AIVO: Advanced Intelligent Virtual Orator

MAIN PROJECT REPORT

Submitted by

DIANA LIZ KURIAKOSE (CHN21CS046)
MOHAMMED HISHAM (CHN21CS085)
MARIA JOSHY (CHN21CS080)
NANDANA VINOD (CHN21CS094)

to

APJ Abdul Kalam Technological University
in partial fulfillment of the requirements for the award of
B.Tech Degree in Computer Science and Engineering

Department of Computer Engineering
College of Engineering Chengannur

2025

College of Engineering Chengannur
Department of Computer Engineering
2024-25

C E R T I F I C A T E

This is to certify that, this report titled ***AIVO: Advanced Intelligent Virtual Orator*** is a bonafide record of the work done by **DIANA LIZ KURIAKOSE (CHN21CS046), MOHAMMED HISHAM (CHN21CS085), NANDANA VINOD (CHN21CS094), MARIA JOSHY (CHN21CS080)**, Eighth Semester B.Tech. Computer Science & Engineering students, for the course work in **CSD416 Project Phase II**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, B. Tech. Computer Science & Engineering of **APJ Abdul Kalam Technological University**.

Sri. Gopakumar G
(Head of the Department)
Associate Professor
Dept. of Computer Engineering
College of Engineering Chengannur

Smt.Syama S
(Project Coordinator)
Assistant Professor
Dept. of Computer Engineering
College of Engineering Chengannur

Smt. Chinchu M Pillai
(Project Coordinator)
Assistant Professor
Dept. of Computer Engineering
College of Engineering Chengannur

Dr.Sabeena K
(Project Coordinator)
Assistant Professor
Dept. of Computer Engineering
College of Engineering Chengannur

External Examiner

Declaration

We, the undersigned, hereby declare that the project report **AIVO: Advanced Intelligent Virtual Orator**, submitted for partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology from the APJ Abdul Kalam Technological University, Kerala, is a bonafide work done by us as a team under the supervision of **Smt. Syama S**, Assistant Professor, Department of Computer Engineering, College of Engineering, Chengannur.

This submission represents our collective ideas and work, and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources.

We also declare that we have adhered to the ethics of academic honesty and integrity and have not misrepresented or fabricated any data, idea, fact, or source in our submission. We understand that any violation of the above will result in disciplinary action by the institute or the University and may also lead to penal action from the sources that have not been properly cited or from whom permission has not been obtained. This report has not previously formed the basis for the award of any degree, diploma, or similar title from any other University.

10-03-2025
Chengannur

DIANA LIZ KURIAKOSE

MOHAMMED HISHAM

MARIA JOSHY

NANDANA VINOD

Acknowledgement

We take this opportunity to express our deepest sense of gratitude and sincere thanks to everyone who helped us to complete this work successfully. We would like to acknowledge and express our gratitude to **Prof.Dr.Hari V S**, the Principal, College of Engineering Chengannur and **Sri. Gopakumar G**, Head of Department, Computer Engineering, College of Engineering Chengannur for providing us with all the necessary facilities and support.

We would like to express our sincere gratitude to our project coordinator **Smt. Chinchu M Pillai**, Assistant Professor, and **Dr.Sabeena K**,Assistant Professor, Department of Computer Engineering, College of Engineering Chengannur for the support and cooperation.

We would like to place on record our sincere gratitude to our project guide, **Smt. Syama S**, Assistant Professor, Department of Computer Engineering, College of Engineering Chengannur for the guidance and mentorship throughout this work.

We would like to give proper credit to the authors of journals, which we used as reference materials for this project. Finally, we thank our family and friends who contributed to the successful fulfilment of this work.

DIANA LIZ KURIAKOSE

MOHAMMED HISHAM

MARIA JOSHY

NANDANA VINOD

Abstract

In the rigorous academic environment of KTU, students face a pressing challenge: navigating a wealth of study materials to extract concise, syllabus-aligned content for exam preparation. The scarcity of time in a short semester exacerbates the difficulty of finding relevant notes and solutions tailored to their curriculum. Despite the availability of resources, many fail to meet the precise demands of the KTU syllabus or the expectations of exam-focused learning. Recognizing this gap, AIVO: Advanced Intelligent Virtual Orator emerges as an innovative solution—a digital professor designed to streamline and enhance the learning process of 3rd year students. AIVO integrates curated notes, past question papers, and verified solutions into a comprehensive and interactive platform. By aligning its teaching methodology with the official KTU syllabus, AIVO ensures that students receive the most relevant, concise, and exam-focused content. Its structured approach to delivering information emulates the clarity and precision of a dedicated professor, enabling students to grasp concepts efficiently. The platform also provides opportunities for targeted practice, helping students consolidate their understanding and excel in exams. AIVO's design philosophy centers on addressing the critical pain points faced by students: lack of time, scattered resources, and the need for exam-oriented preparation. With its innovative approach, AIVO bridges the gap between the available study materials and the specific demands of the curriculum. By fostering clarity, focus, and efficiency, it empowers KTU 3rd year students to maximize their academic potential, setting a new standard for intelligent learning solutions in higher education. AIVO not only aims to ease the academic burden but also to inspire confidence and competence among students, ensuring success in their academic journey.

Contents

List of Figures

List of Abbreviations

AIVO	Advanced Intelligent Virtual Orator
AI	Artificial Intelligence
API	Application Programming Interface
DFD	Data Flow Diagram
LLM	Large Language Model
KTU	Kerala Technological University

Chapter 1

Introduction

1.1 History

In demanding academic settings like Kerala Technological University (KTU), students consistently encounter the significant challenge of efficiently processing extensive study materials. Preparing for examinations requires not just access to information, but the ability to quickly identify concise, syllabus-specific content. The compressed timeframe of semesters intensifies this difficulty, making the search for notes and solutions precisely tailored to the curriculum a critical bottleneck. While numerous resources may be available, they often lack direct alignment with the KTU syllabus or fail to meet the specific needs of exam-focused preparation.

Recognizing this pressing need within the KTU student community, AIVO: Advanced Intelligent Virtual Orator has been conceptualized. AIVO represents an innovative step forward in educational support, functioning as a specialized "digital professor" designed to optimize the learning journey for 3rd-year students. This platform addresses the core problem by integrating carefully curated notes, relevant past question papers, and verified solutions into a single, interactive, and comprehensive system.

What sets AIVO apart is its rigorous alignment with the official KTU syllabus. Its core methodology ensures that students receive content that is not only accurate but also concise and directly applicable to their examination requirements. The platform emulates the structured clarity of an effective professor, presenting information logically to facilitate efficient comprehension. Furthermore, AIVO incorporates features for targeted practice, enabling students to test and reinforce their understanding effectively.

The development of AIVO is centered on alleviating the key frustrations experienced by students: the scarcity of time, the fragmentation of study resources, and the overriding need for exam-oriented learning materials. By bridging the gap between available information and specific curricular demands, AIVO aims to enhance clarity, promote focus, and significantly improve study efficiency. It aspires to be more than just a repository of information; it seeks to empower KTU students, fostering confidence and competence to help them achieve their full academic potential and setting a benchmark for intelligent learning solutions in higher education.

1.2 Project Area

The AIVO platform is specifically designed for the academic environment of Kerala Technological University (KTU), with an initial focus on supporting 3rd-year undergraduate students. KTU is known for its rigorous engineering curriculum and standardized examination system across affiliated colleges. Students within this system operate under significant academic pressure, driven by demanding coursework and semester-based evaluations.

Despite the university's efforts and the availability of various online and offline resources, students face distinct challenges within the KTU context:

- Information Overload: Sifting through vast amounts of textbooks, reference materials, and online notes to find syllabus-relevant information is time-consuming.
- Lack of Exam-Focused Curation: Many available resources are generic or not specifically tailored to the pattern and requirements of KTU examinations.
- Time Constraints: Short semesters limit the time available for extensive searching and consolidation of study materials.
- Difficulty Accessing Verified Solutions: Finding reliable solutions for past examination questions that align with university expectations can be challenging.

AIVO targets these specific issues within the KTU ecosystem. By providing a centralized, digitally accessible platform that offers syllabus-aligned, curated content and practice opportunities, it aims to directly address the study inefficiencies faced by 3rd-year students, helping them navigate their curriculum more effectively and prepare optimally for examinations.

1.3 Objectives

The primary objectives of the AIVO project are:

- Curating Relevant Content: To aggregate, verify, and curate essential study materials, including lecture notes, past question papers, and solutions, specifically for KTU 3rd-year subjects.
- Ensuring Syllabus Alignment: To strictly align all content provided through the platform with the official, current KTU syllabus, ensuring relevance and accuracy.
- Optimizing for Exam Preparation: To structure and present information in a concise, clear, and easily digestible manner optimized for effective exam preparation.

- Providing Structured Guidance: To emulate the structured teaching approach of a professor, guiding students through topics logically and facilitating better comprehension.
- Enhancing Study Efficiency: To streamline the study process by providing a centralized, easily navigable platform, thereby saving students valuable time and effort.
- Facilitating Targeted Practice: To offer students opportunities for focused practice using past KTU examination questions and providing access to verified solutions for self-assessment.
- Developing an Interactive Platform: To create an engaging and interactive user interface that enhances the learning experience beyond static content delivery.

By achieving these objectives, AIVO aims to significantly reduce the academic burden on KTU 3rd-year students, improve their learning efficiency, boost their confidence, and ultimately contribute to their academic success within the university's demanding framework.

Chapter 2

Literature Review

2.1 Text Clustering as Classification with LLMs

Zhang et al. (2024) introduce a novel two-stage approach using large language models (LLMs) that turns text clustering into a classification-based problem[1]. The system first generates potential labels for the dataset and then classifies texts into these labels, rather than relying on conventional clustering algorithms. This method ad- dresses a major challenge of traditional systems that use embeddings like BERT, as it does not require complex fine-tuning and hyperparameter adjustments. The approach significantly improves clustering accu- racy, particularly in complex datasets involving tasks like intent detection and topic mining. Zhang et al. emphasize that few-shot learning enhances the model's performance and facilitates effective in-context learning. Despite the computational cost associated with an API-based imple- mentation, the model effectively handles complex data distributions and yields better results than conventional methods. Overall, this method simplifies text clustering while increasing efficiency and accuracy, representing a significant breakthrough in the field.

2.2 Tailoring Chatbots for Higher Education: Some Insights and Experiences

Kortemeyer offers a comprehensive overview of strategies for adapting Large Language Models (LLMs) for applications in higher education. The author identifies three primary methods: train- ing models from scratch, fine-tuning pre-existing models, and employing augmentation techniques such as Retrieval Augmented Generation (RAG). Training models from the ground up is noted to be overly complex and costly for most institutions, necessitating vast computational power and metic- ulously curated datasets[2]. In contrast, fine-tuning pre-trained open-weight models like Llama 3 emerges as a more viable option, though it still requires considerable effort and a careful approach to maintain the model's overall performance.

The paper positions RAG as potentially the most practical method for numerous higher ed- ucation scenarios. This technique enhances a standard LLM with relevant reference materials at the time of query, enabling customization without altering the foundational model. Kortemeyer discusses a RAG implementation at ETH Zurich, where

course-specific chatbots are developed by embedding course materials and utilizing semantic search to retrieve pertinent information. This method is recognized for its relatively straightforward implementation and adaptability, as it can be applied to various commercial LLMs and easily updated or reconfigured.
A significant point emphasized throughout the paper is the necessity for robust inference infras- tructure. Although cloud-based inference options are available for commercial models, they raise privacy issues and entail ongoing expenses. Institutions aiming to deploy custom models—whether developed from scratch or fine-tuned—encounter challenges in securing the required GPU resources for continuous inference. Kortemeyer observes that most university supercomputing facilities are not optimally equipped for such an always-on service. The paper concludes by highlighting that there is no universal solution for customizing chatbots in higher education, with the optimal choice contingent upon factors such as available resources, privacy concerns, and the specific application at hand.

2.3 Future applications of generative large language models: A data-driven case study on ChatGPT

Analysis of the general diffusion of generative LLMs like ChatGPT in different fields and their possible applications in the future, when using a data-driven approach, it relies on more than 3.8 million tweets between November 2022 and May 2023 to contextualize the tasks assigned by users to ChatGPT. The method incorporates several crucial steps in the process of data processing[3].
Using a rule-based NER system from NLP, the authors downloaded user-described tasks from their collection of tweets. The system extracted 31,747 unique tasks from this data, cleaning it for noise, text normalization, and grouping similar tasks. Using the BERTopic algorithm, which is a topic modeling tool based on NLP techniques to disclose patterns in big text datasets, they then clustered semantically similar tasks. That allowed them to realize six underlying business areas influenced by ChatGPT - namely human resources, programming, social media, office automation, search engines, and education.
The results demonstrate the usability of LLMs such as ChatGPT: it can be a coding assistant, a writing tool, or simply content creator. These features would actually cause vast changes in busi- ness because of the automation of most time-consuming tasks, from code generation and question answering to email composition and much more, resulting in a significantly increased efficiency of all industries. Finally, the authors connect the dots and provoke further research with respect to the integration of LLMs into innovation processes such as idea generation, selection, development, and market adoption. They also argue that such technologies should consider their social and ethical implications, a factor that is enhanced since they are likely to challenge and undermine traditional business models and operations.

2.4 AI-Based Research Companion (ARC): An Innovative Tool for Fostering Research Activities in Undergraduate Engineering Education

AI-Based Research Companion (ARC) is a platform developed to address challenges in under- graduate research by leveraging GPT-4 [4]. The platform aims to enhance student engagement in research through personalized recommendations, helping bridge the gap between academic theory and practical research.ARC serves as a solution by organizing and enhancing research activities us- ing generative AI technology. Through ARC, students can navigate vast academic content, receive research suggestions tailored to their interests, and be guided through each step of their research journey.

ARC situates within the broader context of AI applications in education, emphasizing the growing role of personalized learning systems. ARC integrates collaborative and content-based filtering to provide dynamic and relevant research recommendations, allowing students to work more efficiently. The system adapts to ongoing user interactions, ensuring its recommendations evolve to meet the academic needs of each student. The platform also includes features like manuscript drafting assistance and interactive QA, making the research process more engaging. This is particularly relevant in engineering disciplines where research demands are high.

Feedback highlights ARC's ability to improve research efficiency, though users suggested expand- ing the recommendation system's precision. Overall, ARC holds significant potential to reshape the undergraduate research landscape by providing a more accessible and tailored approach to research activities, promoting innovation, and offering a dynamic tool for students in engineering education.

2.5 "Digital Professor": Interactive Learning with Chatbot Technology

Lubom´ır Jame˘cn´y, Oleksii Yehorchenkov, and Nataliia Yehorchenkova examine how chatbot tech- nology is changing the classroom setting in higher education institutions (HEIs) in their 2023 report [5]. Their study highlights the need to update conventional teaching methods, which frequently fall short of meeting the demands of Generation Z students who seek quick and engaging learning opportunities.

To address these challenges, the authors present the "Digital Professor" chatbot, which uses the Telegram platform to automate processes like quiz grading, course material distribution, and round-the-clock access to learning materials. This invention is indicative of larger patterns in the digital revolution in education, especially in intelligent learning settings that utilize AI-powered tools.

In contrast to typical chatbots, the "Digital Professor" serves as an efficient learning manage- ment system (LMS), giving students access to a variety of resources, homework, assessments, and gamified educational opportunities. Positive feedback from Kyiv National University students high- lights the tool's ease of use and its potential to

enhance educational experiences in HEIs through advanced AI capabilities. Even though the current version relies on button-based navigation, con- versational AI elements are anticipated to be included in future releases.

2.6 AI-assisted learning with ChatGPT and large lan- guage models: Impli- cations for higher education

With an emphasis on ChatGPT, Samuli Laato, Benedikt Morschheuser, Juho Hamari, and Jari Bj¨orne's 2023 study explores the revolutionary effects of large language models (LLMs) in education. It charts the evolution of LLMs from the launch of the Transformer architecture in 2017 to OpenAI's development of GPT models [6]. With features like text and code generation, summarization, and interactive dialogue, systems like ChatGPT—which are built on LLMs—allow students to access expert-level knowledge and participate in reflective learning.
The authors evaluated ChatGPT's impact over a two-month period in order to perform a practical examination of its role in a computer science Bachelor's degree program at a Finnish institution. They list 13 important ramifications, such as the improvement of critical thinking abilities and possible hazards, such as an over-dependence on AI to write code and essays. Although ChatGPT has several benefits, the authors also highlight some drawbacks, like its propensity to "hallucinate" or generate inaccurate information. They address moral dilemmas, such as plagiarism, and advocate for a well-rounded approach to AI-assisted learning, emphasizing the significance of responsible integration in academic settings.

2.7 The effects of artificial intelligence applications in educational settings: Challenges and strategies

Artificial intelligence is revolutionizing the educational environment at very rapid rates, accom- panied by tremendous benefits and many challenges that come with it [7]. There are five areas in which the challenge cuts across: user experience, operational demands, environmental impact, technological limitations, and ethical concerns. Systematic problems can be handled using a review process in formulating an explicit research question, thus leading the criteria for selecting relevant literature. Firstly, will be the preliminary planning which includes finding of keywords to form a strategy for retrieving relevant studies within the largest and most established academic databases including ScienceDirect, IEEE, and Scopus. Using the inclusion/exclusion criteria in filtering arti- cle selections based on focus relevance, and date of publications, an almost impossibly large pool will be shrunk down into a slightly more manageable size of only the highest-quality studies.
Three stages are found in review. The planning phase summarizes the research aims and criteria for inclusion and exclusion of an article. Articles are sifted through a multi-step

protocol in the execution stage: keyword searches followed by screening of titles and abstracts, then full- text reviewing. Quality assessments ensure the study's relevance and rigour, further narrowing the selections to those with high levels of quality to be included. The final stage synthesizes the findings into categorized themes of challenges and strategies, providing a structured overview of the key obstacles and actionable recommendations on AI use in educational settings. Thus, this systematic review framework is robust enough to be explored in AI's role in education and its implications.

2.8 How to Regulate Large Language Models for Responsible AI

A structured approach to safeguarding implementation that deals with the ethical issues in large language models was proposed [8]. In this proposal, it utilizes the MECE principle: it categorizes three broad areas so that it comprehensively and systemically addresses each ethical challenge: it starts with review ethics codes across professions; then assesses ethics awareness within computer science; then pinpoints where the LLM system safeguard points lay.
It breaks the LLM lifecycle down into upstream and downstream, explaining where ethical interventions come most into play. Controls such as input controls that operate through careful data curation are significantly effective for the avoidance of ethical risk but are usually abandoned because they raise transparency issues. Whereas controls that are further downstream like output filtering are being applied much more widely as it is cheaper and less resource-intensive but not so effective at the core of addressing the ethical issues. Such an approach supports the notion of a proactive regulatory framework promoting responsible AI practices as LLMs are gaining fast adoptions.

2.9 Transfer learning with adaptive fine-tuning

Differential Evolution based Fine-Tuning (DEFT) can optimize selection of layers in transfer learn- ing for CNNs [9]. DEFT is used for addressing challenges in fine-tuning efficiently, especially in areas such as medical imaging where large datasets are not available. Transfer learning enables pre-trained models that have been trained on vast datasets to be adapted to new but related tasks. However, choosing which layers to fine-tune in a CNN is not an easy task.
As empirical approaches are mostly not generalizable across tasks, deft employs the DE algo- rithm to do the layer selection automatically. Each candidate solution in the DE algorithm cor- responds to a unique configuration of fine-tunable layers. The DE algorithm iteratively optimizes these configurations by training the cnn using a subset of target dataset and then analyze per- formance based on a fitness function, specifically categorical cross-entropy loss. DEFT's adaptive layer selection mechanism identifies the best combination of layers. It involves striking a compro- mise between keeping generic features of pre-trained layers and customizing them for the intended goal. This automated adaptive approach helps DEFT outperform traditional manual fine-tuning

techniques in performance and reduces trial and error in layer selection.

2.10 Advancing the generative AI in education research agenda: Insights from the Asia-Pacific region

A mixed-method approach combining qualitative as well as quantitative techniques is used to analyse generative ais impact in education [10]. Qualitative methods include case studies that observe how ai is being used for educational purposes in real-life, examining both benefits such as developing critical thinking and challenges such as educators resistance to change. Surveys and interviews are taken along with issues related to reliance on ai and privacy. Concurrently, through sources such as YouTube there is an understanding of peoples views related to the issue of ai in teaching. On the quantitative side, it keeps an eye on current studies in the sphere of ai for educational puposes to pinpoint key areas and track emerging technologies. The analysis by topic modelling further reveals public discourse on ai in education and evolving themes. The study further concludes that ai will make collaboration and critical thinking easy but should complement traditional teaching methods. Although it brings about personalization of learning opportunities, it raises ethical concerns about data collection and biases within the algorithms. Teachers beliefs and technological competency will influence their readiness in adopting ai for learning. AI should be used responsibly in education taking into account ethical concerns, teacher empowerment and the evolving role of AI in classroom.

Chapter 3

Problem Definition

3.1 Existing System

The integration of artificial intelligence (AI) into education has led to the development of various tools and virtual learning assistants aimed at enhancing the learning experience in higher education. These AI-powered tools assist educators and students by personalizing learning paths, automating grading, and providing intelligent tutoring systems. However, many existing solutions are designed for general educational purposes and may not align specifically with the unique curricula of individual institutions like Kerala Technological University (KTU).

Examples of such AI-powered educational tools include Khanmigo, an AI-powered teaching assistant developed by Khan Academy that guides learners to find answers themselves rather than simply providing solutions, fostering deeper understanding. Mindgrasp, which converts lectures, notes, and videos into study tools such as AI-powered flashcards, quizzes, and summaries, acting as a personal AI tutor to enhance the learning process. Cognii, offering intelligent tutoring systems that deliver personalized learning experiences through open-response assessments and pedagogically rich analytics. Iris, an AI-driven virtual tutor integrated into the interactive learning platform Artemis, offering personalized, context-aware assistance to computer science students by guiding them through programming exercises and fostering independent problem-solving skills.

While these AI tools offer significant benefits in education, they often focus on general learning support and may not be specifically tailored to the unique curriculum of a particular university like KTU. General AI models can sometimes provide inaccurate information or biased responses, and concerns exist regarding academic integrity when students might use these tools inappropriately. Over-reliance on general AI could also impede the development of essential critical thinking and problem-solving skills, as students may become accustomed to readily available answers without engaging in the necessary cognitive effort. Moreover, the information provided by general AI might not always be specific enough to the nuances of a university's curriculum or the level of detail required for examinations.

AIVO aims to address these limitations by providing curriculum-specific assistance tailored to the KTU education system. By focusing on the specific needs and challenges faced by KTU students, AIVO seeks to offer accurate, curriculum-aligned information and mitigate the risks associated with the use of general AI models in education. This specialized approach ensures that students receive support that is directly relevant to their coursework and examinations, enhancing their learning experience and academic performance.

3.2 Problem Statement

Third-year BTech students at KTU face significant challenges in accessing and organizing relevant study materials aligned with their syllabus, especially during short semesters. The overwhelming abundance of unstructured resources, coupled with the limited availability of concise and exam-focused content, hampers their ability to prepare effectively for exams. This lack of a centralized and reliable system leads to inefficient use of time and increased academic stress.

To address these issues, there is an urgent need for a solution that can provide students with instant, syllabus-specific, and exam-oriented answers to their queries. A system capable of delivering precise and curated information while simulating the guidance of an experienced professor would greatly enhance their academic preparation and outcomes.

3.3 Proposed Solution

The Advanced Intelligent Virtual Orator (AIVO) is designed to function as a digital professor, offering KTU students personalized, curriculum-specific assistance through advanced voice interaction capabilities. By integrating voice-to-text and text-to-voice technologies, AIVO enables natural, spoken dialogues, allowing students to engage with the system using verbal commands and receive articulate, spoken responses. This approach caters to diverse learning preferences, enhancing accessibility and engagement.

To implement these features, AIVO employs advanced speech recognition and synthesis technologies. By incorporating these technologies, AIVO can function as a digital professor, offering real-time, voice-based interactions that cater to diverse learning preferences and needs. This approach not only enhances engagement but also supports students who may benefit from auditory learning methods, thereby enriching the overall educational experience within the KTU community.

Chapter 4

Project Design

The design of AIVO: Advanced Intelligent Virtual Orator follows a modular approach, integrating multiple components for seamless user interaction, efficient data retrieval, and AI-powered response generation. The system architecture is structured to ensure optimal performance, scalability, and user experience.

4.1 System Architecture

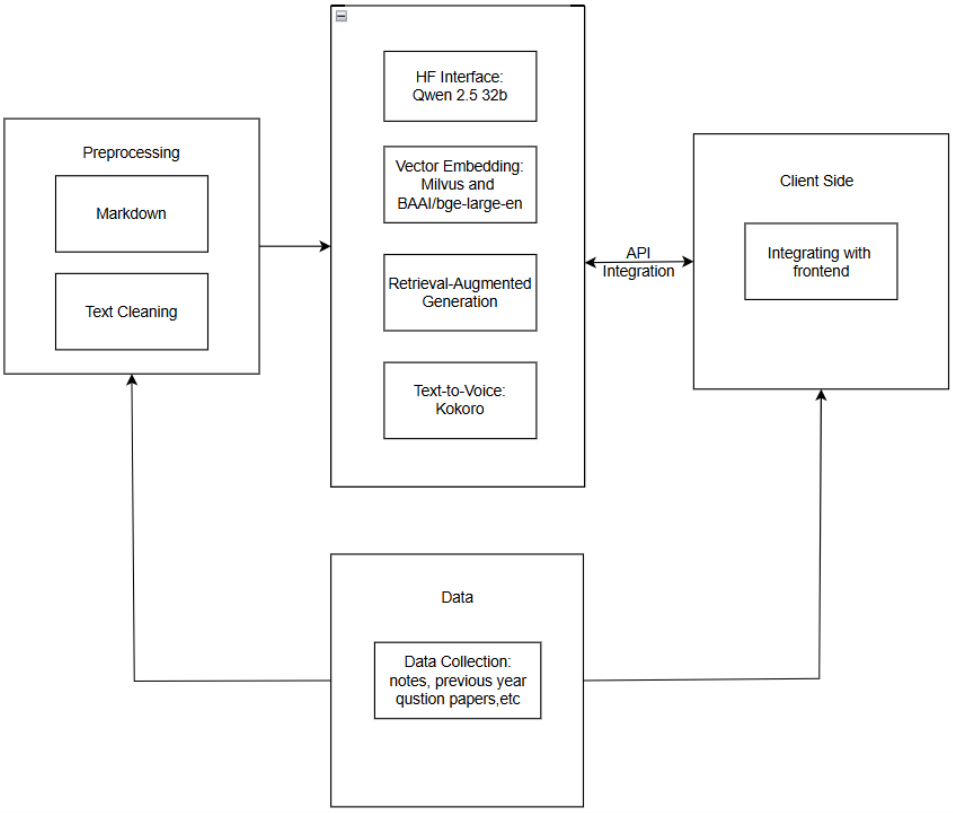

Figure 4.1: System Architecture

The architecture of AIVO introduces a more efficient and structured pipeline for

processing academic materials, training an advanced model, and delivering accurate, syllabus-aligned responses. The system integrates various components to ensure a seamless learning experience for students.

The process begins with data collection, where academic materials such as lecture notes and previous year question papers are gathered. These materials undergo preprocessing, which includes text cleaning to remove redundant content and Markdown formatting to structure the data effectively.

Once preprocessed, the data is used to create vector embeddings using Milvus and BAAI/bge-large-en, enabling efficient document retrieval. When a student submits a query, the system performs a semantic search in the vector database to fetch the most relevant content. This retrieved context is then passed to Qwen 2.5 32b via Hugging Face Inference, ensuring precise and context-aware responses.

To further enhance accessibility, responses are converted into speech using Kokoro, a lightweight text-to-voice model, allowing students to receive answers in both text and voice formats.

The final output is delivered through API integration with a web-based frontend, where students can interact with AIVO effortlessly. Future enhancements aim to incorporate voice-to-text and image-to-text functionalities for a more interactive learning experience.

By leveraging Milvus for fast retrieval, Qwen 2.5 32b for high-quality inference, and Kokoro for voice output, AIVO provides an efficient and user-friendly digital professor tailored for exam-oriented learning.

4.2 Usecase Diagram

The use case diagram represents the interactions between different actors (User, Admin, and System) and the core functionalities of AIVO: Advanced Intelligent Virtual Orator. The system is designed to process user inputs, retrieve relevant information using vector embeddings, and generate responses based on the given prompt.

Actors Involved:

- User: The primary actor who interacts with the system by sending prompts and receiving responses.
- Admin: Responsible for managing data preprocessing, API integration, and overall system maintenance to ensure optimal performance.
- System: The automated entity that processes inputs, retrieves relevant information, and generates responses.

Use Cases:

- Send Prompt (User): The user submits a query to the system, initiating the response generation process.
- Pre-process Data (Admin): The admin ensures that input data is cleaned and structured for efficient processing.
- Vector Embedding (System): The system utilizes Milvus and BAAI/bge-large-en to convert user input into high-dimensional embeddings for efficient retrieval of relevant study materials.
- API Integration (Admin): The admin manages API connections for seamless communication between different components.
- Process User Input (System): The system retrieves relevant content from the embedded database and prepares an appropriate response.
- Generate Response (System): The system formulates a response using the retrieved data and prepares it for output.
- Receive Output (User): The user receives the response in either text or voice format (via Kokoro for text-to-speech conversion).

This use case diagram provides a structured representation of AIVO's workflow, demonstrating how different actors collaborate to deliver an AI-powered, retrieval-augmented learning experience without requiring model fine-tuning.

4.3 Data Flow Diagram

Data Flow Diagrams (DFDs) visually represent how data flows through a system, showing processes, data stores, and the relationships between them. It helps to break down the system into manageable processes and highlights how inputs are transformed into outputs. Below is the explanation of the Level 0 DFD and Level 1 DFD for the AIVO system.

4.3.1 Level 0 DFD

The Level 0 DFD provides a broad view of the AIVO system, showing the interaction between the User and AIVO.

- **Components:**
  1. **User:** Provides a query (prompt) to the system.
  2. **AIVO:** Processes the prompt and generates a relevant response.
- **Data Flow:**
  - **Prompt:** User input such as text queries, images, or voice recordings.

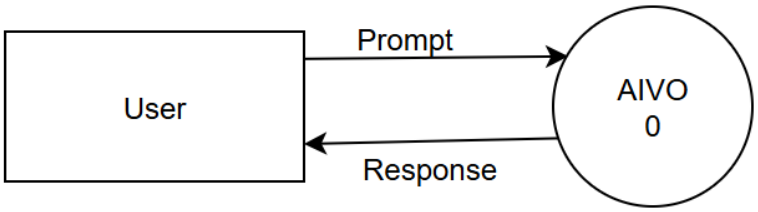

Figure 4.2: Level 0 DFD

– **Response:** AIVO processes the prompt and sends the relevant output back to the user.

This simple structure highlights that AIVO acts as a bridge between the user's needs and its knowledge base, ensuring accurate and efficient answers.

4.3.2 Level 1 DFD

The Level 1 DFD expands on the Level 0 diagram by introducing the internal processes, data stores, and flows within the AIVO system.

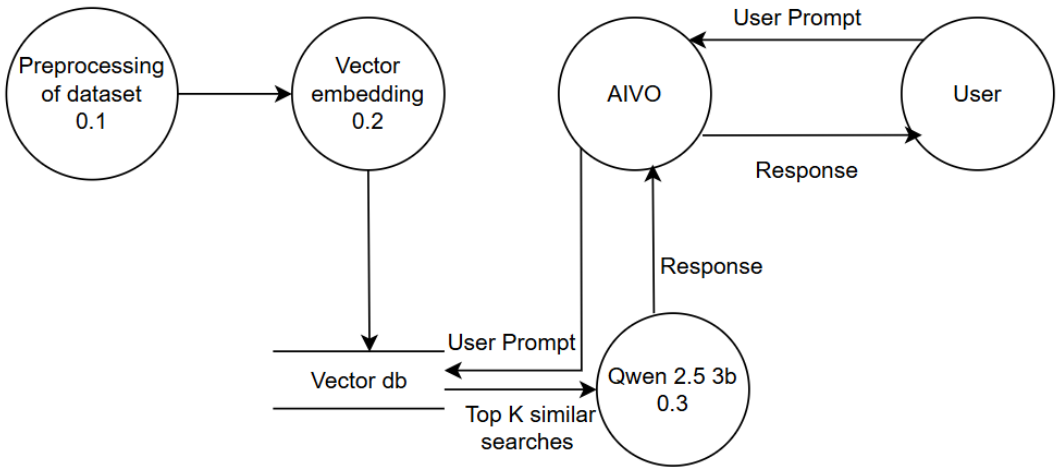

Figure 4.3: Level 1 DFD

Processes and Components:

1. Preprocessing of Dataset:
    - Collects and cleans academic data such as syllabus details, lecture notes, and past question papers.
    - Outputs structured data for embedding.
2. Vector Embedding:
    - AIVO utilizes the BGE large model to generate high-dimensional vector embeddings of KTU curriculum data, enabling efficient semantic search and precise retrieval of relevant study materials.
3. Vector Database:
    - AIVO employs Milvus as its vector database to efficiently store and retrieve high-dimensional embeddings.
4. User Interaction
    - The User provides a User Prompt to the AIVO system.
    - AIVO processes the query and and sends it to the Vector Database.
5. Qwen 2.5 3b
    - After similarity search, the top K similar searches are sent to Qwen 2.5 3b.
    - The model generates a Response and sends it back to AIVO.
6. AIVO
    - User input is retreived either by text or as voice.
    - The generated output can be converted from text to voice using Whisper large.

Key Data Flows:

- **Preprocessed Data:** Flows from the preprocessing module to fine-tuning and embedding.
- **User Prompts and Responses:** The user inputs queries that are processed by AIVO.
- **Model Outputs and Embeddings:** Stored in vector databases for reuse and efficient retrieval.

This detailed view provides insight into how AIVO manages academic data and user inputs to deliver efficient and accurate responses.

4.4 Resource Requirements

The implementation of AIVO requires the following hardware and software resources.

4.4.1 Hardware Requirements

To ensure the efficient performance of AIVO, the following hardware specifications are required:

- Processor: Minimum Intel Xeon or AMD EPYC (Recommended: 16-core or higher)
- RAM: At least 16GB
- Storage:

  SSD (NVMe preferred) with at least 1TB for fast read/write operations

  Additional storage space for dataset expansion
- GPU:

  RTX 3060 for efficient vector embedding operations
- Networking: High-speed internet (1 Gbps or higher) for API integrations and real-time responses

These hardware specifications will ensure seamless vector embedding, retrieval, and text-to-speech generation, enabling a smooth and efficient experience for both users and administrators

4.4.2 Software Requirements

To ensure smooth development, deployment, and execution of AIVO: Advanced Intelligent Virtual Orator, the system requires specific software components for vector embedding, retrieval, inference, and frontend integration. Below are the required software specifications:

- Operating System AIVO can be deployed on multiple operating systems, but for optimal performance, the recommended options are:
  - Ubuntu 22.04 LTS (Preferred for server deployment)
  - Windows 10/11 (For development and testing)
  - MacOS (Optional, suitable for local testing)
- Backend Requirements The backend of AIVO is responsible for handling API requests, processing vector embeddings, and managing retrieval operations. The key software components include:
  - Python 3.10+ for core AI processing and backend logic.
  - FastAPI or Flask for creating efficient RESTful APIs.
  - Milvus as the vector database for storing and retrieving embeddings.
  - BAAI/bge-large-en for generating vector embeddings.
  - PyTorch (CUDA-enabled) for deep learning computations and model inference.

  - Hugging Face Transformers for handling large-scale model interactions, including Qwen 2.5 32B for response generation.
  - Whisper Large for speech-to-text processing.
  - Kokoro for text-to-speech conversion.
  - FFmpeg for processing and managing audio files, which is required for Whisper.

- Frontend Requirements The frontend is responsible for providing a seamless user experience and interaction with AIVO.
  - HTML5 – Structures the web interface.
  - CSS3 – Styles and enhances UI design.
  - JavaScript – Handles user interactions and API requests.
  - Fetch API – Facilitates communication between frontend and backend services.

- Database
  - Milvus for handling and retrieving vector embeddings.

- Deployment
  - Docker for containerized deployment and scalability.

These software requirements ensure seamless execution, scalability, and performance of AIVO, making it efficient for vector embedding retrieval and real-time response generation.

4.5 Work Schedule

August 2024 - September 2024: Data acquisition and preprocessing. This phase involves collecting KTU syllabus documents, lecture notes, and study materials. PDFs and other formats will be converted into Markdown to standardize data, ensuring seamless processing for AI models.

October 2024 - November 2024: Vector embedding model integration and database setup. The BGE large model will be used to generate vector embeddings, and Milvus DB will be implemented for efficient storage and retrieval. Data chunking strategies will be applied to optimize search accuracy.

December 2024 - January 2025: Query processing and AI model integration. The system will be trained to convert user queries into vector embeddings, retrieve relevant curriculum content using Milvus, and generate responses using the Qwen 2.5 32b model. Testing and optimization will ensure accurate results aligned with the KTU syllabus.

February 2025: Frontend and backend development. The frontend interface will be built using HTML, CSS, and JavaScript, while the backend will be developed in Python

to manage API calls and data processing. Voice-based features, including Whisper AI for voice-to-text and Kokoro for text-to-speech, will be integrated.

March 2025: System testing, refinement, and deployment. Comprehensive testing will be conducted to ensure functionality, security, and performance. User feedback will be incorporated for final refinements. The project will be documented with technical specifications and user manuals. A final presentation will showcase the system's capabilities before the official launch.

4.6 Technologies Used

The AIVO website is designed with several key features to provide a comprehensive and user-friendly learning environment for KTU students.

A primary feature is the integration of voice-to-text functionality powered by the Whisper AI model from Hugging Face. This allows students to verbally ask their questions, providing an alternative to typing and enhancing accessibility. Complementing this is the text-to-voice feature, which uses the Kokoro model to read out study materials and generated answers. This can aid students with different learning styles and provide a hands-free way to engage with the content.

Feature	BGE large-en-v1.5 (BAAI)	Qwen 2.5 32b (Alibaba Cloud)
Number of Parameters	335 Million	32.5 Billion
Embedding Dimensions	1024	N/A
Maximum Input Tokens	512	131,072
Key Capabilities	Text Embedding, Retrieval, Semantic Search	Language Generation, Reasoning, Code Generation

Figure 4.4: Technical Specifications of Key AI Models

The website also contains study materials that are separated by year and branch. This organizational structure enables students to easily navigate and find the specific content relevant to their academic program and current year of study, streamlining the process of locating necessary resources. Furthermore, AIVO features a dedicated chatbot area where students can input their queries and interact with the AI professor. This central hub facilitates a direct and interactive learning experience, allowing students to ask questions and receive immediate, curriculum-specific responses. These features collectively aim to create an intuitive and effective platform for KTU students to access academic support.

The development of the Advanced Intelligent Virtual Orator (AIVO) system incorporates a range of technologies to deliver an interactive and efficient academic assistant for KTU students. The key technologies utilized include:

- Qwen 2.5 32B: Generates contextually relevant responses based on user queries, enhancing the conversational experience.
- Whisper Large: Converts speech input into text, enabling accurate processing of voice commands.
- Kokoro: Synthesizes human-like speech from text, providing natural and clear audio responses.
- BAAI/bge-large-en : An advanced English language embedding model developed by the Beijing Academy of Artificial Intelligence (BAAI), to generate vector embeddings that enhance AIVO's natural language processing capabilities.
- Milvus: Manages and retrieves vector embeddings, supporting efficient similarity searches and data retrieval.
- Docker: Containerizes the application, promoting consistency across various development and deployment environments.

  By integrating these technologies, AIVO delivers a seamless, responsive, and intelligent user experience tailored to the academic needs of KTU students.

Chapter 5

Report of Project Implementation

The AIVO system is designed with a multi-layered architecture to provide curriculum-specific information and support to KTU students. The process begins with the acquisition and preprocessing of relevant data.

5.1 Data Acquisition and Preprocessing

The foundation of AIVO lies in the collection of KTU syllabus documents and supplementary notes from various available resources. These documents, often in PDF format, undergo a crucial preprocessing step where they are extracted and converted into Markdown format. This transformation into a standardized, easily parsable format is essential for improving the understanding and subsequent processing of the information by the AI models. By converting diverse data sources into a uniform Markdown structure, AIVO ensures a consistent input for the downstream processes of information retrieval and vector embedding.

5.2 Vector Embedding Model

At the heart of AIVO's information retrieval mechanism is the BGE large model (BAAI/bge-large-en-v1.5), which generates dense vector representations of the preprocessed KTU curriculum data. This model transforms textual data into high-dimensional vectors (1024 dimensions), capturing the semantic meaning of the text. The BGE large model is particularly well-suited for tasks such as information retrieval and semantic search, allowing AIVO to understand the context and meaning of student queries and match them with relevant sections of the KTU syllabus and notes. Given the model's maximum input token limit of 512, careful chunking of the curriculum data is necessary during the embedding process.

METRIC	BGE LARGE EN (VALUE)	NORMIC EMBED TEXT (VALUE)
Embedding Quality	0.92	0.85
Dimensionality	1024	768
Accuracy on Downstream Tasks	0.88	0.82

Figure 5.1: Comparison between BAAI/bge-large-en and nomic-embed-text

5.3 Vector Database

To efficiently store and retrieve the vector embeddings generated by the BGE large model, AIVO utilizes the Milvus DB. Milvus is a cloud-native vector database designed to handle vast quantities of vector data, making it suitable for managing extensive curriculum information. It supports high-performance vector similarity searches, enabling AIVO to quickly identify the most relevant curriculum content in response to student queries. Milvus offers various indexing techniques, such as Hierarchical Navigable Small World (HNSW) and Inverted File (IVF), which significantly speed up the search process. The selection of Milvus over ChromaDB suggests a requirement for a more robust and scalable solution for curriculum-specific information retrieval.

METRIC	CHROMADB(VALUE)	MILVUS(VALUE)
Precision	0.55	0.72
Recall	0.58	0.79
F1 Score	0.51	0.70

Figure 5.2: Comparison between ChromsDB and Milvus

5.4 Query Processing and Answer Generation

When a student poses a question through the AIVO interface, the system first converts the query into a vector embedding using the same BGE large model. This query vector is used to perform a similarity search within the Milvus database to extract the top-k most relevant results. For AIVO, the top 2 results are extracted to provide context for the subsequent answer generation phase. These retrieved results are then fed into the Qwen 2.5 32b language model. Qwen 2.5 32b is a large

language model developed by Alibaba Cloud, featuring 32.5 billion parameters and an extensive context window of 128K tokens. This model enhances the retrieved information from Milvus, synthesizing it into a comprehensive and contextually relevant answer aligned with the KTU curriculum and examination expectations.

5.5 Frontend and Backend Technologies

The AIVO website employs a standard web application architecture. The frontend, responsible for user interaction, is built using HTML, CSS, and JavaScript. The backend, handling server-side logic and data processing, is implemented in Python. Integration with the Qwen 2.5 32b model is facilitated through the Hugging Face Inference API, allowing AIVO to leverage the model's capabilities via API calls. For voice-related features, AIVO incorporates the Whisper AI model from Hugging Face for voice-to-text conversion and the Kokoro model for text-to-voice synthesis. These voice models operate entirely offline, ensuring their availability regardless of internet connectivity.

5.6 Voice Features

To enhance accessibility and usability, AIVO integrates voice-to-text and text-to-voice functionalities. The voice-to-text feature utilizes the Whisper AI model, enabling students to input queries using voice commands. This is particularly beneficial for students who prefer speaking over typing or those with accessibility needs. The text-to-voice feature employs the Kokoro model to convert generated answers and study materials into spoken words, helping auditory learners or multitasking students. Running both Whisper and Kokoro models offline ensures consistent availability, even in low-connectivity situations, thereby improving the reliability and convenience of the AIVO system.

5.7 Development Challenges and Resolutions

The development of the AIVO system involved several challenges, particularly in achieving the desired level of accuracy and performance.
One of the initial approaches involved fine-tuning a Llama 3.2 3b model with the aim of enhancing its accuracy for the specific task of providing KTU-related answers. However, this approach yielded very low accuracy, prompting a re-evaluation of the language model choice. Benchmarking data suggests that Qwen 2.5 32b generally demonstrates superior performance compared to Llama 3.2 3b across various tasks, including reasoning and code generation, which are relevant to understanding and generating curriculum-specific content. The smaller size and potentially less comprehensive training data of Llama 3.2 3b likely contributed to its inability to provide the nuanced and accurate responses

required for the KTU curriculum. Consequently, this initial fine-tuning approach was discontinued in favor of a more robust language model.

Another challenge encountered during the development process involved the initial selection of ChromaDB for vector database storage and the Nomic Embed text model for generating embeddings. The Nomic Embed text model produces embeddings with a dimensionality of 768. This lower dimensionality, compared to the 1024 dimensions of the BGE large model, was found to result in low accuracy in retrieving relevant information. The higher dimensionality offered by the BGE large model likely allows for a richer and more detailed representation of the semantic information within the KTU curriculum data, leading to more effective matching with student queries.

To overcome these limitations, a successful transition was made to Milvus DB for vector storage and the BGE model for embedding. The BGE model, with its 1024-dimensional embeddings, significantly improved the accuracy of information retrieval. Furthermore, Milvus DB proved to be a more suitable vector database solution, likely due to its optimized architecture for handling large-scale vector data and performing efficient similarity searches. This combination of a higher-dimensional embedding model and a more robust vector database was crucial in achieving the desired performance for the AIVO system.

Additionally, the fine-tuning part of the initial approach, which involved Llama 3.2 3b, was ultimately dropped. This decision suggests that the pre-trained Qwen 2.5 32b model was found to be sufficiently capable for the task without the need for further fine-tuning, or that the fine-tuning process itself was not yielding the desired improvements in accuracy.

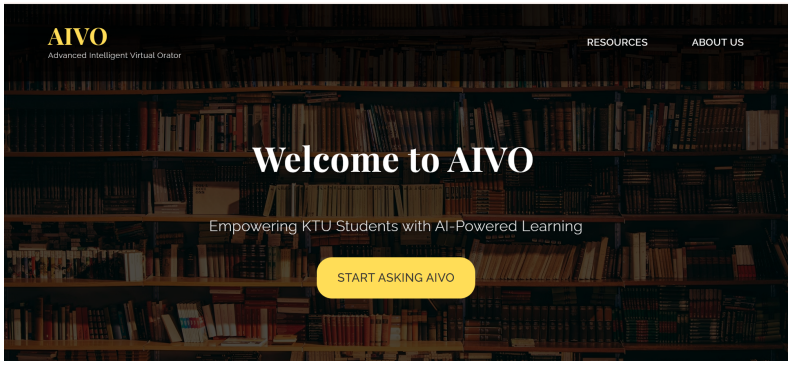

Figure 5.3: Home page

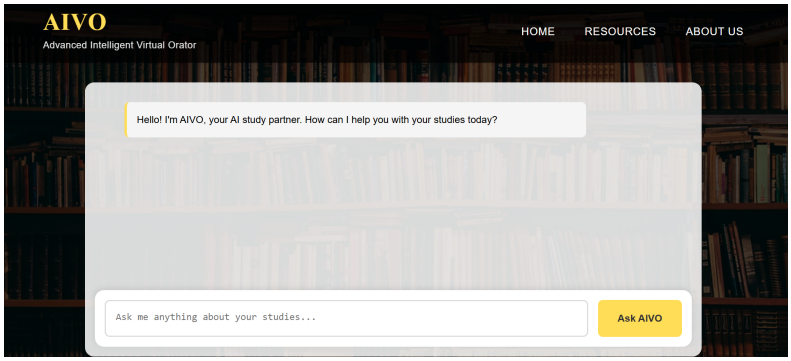

Figure 5.4: Ask AIVO

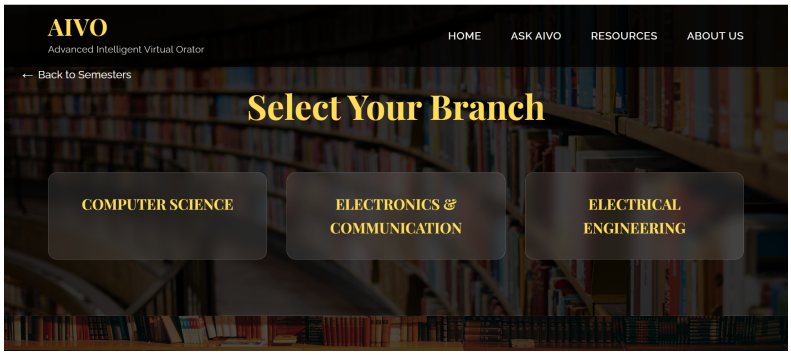

Figure 5.5: Branches

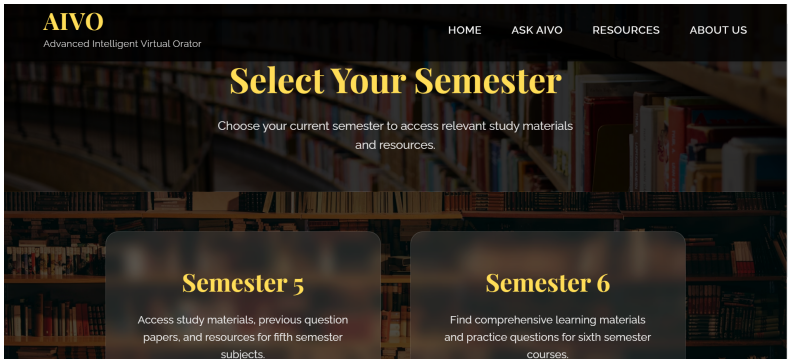

Figure 5.6: Semester

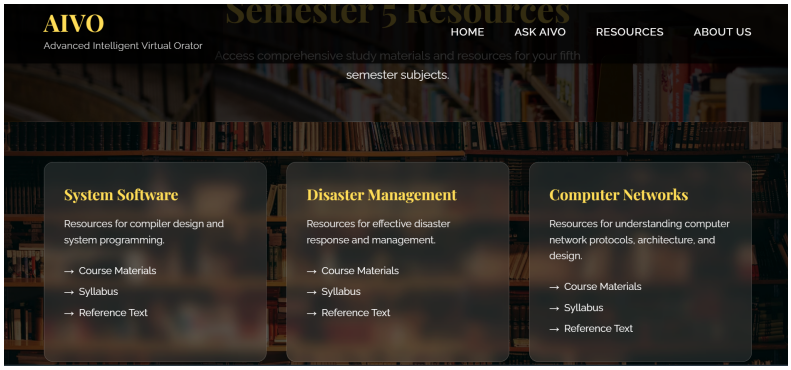

Figure 5.7: Subjects

Chapter 6

Results & Conclusion

6.1 Result

The development of AIVO presents several potential benefits for students at Kerala Technological University (KTU). By providing access to syllabus-specific information, AIVO can help students move beyond the generalized answers offered by general AI tools, ensuring that their study efforts are focused on the content directly relevant to their curriculum. This targeted approach can lead to an enhanced understanding of course material that is aligned with the specific examination patterns of KTU, potentially improving students' preparedness for assessments. The ability to quickly access relevant and curated information can also significantly increase study efficiency, allowing students to focus on key concepts without sifting through extraneous material. Furthermore, AIVO holds the promise of delivering personalized learning experiences tailored to the KTU curriculum, adapting to the specific needs and queries of individual students. The integration of voice features further enhances the accessibility of the system, catering to diverse learners and providing flexibility in how students interact with the AI professor. Finally, the organized structure of study materials on the website simplifies navigation and information retrieval, making it easier for students to find the resources they need.

Compared to existing general AI tools, AIVO offers a distinct advantage through its curriculum-specific focus. While tools like ChatGPT provide broad knowledge across various domains, AIVO's methodology is centered on KTU-specific data, ensuring a higher degree of relevance and accuracy within the context of the university's academic programs. The entire system, from data processing to answer generation, is designed to align with the KTU syllabus and examination expectations, a level of specificity that general AI tools cannot inherently provide. Moreover, the inclusion of voice features and the organization of study materials are tailored to the needs of KTU students, creating a more integrated and supportive learning environment. This specialization positions AIVO as a more valuable resource for KTU students seeking academic assistance directly related to

their coursework.

Despite the potential benefits, it is important to acknowledge certain potential limitations and ethical considerations associated with AIVO. The effectiveness of the system is inherently dependent on the completeness and accuracy of the KTU syllabus and notes collected and processed. Any gaps or inaccuracies in these source materials could impact the quality of the information provided by AIVO. Additionally, ethical considerations surrounding the use of AI in education, such as the potential for over-reliance on the system and issues related to academic integrity, need to be carefully addressed. Future research should focus on empirically evaluating the effectiveness of AIVO with KTU students, gathering feedback on its usability and accuracy, and further refining the system based on these findings. Exploring the seamless integration of AIVO with the learning management systems currently used by KTU could also enhance its utility. Furthermore, ongoing efforts should be directed towards identifying and mitigating any potential biases in the training data or the responses generated by the AI models to ensure fairness and equity in its application.

6.2 Conclusion

The development of the Advanced Intelligent Virtual Orator (AIVO) represents a significant advancement in the application of artificial intelligence to address the specific needs of students within the Kerala Technological University (KTU) education system. By employing a methodology that prioritizes curriculum-specific data, utilizing the BGE large model for vector embeddings, Milvus DB for efficient storage, and the Qwen 2.5 32b language model for enhanced answer generation, AIVO offers a targeted alternative to general-purpose AI tools. The system's architecture and key features, including voice-to-text, text-to-voice, and organized study materials, are designed to create an accessible and effective learning environment.

The challenges encountered during the development process, particularly with earlier models and embedding techniques, highlight the importance of careful selection and optimization of AI components. Ultimately, AIVO demonstrates the potential of curriculum-specific AI tools to transform learning outcomes in specialized educational contexts like KTU, offering a more relevant and tailored approach to academic support compared to the broader capabilities of general AI.

Chapter 7

Future Scope

To further enhance AIVO's capabilities and improve its effectiveness for KTU students, several key advancements are planned for future development:

- Integration of Image-to-Text Functionality
  AIVO will incorporate optical character recognition (OCR) technology to analyze and convert images of diagrams, equations, and handwritten notes into text. This will allow students to upload study materials in image format and receive corresponding explanations, expanding the system's usability.
- Support for Handwritten and Complex Equations
  Recognizing the challenges students face with complex mathematical and engineering equations, AIVO will be enhanced to interpret and process handwritten content accurately, offering step-by-step explanations for problem-solving.
- Expansion of the Knowledge Base
  The system will integrate additional academic resources such as previous university question papers, research papers, subject-specific textbooks, and interactive tutorials. By broadening its dataset, AIVO will provide a more comprehensive and in-depth understanding of KTU's syllabus.
- Multimodal Input and Processing
  Future updates will enable AIVO to process queries through multiple input formats, including text and uploaded documents. This will allow students to interact with the system in a way that best suits their study preferences.
- Adaptive Learning and Personalization
  AIVO will track user interactions to personalize responses based on individual learning styles and preferences. It will provide tailored suggestions, adaptive quizzes, and topic-based recommendations to enhance the student learning experience.
- Seamless Integration with KTU's Learning Management System (LMS)
  AIVO will be directly embedded into KTU's existing online learning platform, allowing students to access it within their course modules. This integration will ensure that AIVO becomes a seamless part of students' daily academic activities.

- Offline Accessibility for Key Features
  To support students in areas with limited internet connectivity, offline capabilities will be introduced for essential features like note retrieval, quiz-based learning, and syllabus-specific recommendations.
- Performance Optimization and Scalability
  Future upgrades will focus on optimizing AIVO's response time and scalability, ensuring it can efficiently handle an increasing number of users and larger datasets without compromising performance.
- AI-Powered Doubt Resolution and Discussion Forums
  AIVO will introduce an AI-assisted forum where students can post academic queries and receive AI-generated responses alongside peer discussions. This will encourage collaborative learning and faster doubt resolution.

By implementing these advancements, AIVO will evolve into a more intelligent, personalized, and interactive academic assistant, revolutionizing how KTU students prepare for their exams and engage with study materials.

Chapter 8

Publication

Certified

Certificate of Publication

IJAEM

This is to confirm the

Mohammed Hisham, Nandana Vinod, Diana Liz Kuriakose, Maria Joshy, Syama S

Published following article

Survey on Generative AI in Education

Volume 6, Issue 11, pp: 615-620

www.ijaem.net

A Peer Reviewed Journal

International journal of Advances in Engineering and Management (IJAEM)

ISSN: 2395-5252

Publication Head

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

Survey on Generative AI in Education

Mohammed Hisham, Nandana Vinod, Diana Liz Kuriakose, Maria Joshy, Syama S
B.Tech Computer Science Engineering
College of Engineering Chengannur

Date of Submission: 25-11-2024 Date of Acceptance: 05-12-2024

ABSTRACT—This paper explores how large language models are being used in educational institutions, after which it goes on to examine their applications, challenges, and ethical implications. By synthesizing recent insights, the paper will delve into LLM text clustering, customization of higher education, and personalized learning. The presence of generative AI tools, such as ChatGPT, will transform the arena of research efficiency and adaptive learning. However, issues related to privacy, data security, and ethics such as algorithmic bias are still present and have to be sorted out. The paper presents suggestions for the responsible integration of LLMs into learning structures to nullify risk issues that may pose a threat to augmenting positive learning outcomes.

Index terms - Academic Resources, Generative AI, Student Support, Lecture Notes, Learning Enhancement.

I. INTRODUCTION

LLMssuchasGPT4andChatgptarechanging the way we think about education. These are powerful models that takes and generates human readable texts, opening doors for students teachers and researchers to get a lot done from automating routine tasks to creating more personalized learning experiences. LLMs are changing classrooms and making education more accessible, efficient and engaging. This paper explores the many ways LLMs are being used in education today. The challenges that go along with its use and what in the future is likely to be for it in academic settings.

A. Why LLMs Matter in Education

LLMs are now playing a significant role in the world of education, and there is a good reason for this. These models can understand and generate language in ways that make them useful across a wide range of tasks. Whether it is answering student questions, grading assignments, or helping with research, LLMs are making life easier for both students and educators.

One of the biggest advantages offered by LLMs is that they can process large sums of information quickly and accurately. So, these are especially helpful when doing personalized learning instruction, allowing students to obtain a personal sense of feedback. For example, an LLM might explain a math concept in a way that makes sense to a struggling student while providing another student with more challenging work. Such individual- ized support is not something that traditional teaching methods can easily provide on a large scale.

With more schools and universities shifting towards digital and distant learning, LLMs also make education more fluid and accessible. They enable access to learning materials at any moment in time, from anyplace, and in a manner that best suits personal learning tastes. This is especially imperative given the evolving nature of education. Modern society has been shaped through rapid technological advancement in just about every area of living.

B. Real-World Applications of LLMs in Education

LLMs are already making a big impact in classrooms, research labs, and beyond. One exciting example is how they're being used to improve the way we classify and understand large sets of text. For instance, instead of relying on traditional algorithms to cluster related pieces of text together, LLMs can categorize documents based on specific criteria, like intent or topic. This is particularly useful for tasks like grading essays, sorting through academic papers, or helping students find the information they need.

Another area where LLMs are making a difference is in research. Tools like the AI-Based Research Companion (ARC) use LLMs for assisting students and researchers by suggesting appropriate sources, helping with data analysis, and even guiding the writing process. This would save hours of work, especially in fields like engineering

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

or medicine, where the research process can be both data-heavy and time-consuming. Similarly, AI- powered chatbots like "Digital Professor" are automating routine academic tasks, from grading quizzes to answering student inquiries, freeing up educators to focus on more meaningful aspects of teaching.

Personalized learning is another area where LLMs are shining. Techniques like Retrieval Augmented Generation (RAG) are enhancing how students interact with course materials. RAG allows LLMs to pull in relevant information from external sources, providing students with more accurate and context-specific answers to their questions. So, instead of just getting a generic response, students are receiving insights that are directly tied to their curriculum, making learning more engaging and effective.

C. Challenges and Ethical Questions

While the scope of LLMs in education is clear, their use also raises some important challenges. One of the biggest concerns is the issue of bias. Because LLMs learn from large datasets that may contain biased or incomplete information, they can sometimes reinforce harmful stereotypes or provide inaccurate answers. This is especially troubling in education, where fairness and equity are so important. If not carefully monitored, LLMs could end up giving certain students an unfair advantage or misrepresenting important information.

Another challenge is the risk of students becoming too reliant on AI-generated content. While LLMs can be great tools for learning, there's a concern that students might use them to complete assignments without fully engaging with the material themselves. This could lead to a decline in critical thinking and creativity, as students might be tempted to take shortcuts rather than grappling with complex problems on their own.

Privacy is another important issue. Since LLMs process large amounts of data, including personal and sensitive information, it's crucial that schools and universities have strong data protection policies in place. If not handled properly, student data could be at risk of being accessed or misused. Additionally, running these large models requires significant computing power, which can be a barrier for schools with limited resources. This raises questions about the environmental impact of LLMs, as the energy required to train and operate these models is substantial.

D. Looking Ahead: The Future of LLMs in Education

But despite all of these challenges, the future for LLMs in education does not seem in the least dire. Some truly very promising work involves utilizing transfer learning, in which adaptation to a novel task would be possible only with minimal need for retraining. It would thus be very specialized for educational establishments and fine-tuned either to subjects or areas of interest, so that they could prove highly useful both for students and teachers.

In the near future, we will more and more witness LLMs applied to personalized learning. Models are adaptive with real-time progression of the student, and the recommendations adjust according to this. One can imagine an LLM not only helping students with homework but also providing topics of interest or areas of strength for the student to learn. It would then make studying much more dynamic and interactive because it encourages students to go deeper into their studies.

While educational institutions need to be very careful with the use of LLMs, the technology is full of promise and should support but not replace the traditional method of teaching. Teachers will need to find a balance in the future between exercising the potential of AI and still retaining the humanness of learning through such skills as critical thinking, creativity, and social skills.

LLMs are changing the face of education. They have opened new ways of learning, teaching, and researching. The benefits can be very significant. While challenges persistinethics, privacy, and access, we can make learning environments more inclusive, efficient, and engaging for students of all backgrounds and abilities. By carefully integrating LLMs into the educational process, the future of education powered by AI is just beginning, and it promises exciting possibilities for everyone.

II. LITERATURE REVIEW

[1] Zhang et al. (2024) introduce a novel two-stage approach using large language models (LLMs) that turns text clustering into a classification-based problem. The system first generates potential labels for the dataset and then classifies texts into these labels, rather than relying on conventional clustering algorithms. This method ad- dresses a major challenge of traditional systems that use embeddings like BERT, as it does not require complex fine-tuning and hyper parameter adjustments.

The approach significantly improves

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

clustering accuracy, particularly in complex data sets involving tasks like intent detection and topic mining. Zhangetal. emphasize that few-shot learning enhances the model's performance and facilitates effective in-context learning. Despite the computational cost associated with an API-based implementation, the model effectively handles complex data distributions and yields better results than conventional methods. Overall, this method simplifies text clustering while increasing efficiency and accuracy, representing a significant breakthrough in the field.

[2] Kortemeyer offers a comprehensive overview of strategies for adapting Large Language Models (LLMs) for applications in higher education. The author identifies three primary methods: training models from scratch, fine-tuning preexisting models, and employing augmentation techniques such as Retrieval Augmented Generation(RAG).Training models from the ground up is noted to be overly complex and costly for most institutions, necessitating vast computational power and meticulously curated datasets. In contrast, fine-tuning pre-trainedopen-weightmodelslikeLlama3emergesas a more viable option, though it still requires considerable effort and a careful approach to maintain the model's overall performance.

The paper positions RAG as potentially the most practical method for numerous higher education scenarios. This technique enhances a standard LLM with relevant reference materials at the time of query, enabling customization without altering the foundational model. Kortemeyer discusses a RAG implementation at ETH Zurich, where course-specific chatbots are developed by embedding course materials and utilizing semantic search to retrieve pertinent information. This method is recognized for its relatively straightforward implementation and adaptability, as it can be applied to various commercial LLMs and easily updated or reconfigured.

A significant point emphasized throughout the paper is the necessity for robust inference infrastructure. Although cloud-based inference options are available for commercial models, they raise privacy issues and entail ongoing expenses. Institutions aiming to deploy custom models—whether developed from scratch or fine- tuned—encounter challenges in securing the required GPU resources for continuous inference. Kortemeyer observes that most university supercomputing facilities are not optimally equipped for such an always-on service. The paper concludes by highlighting that there is no universal solution for customizing chatbots in higher education, with the optimal choice contingent upon factors such as available resources, privacy concerns, and the specific application at hand.

[3] Analysis of the general diffusion of generative LLMs like ChatGPT in different fields and their possible applications in the future, when using a data-driven approach, it relies on more than 3.8 million tweets between November2022andMay2023tocontextualize the tasks assigned by users to ChatGPT. The method incorporates several crucial steps in the process of data processing.

Using a rule-based NER system from NLP, the authors downloaded user-described tasks from their collection of tweets. The system extracted 31,747 unique tasks from this data, cleaning it for noise, text normalization, and grouping similar tasks. Using the BERT opic algorithm, which is a topic modeling tool based on NLP techniques to disclose patterns in big text datasets, they then clustered semantically similar tasks. That allowed them to realize six underlying business areas influenced by ChatGPT - namely human resources, programming, social media, office automation, search engines, and education.

The results demonstrate the usability of LLMs such as ChatGPT: it can be a coding assistant, a writing tool, or simply content creator. These features would actually cause vast changes in business because of the automation of most time-consuming tasks, from code generation and question answering to email composition and much more, resulting in a significantly increased efficiency of all industries. Finally, the authors connect the dots and provoke further research with respect to the integration of LLMs into innovation processes such as idea generation, selection, development, and market adoption. They also argue that such technologies should consider their social and ethical implications, a factor that is enhanced since they are likely to challenge and undermine traditional business models and operations.

[4] AI-Based Research Companion (ARC) is a plat- form developed to address challenges in undergraduate research by leveraging GPT-4. The platform aims to enhance student engagement in research through personalized recommendations, helping bridge the gap between academic theory and practical research.ARC serves as a solution by organizing and enhancing research activities using generative AI technology. Through ARC, students can navigate vast academic content, receive research suggestions tailored to their interests, and be guided through each step of their research journey.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

ARC situates within the broader context of AI applications in education, emphasizing the growing role of personalized learning systems. ARC integrates collaborative and content-based filtering to provide dynamic and relevant research recommendations, allowing students to work more efficiently. The system adapts to ongoing user interactions, ensuring its recommendations evolve to meet the academic needs of each student. The platform also includes features like manuscript drafting assistance and interactive Q&A, making the research process more engaging. This is particularly relevant in engineering disciplines where research demands are high.

Feedback highlights ARC's ability to improve research efficiency, though users suggested expanding the recommendation system's precision. Overall, ARC holds significant potential to reshape the undergraduate research landscape by providing a more accessible and tailored approach to research activities, promoting innovation, and offering dynamic tool for students in engineering education.

[5] Lubom´ır Jamečný, Oleksii Yehorchenkov, and Nataliia Yehorchenkova examine how chatbot technology is changing the classroom setting in higher education institutions (HEIs) in their 2023 report. Their study highlights the need to update conventional teaching methods, which frequently fall short of meeting the demands of Generation Z students who seek quick and engaging learning opportunities.

To address these challenges, the authors present the "Digital Professor" chatbot, which uses the Telegram platform to automate processes like quiz grading, course material distribution, and round-the-clock access to learning materials. This invention is indicative of larger patterns in the digital revolution in education, especially in intelligent learning settings that utilize AI-powered tools.

In contrast to typical chatbots, the "Digital Professor" serves as an efficient learning management system (LMS), giving students access to a variety of resources, homework, assessments, and gamified educational opportunities. Positive feedback from Kyiv National University students highlights the tool's ease of use and its potential to enhance educational experiences in HEIs through advanced AI capabilities. Even though the cur- rent version relies on button-based navigation, conversational AI elements are anticipated to be included in future releases.

[6] With an emphasis on ChatGPT, Samuli Laato, Benedikt Morschheuser, Juho Hamari, and Jari Björne's2023 study explores the revolutionary effects of large language models (LLMs) in education. It charts the evolution of LLMs from the launch of the Transformer architecture in 2017 to Open AI's development of GPT models. With features like text and code generation, summarization, and interactive dialogue, systems like ChatGPT—which are built on LLMs—allow students to access expert-level knowledge and participate in reflective learning.

The authors evaluated ChatGPT's impact over a two- month period in order to perform practical examination of its role in a computer science Bachelor's degree program at a Finnish institution. They list 13 important ramifications, such as the improvement of critical thinking abilities and possible hazards, such as an over-dependence on AI to write code and essays. Although ChatGPT has several benefits, the authors also highlight some drawbacks, like its propensity to" hallucinate" or generate inaccurate information. They address moral dilemmas, such as plagiarism, and advocate for a well- rounded approach to AI-assisted learning, emphasizing the significance of responsible integration in academic settings.

[7] Artificial intelligence is revolutionizing the educational environment at very rapid rates, accompanied by tremendous benefits and many challenges that come with it. There are five areas in which the challenge cuts across: user experience, operational demands, environmental impact, technological limitations, and ethical concerns. Systematic problems can be handled using a review process in formulating an explicit research question, thus leading the criteria for selecting relevant literature. Firstly, will be the preliminary planning which includes finding of keywords to form a strategy for retrieving relevant studies within the largest and most established academic databases including Science Direct, IEEE, and Scopus .Using the inclusion/ exclusion criteria in filtering article selections based on focus relevance, and date of publications, anal most impossibly large pool will be shrunk down into aslightly more manageable size of only the highest-quality studies.

Three stages are found in review. The planning phase summarizes the research aims and criteria for inclusion and exclusion of an article. Articles are sifted through a multi-step protocol in the execution stage: keyword searches followed by screening of titles and abstracts, then full-text reviewing. Quality assessments ensure the study's relevance and rigour, further narrowing the selections to those with high levels of quality to be included. The final stage synthesizes the findings into categorized themes of challenges and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

strategies, providing a structured overview of the key obstacles and actionable recommendations on AI use in educational settings. Thus, this systematic review framework is robust enough to be explored in AI's role in education and its implications.

[8] A structured approach to safeguarding implementation that deals with the ethical issues in large language models was proposed. In this proposal, it utilizes the MECE principle: it categorizes three broad areas so that it comprehensively and systemically addresses each ethical challenge: it starts with review ethics codes across professions; the nassessesethics awareness within computer science; then pin points where the LLM system safeguard points lay.

It breaks the LLM lifecycle down into upstream and downstream, explaining where ethical interventions come most into play. Controls such as input controls that operate through careful data curation are significantly effective for the avoidance of ethical risk but are usually abandoned because they raise transparency issues. Whereas controls that are further downstream like output filtering are being applied much more widely as it is cheaper and less resource-intensive but not so effective atthecore of addressing the ethical issues. Such an approach supports the notion of a proactive regulatory framework promoting responsible AI practices as LLMs are gaining fast adoptions.

[9] Differential Evolution based Fine-Tuning (DEFT) can optimize selection of layers in transfer learning for CNNs. DEFT is used for addressing challenges in fine- tuning efficiently, especially in areas such as medical imaging where large datasets are not available. Trans-fer learning enables pre-trained models that have been trained on vast datasets to be adapted to new but related tasks. However, choosing which layers to fine-tune in a CNN is not an easy task.

As empirical approaches are mostly not generalizable across tasks, deft employs the DE algorithm to do the layer selection automatically. Each candidate solution in the DE algorithm corresponds to a unique configuration of fine-tunable layers. The DE algorithm iteratively optimizes these configurations by training the CNN using a subset of target dataset and then analyze performance based on a fitness function, specifically categorical crossentropyloss. DEFT's adaptive layer selection mechanism identifies the best combination of layers. It involves striking a compromise between keeping generic features of pretrained layers and customizing them for the intended goal. This automated adaptive approach helps DEFT outperform traditional manual fine-tuning techniques in performance and reduces trial and error in layer selection.

[10] A mixed-method approach combining qualitative as well as quantitative techniques is used to analyse generative ais impact in education. Qualitative methods include case studies that observe how ai is being used for educational purposes in real-life, examining both benefits such as developing critical thinking and challenges such as educators resistance to change. Surveys and interviews are taken along with issues related to reliance on ai and privacy. Concurrently, through sources such as YouTube there is an understanding of peoples views related to the issue of ai in teaching. On the quantitative side, it keeps an eye on current studies in the sphere of ai for educational puposes to pinpoint key areas and track emerging technologies. The analysis by topic modelling further reveals public discourse on ai in education and evolving themes. The study further concludes that ai will make collaboration and critical thinking easy but should complement traditional teaching methods. Although it brings about personalization of learning opportunities, it raises ethical concerns about data collection and biases within the algorithms. Teachers beliefs and technological competency will influence their readiness in adopting ai for learning. AI should be used responsibly in education taking into account ethical concerns, teacher empowerment and the evolving role of AI inclassroom.

III. DISCUSSION

The ten examined papers show notable advancements in the use of AI-driven technologies and large language models (LLMs) in research and teaching. The creation of generative LLMs like ChatGPT, which are revolutionizing higher education by improving student learning experiences and offering educators cutting-edge resources, is one of the major developments. Research on the use of chatbots and LLMs in academic contexts, such as those conducted by Yehorchenkov etal. and Laato et al., highlights the tools' capacity to automate chores, promote critical thinking, and increase accessibility to learning materials.

The literature is replete with issues like an excessive dependence on AI-generated content and moral dilemmas. The necessity of responsible AI use and regulatory frameworks is covered in papers like Berengueres (2024) and Wong and Looi (2024), which emphasize the significance of data quality, system openness, and ethical safeguards. Furthermore, Vrbanci c and Podgorelec's exploration of adaptive fine-tuning is recognized as a crucial strategy for enhancing LLM performance

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 11 Nov. 2024, pp: 615-620 www.ijaem.net ISSN: 2395-5252

in specialized fields like medical imaging and teaching. These developments suggest that LLMs will eventually be extensively in corporate into research and educational settings, encouraging creativity while also needing rig- orous evaluation of practical and ethical issues.

IV. CONCLUSION

The developments in LLMs and AI-powered learning tools covered in this review demonstrate the increasing influence of these resources on research and higher education. AI-assisted learning platforms, chatbots, and adaptive fine-tuning are improving student engagement and expediting academic procedures. But there are important issues that need to be addressed, including moral dilemmas, the danger of relying too much on AI-generated material, and the requirement for strong regulatory frameworks.

Future research should concentrate on improving LLM performance, especially in domain-specific applications, ethical protections, and responsible AI use. Maximizing the advantages of these technologies in education and beyond will also need strengthening integration with current educational systems and creating plans to balance AI and human involvement.

REFERENCES

[1] Huang, Chen, and Guoxiu He. "Text Clustering as Classification with LLMs." arXiv preprint arXiv:2410.00927 (2024).

[2] Kortemeyer,Gerd."TailoringChatbotsforHigherEdu- cation: Some Insights and Experiences." arXiv preprint arXiv:2409.06717 (2024).

[3] Filippo, Chiarello, Giordano Vito, Spada Irene, Barandoni Si- mone, and Fantoni Gualtiero. "Future applications of generative large language models: A data-driven case study on ChatGPT." Technovation 133 (2024): 103002.

[4] Vishnumolakala,SaiKrishna,C.C.Sobin,N.P. Subheesh,Prab- hatKumar,andRandhirKumar."AI-BasedResearchCompanion (ARC): An Innovative Tool for Fostering Research Activities in Undergraduate Engineering Education." In 2024 IEEE Global Engineering Education Conference (EDUCON), pp. 1-5. IEEE, 2024.

[5] Yehorchenkov, Oleksii, Nataliia Yehorchenkova, and L'ubom´ır Jamec˘ny´."DigitalProfessor":InteractiveLearningwithChatbot Technology." In 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST), pp. 79-83. IEEE, 2023.

[6] Laato, Samuli, Benedikt Morschheuser, Juho Hamari, and Jari Bjo¨rne. "AI-assisted learning with ChatGPT and large lan- guage models: Implications for higher education." In 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 226-230. IEEE, 2023.

[7] Ali, Omar, Peter A. Murray, Mujtaba Momin, Yogesh K. Dwivedi, and Tegwen Malik. "The effects of artificial intelligenceapplicationsineducationalsettings:Challengesandstrate- gies." Technological Forecasting and Social Change 199 (2024): 123076.

[8] Berengueres,Jose."HowtoRegulateLargeLanguageModelsfor Responsible AI." IEEE Transactions on Technology and Society (2024).

[9] Vrbanc˘ic´, Grega, and Vili Podgorelec. "Transfer learning with adaptive fine-tuning." IEEE Access 8 (2020): 196197-196211.

[10] Wong, Lung-Hsiang, and Chee-Kit Looi. "Advancing the gen- erative AI in education research agenda: Insights from the Asia-Pacific region." Asia Pacific Journal of Education 44, no. 1 (2024): 1-7.

References

[1] Huang, Chen, and Guoxiu He. "Text Clustering as Classification with LLMs." arXiv preprint arXiv:2410.00927 (2024).

[2] Kortemeyer, Gerd. "Tailoring Chatbots for Higher Education: Some Insights and Experiences." arXiv preprint arXiv:2409.06717 (2024).

[3] Filippo, Chiarello, Giordano Vito, Spada Irene, Barandoni Simone, and Fantoni Gualtiero. "Future applications of generative large language models: A data-driven case study on ChatGPT." Technovation 133 (2024): 103002.

[4] Vishnumolakala, Sai Krishna, C. C. Sobin, N. P. Subheesh, Prabhat Kumar, and Randhir Kumar. "AI-Based Research Companion (ARC): An Innovative Tool for Fostering Research Activities in Undergraduate Engineering Education." In 2024 IEEE Global Engineering Education Conference (EDUCON), pp. 1-5. IEEE, 2024.

[5] Yehorchenkov, Oleksii, Nataliia Yehorchenkova, and L'ubomír Jamečný. "'Digital Professor': Interactive Learning with Chatbot Technology." In 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST), pp. 79-83. IEEE, 2023.

[6] Laato, Samuli, Benedikt Morschheuser, Juho Hamari, and Jari Björne. "AI-assisted learning with ChatGPT and large language models: Implications for higher education." In 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 226-230. IEEE, 2023.

[7] Ali, Omar, Peter A. Murray, Mujtaba Momin, Yogesh K. Dwivedi, and Tegwen Malik. "The effects of artificial intelligence applications in educational settings: Challenges and strategies." Technological Forecasting and Social Change 199 (2024): 123076.

[8] Berengueres, Jose. "How to Regulate Large Language Models for Responsible AI." IEEE Transactions on Technology and Society (2024).

[9] Vrbančič, Grega, and Vili Podgorelec. "Transfer learning with adaptive fine-tuning." IEEE Access 8 (2020): 196197-196211.

[10] Wong, Lung-Hsiang, and Chee-Kit Looi. "Advancing the generative AI in education research agenda: Insights from the Asia-Pacific region." Asia Pacific Journal of Education 44, no. 1 (2024): 1-7.

[11] Diana Liz Kuriakose, Mohammed Hisham, Maria Joshy, Syama S "Survey on Generative AI in Education", International Journal of Advances in Engineering and Management (IJAEM), Volume 6, Issue 11, Nov. 2024, pp: 615-620, https://www.ijaem.net/issue_dcp/Survey%20on%20Generative%20AI%20in%20Education.pdf