEGASUS, and the transformer-based models have all achieved state-of-the-art results on a variety of summarization tasks. These models are likely to play an important role in the development of future text summarization systems. The field of text summarization is rapidly evolving, and there are a number of exciting research directions that are being explored. As these research directions continue to be developed, it is likely that text summarization systems will become even more powerful and capable.

Chapter 3

System Analysis

3.1 Functional Requirements

1. **Search and Discovery:** The system will allow users to search for academic papers, research articles, and related content The search results will be relevant and presented in a user-friendly format.

2. **Web Scraping:** Web scraping is the process of extracting content and data from a website. Using automated tools such as Selenium and requests_html libraries this task can be performed seamlessly.

3. **Natural Language Processing:**The system will employ advanced natural language processing techniques to understand user queries, interpret content, and generate summaries.

4. **User Interface:** An intuitive and user-friendly interface allows users to interact with the assistant easily. The interface is responsive, accessible, and adaptable to different devices.

5. **Abstractive Summarization:** Task of generating a concise summary that captures the salient ideas of the source text

3.2 Non-Functional Requirements

1. **Performance:** The system should be capable of handling a large volume of data without experiencing significant delays. Response times for various

tasks (e.g., search, summarization) should be within acceptable limits.

2. **Usability and User Experience:** The user interface would be intuitive, user-friendly, and accessible. It will be designed to accommodate researchers with varying levels of technical expertise and adhere to established design principles for effective user experience.

3. **Compatibility:** The system should be compatible with a variety of devices and platforms, including desktops, laptops, tablets, and mobile phones, and support multiple web browsers.

4. **Ethical Considerations:** The system should adhere to ethical guidelines, ensuring transparency in how recommendations are generated and handling user data for privacy regulations.

3.3 Specific Requirements

Hardware Specifications:

- 4 GB RAM minimum, 8 GB RAM recommended.
- 1.6 GHz or faster processor.
- 200 MB of available disk space minimum, 500 MB Recommended(for Visual Studio Code).

Software Specifications:

- Operating System: Windows
- Backend: Flask
- Machine learning model: LongT5
- Frontend: HTML, CSS, Javascript
- Editor: Visual Studio Code

3.4 Use-Case Diagrams and Description

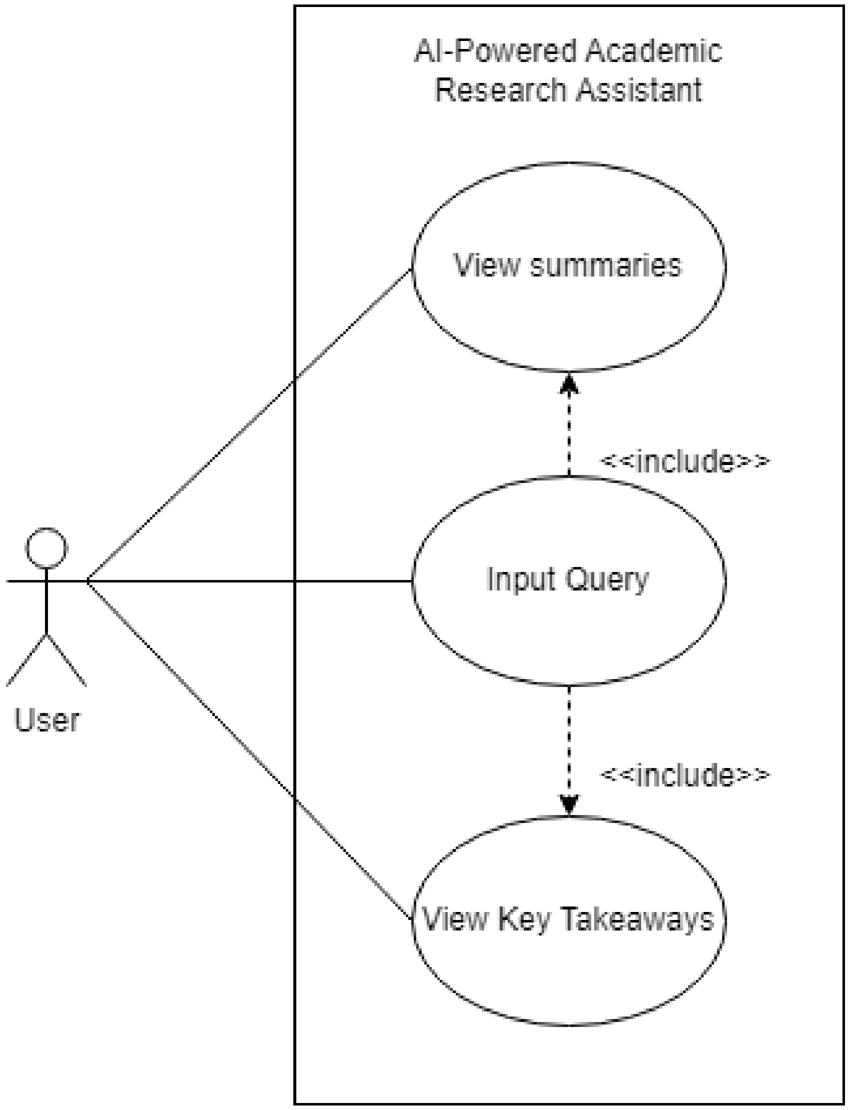

Figure 3.1: Use-Case Diagram for AI-Powered Academic Research Assistant

Use Case Description:

The above use case diagram (Figure 3.1) shows the high-level functions and scope of the system. A use case describes how a system behaves in response to a request from the perspective of the user. Users of our system can carry out the following tasks: The user enters their query using natural language. They can then see the summary that was created and the key takeaways.

Chapter 4

Analysis Modeling

4.1 Activity Diagram

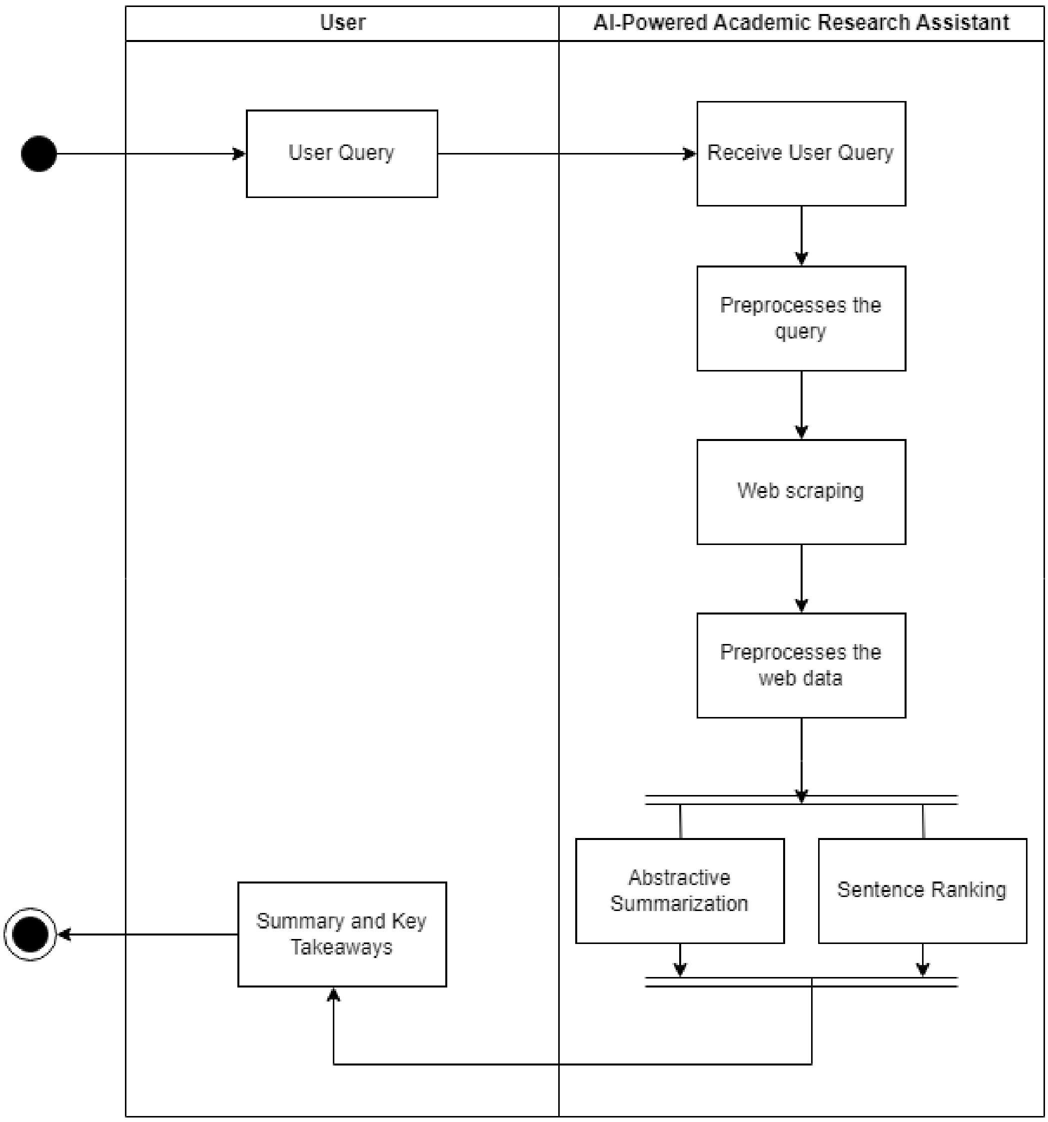

Figure 4.1: Activity diagram for AI-Powered Academic Research Assistant

The Activity Diagram 4.1 shows the process of generating a summary and key takeaways from an AI-powered model. The system uses techniques such as Data Preprocessing, Web-scraping, and a transformer model to generate the output.

The activity diagram shows the following steps:

1. The user inputs the topic they want to research.
2. The system preprocesses the query and finds out relevant topics.
3. The model performs web-scraping for the given topic.
4. The scraped data is further preprocessed and fed to the summarization and sentence ranking model.
5. The model generates the Summarization and Key Takeaways for the queried topic.

4.2 Functional Modeling

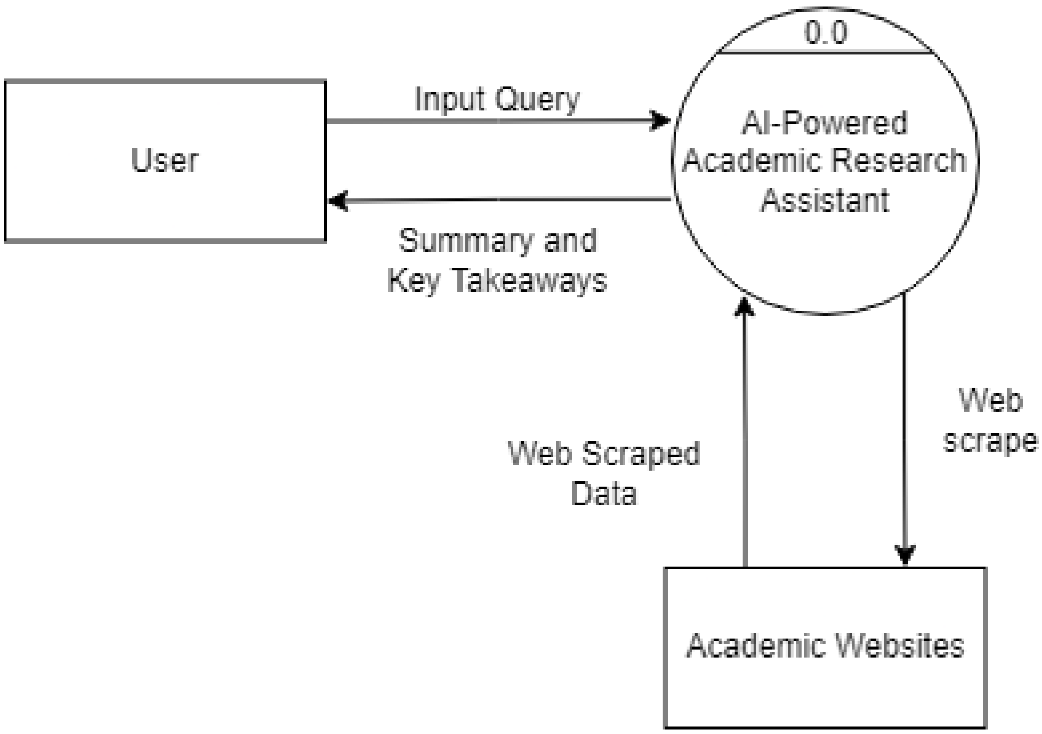

Figure 4.2: DFD Level 0 for AI-Powered Academic Research Assistant

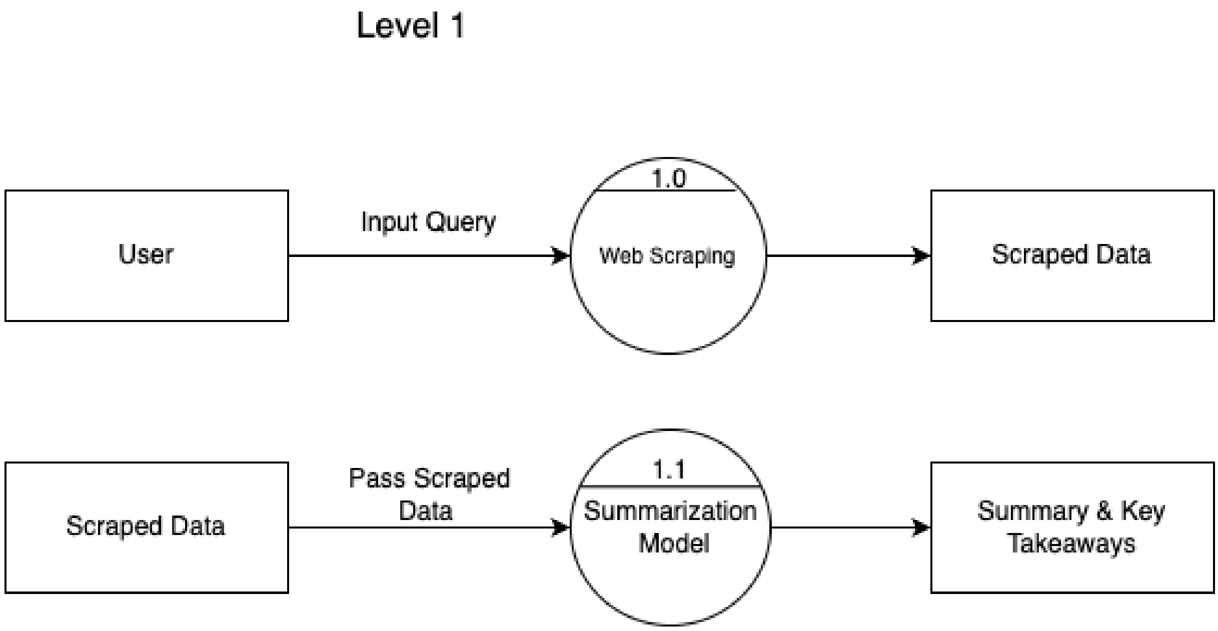

Figure 4.3: DFD Level 1 for AI-Powered Academic Research Assistant

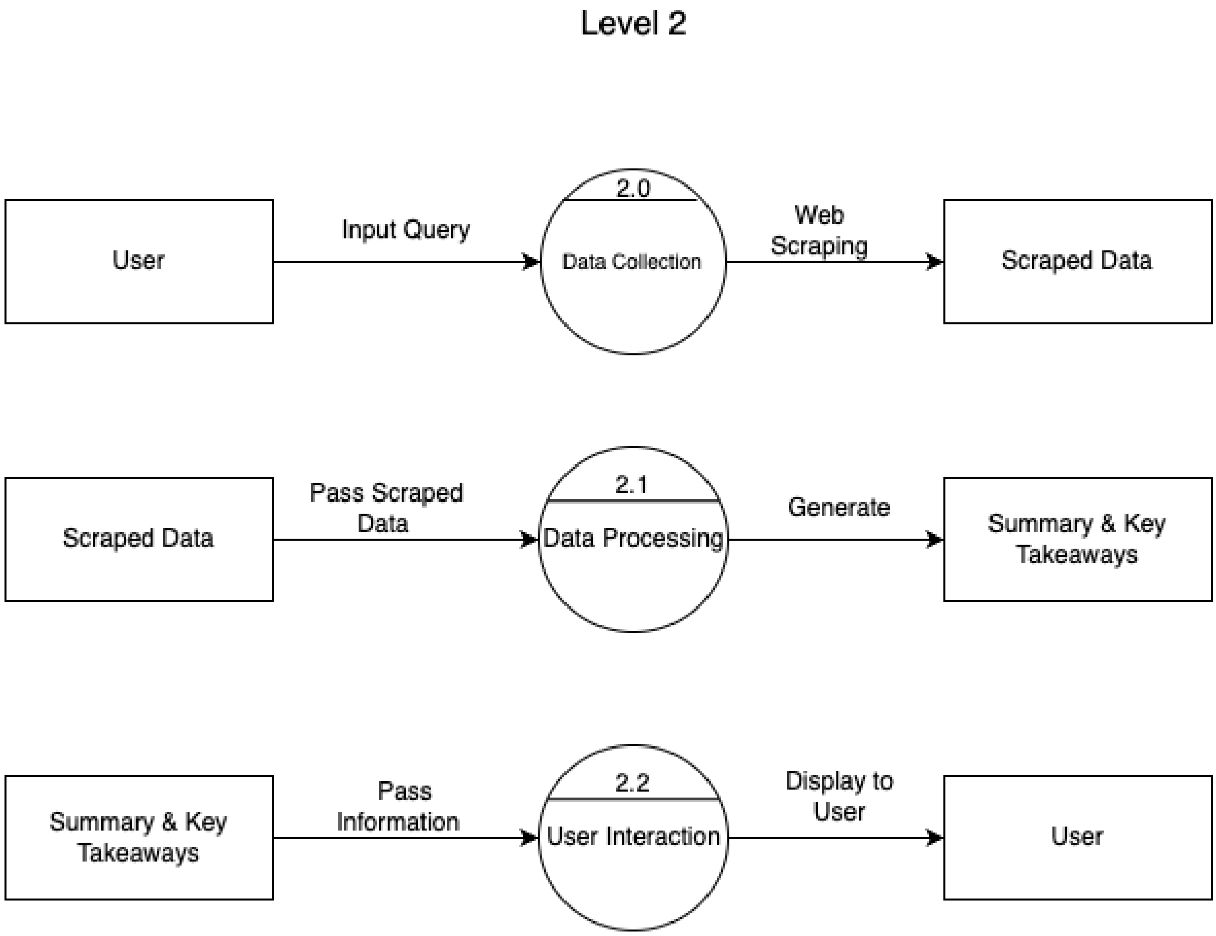

Figure 4.4: DFD Level 2 for AI-Powered Academic Research Assistant

The above diagrams show the data flow of the system in three levels which are Level 0, Level 1, and Level 2.

Level 0 shows the overall interaction between the user and the system. The User inputs the Query to the model and gets a Summary and Key Takeaways as the output. The model scraps the data from the academic websites for generating the output.

Level 1 is further divided into 2 sublevels focusing in more detail about the individual actors separately, giving more information about the data flow in the system.

Level 2 focuses on each module, starting from the User input which is given to the model for web-scrapping. This Scraped data is used to generate Summaries and Key takeaways which are displayed to the user as the final output.

Chapter 5

Design

5.1 Architecture Design

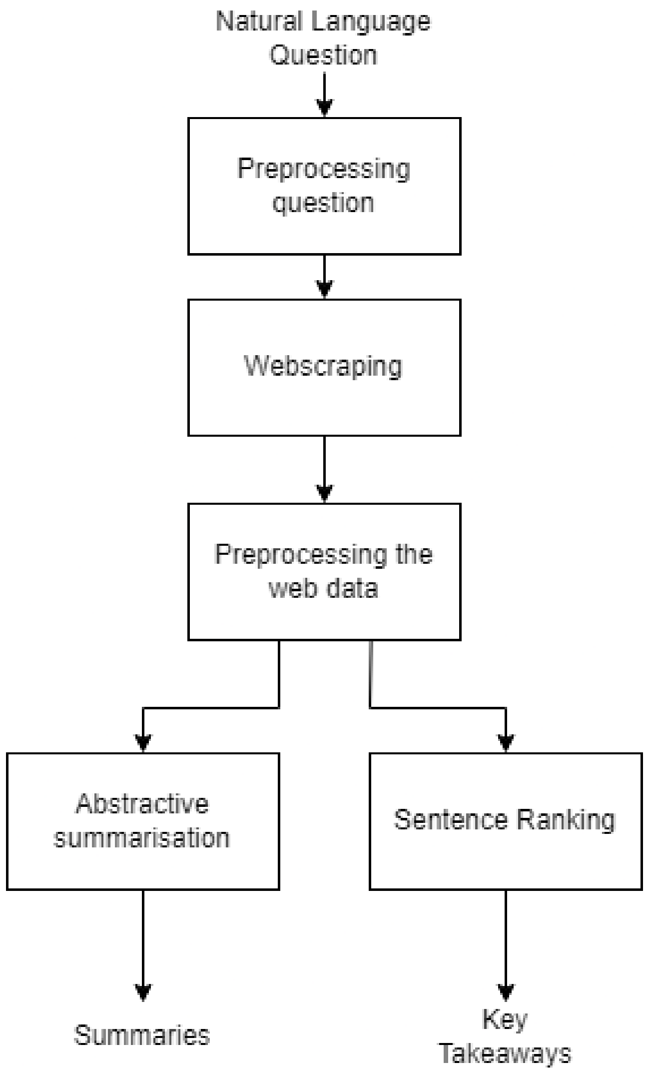

Figure 5.1: Block Diagram for AI-Powered Academic Research Assistant

5.1.1 Preprocessing the question

Tokenization: Split the question into individual words or tokens. This can be done using libraries like NLTK.
Stopword Removal: Remove common stop words (e.g., ”the,” ”is,” ”in”) from the list of tokens. These words often don’t contribute significantly to the meaning and are unlikely to be keywords.

5.1.2 Web scraping

The sample data was preprocessed by: For web scraping the research papers, the requests_html library is used in Python. It has an HTMLSession object that opens a browser window (Chromium) for scraping. After creating a session, a get request can be sent to the desired link using the get function in the HTMLSession class. It calls the request function to create a request for a given URL which then returns a response. The render function is used to load the Javascript on the website before scraping. After the page is loaded, we use the find function to extract the desired data.

5.1.3 Preprocessing the web data

The data that we extracted from the web may not be in a format that is suitable for further processes. In this step, we will need to clean the data by removing any unwanted characters, correcting any errors, and converting the data into a consistent format(Ex. UTF-8)

5.1.4 Abstractive Summarisation

A Transformer is used for abstractive summarisation, specifically the LongT5 model. The LongT5 model works by first encoding the input text into a sequence of tokens. These tokens are then passed through the Transformer encoder, which learns to identify the relationships between the tokens. The encoded tokens are then passed through the Transformer decoder, which generates the output summary.

5.1.5 Sentence Ranking

Find out key takeaways from all the papers. These key takeaways give us an idea about the essential sentences from the papers. TextRank turns out to be well suited for this type of application, since it allows for a ranking over text units that is recursively computed based on information drawn from the entire text.

[5]Formally, given two sentences S_i and S_j, with a sentence being represented by the set of N_i words that appear in the sentence: $S_i = w_1^i, w_2^i, ..., w_{N_i}^i$,

$$Similarity(S_i, S_j) = \frac{|w_k|w_k \in S_i \& w_k \in S_j|}{log(|S_i|) + log(|S_j|)} \tag{5.1}$$

The score is calculated for all sentences using,

$$Score(s) = d + sum(Score(t) * Similarity(s,t)) \tag{5.2}$$

Where,
$Score(s)$ is the score of sentence s,
d is a damping factor, typically set to 0.85,
$Score(t)$ is the score of sentence t,
$Similarity(s,t)$ is the similarity between sentences s and t.

5.2 Performance Evaluation Parameters

5.2.1 Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score

The ROUGE score is used to evaluate the quality of machine translation outputs. It is calculated by comparing the machine translation output to a set of human-generated reference summaries. A higher ROUGE score indicates that the machine translation output contains more information from the reference summaries.

5.2.2 GLUE Benchmark

GLUE benchmark is a natural language processing (NLP) benchmark that consists of a collection of nine tasks, including text summarization, question answering, and natural language inference. A higher score on the GLUE benchmark indicates that the model is better at performing a variety of NLP tasks.

5.2.3 SQuAD Benchmark

SQuAD benchmark is a question-answering benchmark that consists of a collection of over 100,000 questions that are answered by paragraphs from Wikipedia. A higher score on the SQuAD benchmark indicates that the model is better at answering text questions.

Chapter 6

Implementation

6.1 Algorithm

Transformer Architecture

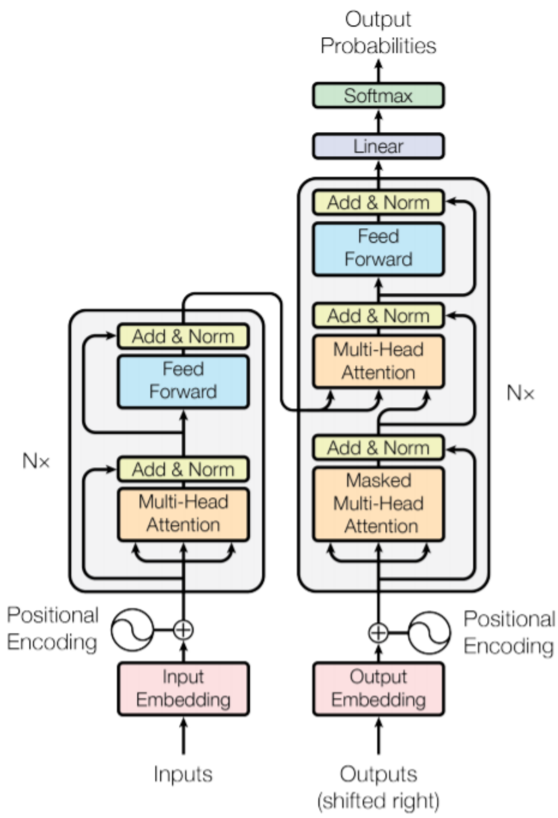

Figure 6.1: Transformer architecture.[4]

Encoders in Transformers, also known as the transformer encoder, are a key component of the transformer encoder-decoder architecture. They are respon-

sible for analyzing and representing the input sequence in a way the model can understand. The encoder processes the input sequence and produces a continuous representation, or embedding, of the input. These embeddings are then passed to the decoder to generate the output sequence. Each layer of the encoder contains a self-attention mechanism that allows the model to weigh the importance of different input sequence parts by calculating the embeddings' dot product. This mechanism is also known as multi-head attention.
The decoder also typically consists of multiple layers, including a self-attention mechanism and a feed-forward network. The decoder uses the embeddings produced by the encoder and its internal states to generate the output sequence.

Attention Mechanism of LongT5

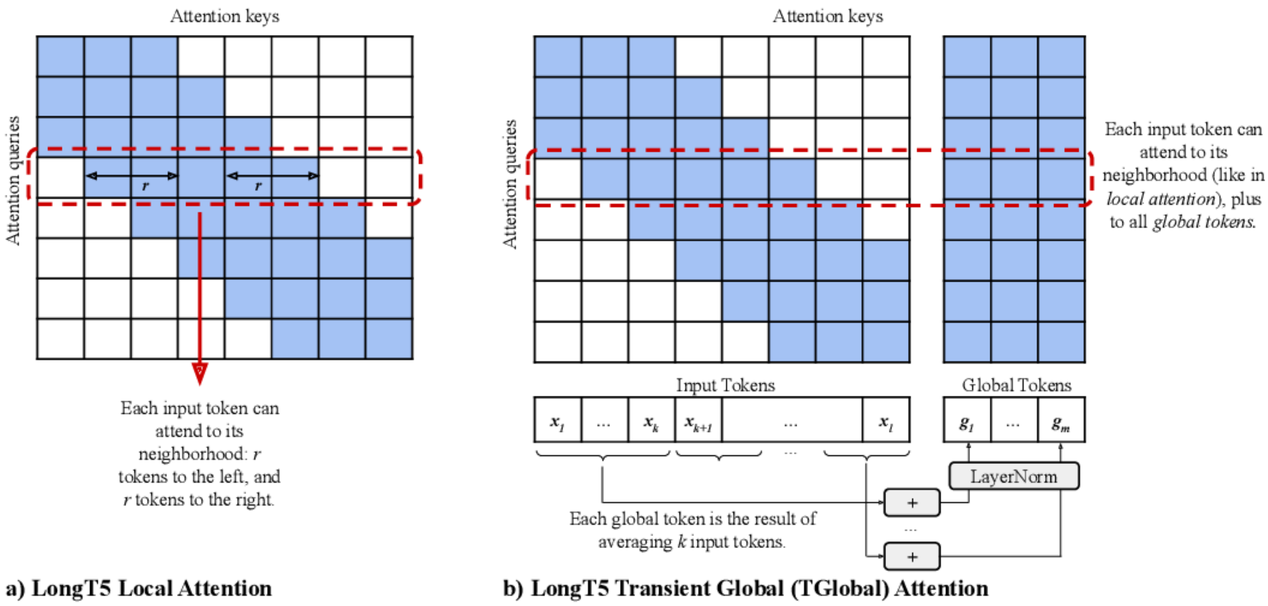

Figure 6.2: Illustration of the two attention mechanisms experimented with in LongT5.[1]

For Local Attention, the sparse sliding-window local attention operation allows a given token to attend only r tokens to the left and right of it (with r=127 by default). Local Attention does not introduce any new parameters to the model. The complexity of the mechanism is linear in input sequence length l: $O(l * r)$. Transient Global Attention is an extension of Local Attention. It, furthermore, allows each input token to interact with all other tokens in the layer.

This is achieved via splitting an input sequence into blocks of a fixed length k (with a default k=16). Then, a global token for such a block is obtained via summing and normalizing the embeddings of every token in the block. Thanks to this, the attention allows each token to attend to both nearby tokens like in Local attention, and also every global token like in the case of standard global attention (transient represents the fact the global tokens are constructed dynamically within each attention operation). As a consequence, TGlobal attention introduces a few new parameters — global relative position biases and a layer normalization for global token's embedding. The complexity of this mechanism is $O(l(r+l/k))$.

6.2 Experimentation

The experimental results of preprocessing of question and web scraping of a single research paper is shown in Fig 6.3 and Fig 6.4, respectively.

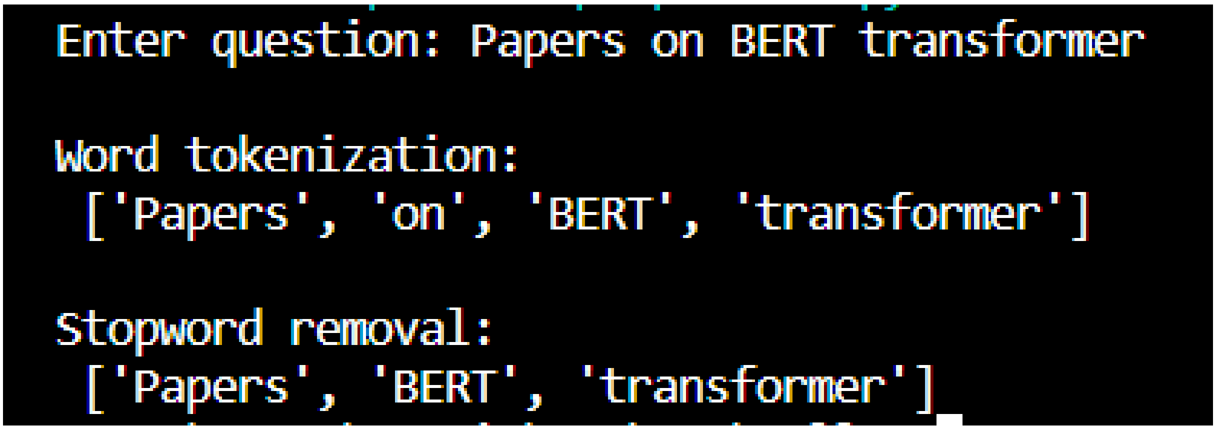

Figure 6.3: Preprocessing

```
Sentence-T5: Scalable Sentence Encoders
from Pre-trained Text-to-Text Models
Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma,
Keith B. Hall, Daniel Cer, Yinfei Yang
Google Research
Mountain View, CA
Abstract
We provide the first exploration of text-to-text transformers (T5) sentence embeddings. Sentence embeddings are broadly useful for
language processing tasks. While T5 achieves impressive performance on language tasks cast as sequence-to-sequence mapping problems, it
is unclear how to produce sentence embeddings from encoder-decoder models. We investigate three methods for extracting T5 sentence
embeddings: two utilize only the T5 encoder and one uses the full T5 encoder-decoder model. Our encoder-only models outperforms
BERT-based sentence embeddings on both transfer tasks and semantic textual similarity (STS). Our encoder-decoder method achieves
further improvement on STS. Scaling up T5 from millions to billions of parameters is found to produce consistent improvements on
downstream tasks. Finally, we introduce a two-stage contrastive learning approach that achieves a new state-of-art on STS using
sentence embeddings, outperforming both Sentence BERT (Reimers and Gurevych, 2019) and SimCSE (Gao et al., 2021).
††Sentence T5 preprint.†
1 Introduction
Sentence embeddings providing compact meaning representations that are broadly useful for a variety of language processing tasks
include classification, question-answering, semantic retrieval, bitext mining, and semantic similarity tasks. Sentence embedding models
have been trained using a variety of methods including: supervised tasks such as natural language inference Conneau et al. (2017); Gao
et al. (2021) or with semi-structured data such as question-answer pairs Cer et al. (2018); translation pairs Yang et al. (2020a); Feng
et al. (2020); and adjacent sentence pairs Kiros et al. (2015); Logeswaran and Lee (2018). Recent work has shown that scaling up model
parameters and leveraging pre-trained models Devlin et al. (2019); Liu et al. (2019) are two effective approaches to improve
performance Reimers and Gurevych (2019, 2020); Yang et al. (2020b); Gao et al. (2021).
Figure 1: Scaling up our ST5 model size improves performance on SentEval (left) and STS (right).
Transfer
STS-B
ST5-EncDec (3B params)
90.46
84.94
ST5-Enc (11B params)
91.48
84.59
SimCSE-RoBERTa (large) Gao et al. (2021)
90.23111SimCSE-RoBERTa achieves the best performance on transfer tasks by adding an additional masked language model loss during
training while ST5 and other models don't.1
83.76
```

Figure 6.4: Webscraping

Chapter 7

Conclusions

7.1 Conclusion

In conclusion, The AI-powered academic research assistant with web scraping capabilities has the potential to be an indispensable tool for researchers, significantly reducing the time and effort spent on information retrieval, data analysis, and literature review. By leveraging AI/ML techniques and web scraping, this research assistant aims to empower researchers and contribute to advancements in various fields of study.

7.2 Future Scope

The AI-powered academic research assistant with web scraping capabilities holds great potential for further development and integration into the research community. As technology advances, we can expect this tool to become increasingly sophisticated and user-friendly. It will likely evolve to support a broader range of research tasks, such as automated data synthesis, and even assisting in hypothesis formulation. Additionally, improvements in natural language processing and machine learning techniques will enhance its ability to understand and interpret complex research questions. Furthermore, as ethical and privacy concerns related to web scraping are addressed and regulated, this assistant can become a trusted and widely adopted resource for researchers across diverse domains, fostering more efficient and impactful scientific discoveries.

References

[1] Guo, Mandy & Ainslie, Joshua & Uthus, David & Ontanon, Santiago & Ni, Jianmo & Sung, Yun-Hsuan & Yang, Yinfei. “LongT5: Efficient Text-To-Text Transformer for Long Sequences.” *arXiv:2112.07916 [cs.CL]* 2021.

[2] Zhang, Jingqing & Zhao, Yao & Saleh, Mohammad & Liu, Peter. “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization.” *arXiv:1912.08777 [cs.CL]* 2020.

[3] Subramanian, Sandeep & Li, Raymond & Pilault, Jonathan & Pal, Christopher. “On Extractive and Abstractive Neural Document Summarization with Transformer Language Models.”, *arXiv:1909.03186 [cs.CL]* 2020.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need.” *In Advances in Neural Information Processing Systems, Volume 30 arXiv:1706.03762v7 [cs.CL]* 2023

[5] Mihalcea, Rada & Tarau, Paul. ”TextRank: Bringing Order into Texts”, *Association for Computational Linguistics* (2004)

Acknowledgements

We are thankful to a number of individuals who have contributed to our final year project and without their help; it would not have been possible. First of all, we would like to express our gratitude to Prof. Pradnya Sawant our project guide, for their prompt suggestions, guidance, and encouragement over the course of our entire project.

We sincerely appreciate all our project coordinators, for their tremendous assistance to our project. We also appreciate the support of the faculty in our department.

We express our sincere gratitude to our respected Director Bro. Shantilal Kujur, our Principal Dr. Sincy George, and our Head of Department (CMPN) Dr. Kavita Sonawane for providing the facilities, encouragement as well as a conducive environment for learning.

Lastly, we would like to thank our parents and friends who played a crucial role in keeping us motivated throughout with their constant nurture and support.

Sincerely,

Daniel Ferreira

Prem Tatkari

Divyesh Mistry

Kyran Almeida