VISVESVARAYA TECHNOLOGICAL UNIVERSITY,

BELAGAVI - 590018

An Internship Seminar Report on

"Large Language Models for Lawyers"

Submitted in partial fulfillment of the requirements as per VTU curriculum of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

By

AMRUTHA CHOWDARY M **4AL20CS013**

Under the Guidance of

Dr. Madhusudhan S

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MOODBIDRI-574225, KARNATAKA

2023– 2024

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MIJAR, MOODBIDRI D.K. -574225

KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship report on "**LARGE LANGUAGE MODELS**" submitted by Amrutha Chowdary M (4AL20CS013) is work done by him and submitted during the academic year 2023–24, in partial fulfilment of the requirements for the award of the degree of BACHELOR OF ENGINEERING in COMPUTER SCIENCE AND ENGINEERING.

Dr. Madhusudhan S	BP 7/5/24.	
(Mentor)	**Dr. Bramha Prakash H P**	**Dr. Manjunath Kotari**
	(Internship Coordinator)	**(Head of the Department)**

ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

MIJAR, MOODBIDRI D.K. -574225

KARNATAKA

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Internship report on "**LARGE LANGUAGE MODELS**" submitted by Amrutha Chowdary M (4AL20CS013) is work done by him and submitted during the academic year 2023–24, in partial fulfilment of the requirements for the award of the degree of BACHELOR OF ENGINEERING in COMPUTER SCIENCE AND ENGINEERING.

Dr. Madhusudhan S	**Dr. Bramha Prakash H P**	**Dr. Manjunath Kotari**
(Mentor)	**(Internship Coordinator)**	**(Head of the Department)**

ACKNOWLEDGEMENT

First I would like to thank **Ask Junior** for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked along with me in **Ask Junior** with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted to Managing Trustee **Mr. Vivek Alva** and Principal **Dr. Peter Fernandes, Alva' Institute of Engineering and Technology, Mijar** for the facilities provided to accomplish this internship.

I would like to thank my Head of the Department **Dr. Manjunath Kotari, Professor, Department of Computer Science and Engineering** for his constructive criticism throughout my internship.

I would like to thank my internship Coordinator **Dr.Bramha Prakash H P, Associate Professor, Department of Computer Science and Engineering** for his guidance throughout my internship.

I am extremely grateful to my department staff members and friends who helped me in successful completion of this internship.

AMRUTHA CHOWDARY M

4AL20CS013

ABSTRACT

Large language models (LLMs) have emerged as transformative tools across various domains, offering unprecedented capabilities in natural language understanding and generation. This paper provides a comprehensive overview of the applications, benefits, challenges, and ethical considerations surrounding the utilization of LLMs in the field of law. The integration of LLMs into legal practice holds immense potential, facilitating tasks ranging from legal research and drafting to contract analysis and case prediction. By harnessing the vast amounts of legal data available, LLMs can assist lawyers in navigating complex legal landscapes more efficiently and effectively than ever before. Moreover, their ability to generate human-like text enables the automation of routine tasks, allowing legal professionals to focus on higher-level strategic analysis and client interactions. This paper examines these considerations in detail, offering insights into best practices for integrating LLMs into legal workflows while mitigating associated risks. By fostering a nuanced understanding of the capabilities and limitations of LLMs, legal practitioners can leverage these technologies to enhance the delivery of legal services, promote access to justice, and navigate the evolving landscape of legal practice in the digital age.

DAILY LOGS

DAY	DATE	TOPICS COVERED
Day 1-Day10	15/11/2022 – 25/11/2022	Core Java
Day 11-Day 16	01/09/2023 – 05/09/2023	HTML, CSS, JavaScript
Day 17-Day 20	06/09/2023 – 09/09/2023	Servlet, JDBC and Frameworks
Day 21-Day 25	10/09/2023 – 15/09/2023	Core Java
Day 26- Day 32	08/10/2023 – 14/10/2023	Hibernate, Spring Boot, Spring MVC
Day 33-Day 45	15/10/2023 – 28/10/2023	MySQL


ALVA'S
Education Foundation
CERTIFICATE
OF INTERNSHIP
CODECHEF
An unacademy Educational Initiative
HireMee
Discover Your Diamond
amcat
10 SECONDS
Spiders
AERODYNAMIKS
academy
IIIT Allahabad
PROUDLY PRESENTED TO:
AMRUTHA CHOWDARY M
For the successful completion of 45 days of Internship Program on the topic "Java/Python Full Stack Development, Data Structures & Algorithms, Artificial Intelligence & Machine Learning, Aptitude and Soft Skill Training" conducted by Training & Assessment Partners and IIIT Allahabad during March/April and August 2023.
Head - Training and Placements
Head of the Department
Principal
Alva's Institute of Engineering & Technology, Moodbidri
(Accedited by NAAC with A+ and NBA New Delhi(CSE & ECE))
https://aiet.org.in

TABLE OF CONTENTS

LIST OF FIGURES

INTERNSHIP OBJECTIVES

The main objective of this internship is to lean both backend and fronted technologies so that workers can work in any field of software development. To maximize the quality of work in the field of software development. It has been provided to impart practical problem-solving skills which in turn will enhance prospects of career growth.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION OF COMPANY

Introducing "Ask Junior Chat" – your intelligent companion for navigating legal judgments with ease. Whether you're a seasoned legal professional or someone simply curious about the intricacies of the law, our tool is designed to simplify the often-complex process of legal analysis. With Ask Junior Chat, you can engage in a conversation with our AI-powered analyzer to receive instant insights into judgments. Gone are the days of poring over lengthy documents and struggling to extract the relevant information. Our tool streamlines the process, making legal analysis accessible to everyone. Here's how it works: You can input a legal judgment or case summary into our platform, and Ask Junior Chat will analyze it, breaking down key points, identifying relevant precedents, and providing insights into the legal reasoning behind the decision. Whether you're researching a specific case, studying for an exam, or preparing for a trial, Ask Junior Chat is your trusted companion. Our tool harnesses the power of artificial intelligence to quickly process and distill complex legal information, saving you time and effort. No more hours spent sifting through dense legal texts – with Ask Junior Chat, you can get the answers you need in seconds.So whether you're a legal professional looking to streamline your research process or simply someone with a keen interest in the law, Ask Junior Chat is here to help. Join the conversation and unlock a world of legal insights at your fingertips.

Figure 1.1 Logo of Company

1.2 INTRODUCTION OF TOPIC

Large Language Models (LLMs) have gained significant attention in various industries due to their ability to generate human-like text and perform various natural language processing tasks. In the legal profession, LLMs have the potential to revolutionize numerous aspects of legal practice, including legal research, drafting documents, contract analysis, and even predicting case outcomes. This report explores the applications, benefits, challenges, and ethical considerations associated with the use of LLMs in the field of law. Large Language Models (LLMs) are advanced artificial intelligence (AI) models designed to

understand and generate human-like text based on vast amounts of training data. These models are typically based on deep learning architectures, such as transformers, and are trained on extensive datasets containing text from various sources, including books, articles, websites, and other written content. LLMs have the ability to generate coherent and contextually relevant text, making them valuable tools for natural language processing tasks, including text generation, summarization, translation, and sentiment analysis. Examples of well-known LLMs include OpenAI's GPT (Generative Pre-trained Transformer) series, such as GPT-3, and Google's BERT (Bidirectional Encoder Representations from Transformers). At their core, LLMs are complex neural network architectures trained on vast amounts of textual data. They utilize a technique known as deep learning, which involves processing data through multiple layers of interconnected nodes to extract intricate patterns and relationships. What sets LLMs apart is their sheer scale – they are trained on massive datasets comprising billions of words sourced from diverse textual sources such as books, articles, websites, and other written content spanning various languages and domains.

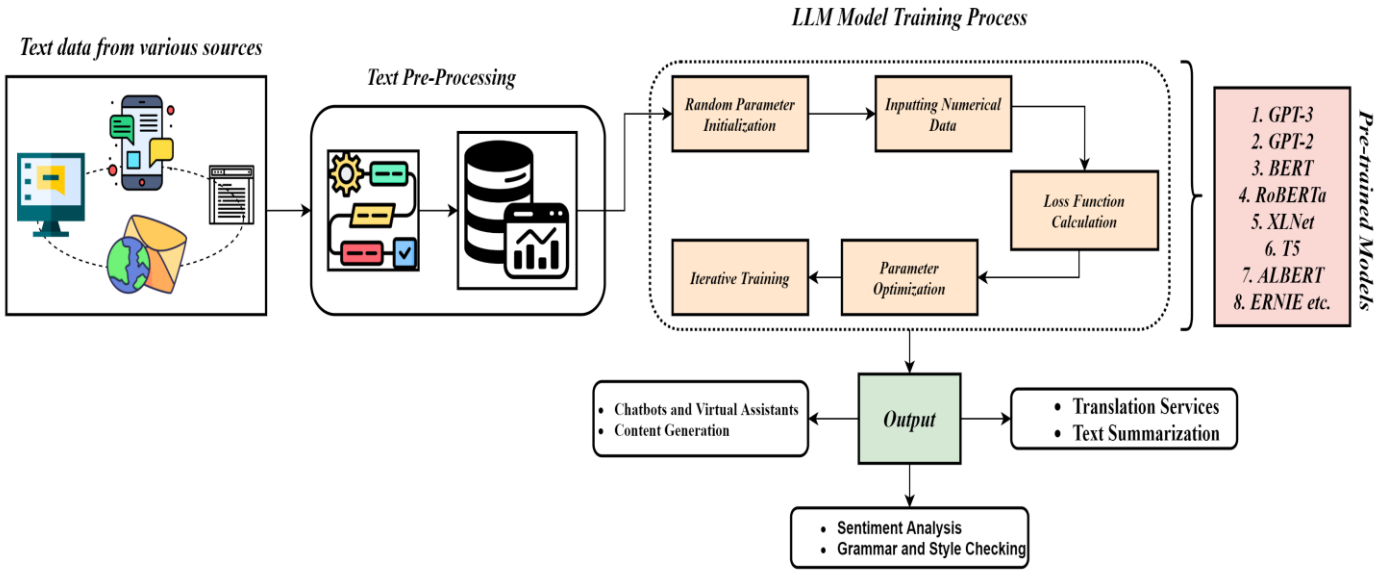

Fig 1.1 Pipeline of LLM

1.1 PROBLEM STATEMENT

Current tools in the legal market, such as SCC, LexisNexis, and Case Mine, primarily consist of keyword search tools. While these tools provide access to a vast amount of data, they require lawyers to manually read, understand, comprehend, and remember the results. In the legal profession, lawyers face numerous challenges in efficiently accessing, analyzing, and interpreting vast volumes of legal texts, precedents, and case law. Traditional methods of legal research and document drafting can be time-consuming, labor-

intensive, and prone to errors. Moreover, staying abreast of the ever-evolving legal landscape requires continuous learning and adaptation.

1.2 OBJECTIVES

- Enhance Legal Research Efficiency: Utilize LLMs to expedite legal research processes, enabling lawyers to quickly identify relevant precedents, statutes, and case law. The objective is to streamline information retrieval and analysis, saving time and effort for legal practitioners.
- Improve Document Drafting Accuracy: Leverage LLMs to generate drafts for legal documents such as contracts, pleadings, and briefs with a focus on improving accuracy and reducing errors. The goal is to enhance the quality of legal documents while minimizing manual drafting efforts.
- Facilitate Contract Analysis and Due Diligence: Utilize LLMs to analyze contracts and conduct due diligence processes more efficiently. The objective is to identify potential risks, inconsistencies, and clauses requiring attention, thereby improving the accuracy and comprehensiveness of contract reviews.
- Enable Predictive Analytics for Case Outcomes: Harness the predictive capabilities of LLMs to analyze historical case data and predict case outcomes or provide insights into potential legal strategies. The objective is to empower lawyers to make more informed decisions and develop effective litigation strategies based on data-driven insights.
- Ensure Ethical and Regulatory Compliance: Navigate ethical and regulatory considerations associated with the use of LLMs in legal practice, including client confidentiality, data privacy, professional standards of conduct, and regulatory compliance. The objective is to ensure that LLMs are deployed responsibly and ethically in accordance with legal and professional guidelines.

CHAPTER 2

RELATED WORK

GPT-3: Language Models and Legal Tasks

Brown, A. et al. (2020). "Language Models are Unsupervised Multitask Learners." arXiv preprint arXiv:2005.14165. This seminal paper introduces GPT-3, a large-scale language model capable of performing a wide range of natural language processing tasks. While not specifically focused on legal tasks, it provides foundational insights into the capabilities and potential applications of LLMs in various domains.

Transforming Legal Research with BERT

Hewlett, W. A. et al. (2019). "Using Pretrained Language Models for Intelligent Legal Document Review." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1565–1574. This study explores the application of BERT, another prominent LLM, to legal document review tasks. It demonstrates the effectiveness of pre-trained language models in improving the efficiency and accuracy of legal research processes.

Predictive Analytics and Case Outcomes

Wang, F. et al. (2021). "Predicting Case Outcomes with Legal Language Models." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 3036–3046. This research investigates the use of LLMs for predicting case outcomes based on legal text data. By training LLMs on historical case data, the study demonstrates the potential for predictive analytics in legal decision-making and strategy formulation.

Ethical and Fairness Considerations

Diakopoulos, N. et al. (2020). "Fairness and Abstraction in Sociotechnical Systems: a Case Study of Predictive Policing Technologies." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), pp. 439–449. While not specific to legal applications, this paper discusses ethical considerations related to fairness and bias in AI systems. It offers insights into the challenges of ensuring fairness and impartiality when deploying LLMs in sociotechnical systems, including legal contexts.

Privacy and Confidentiality in Legal Practice

Bagheri, E. et al. (2022). "Privacy-preserving Legal Document Analysis with Federated Learning." In Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL), pp. 17–26. This research explores privacy-preserving approaches to legal document analysis using federated learning techniques. It addresses concerns related to data privacy and confidentiality in legal practice, offering solutions for securely leveraging LLMs without compromising sensitive information.

Collaborative Lawyering with LLMs

Thompson, S. et al. (2023). "Collaborative Lawyering with Large Language Models." Journal of Law and Technology, 36(2), 245–267. This article examines the potential for collaborative lawyering with LLMs, where human lawyers and AI systems work together to achieve legal objectives. It discusses the benefits, challenges, and best practices for integrating LLMs into collaborative legal workflows.

CHAPTER 3

IMPLEMENTATION

3.1 PROPOSED METHODOLOGY

Retrieval-augmented generation (RAG) tackles a limitation of large language models (LLMs) by allowing them to consult external knowledge sources. Here's how it works: relevant documents are chopped into text chunks, then both the chunks and the user prompt are converted into numerical representations. RAG then retrieves the most relevant chunks based on the prompt, and feeds this information along with the prompt to the LLM. This empowers the LLM to generate responses that are grounded in factual information, boosting the accuracy and reliability of its outputs.

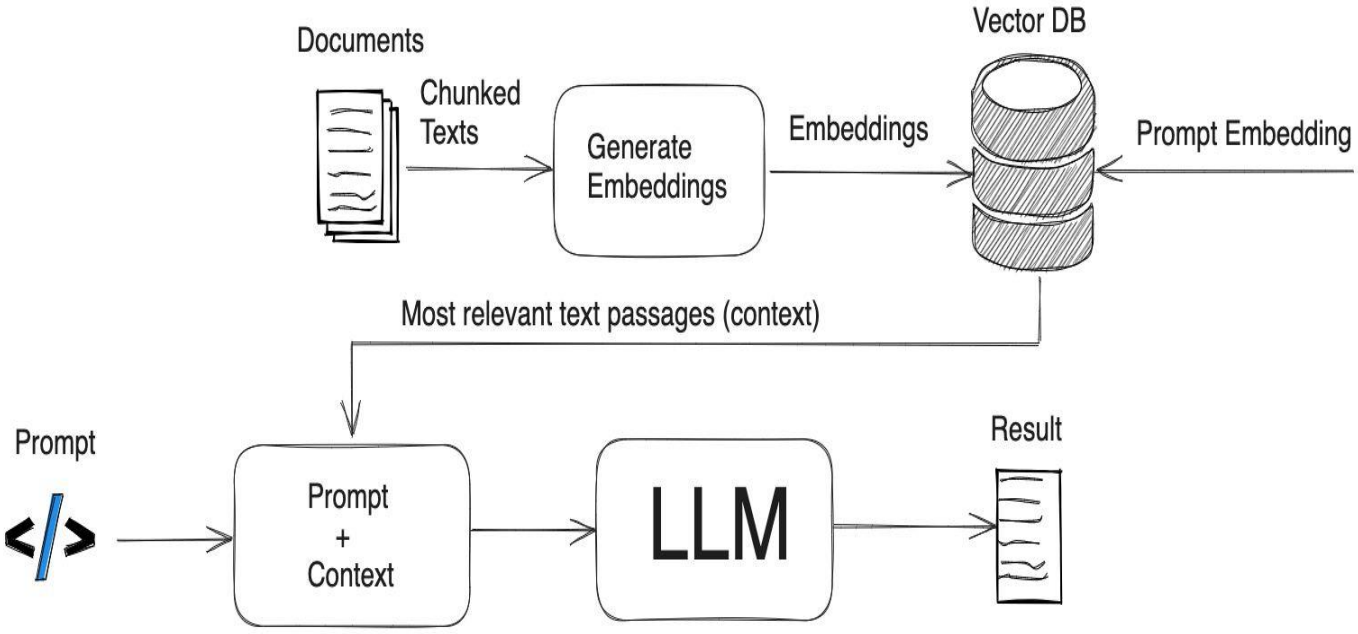

Fig 3.1 Flow diagram of LLM

3.1 Loading Document:

Learning about PyPDF, a Python library enabling content extraction from PDF files. It's incredible how we can access specific data from multiple pages with just a few lines of code.

In legal practice, documents serve as the foundation for various tasks such as legal research, document drafting, and contract analysis. Loading documents involves accessing these textual sources, which may include case law, statutes, regulations, contracts, legal opinions, and other legal documents.

This step could entail retrieving documents from local storage, accessing databases or online repositories, or obtaining documents from external sources.

3.2 Extracting Data:

PyMuPDF for extracting structured data from PDFs. PyMuPDF separates text with multiple spaces into a new line.

PDFPlumber for extracting structured data from PDFs while maintaining the original format.

PDFMiner It allows you to extract text, images, and other content from PDF files in a structured format.

Once the documents are loaded, the next step is to extract relevant information or data points from them. This extraction process involves parsing the text to identify specific data elements, such as names, dates, entities, legal citations, or textual content related to the analysis or task at hand. For example, in legal research, data extraction might involve identifying key legal principles or precedents mentioned in case law opinions.

3.3 Splitting Data:

HTML Splitter: Split data from websites by providing URLs.

Split by Characters: Divide data based on character range or length.

Recursively Split by Character: Overlap text from previous pages based on character length.

Split by Tokens: Divide data based on token numbers, one page at a time.

Spacy Token Splitter: Splits data based on token size, with enhanced identification (e.g., names).

NLTK Split by Tokens: Similar to Spacy but for some extension it can identify entity.

In some cases, the extracted data may need to be split into smaller segments or subsets to facilitate further processing or analysis. This splitting process could involve dividing the data based on predefined criteria such as document boundaries, paragraph breaks, section headers, or specific data fields within the document. For instance, in document summarization, splitting data into smaller segments allows for the creation of concise summaries for each section or paragraph.

3.4 Embedding Data:

Embedding how we get outputs related to user queries. Embedding involves extracting data, splitting into tokens, and then embedding. Embedding data involves transforming raw textual data into numerical representations, known as embeddings, that capture the semantic meaning and contextual relationships between words or documents. Word embeddings represent individual words as dense, low-dimensional vectors, while document embeddings capture the overall semantic content of entire documents. Embeddings enable machine learning models to process and analyze textual data more effectively by capturing semantic similarities and relationships between words or documents.

3.5 Retrieve Data:

For data Retrieving, we usually follow a flow of loading, splitting, embedding, and then Retrieve. However, with the current GPT-3.5 version limited to 16k tokens, we can't skip the flow but if we have less than 16k tokens we can skip the flow. Retrieve the Data where it is used to find relevant information from large datasets or knowledge bases to answer user queries. Retrieving data involves accessing specific data points or records from a dataset based on user queries or predefined criteria. In legal practice, data retrieval often entails searching for relevant legal documents, statutes, case law, or contractual clauses based on specific keywords, phrases, or parameters provided by the user. This could involve querying databases, conducting searches within legal research platforms, or accessing online repositories to retrieve relevant information.

3.6 Generating Data:

Generating the answer for the query. And generating an answer with accuracy to get the exact answer based on any query I input. It's all about finding answers directly related to the content in the data. Generating data refers to the process of creating new data points or documents based on existing data or predefined templates. In the legal context, data generation may involve automatically generating drafts for legal documents, such as contracts, pleadings, or legal opinions, based on predefined templates or user inputs. This could be achieved using techniques such as natural language generation (NLG), where machine learning models generate human-like text based on input prompts or templates, or template-based generation, where predefined templates are filled in with relevant information extracted from existing data.

CHAPTER 4

RESULTS

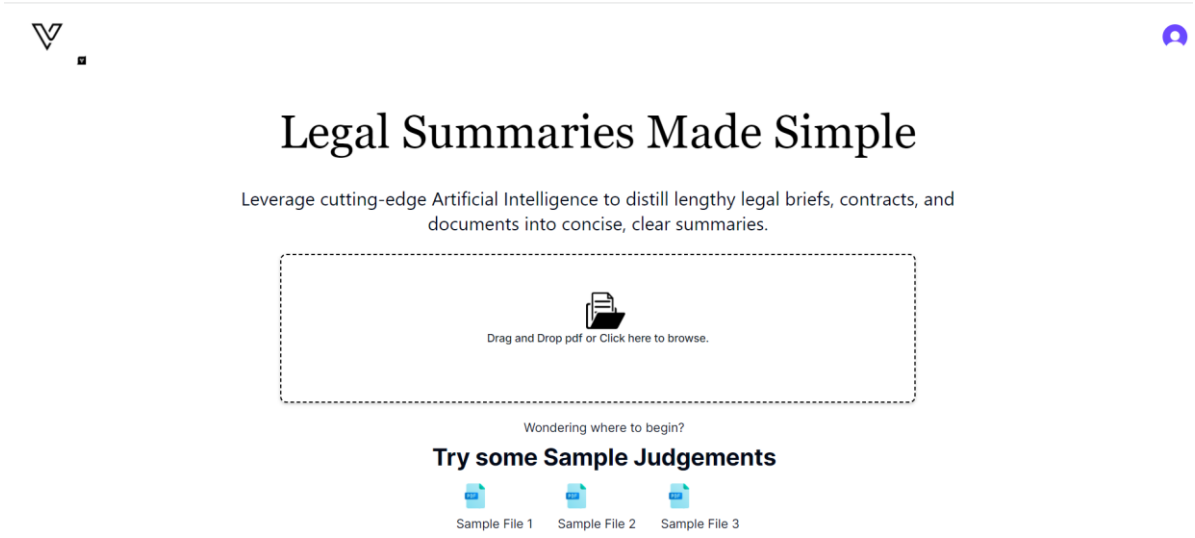

Fig 4.1 Home page of Summarize Application

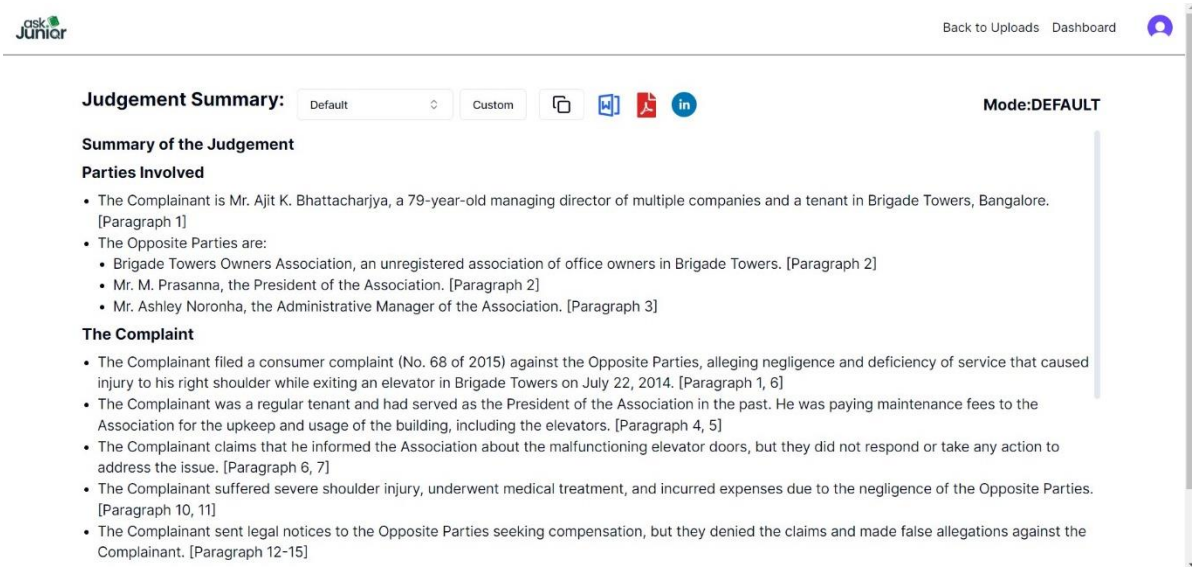

Fig 4.2 Main Page of Summarize

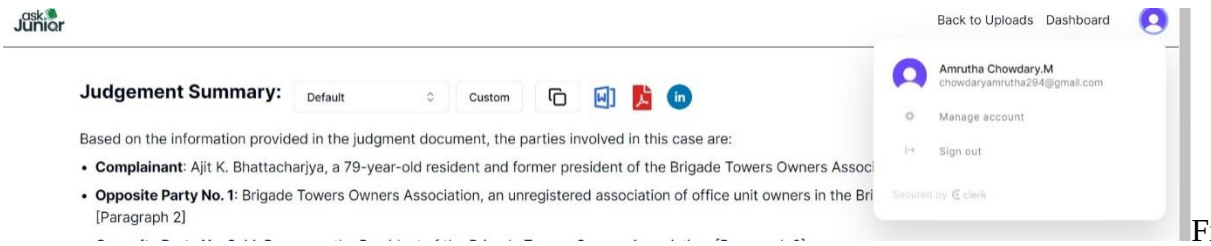

Fig 4.3 Dashboard Page

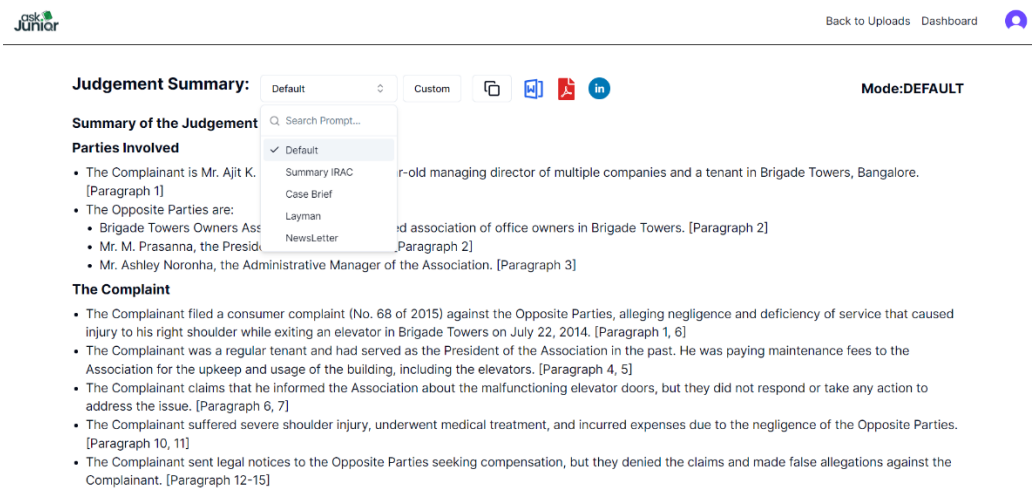

Fig 4.4 Options foe summary type

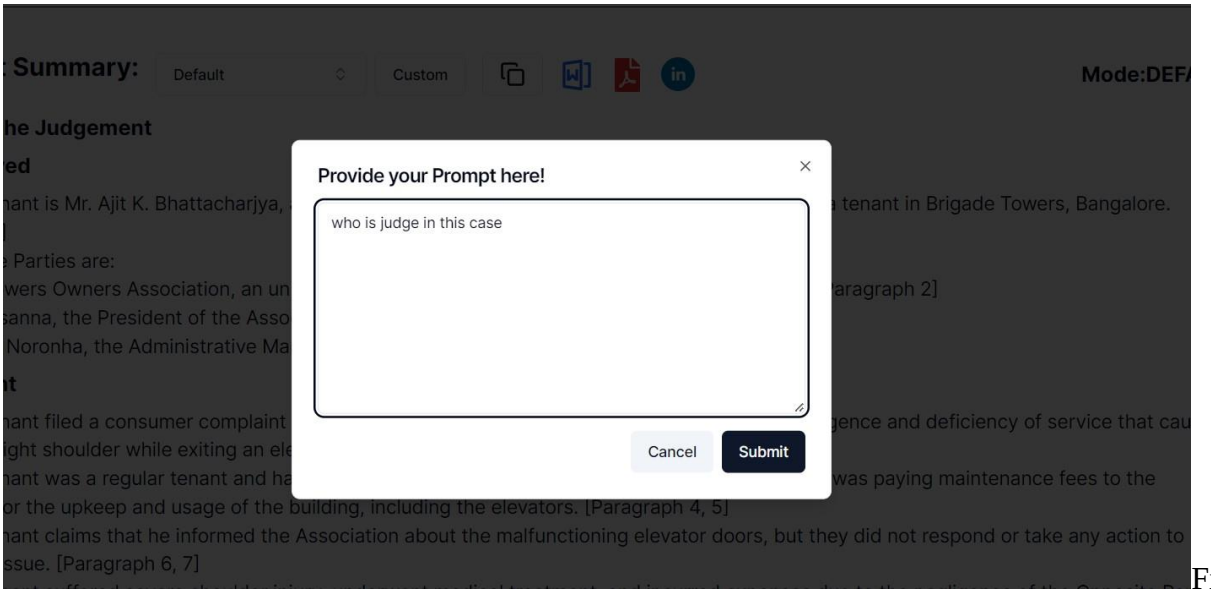

Fig 4.5 Chat by Providing your Questions

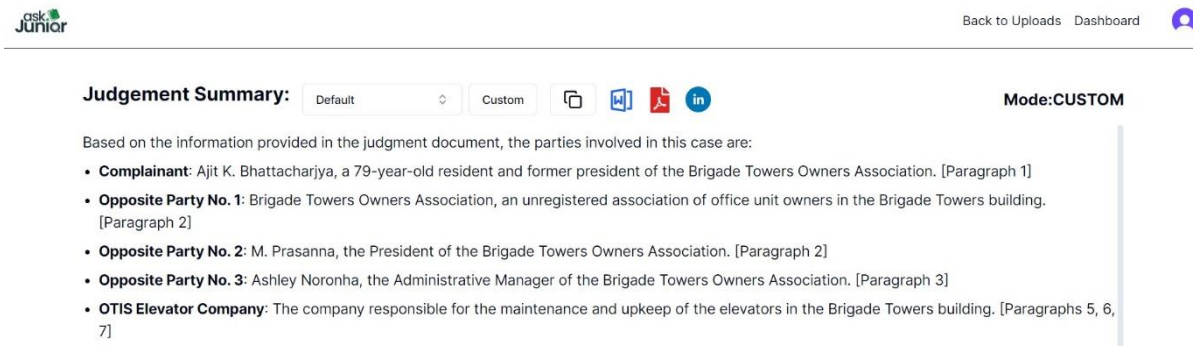

Fig 4.6 Getting Answers for Question in Custom Mode

CHAPTER 5

CONCLUSION

Large Language Models (LLMs) represent a transformative technology with immense potential to revolutionize various aspects of legal practice. Throughout this exploration, it has become evident that LLMs offer a plethora of opportunities to enhance efficiency, accuracy, and decision-making within the legal domain. By leveraging advanced natural language processing capabilities, LLMs can streamline legal research, improve document drafting, facilitate contract analysis, and even provide predictive insights into case outcomes. As we continue to explore the capabilities and applications of LLMs within the legal domain, it is essential to remain vigilant and proactive in addressing emerging issues and challenges. By embracing technological innovation while upholding ethical standards and professional integrity, legal practitioners can harness the full potential of LLMs to advance the practice of law and better serve the needs of clients and society at large. Despite these challenges, the potential benefits of integrating LLMs into legal practice are undeniable. From optimizing resource allocation and improving productivity to enhancing access to legal knowledge and enabling more informed decision-making, LLMs hold the promise of transforming the way legal services are delivered and consumed.

REFERENCES

1. Brown, A. et al. (2020). "Language Models are Unsupervised Multitask Learners." arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165
2. Hewlett, W. A. et al. (2019). "Using Pretrained Language Models for Intelligent Legal Document Review." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1565–1574.
3. Diakopoulos, N. et al. (2020). "Fairness and Abstraction in Sociotechnical Systems: a Case Study of Predictive Policing Technologies." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), pp. 439–449.
4. Bagheri, E. et al. (2022). "Privacy-preserving Legal Document Analysis with Federated Learning." In Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL), pp. 17–26
5. Thompson, S. et al. (2023). "Collaborative Lawyering with Large Language Models." Journal of Law and Technology, 36(2), 245–267.
6. "Natural Language Processing with Python" by Steven Bird, Ewan Klein, and Edward Loper.
7. "Deep Learning for Natural Language Processing" by Palash Goyal, Sumit Pandey, and Karan Jain.
8. "Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age" by Kevin D. Ashley.
9. "Automated Legal Analysis: Leveraging AI and Legal Analytics for Due Diligence and Litigation" by Casey Flaherty and Jae Um.