International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Subject-Aware Generative Educational Agent

Mrs. Jayavani R/ME [1], Bandi Sathwika [2], Ganapathi Rithika [3], Nidhi Tiwari S [4], Yukesh S [5]

[1,2,3,4] Dept. of AI & DS
T.J.S. Engineering College Chennai, India

Abstract- Personalised learning systems are becoming essential in modern education due to the increasing demand for adaptive and context-aware knowledge delivery. However, most traditional learning platforms lack subject-level isolation and fail to effec-tively utilize user-provided study materials. To address these limitations, this paper proposes SAGE.AI (Subject-Aware Gen-erative Educational Agent), an intelligent system that integrates Retrieval-Augmented Generation (RAG) with Large Language Models (LLaMA), guided through structured prompt engineer-ing, to deliver subject- specific and context-aware responses. The system allows users to dynamically create subjects, upload PDF-based learning resources, and interact through a dual-interface design consisting of a PDF analysis panel and a chat interface. The RAG pipeline extracts and processes relevant content from uploaded documents, while carefully designed prompts ensure that the LLM generates responses strictly constrained to the selected subject. This approach enforces subject isolation and prevents cross-domain response leakage. By combining document understanding, retrieval-based context injection, and controlled generative responses, the proposed system enhances learning efficiency and personalisation. SAGE.AI can be deployed as a scalable web-based educational assistant for students, profession-als, and self-learners.

Index Terms—Retrieval-Augmented Generation (RAG), LLaMA, Prompt Engineering, Subject-Aware AI, PDF-Based Learning, Generative AI.

I. INTRODUCTION

Modern educational systems increasingly demand person-alised, adaptive, and interactive learning environments. With the rapid advancement of artificial intelligence, learners expect systems that not only provide accurate information but also adapt to specific subjects and individual learning resources. However, most existing AI-based learning platforms generate.

Identify applicable funding agency here. If none, delete this. generalised responses without enforcing subject boundaries or incorporating user-provided materials such as notes or textbooks. Traditional approaches to digital learning often rely on static content delivery or manual instructional methods, which lack scalability and adaptability. Even recent AI-driven systems tend to function as generic assistants, leading to context dilution, where responses may mix information across different domains without maintaining subject- specific rel-evance. To overcome these limitations, this paper proposes SAGE.AI (Subject-Aware Generative Educational Agent), a system designed to deliver subject-restricted, context-aware learning experiences. The system introduces a structured ap-proach that combines:
label=◦

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Subject-specific interaction with strict isolation
- Dynamic utilisation of user-uploaded PDF learning ma-terials
- Retrieval-Augmented Generation (RAG) for contextual knowledge retrieval
- Controlled response generation using LLaMA guided by prompt engineering

Unlike conventional chat-based AI systems, SAGE.AI en-forces a subject isolation mechanism, ensuring that responses remain confined to the selected domain. At the same time, it leverages both retrieved document context and model knowl-edge to provide accurate and meaningful explanations. This approach enables a more focused and reliable learning experi-ence, where users can interact with AI in a manner that closely aligns with structured academic study. By integrating retrieval, generation, and subject control into a unified framework, SAGE.AI represents a significant step toward intelligent and personalized educational assistants.

II. PROJECT OVERVIEW

SAGE.AI (Subject-Aware Generative Educational Agent) is an intelligent educational system designed to provide per-sonalised, subject-specific learning assistance by integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLaMA). The system focuses on delivering accu-rate and context-aware responses while strictly maintaining subject isolation. The core functionality of SAGE.AI revolves around enabling users to dynamically create and interact with individual subjects. Each subject acts as an isolated learn-ing environment, ensuring that queries and responses remain confined to the selected domain. This eliminates the problem of cross-topic interference commonly observed in general-purpose AI systems. A key feature of the system is its ability to process user-uploaded PDF documents. When a user uploads a document, the system extracts and analyzes its content, which is then used as a contextual knowledge source. Through the RAG mechanism, relevant portions of the document are retrieved and injected into the response generation process, enabling more precise and meaningful answers. The system is designed with a dual-interface architecture, consisting of: label=◦

- A PDF analysis panel for document upload and content visualization
- A chat interface for interactive query-based learningTo ensure accuracy and control, prompt engineering tech-niques are applied to guide the LLaMA model.
- Responses are strictly limited to the selected subject
- Irrelevant or out-of-domain queries are restricted
- Context from uploaded documents is effectively utilized

The overall workflow of the system includes subject selec-tion, document ingestion, contextual retrieval, and controlled response generation. This structured pipeline enhances both learning efficiency and user experience.

SAGE.AI is implemented as a web-based application using a lightweight frontend (HTML, CSS, JavaScript) and a Flask-based backend, making it scalable and easy to deploy. The system is designed to support students, educators, and pro-fessionals by providing a focused and intelligent learning environment.

[User Input] → [Subject Selection] → [PDF Processing]
→ [RAG Module] → [Prompt Engineering] → [LLaMA Model] → [Subject Isolation] → [Response Output]

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

III. EXISTING SYSTEM

Existing educational and AI-based learning systems primar-ily focus on providing generalized responses without enforcing subject-specific boundaries. Traditional e-learning platforms rely on static content such as pre-recorded lectures, textbooks, and manually curated materials, which lack adaptability and real-time interaction. With the advancement of artificial intel-ligence, conversational agents and chatbot-based systems have been introduced to enhance user interaction. However, most of these systems function as general- purpose assistants, generat-ing responses based on broad knowledge without considering subject-level constraints. As a result, they often produce mixed or irrelevant answers when handling queries from different domains. Some modern systems incorporate document-based question answering, where users can upload files and ex-tract information. While these systems improve accessibility, they typically lack a structured retrieval mechanism and do not enforce strict contextual control. The absence of subject isolation leads to context leakage, where information from one topic may influence responses in another. Additionally, many existing solutions do not effectively utilise user-provided learning materials in a dynamic manner. They either: ◦

- Depend heavily on pre-trained knowledge
- Use basic keyword-based retrieval techniques
- Lack integration between document understanding and response generation

Unlike the deep learning-based architecture discussed in the reference paper, which focuses on image-based classification using CNNs , current educational AI systems often fail to com-bine retrieval, contextual reasoning, and controlled generation into a unified framework.

IV. DRAWBACKS OF EXISTING SYSTEM

- Lack of subject isolation, leading to mixed-domain re-sponses
- Limited utilisation of user-uploaded learning materials
- Absence of structured retrieval mechanisms (basic or no RAG)
- Over-reliance on generic pre-trained knowledge
- Poor contextual relevance in responses
- No controlled response generation (no prompt con-straints)
- Inefficient handling of multi-subject learning environ-ments

V. LITERATURE SURVEY

The rapid evolution of artificial intelligence has significantly influenced the development of intelligent educational systems. Various research efforts have focused on improving learning efficiency through machine learning, deep learning, and con-versational AI technologies. Early approaches in intelligent learning systems relied on rule-based systems and traditional machine learning algorithms, which required manual feature extraction and predefined knowledge structures. While these systems provided basic automation, they lacked adaptability and were unable to handle complex or unstructured data effectively. Recent advancements have introduced deep learn-ing models and Large Language Models (LLMs), which are capable of understanding and generating human-like text. These models are trained on large-scale datasets and can provide informative responses across a wide range of domains. However, as highlighted in existing studies, such models often suffer from lack of contextual grounding, leading to inaccurate or hallucinated responses when external knowledge is not properly integrated. To address this limitation, the concept of Retrieval-Augmented Generation (RAG) has been proposed. RAG combines information retrieval techniques with gener-ative models by fetching relevant data from

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

external sources and incorporating it into the response generation process. This approach improves accuracy and ensures that responses are grounded in real data rather than solely relying on pre-trained knowledge. Several research works have also explored the use of document-based question answering systems, where user-provided files such as PDFs are processed to extract meaning-ful information. While these systems enhance personalisation, many of them lack structured retrieval pipelines and do not en-force strict contextual boundaries, resulting in inconsistent out-puts.In parallel, studies in educational technology emphasise the importance of personalised and subject- specific learning environments. However, most existing AI-based educational assistants do not implement a mechanism to maintain strict subject isolation, often leading to cross-domain interference in responses. The reference system discussed in the provided paper demonstrates the effectiveness of deep learning in a different domain (plant disease detection using CNNs), where automated feature extraction improves accuracy and scalability.

Inspired by such advancements, modern AI systems aim to reduce manual intervention and enhance automation in knowledge processing tasks. Despite these developments, there remains a gap in integrating subject isolation, document-based retrieval, and controlled response generation into a unified educational platform. This gap motivates the development of SAGE.AI, which combines RAG, prompt engineering, and LLM-based reasoning to deliver accurate, context-aware, and subject-restricted learning experiences.

VI. OBJECTIVE OF THE PROJECT

The primary objective of the SAGE.AI (Subject-Aware Gen-erative Educational Agent) project is to develop an intelligent and adaptive learning system that delivers accurate, subject-specific, and context-aware responses by integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLaMA).The system is designed to overcome the limitations of existing AI-based learning platforms by introducing a sub-ject isolation mechanism, ensuring that each subject operates independently without interference from other domains. This enables users to focus on a single area of study and receive highly relevant responses. The key objectives of the project are as follows:

- To design a system that allows dynamic creation and management of subjects, providing a structured learning environment
- To implement strict subject-based query handling, pre-venting cross-domain responses and maintaining contex-tual relevance
- To enable PDF-based learning, where users can upload documents and extract meaningful information
- To integrate Retrieval-Augmented Generation (RAG) for retrieving relevant content from user-provided materials
- To utilize prompt engineering techniques to control and guide LLaMA in generating accurate and subject-restricted responses
- To develop a dual-interface system consisting of a PDF analysis panel and a chat interface for improved usability
- To enhance learning efficiency and personalisation through contextual and interactive AI-based assistance
- To build a scalable and user-friendly web application suitable for students, educators, and professionals

Overall, the project aims to create a reliable, intelligent, and focused educational assistant that bridges the gap between traditional learning methods and modern AI-driven systems. Why this is strong:

- Clearly lists measurable objectives

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Matches your actual implementation
- Highlights your innovation (subject isolation + RAG + prompt control)
- Clean academic structure (good for marks)

VII. SCOPE OF THE PROJECT

The scope of SAGE.AI (Subject-Aware Generative Educa-tional Agent) focuses on developing a scalable and intelli-gent educational platform that enhances personalised learning through subject isolation, document- based knowledge integra-tion, and AI-driven interaction.

VIII. METHODOLOGY

The SAGE.AI system follows a modular methodology that integrates subject management, document processing, retrieval mechanisms, and controlled response generation to provide a structured and intelligent learning experience. The overall workflow is divided into multiple functional modules, each responsible for a specific stage of the system.

VIII.1. Subject Management Module

This module is responsible for handling the creation, se-lection, and maintenance of subjects. Users can dynamically create subjects based on their learning needs, and each subject operates as an independent and isolated environment. Once a subject is selected, it is stored and used as a constraint for all subsequent interactions. This ensures that:

- Queries are restricted to the selected subject
- Responses remain contextually relevant
- Cross-domain interference is avoided

This module forms the foundation of the system by enforcing the subject isolation mechanism, which is the core innovation of SAGE.AI.

VIII.2. PDF Processing Module

The PDF Processing Module enables users to upload their own learning materials in the form of PDF documents. Upon upload, the system performs:

- File validation and storage
- Text extraction using PDF parsing techniques
- Cleaning and structuring of extracted content

The extracted text is stored temporarily and serves as a knowledge source for further processing. This module allows the system to move beyond static knowledge and incorporate user-specific learning resources.

VIII.3. Retrieval-Augmented Generation (RAG) Module

The RAG module enhances the system's ability to generate accurate and context-aware responses by integrating retrieval mechanisms with generative models. The process includes:

- Breaking extracted document text into manageable seg-ments
- Identifying relevant portions based on user queries
- Supplying retrieved context to the language model

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

This ensures that responses are grounded in actual document content rather than relying solely on pre-trained knowledge. Although the current implementation uses simplified retrieval, it establishes the foundation for advanced vector-based re-trieval systems.

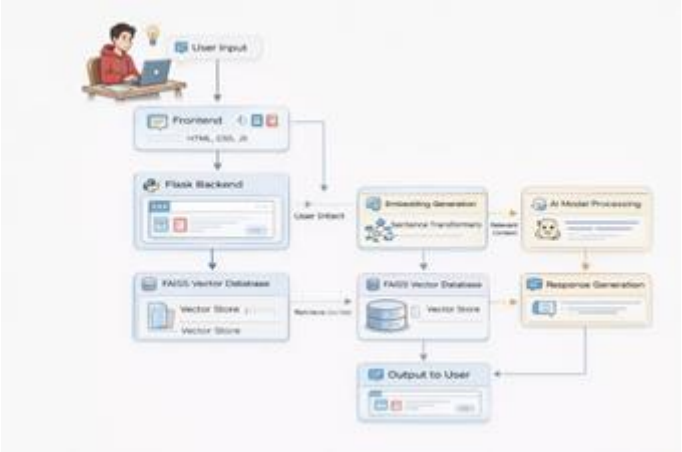

Fig. 1. RAG Module Workflow

VIII.4. Prompt Engineering and Response Generation Module

This module controls how the Large Language Model (LLaMA) generates responses. Instead of allowing unrestricted generation, structured prompts are used to guide the model. The prompt includes:

- Selected subject
- User query
- Retrieved contextual information (if available)
- Instructions to restrict responses within the subject This approach ensures:
- Accurate and relevant answers
- Prevention of out-of-domain responses
- Consistent explanation style

The use of prompt engineering plays a crucial role in main-taining response quality and subject discipline.

VIII.5. Dual Interface Interaction Module

The system is designed with a dual-interface architecture to improve usability and learning efficiency. It consists of:

- PDF Analysis Panel: Displays uploaded document con-tent and generated explanations
- Chat Interface: Enables interactive query-based learning This design allows users to:
- Analyze documents and ask questions simultaneously
- Gain both structured and conversational understanding
- Improve comprehension through interactive engagement

VIII.6. System Workflow

The overall methodology follows a structured workflow:

- User creates or selects a subject
- Subject is stored and enforced as a constraint
- User uploads a PDF document (optional)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Text is extracted and processed
- User submits a query
- Relevant context is retrieved (RAG)
- Prompt is constructed using subject + query + context
- LLaMA generates the response
- Output is displayed in the chat interface

IX. SOFTWARE REQUIREMENTS

IX.1. Operating System

- Windows 10 / Windows 11 / Linux / macOS
- The system is platform-independent and can run on any operating system that supports modern web browsers and Python environments.

IX.2. Programming Languages

- Python – Used for backend development, API handling, PDF processing, and AI model integration
- HTML, CSS, JavaScript – Used for designing the fron-tend user interface and enabling interactive user experi-ence

IX.3. Development Tools

- Visual Studio Code (VS Code) – Used as the primary code editor for both frontend and backend development
- Web Browser (Chrome/Edge) – Used for testing and interacting with the application
- Git & GitHub (Optional) – Used for version control and project management

IX.4. Backend Framework

- Flask – A lightweight Python web framework used to build RESTful APIs and handle communication between frontend and backend

IX.5. AI and Model Integration

- LLaMA (via Ollama) – Used as the Large Language Model for generating responses
- Prompt Engineering Techniques – Used to control and guide the behavior of the LLM for subject-specific re-sponses

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

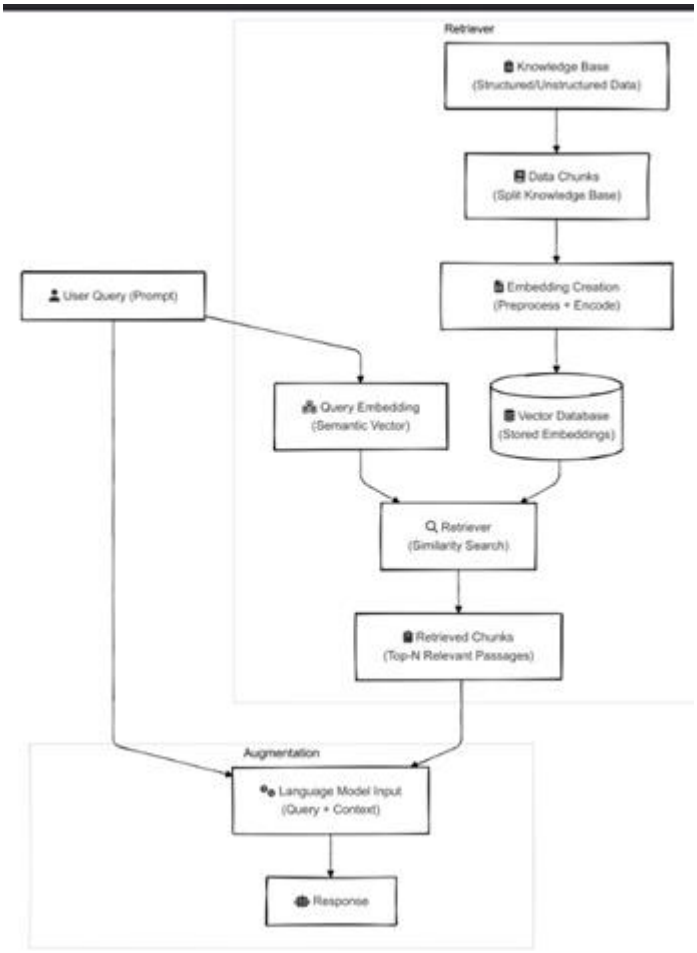

Fig. 2. System Workflow

IX.6. Document Processing

- PyPDF2 – Used for extracting text from uploaded PDF documents
- File Handling Modules (OS, Werkzeug) – Used for file storage and management

IX.7. API and Communication

- REST API – Enables communication between frontend and backend
- JSON Format – Used for structured data exchange

IX.8. Supporting Libraries

- Flask-CORS – Handles cross-origin requests between frontend and backend
- Requests Library – Used for communicating with the LLaMA API (Ollama server)

IX.9. Future Enhancements (Optional Software)

- FAISS / Pinecone – For implementing advanced vector-based retrieval in RAG
- Streamlit / React – For enhanced UI development
- Docker – For containerized deployment

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

X. MODULE DEVELOPMENT

This module handles the visual design, navigation flow, and user interaction of the SAGE application. The goal of this module is to provide a calm, distraction-free, and student-friendly learning environment.

X.1. Landing Page (Entry and Trust Page)

Firstly, the application must have a landing page for SAGE.AI. This page acts as the entry point and trust-building page of the application. It should clearly answer what this application is and why users should trust it. This page should introduce SAGE.AI, clearly explaining its purpose and vision. It should highlight the problem faced by users when using open-domain AI chatbots, such as confusion and hallucination due to mixed subjects, and then present SAGE.AI as the solution through subject-isolated learning. The landing page should invite users to start using the application with a clear call-to-action button like "Get Started," which navigates users smoothly to the next page.

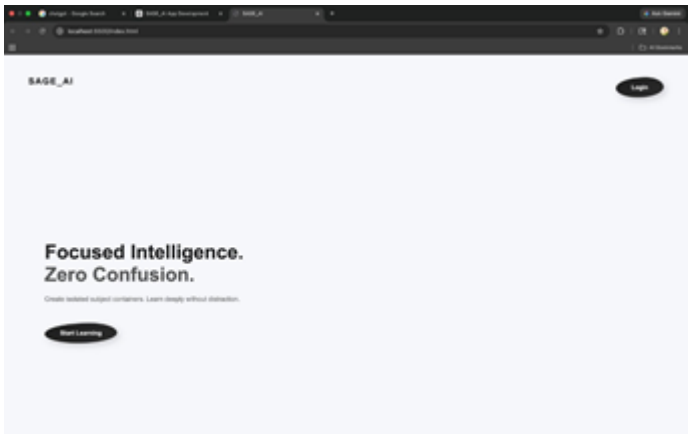

Fig. 3. Login Page

X.2. Understanding and Guidance Page for SAGE.AI

After the landing page, the application should guide users to a page that explains SAGE.AI in detail. This page should be understandable to students, working professionals, and anyone who uses the application. This page should clearly describe the problem statement, explain why open AI chatbots hallucinate, and how mixing multiple domains causes confusion. It should then explain how SAGE.AI solves this problem using subject isolation, RAG, and prompt engineering. Additionally, this page should describe the technologies used in the application so users understand how the system works behind the scenes. A continue button should guide users to authentication.

X.3. Access and Identity Page (Login & Signup)

Next, the application should move to an access and identity page where both login and signup are available on a single page. This page should allow users to log in using their email and password with proper validation. Optionally, the page can also include login using a phone number with a valid OTP

mechanism (this can be implemented using placeholder logic for now). Once authenticated successfully, the user should be redirected to the subject selection page. Any login or authentication errors should be handled politely with clear messages.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

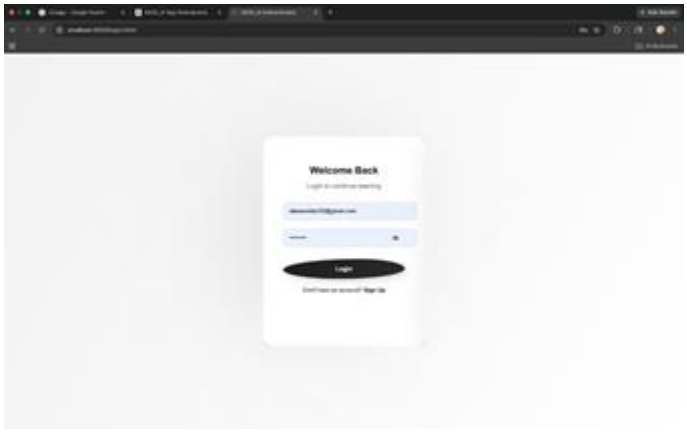

Fig. 4. Access and Identity page

X.4. Authentication, Privacy, and Safety Measures

The application must follow strong authentication and safety measures. User privacy must be protected, and personal data such as email or phone number should be safeguarded. The app should clearly define its limitations and prevent misuse. It must follow academic compliance by explaining what SAGE.AI can answer, what it cannot answer, why refusals happen, and include ethical usage statements. These rules must be enforced throughout the application.

X.5. Chat / Assistant Page (Core AI Interaction)

The next and most important page is the Chat / Assistant page, which is the core of SAGE.AI. This page represents the main AI interaction system. It must contain a subject selector, an assistant response area, a clear refusal message area, and buttons to reset or clear the chat. This page demonstrates the core motto of SAGE.AI, which is subject restriction and isolation. If a user asks a question outside the selected subject, the assistant must respond politely that the question is out of scope or not part of the selected subject. This page must enforce guardrails, reduce hallucinations, and ensure the assistant responds only within the allowed subject using prompt engineering and RAG-based logic. The subject should remain locked during the session unless the user resets or changes it intentionally.

X.6. Subject Selection Page

The application should include a subject selection page where users can choose the subject they are interested in learning. Each subject should act as an isolated compartment, ensuring that no information from other subjects interferes with the selected one. Once a subject is selected, the user should be redirected to the chat assistant page, and all re-sponses must strictly follow the selected subject context.

Fig. 5. chat page

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

X.7. Retrieval-Augmented Generation (RAG) Module

This module retrieves relevant information from the subject-specific knowledge base before generating responses.

- Query embedding
- Document retrieval
- Context injection
- Response grounding

X.8. Prompt Engineering and Guardrail Module

This module defines how the AI is instructed to behave, ensuring strict subject adherence and ethical usage.

- Subject-bound prompts
- System-level instructions
- Refusal policies
- Response formatting

X.9. AI Response Generation Module

This module integrates the Large Language Model (LLM) to generate responses based on retrieved subject data and prompt constraints.

- Context-aware response generation
- Academic tone control
- Subject-restricted answers

X.10. Error and Refusal Handling

The application must include a proper error and refusal han-dling system. Whenever a user asks an out-of-scope question or an error occurs, the response should be modified efficiently and presented in a polite, respectful, and user-friendly manner. The system should never respond harshly or expose technical errors.

X.11. Feedback Page

Finally, the application should include a feedback page where users can rate the application and provide feedback. Users should be able to suggest additional features, report is-sues, or share improvement ideas. This feedback system helps make SAGE.AI more dynamic, user-centric, and continuously improving.

Acknowledgment

We would like to express our sincere gratitude to Jayavani R/M.E (srij088@gmail.com) for her invaluable guidance and technical support throughout this project. Her insights played a crucial role in the successful development and refinement of the SAGE.AI system.

REFERENCES

1. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020. This paper introduced the RAG framework which allows Large Language Models to retrieve relevant knowledge from external documents before generating responses, improving factual accuracy and contextual relevance in domain-specific applications such as educational tutoring systems.
2. J.-W. Wu and M.-H. Tseng, "Developing a Local Generative AI Teaching Assistant System Using RAG," 2025. This study developed an AI teaching assistant system based on Retrieval-Augmented

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Generation that enables instructors to upload academic content and generate subject-specific question answering support, achieving improved response accuracy of up to 85.6% in educational environments.

3. Vu-Minh et al., "Designing a Course-Grounded AI Tutor with Retrieval-Augmented Generation," 2025. This research proposes design principles for developing intelligent tutoring systems using generative AI integrated with retrieval mechanisms to provide personalised learning support across technical subjects.
4. Swacha et al., "Retrieval-Augmented Generation Chatbots for Education: A Survey," 2025. This survey analyses more than 47 research works on educational chatbots using RAG technology and highlights their effec-tiveness in improving learning outcomes and domain-specific tutoring.
5. Akheel et al., "AI Tutors in E-Learning: Analyzing Personalized Learning Pathways," 2025. This study explores AI-powered personalised tutoring systems that adapt learning pathways dynamically based on student needs using NLP and machine learning techniques.