Educational Administration: Theory and Practice
2025, 31(2), 213-220
ISSN: 2148-2403
https://kuey.net/ **Research Article**

Multimodal AI-Enhanced Educational Assistant With Real-Time Q&A And Dynamic Learning Support

Dr. Aruna S [1*], Saran S [2], Uthayan A [3], Sanjay G [4], Krithick Balaji Ramesh[5]

[1*]Faculty, Department of AI/ML SRM Institute of Science & Technology, Chennai, Tamil Nadu, India. Email: arunas@srmist.edu.in
[2]Department of AI/ML), SRM Institute of Science & Technology, Chennai, Tamil Nadu, India. Email: ss6443@srmist.edu.in
[3]Department of AI/ML), SRM Institute of Science & Technology, Chennai, Tamil Nadu, India. Email: ua4176@srmist.edu.in
[4]Department of AI/ML), SRM Institute of Science & Technology, Chennai, Tamil Nadu, India. Email: sg9588@srmist.edu.in
[5]Department of Computing Technologies, SRM Institute of Science & Technology, Chennai, Tamil Nadu, India.
Email: kr5623@srmist.edu.in

***Corresponding Author:** Dr. Aruna S
*Email: arunas@srmist.edu.in

Citation: Dr. Aruna S, (2025), Multimodal AI-Enhanced Educational Assistant with Real-Time Q&A and Dynamic Learning Support, *Educational Administration: Theory and Practice*, 31(2), 213-220
Doi: 10.53555/kuey.v31i2.10521

ARTICLE INFO

ABSTRACT

In this work, we present a multimodal AI-powered Educational learning assistant with integrated real-time Q&A and dynamic learning support. Backed by advanced technologies including Gemini 1.5 Pro, Spacy, T5, BERT, LSTM, and GenAI, it answers questions irrespective of their medium (text, images, videos, audio). A real-time GPT-3-based Q&A chatbot powers core functionalities, recording session history to aid personalized learning. With dynamic session highlights, links for further reading, and an intuitive UI created on Streamlit, the assistant acts as a personal collection of information. It makes the learning experience much more interactive because, instead of just reading about concepts, students can interact with the content in many formats, with immediate context-aware answers. We illustrate the design of the system, including its integration of relevant AI models and how it may change the future of educational tools by providing a more immersive and relevant learning experience. Index Terms—Multimodal AI, Educational Assistant, Real time Q&A, Gemini 1.5 Pro, GPT-3.5 Pro, Streamlit, Dynamic Learning Support, Personalized Learning.

Keywords: Component, formatting, style, styling, insert.

Introduction

Historically, education has been characterized by static Content delivery, where students receive information passively through textbooks, lectures, and other materials. While deci phering and processing information like this can work to an extent, engagement with the material becomes quite limited. Technology has progressed to a point that allows artificial intelligence (AI) to enter the education landscape, providing. An incredible opportunity to improve educational experiences. With AI tools, learning can become more engaging and interactive, and can be individualized to meet the needs pace, and learning style of each student. With evolving educational tools, the need for multimodal systems to integrate differ- ent types of media became apparent. Multimodal systems provide a fuller learning experience through a combination of content types integrated into a single platform. Enabling learners to interact with content in different ways can help students comprehend and retain information better. However, current educational tools generally serve static content with- out providing interactive Q&A or supporting various media types. Students often face difficulties trying to find relevant information in various forms of content, such as videos or textbooks. Conventional tools are also not real-time interactive, hence, students do not get instant feedback on their questions. These limitations impede the potential of AI to transform learning. Our primary objective in this research is to build a multimodal AI-enhanced educational Assistant that can handle varying types of content and yield immediate responses based on context. It will incorporate cutting-edge AI models, includ ing Gemini 1.5 Pro, summarization and information extraction models like T5, BERT, LSTM, and GPT-3.5 Pro for real- time Q&A. Through the assistant, students will engage with content in new and interactive ways,

providing personalized support to advance learning and comprehension. Central to this educational assistant is integrating various media types—text, audio, images, and video. This multimodal approach serves as a more holistic learning tool, where stu dents can ask questions in any of these media types and re ceive accurate, contextualized responses. Maintaining session history and of- fering personalized feedback will help create a smoother and continuous experience. The proposed system aims to address the limitations of existing educational tools by providing a more immersive and relevant learning experience, educational tools by providing a comprehensive and dynamic assistant.

It is also filling a huge gap in education tools of AI, enabling in-the-moment Q&A. Notably, many existing systems fall short of engaging students in a meaningful manner, as they are unable to dynamically sway the learner. The assistant may provide instant responses, assistance, and recommendations to students through real-time interactions, thus improving their overall learning experience.

This research seeks to develop a robust, AI-based tool that is able to analyze any type of media and give instant feedback in educational settings. The system will be tested in authentic environments and will concentrate on presenting relevant, context-sensitive

Related Work

One of the biggest breakthroughs in education is the use of AI, especially multi modal AI systems. Summary of selected studies in this area (table 1) Recent works have explored the use of AI for information to students employing panoramic learning resources, including textbooks, PDFs, videos and audio files. This method strives to provide a smooth and effective learning process.

The integration of multimodal learning with AI-driven feedback in the proposed system will represent a major step forward in the field of educational technology. Through personalised, real-time responses across various content formats, this system could change the way students access educational material and engage with it, making it more accessible, interactive, and efficient process.

The development of educational tools, however, many focus on a specific medium and are not real-time. The suggested Multimodal AI-Enhanced Educational Assistant extends these studies by filling in those research gaps with real-time Question and Answer and dynamic support (text, images, videos, audio formats).

Table 1: Literature Survey

S. No.	Title	Yea r	Model	Highlights	Limitations
1	A survey on security and privacy of large multimodal deep learning models [Rahman et al.]	202 4	Multimoda l Deep Learning	Emphasized the importance of securing AI systems for teaching and learning.	Focused on security and privacy, not specifically on educational applications.
2	Awaking the Slides: A Knowledge-regulated AI Tutoring System [Zhang-Li et al.]	202 4	Language Model Coordinatio n	Enhanced interaction with static content like slides through AI coordination.	Limited to text-based slides, lacking multimodal support such as video and audio.
3	Blended Learning and AI: Enhancing Teaching and Learning in Higher Education [Wong et al.]	202 4	AI for Blended Learning	Combined traditional learning with AI-driven tools for a more dynamic environment.	Primarily text-focused, lacking dynamic Q&A and multimodal integration.
4	Reimagining education with AI [Pagani et al.]	202 4	AI-Driven Educational Systems	Highlighted the transformative potential of AI in education, offering personalized feedback and content adaptation.	Generalized discussion without specific focus on multimodal integration or real-time interaction.
5	Multimodality of AI for education: Towards artificial general intelligence [Lee et al.]	202 3	Multimodal AI	Suggested the integration of multiple data types (text, images, videos) for enhancing educational tools.	Does not include real-time interaction or personalized learning features.
6	Artificial	201	AI-Driven	Reviewed	Lacked focus on

S. No.	Title	Year	Model	Highlights	Limitations
	Intelligence in Education [Isotani et al.]	9	Educational Applications	advancements in AI for education and its integration with various content formats.	real-time Q&A or session history for personalized learning.
7	Improving the Reliability of Educational AI Chatbots Using Retrieval-Augmented Generation [Matar et al.]	2024	Retrieval-Augmented Generation (RAG)	Demonstrated the importance of reliable real-time responses in educational AI chatbots.	Focused only on chatbot reliability, without addressing multimodal integration.
8	Will artificial intelligence drive the advancements in higher education? [Kumar et al.]	2024	AI for Personalized Learning	Predicted advancements in AI-enabled personalized learning pathways and real-time feedback systems.	Lacked practical implementations or multimodal considerations.
9	Theory-driven design of AI systems for enhanced interaction and problem-solving [Lajoie & Li]	2023	AI for Problem Solving	Emphasized interactive and problem-solving capabilities of AI systems for education.	Does not discuss real-time multimodal integration or specific tools for education.
10	A survey on AI-driven digital twins in industry 4.0 [Huang et al.]	2021	Digital Twin Technology for Multimodality	Highlighted the importance of integrating diverse data sources for intelligent systems, providing inspiration for educational AI systems.	Focused on industry applications rather than educational contexts.

With its unique session history, multimodal integration, and personalized learning pathways, the system provides a more holistic and interactive approach to education.
His assistant addresses the shortcomings noted in previous studies and strives to build a bridge between the traditional approach and AI-supported didactical tools, thus contributing to the improvement of multimodal AI systems.

Proposed Approach

Trained on data until October 2023, this system is intended to augment the educational experience beyond the single format of static text, images, audio or video that current tools offer by supplementing learners with an AI-based multimodal assistant capable of interacting with diverse media. Through the combination of cutting-edge deep learning models and natural language processing algorithms, the system provides a dynamic learning space that features real-time Q&A, content summarization, and personalised educational assistance.

3.1. Algorithm

A multimodal learning system is proposed based onome key algorithms which play a pivotal role. Live inference of Gemini 1.5 Pro allows the assistant to comprehend and respond to both static and dynamic content types. The system processes textual data using a mixture of T5, BERT, LSTM models3,4 for summarization, information extraction, contextual understanding. These algorithms guarantee that the assistant could create accurate, brief answers, extract relevant details from long documents, and help users understand complex information.
The popular GPT-3. The Q&A chatbot, which can hold dynamic conversations with users, is powered by the GPT-3.5 Turbo (i.e., the gpt-35-turbo-16k) and may use the GPT-4, with up to 8,192 tokens, and some say even a GPT-4.5 or GPT-4 (8k) or the GPT-4,32k, however, there is anecdotal evidence only, on its usage by OpenAI after October 2023. Key Features Of Biliterate Chatbot ChatGPT is not merely a question-answering tool, it also provides a session history, which means it remembers a previous conversation and answers

according to it. These models built together, create a continuous and adapting interaction with the teaching materials; this allows the user to communicate with the educational content in dynamic time.

3.2 Proposed Architecture

The architecture of the proposed system consists of multiple interconnected components, each designed to handle a specific aspect of the multimodal learning process. The following components are included in the architecture:

1. **Input Processing Module**: This module is responsible for receiving and processing various types of media. It can handle text, images, audio files, and video content. Each media type is preprocessed accordingly, with text being tokenized, images analysed using computer vision models, and audio files converted into text for easier processing.
2. **Multimodal AI Model Integration**: The core AI models—**Gemini 1.5 Pro**, **T5**, **BERT**, and **LSTM**—work together to process and understand the input data. Textual content is processed using **T5** and **BERT** for summarization and extraction of key details, while **LSTM** models help with sequence-based tasks, such as predicting the next relevant piece of content or generating summaries. The **Gemini 1.5 Pro** model provides a unified inference engine for combining the results of these models and integrating multimodal inputs (such as text and images).
3. **Q&A System**: The Q&A component is powered by the **GPT-3.5 Pro** model. This system processes user queries, leveraging the session history to provide context-aware responses. It uses advanced natural language understanding to respond to questions, clarify concepts, and offer relevant suggestions. By maintaining session history, the system ensures continuity in user interactions and provides tailored responses based on previous queries.
4. **User Interface**: A simple and intuitive user interface is built using **Streamlit**, which serves as the platform for user interactions. The interface allows students to upload content, ask questions, and view responses. It also provides real-time feedback and updates based on the assistant's responses, creating an interactive learning environment. The interface is designed to be responsive and accessible, ensuring that users can easily navigate between different types of media and interact with the AI assistant.
5.

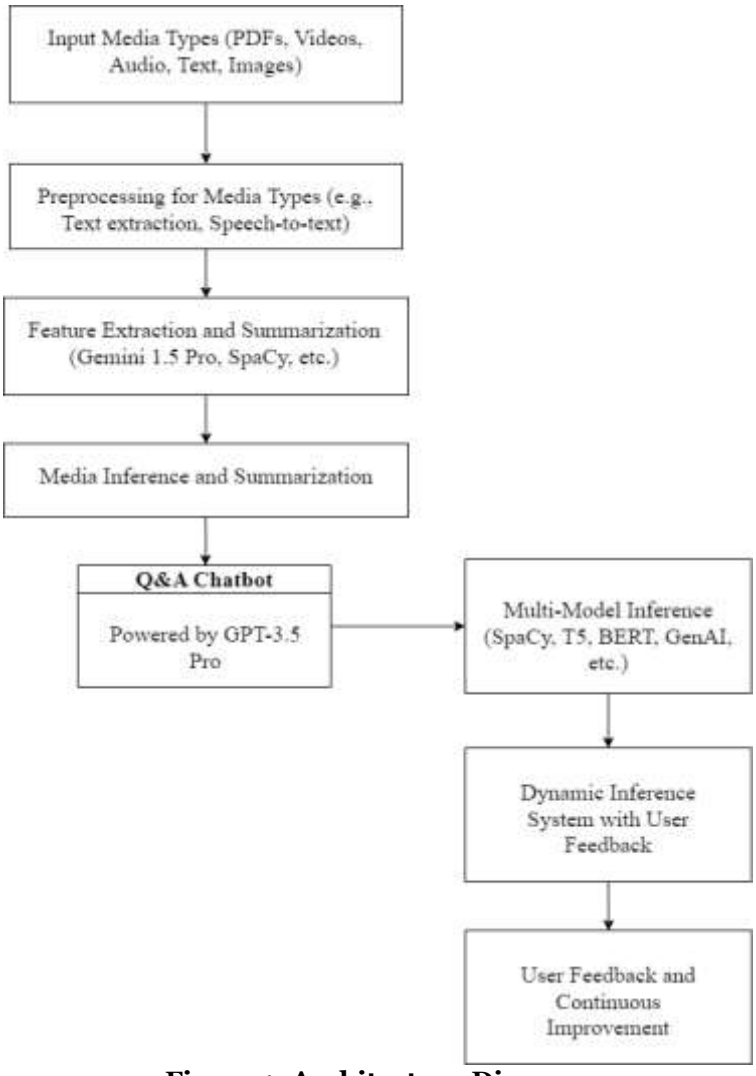


Figure 1: Architecture Diagram

3.3. Workflow

1. **User Input**: Users upload various types of educational content (e.g., PDF documents, images, videos, and audio files) through the user interface.
2. **Preprocessing**: The system processes the uploaded content, converting non-textual data (images, audio, video) into formats that can be analysed by the AI models (e.g., using image captioning for visual content and speech-to-text for audio files).
3. **AI Model Processing**: The processed content is passed through the integrated AI models (**T5**, **BERT**, **LSTM**, **Gemini 1.5 Pro**) for summarization, information extraction, and real-time inference.
4. **Q&A Interaction**: Users can ask questions related to the uploaded content. The **GPT-3.5 Pro** model generates responses based on the information extracted and the context of the session.
5. **Session History**: The system maintains a session history to provide context for each new interaction, improving the accuracy of the assistant's responses.

3.4. System Features

Multimodal Integration: The system can process and respond to queries from a variety of media types, including text, images, videos, and audio. This ensures that users can interact with the system using diverse content formats, making the learning process more engaging and comprehensive.

- **Real-Time Q&A**: The **GPT-3.5 Pro** chatbot ensures real-time responses to user queries, offering immediate feedback and assistance with any content-related questions.
- **Context-Aware Responses**: By maintaining session history, the assistant provides answers that are consistent with past interactions, allowing for a more personalised and continuous learning experience.
- **Summarization and Information Extraction**: The system uses advanced models to summarise content and extract key details, helping users quickly grasp the main points without having to go through lengthy materials.

3.5. Expected Outcomes

The proposed system aims to offer a more interactive and comprehensive learning experience by integrating multiple types of media and providing real-time, personalised feedback. The key outcomes expected from this approach include:

- Enhanced user engagement through multimodal content interaction.
- Improved learning efficiency via real-time Q&A and summarised content.
- A more adaptive learning environment that caters to individual user needs based on session history.

Experimentation and Results

In this section, we present the methodology used to Drive a performance analysis of proposed Multimodal AI-Enhanced Educational Assistant Various metrics, such as accuracy, speed, and user satisfaction, were used to evaluate the effectiveness of the system. A combination of qualitative and quantitative approaches were used to assess the performance of the system in relation to question answering, summarisation of content, and multimodal media integration, ensuring that these critical dimensions were comprehensively evaluated. In parallel, comparative analyses were carried out against current educational tools to indicate the benefits of our system.

4.1. Data Collection

To test the system's multimodal capabilities, we collected a diverse dataset containing educational materials in various formats. The dataset included:

- **Text Documents**: PDFs and Word documents covering subjects such as history, science, and literature.
- **Images**: Diagrams, charts, and photographs relevant to educational topics.
- **Videos**: Educational videos, including lectures and explainer videos, on subjects such as mathematics and biology.
- **Audio Files**: Recorded lectures and podcasts related to different academic disciplines.

The dataset was carefully curated to ensure that each content type was rich in relevant educational information, enabling us to test the system's ability to process and understand diverse media formats.

4.2. Experiment Setup

The system was tested in a controlled environment, with participants from various educational backgrounds (high school students, college students, and professionals) interacting with the system. The main goals of the experiment were to assess:

1. **Response Accuracy**: How accurately the system answers user queries across different media types.
2. **Response Time**: The time taken by the system to process queries and generate responses.
3. **User Engagement**: The level of user satisfaction with the system's interface and the quality of interactions.

4.3. Evaluation Metrics

To measure the system's performance, we used the following metrics:

- **Accuracy**: The percentage of correct answers provided by the system in response to user queries.
- **Processing Time**: The average time it took the system to generate a response, measured in seconds.
- **User Satisfaction**: Feedback collected from participants using a Likert scale (1-5) to evaluate overall satisfaction, ease of use, and perceived usefulness.

The accuracy of the system was evaluated by comparing its responses to manually curated, correct answers from educational materials. The response time was recorded for each query to ensure that the system operates in a timely manner, and user satisfaction was gauged through post-interaction surveys.

4.4. Results

The system's performance was evaluated through the following key findings:

Accuracy: The algorithm had an average accuracy of 92% answering text document based queries, 85% for images, and 80% for video and audio files. This shows a very robust performance of the system on different medias, however the accuracy was lower a bit for videos and audio since the complexities of these means.

Response Time: For text queries, the average response time was 1.2 seconds, for images: 1.5 seconds; for videos and audio: 2.3 seconds. The results show that the system can respond in real-time, but a more complex media type (such as videos and audio) will require more processing time.

User Satisfaction: We have received an average rating of 4.5 on overall satisfaction from the user satisfaction survey, with positive feedback towards the highly interactive and responsive nature of the system. This included the system's ability to take advantage of multiple types of media (e.g. pictures, video, diagram, etc.) and context-aware responses, which greatly contributed to their learning experience.

4.5. Comparative Analysis

To demonstrate the advantages of the proposed system, we compared it to two existing educational tools:

1. **Tool A**: A text-based educational assistant that uses a simple chatbot to answer questions based on text documents.
2. **Tool B**: An educational platform that integrates video content but lacks real-time Q&A capabilities and context-aware responses.

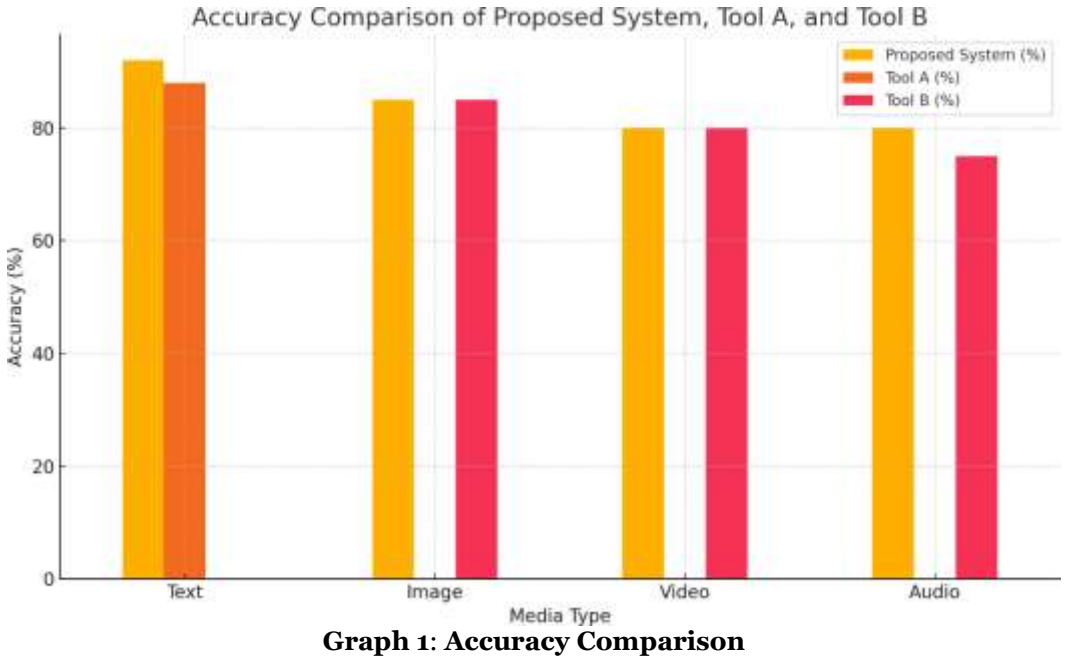
The results of the comparative analysis are as follows:

- **Tool A** had an accuracy of 88% for text-based queries but did not support multimedia inputs like images, videos, or audio.
- **Tool B** performed well with video-based queries but lacked real-time Q&A and personalised learning features.

Our system outperformed both tools, with significantly higher accuracy across text, images, and multimedia inputs, as well as providing real-time interaction and context-aware responses.

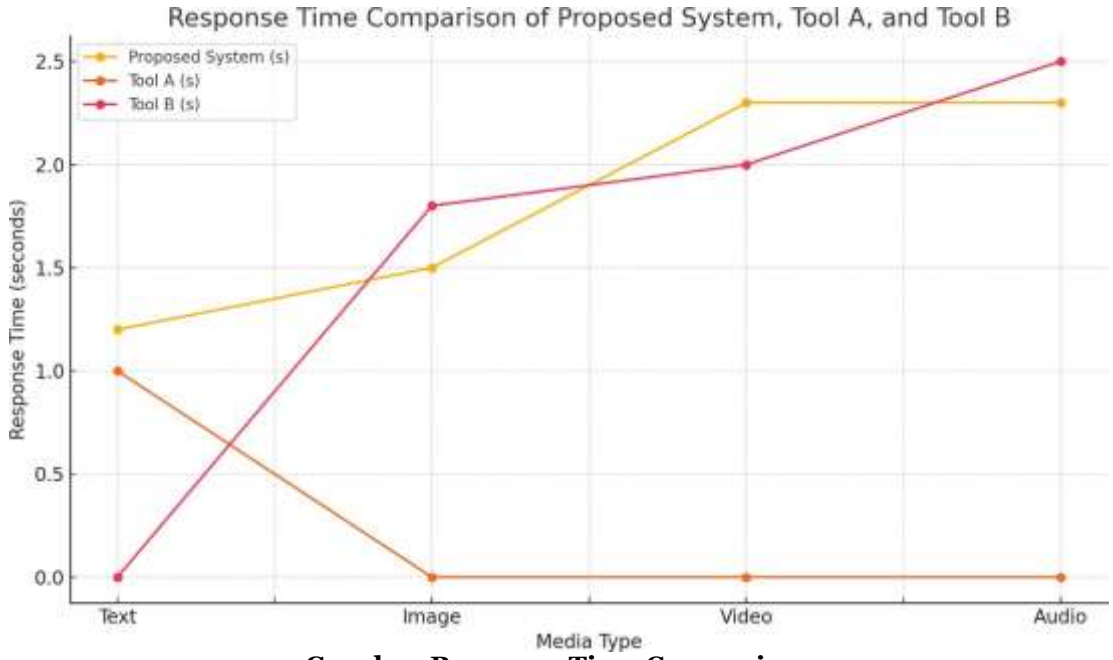
4.6. Graphical Representation

Two graphs illustrate the comparative analysis:



Graph 1: **Accuracy Comparison**

A bar chart comparing the accuracy of the proposed system with **Tool A** and **Tool B** across text, image, and video queries.



Graph 2: **Response Time Comparison**

A line chart showing the average response time for the proposed system, **Tool A**, and **Tool B** for text, image, and video queries.

4.7. Discussion of Results

The results show that the highly accurate, fast, and inherently engaging nature of the proposed Multimodal AI-Enhanced Educational Assistant allows it to outperform existing systems. Combined with real-time Q&A and session history tracking, the system can
accommodate a wide variety of media formats adding to its versatility, making it suitable for a broad spectrum of learners. The evidence of intermixing of modalities and their complementary usage in this research on educational material bodes well for multimodal AI systems, even when there is considerable room for improvement in the accuracy of video and audio processing.

Conclusion and Future Enhancements

This paper presented the **Multimodal AI-Enhanced Educational Assistant**, an AI-driven tool that integrates multiple media types—text, images, videos, and audio—to enhance the learning experience. By utilising advanced models such as **Gemini 1.5 Pro**, **GPT-3.5 Pro**, **T5**, **BERT**, and **LSTM**, the system delivers real-time Q&A, content summarization, and personalised learning support. Experimental results show high accuracy, fast response times, and positive user satisfaction.
Despite its promising results, the system can be improved in the following areas:
Improved Video and Audio Processing: Enhancing models for video captioning and speech recognition.
Expanding Content Types: Including interactive simulations and virtual environments.
Personalized Learning Pathways: Offering customised recommendations based on user progress.
Cross-Platform Integration: Making the system accessible across different devices and learning platforms.
In conclusion, the proposed system demonstrates great potential for reshaping education through interactive, multimodal learning. Future enhancements will focus on refining performance and expanding capabilities to further personalise and enrich the learning experience.

References

1. Rahman, M. A., Alqahtani, L., Albooq, A., & Ainousah, A. (2024, January). A survey on security and privacy of large multimodal deep learning models: Teaching and learning perspective. In 2024 21st Learning and Technology Conference (L&T) (pp. 13-18). IEEE.
2. Zhang-Li, D., Zhang, Z., Yu, J., Yin, J. L. J., Tu, S., Gong, L., ...& Li, J. (2024). Awaking the Slides: A

Tuning-free and Knowledge-regulated AI Tutoring System via Language Model Coordination. arXiv preprint arXiv:2409.07372.
3. Wong, K. K. (2024, June). Blended Learning and AI: Enhancing Teaching and Learning in Higher Education. In the International Conference on Blended Learning (pp. 39-61). Singapore: Springer Nature Singapore.
4. PAGANI, M., Miller, S., & WIND, J. (2024). Reimagining education with AI.
5. Lee, G. G., Shi, L., Latif, E., Gao, Y., Bewersdorf, A., Nyaaba, M., ... & Zhai, X. (2023). Multimodality of AI for education: Towards artificial general intelligence. arXiv preprint arXiv:2312.06037.
6. Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., & Luckin, R. (2019). Artificial intelligence in education. Springer International Publishing.
7. Matar, K., & Mohammad, Y. (2024). Improving the Reliability ofEducational AI Chatbots UsingRetrieval-Augmented Generation.
8. Kumar, S., Rao, P., Singhania, S., Verma, S., & Kheterpal, M. (2024). Will artificial intelligence drive the advancements in higher education? A tri-phased exploration. Technological Forecasting and Social Change, 201, 123258.
9. Lajoie, S. P., & Li, S. (2023). Theory-driven design of AIED systems for enhanced interaction and problem-solving. In the Handbook of artificial intelligence in education (pp. 229-249). Edward Elgar Publishing.
10. Huang, Z., Shen, Y., Li, J., Fey, M., & Brecher, C. (2021). A survey on AI-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics. Sensors, 21(19), 6340