Abstract

The advancements in artificial intelligence (AI) have opened new path ways in education system, enabling transformative changes in how learning is personalized, content is delivered, and engagement is enhanced. The research focuses on the innovative application of adaptive AI systems and large language models (LLMs) in education, specifically targeting their ability to improve personalized learning experiences, automate content generation, and enhance educational outcomes. By leveraging state-of-the-art technologies such as Retrieval-Augmented Generation (RAG), unified text-to-text transformers like T5, and advanced text summarization models such as PEGASUS, this study proposes scalable and intelligent frameworks for modern educational platforms.

The novelty of this research lies in its holistic approach to integrating cutting-edge AI technologies into educational systems, addressing key challenges such as contextual relevance in content delivery, dynamic question generation, and automated feedback systems. The study introduces innovative methodologies for applying AI in dialogue systems and domain-specific tasks provide a foundation for creating adaptive and inclusive learning environments that cater to diverse student needs while optimizing the teaching process.

The findings of this research will contribute significantly to the technological, and engineering domains by advancing the understanding and application of AI in education. This work highlights the transformative potential of AI-driven adaptive learning systems and provides actionable insights for educators, and developers seeking to revolutionize study experiences. By offering a roadmap for the integration of advanced AI technologies into education, this research paves the way for future innovations that will shape the learning.

Table of Contents

I. Introduction

I.1.Problem

For a long time, students and professionals have been frequently engaging with extensive study materials, research papers, and technical reports that require extensive learning, reading, and comprehension. However, manual processing and extraction of key insights from these documents is often time-consuming and inefficient. Traditional summarization tools offer static overviews but lack the ability to dynamically adapt to individual learning needs.

I.2. Motivations

Due to the increase in study materials, research papers, and technical reports, extracting information has become a challenge. In the current advances in Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) brings an opportunity to enhance learning efficiency using Artificial Intelligence (AI)-driven summarization, self-assessment, and interactive Q&A, helping all the users to learn faster.

I.3. Significance

The study aims to contribute to the advancement of education using AI, by integrating RAG and LLMs into an interactive learning framework. Unlike traditional summarization and question-generation tools, LearnDoc dynamically adapts to user needs, providing personalized, context-aware responses that enhance comprehension and retention of the data.

Existing solutions either provide static summaries or lack adaptive self-assessment features, making them less effective for self-learning. In addition, online tools don’t provide privacy and require membership to use their resources, and no tool exists that integrates Q&A and AI-generated quizzes with text summarization.

LearnDoc bridges the gap by combining text summarization, interactive Q&A, and AI-generated quizzes, providing user engagement and learning experiences. Its ability to process documents and generate tailored assessments makes it an alternative to conventional study tools and will be free to access.

I.4. Challenges

Despite advancements in AI, existing tools lack adaptive learning features that dynamically personalize summaries and self-assessments, limiting their effectiveness in diverse learning environments. This study is essential to bridge the gap between static summarization tools and truly interactive AI-driven learning systems that improve comprehension and retention. The problem is non-trivial as it requires an intelligent, context-aware approach that current solutions fail to provide –

1. **Accuracy and Efficiency**– Ensuring AI-generated summaries and questions are contextually relevant, factually accurate, and computationally efficient while processing large volumes of text in real time. Managing retrieval precision and minimizing AI-generated errors (hallucinations) remains a significant challenge.
2. **User Adaptability and Engagement** Developing an interactive learning system that responds to user inputs and adapts to different learning styles. Balancing automation with meaningful personalization while keeping users engaged and avoiding cognitive overload is a critical challenge.

I.5. Objectives

The study aims to develop an application called LearnDoc, an AI-driven system which uses RAG approach. Its primary objectives are:

- Develop an interactive AI-powered platform for processing and analyzing uploaded documents.
- Implement RAG-based document retrieval and summarization to generate concise and contextually relevant insights.
- Enable dynamic question generation to create personalized quizzes and facilitate self-assessment.

II. Overview

II.1. History of the problem

Early solutions relied on manual summarization and note-taking, which were time-consuming and lacked scalability. In the 2000s, Natural Language Processing (NLP) introduced rule-based and statistical summarization methods, but they struggled with contextual understanding.

The emergence of transformer-based AI models like BERT, GPT-3, and GPT-4 revolutionized text comprehension, enabling AI-generated summaries and automated question-answering. Despite the advancements in AI, existing learning solutions remain static, lacking adaptability and interactive learning features.

II.2. State of the Art

In the domain of processing and learning from text has seen significant advancements in recent years with the development of Natural Language Processing (NLP) and Artificial Intelligence (AI). However, while automated text summarization and question-answering systems have improved, they still lack the ability to dynamically adapt to user learning needs and provide interactive self-assessment features. The status of the problem, areas that were solved are -

- **Extractive and Abstractive Summarization** – NLP models such as BERTSUM (Liu & Lapata, 2019) and PEGASUS (Zhang et al., 2020) have made significant progress in generating coherent and meaningful summaries.
- **Question-Answering (QA) Systems** – Open-domain QA models, such as T5 (Raffel et al., 2020) and GPT-4, can generate responses based on input context with high accuracy.
- **Retrieval-Augmented Generation (RAG)** – Hybrid models that combine vector-based retrieval and generative AI (Lewis et al., 2020) improve knowledge extraction for QA tasks.

In terms of user learning and education, areas such as personalized and adaptive learning remain a challenge, as current models fail to dynamically tailor responses based on user behavior, leading to inconsistent self-assessment and knowledge retention. While AI-driven question generation has made progress [Shuster et al., 2022], models often struggle with ensuring contextual relevance and difficulty alignment, reducing their effectiveness in personalized learning experiences. Recent discussions in AI

education systems [Khan and Huang, 2023] and industry insights [Infosys BPM, 2025] highlight the need for more adaptable and engaging AI systems for learning. Additionally, advancements in automated question generation [ArXiv, 2025] provide foundational progress but require further innovation to meet diverse user requirements.

III. Techniques

III.1. Principles, Concepts, and Theoretical Foundations of the research problem

The foundation of LearnDoc is built on Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), which combines information retrieval with text generation to enhance learning. The mathematical formulation of the problem involves vector embeddings for **Document Embedding and Retrieval**, token-based chunking for **Text Generation via Language Models** , and retrieval mechanisms for **Quiz Generation via Structured Prompting**

III.2. Techniques that have been used by other researchers for the research problem

Several techniques have been explored in document summarization, question-answering, and adaptive learning using Artificial Intelligence (AI) and Natural Language Processing (NLP). Following are the key techniques -

1. **Retrieval-Augmented Generation (RAG)** - RAG is an architecture that combines retrieval-based search with LLM-based text generation (Lewis et al., 2020). The retrieval mechanism fetches the most relevant document chunks before passing them as context to an LLM.

    - Embedding Representation**:** Documents and queries are encoded as high-dimensional vectors using transformer-based embeddings (e.g., **OpenAI Ada, Sentence-BERT**):

$$E(X) = \text{Transformer}(X)$$

    - Similarity Search**:** Relevant document chunks are retrieved based on cosine similarity -

$$\text{Similarity}(Q, C_i) = \frac{E(Q) \cdot E(C_i)}{\|E(Q)\| \|E(C_i)\|}$$

    - Generation via Conditional Probability**:** The LLM conditions its output on the retrieved context:

$$P(R|C_{top-k}, Q) = \text{softmax}(W \cdot h)$$

2. **Text Generation via Language Models:** The document chunks are concatenated into context and fed into a generative LLM (e.g., GPT-4) for summarization or Q&A. The model generates an output R based on a conditional probability distribution:

$$P(R|C_{top-k}, Q) = \text{softmax}(W \cdot h)$$

where W is the model's learned weights, and h represents the contextual embeddings of the retrieved document chunks.

3. **Quiz Generation via Structured Prompting**: Given an extracted key concept K from the text, a question Q' is generated with distractor options O using:

$$P(Q', O|K) = \text{LLM}_{\text{prompt}}(K)$$

III.3. Relevant technologies that would be useful to this research

The study relies on a combination of AI-driven document processing, retrieval systems, and large-scale language models to improve summarization, question-answering, and quiz generation. Below are the key techniques and potential technologies applicable to this study.

1. **Retrieval-Augmented Generation (RAG):** Combines document retrieval with LLM-based generation to ensure contextually relevant responses.
2. **Document Embeddings and Vector Search:** Converts documents into high-dimensional vector representations for semantic search.
3. **Abstractive Summarization:** Uses transformer models (PEGASUS, GPT-4) to rephrase and generate concise summaries.
4. **Context-Aware Question Generation:** Uses AI models to create multiple-choice, fill-in-the-blank, and short-answer questions based on document content.

The following technology Stack will be used to implement –

- Python FAST API for backend service and React.JS for frontend.
- Libraries that will be used - Transformers (Hugging Face) and Langchain(for RAG).

- Vector Database: FAISS / Pinecone – For RAG model.
- MongoDB – To store the documents

IV. Approach(s)

IV.1. Methodologies I am going to apply in this research

To develop LearnDoc, a combination of research, design, implementation, and evaluation methodologies will be applied. The research will follow a structured AI development lifecycle for document processing, interactive learning, and adaptive quiz generation. The key methodologies include:

- Literature Review & Technology Survey: Prior work includes Lewis et al. (2020) on RAG models, Liu & Lapata (2019) on abstractive summarization, and Raffel et al. (2020) on text-to-text transformers for QA tasks, can help towards developing the project.
- Algorithm Design & Implementation: Implement Naive RAG for initial retrieval-augmented learning and transition to Modular RAG for better document chunking, ranking, and response refinement.
- Usage of retrieval strategies (exact match, semantic similarity, modular RAG) to improve document processing efficiency.
- Use Reinforcement Learning with Human Feedback (RLHF) to fine-tune question difficulty and response relevance.
- Software Tools & Facilities Used
    - Python, FastAPI – Backend development.
    - FAISS, Pinecone – Vector search for document retrieval.
    - OpenAI API, Hugging Face – LLM-based summarization and quiz generation.
    - MongoDB – Storing user data and document metadata.

IV.2. Techniques I am going to use to solve the problem

To perform document summarization, question-answering, and quiz generation, this study will modify existing AI techniques and introduce approaches to improve learning experience. The key modifications and innovations include:

- Enhancing Retrieval-Augmented Generation (RAG) : Traditional Naïve RAG uses a simple retrieve-and-generate approach leading to less relevant responses. Whereas, in the Proposed

Technique for the study, I will implement Naïve RAG as the baseline, followed by a transition to Modular RAG.

- Context-Aware question Generation: The study will implement structured prompting in LLMs to generate context relevant quizzes using:
    - Concept Extraction via Named Entity Recognition (NER)
    - Reinforcement Learning with Human Feedback (RLHF)
    - Distractor Selection for Multiple Choice Questions (MCQs)

IV.3. Processes I am going to engage in this research

The research and development of LearnDoc will follow a structured AI development workflow, involving programming, experimentation, testing, and evaluation. Once these criteria are met, the project will be successful –

- **Development Process -** The system will be developed iteratively with improvements through experimentation and testing. An initial prototype will be developed using Naïve RAG and later will be transitioned to Modular RAG followed by Quiz generation enhancements and Optimizations for large documents.
    - Development tools: Python FastAPI, PostgreSQL, FAISS, OpenAI API, and React.js for UI.
- **Experimentation Process -** Experiments will be conducted by comparing different LLMs and finalize a LLM to perform summarization and Q&A accuracy and assess the Quiz Generation Quality such as question relevance, difficulty scaling, and distractor effectiveness.
- **Experimentation and Test Result Collection -** Data will be collected through qualitative evaluation methods such as User Evaluation to conduct user testing to evaluate AI-generated summaries, Q&A responses, and quizzes. Metrics such as ROGUE and BLEU will be used.

IV.4. How the Results (Outcomes) of the Research Will Be Demonstrated

The research outcomes for LearnDoc will be demonstrated through a user engagement workflow. The demonstration will highlight AI-driven document summarization, interactive Q&A, and automated quiz generation, ensuring the system effectively enhances learning experiences.

The following core functionalities will be showcased:

- **Document Summarization:** Users will upload a document (PDF, DOCX, or TXT), and the system will generate a concise, AI-powered summary using abstractive summarization.
- **Interactive Q&A:** Users will ask questions about an uploaded document, and LearnDoc will provide real-time AI-generated answers using RAG-based retrieval.
- **Automated Quiz Generation:** The system will generate multiple-choice, fill-in-the-blank, and short-answer questions from a document.

The expected experimental & demonstration setting would be - The system will be tested in a controlled user environment with a web-based interface allowing participants to:

1. **Upload documents** of varying complexity and observe AI-generated summaries.
2. **Ask context-aware questions** and compare retrieval effectiveness.
3. **Take AI-generated quizzes** and provide feedback on question relevance.

IV.5. How the Results (Outcomes) of the Research Will Be Evaluated, Analyzed, and Compared with Previous Research?

Since the study aims at focusing on using existing RAG and LLM approach and provide better learning experience hence it will be evaluated based on -

- **User Experience and Evaluation Metrics**- LearnDoc will be evaluated based on ease of use, response clarity, and user satisfaction with document summarization, Q&A, and quiz generation.
- **Accuracy & Relevance Assessment** - User feedback will assess whether AI-generated summaries, answers, and quizzes are coherent, contextually relevant, and useful for document comprehension.
- **Comparative Usability Testing** - LearnDoc's performance will be compared to existing tools based on retrieval accuracy, system responsiveness, and interface intuitiveness to highlight its advantages in document processing and assessment.

IV.6. Facilities and Supplies Needed for This Research

The successful implementation of LearnDoc requires a combination of hardware, software,datasets for development, testing, and deployment. Below is a breakdown of the required resources and their availability.

- Desktop – Personal desktop to design, develop and test the application
- Software development tools suchs as Python, FastAPI, React.js will be used.
- AI & NLP Libraries such as
    - OpenAI API – For LLM-based summarization and Q&A.
    - Transformers (Hugging Face) – For running models like GPT-4, PEGASUS, and BERTSUM.
- For Database & Storage MongoDB will be used.
- Datasets Required –
    - SQuAD (Stanford Question Answering Dataset) – For testing question-answering capabilities.
    - CNN/DailyMail Summarization Dataset – For evaluating AI-generated summaries.
    - HotpotQA / Natural Questions (NQ) – For benchmarking retrieval performance.
    - User-Uploaded Documents (Real-World Testing Data) - Users will upload their own documents, and the system will process them in real-time.

All the facilities are readily available, only non-available resources are paid AI API Access (OpenAI, Hugging Face Pro) – If necessary, initial testing will use free-tier APIs, with a budget plan for premium access if required.

V. Work Plan

V.1 Tasks to Be Performed in This Research

The project begins with a Literature Review & Technology Survey (1-2 weeks), where research papers, AI model benchmarks, and existing techniques in RAG, summarization, and quiz generation will be analyzed to identify current limitations and areas for improvement.

This will be followed by Prototype Development (2-3 weeks), where Naive RAG will be implemented using basic document retrieval, OpenAI embeddings, and FAISS/Pinecone vector search, leading to a working system with initial Q&A and summarization capabilities.

Simultaneously, an AI-Generated Quiz System (1 week) will be developed using Named Entity Recognition (NER), TF-IDF, and structured prompting to generate context-aware quizzes with difficulty scaling. Next, the system will transition to Modular RAG (1-2 weeks), integrating chunk ranking, multi-query expansion, and response filtering to improve retrieval accuracy and contextual relevance.

The system will then undergo Experimentation & Performance Evaluation (2-3 weeks), where it will be benchmarked against GPT-4, BERTSUM, and PEGASUS to assess retrieval accuracy, latency, and quiz relevance, generating quantitative and qualitative insights. User Testing & Feedback Integration (1-2 weeks) will follow, involving user response collection, and Reinforcement Learning with Human Feedback (RLHF) to enhance AI-generated responses and assessments. Finally, the Final Demonstration (1 week) with frontend, and API optimization, culminating in a fully functional, AI-driven learning assistant ready for real-world use.

V.2. Schedule, timeline, and milestones

Table 1. Tentative schedule of research and thesis

Order	Dates	Task/Activity	Prerequisites (Knowledge, Skill, or Tools)	Expected Results
1	From Jan 25 to Feb 5	Literature Review & Technology Survey	Research papers, AI/ML model benchmarks, NLP tools	Identified gaps in current AI models and techniques
2	From Feb 6 to Feb 26	Prototype Development: Naïve RAG	Python, OpenAI API, FAISS, vector databases	Initial working retrieval-based system

3	From Feb 26 to March 5	Quiz System Implementation	NER, TF-IDF, structured prompting	Context-aware quizzes with difficulty scaling
4	From March 5 to March 26	Transition to Modular RAG	Multi-query expansion, ranking algorithms, embedding models	Enhanced retrieval accuracy and contextual relevance
6	From March 26 to April 16	Experimentation & Performance Evaluation & Fine-Tuning	Benchmark datasets (SQuAD, CNN/DailyMail), evaluation metrics (ROUGE, BLEU)	Quantitative results on performance and efficiency
7	From April 16 to April 23	User Testing & Feedback	User interface, RLHF.	System refinement based on user interaction data
8	From April 23 to April 30	Final Documentation & Demonstration	Document the findings	Final Documentation and real-time demonstration

VI. Mitigations

VI.1. Anticipated Problems and Issues

- **Technical Challenges with Proposed Techniques -** One major technical challenge is that the proposed Naïve RAG and Modular RAG approaches may fail to accurately retrieve document chunks for highly complex queries, leading to incomplete or irrelevant answers. To mitigate this, extensive benchmarking on diverse datasets will be conducted to refine chunking and retrieval strategies. Additionally, a fallback mechanism will be implemented to rephrase queries or retrieve larger document chunks for better context. Another issue is the possibility of generating overly generic or irrelevant quiz questions due to the limitations of AI prompt engineering. This can be mitigated by using Reinforcement Learning with Human Feedback (RLHF) to fine-tune the question-generation process and by including a manual validation step during the testing phase.
- **Experimental Setting and Scalability Issues -** Processing large documents could become a bottleneck due to token limitations in LLMs like GPT-4, making it difficult to handle long contexts effectively. To overcome this, dynamic chunking and hierarchical summarization techniques will be employed to break documents into manageable segments, and sliding window attention models will be used for better long-context handling.
- **Data and Resource Availability Issues -** Another challenge is the lack of sufficient real-world datasets for testing retrieval, summarization, and quiz generation performance. This will be mitigated by utilizing publicly available datasets like SQuAD, CNN/DailyMail, and HotpotQA for initial testing phases. To address real-world applicability, a custom dataset will be created using user-uploaded documents collected during testing.
- **User Feedback and Usability Challenges -** User feedback during testing may reveal significant gaps in system usability, question quality, or response accuracy, which could impact the overall effectiveness of LearnDoc. To mitigate this, iterative user testing will be conducted, and feedback will be addressed through frequent system updates.

VI.2. Limitations and Constraints of the Research

- **Technical Limitations -** The techniques used in this research, such as Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), come with inherent limitations. Naïve RAG may struggle with retrieving highly contextualized and precise information for complex

queries, especially in documents with dense or ambiguous content. Additionally, token limitations in LLMs restrict the ability to process large documents in a single context window, requiring dynamic chunking, which can lead to loss of context between segments. Moreover, while RLHF can improve quiz generation and question relevance, it is resource-intensive and may require significant computational and data overhead.

- **Pre-Conditions and Constraints on Techniques -** The research heavily relies on external AI APIs (e.g., OpenAI GPT-4, Hugging Face), which may introduce constraints such as rate limits, dependency on internet connectivity, and high operational costs for large-scale deployment. The reliance on cloud platforms for computational resources could also pose challenges in environments with limited access to cloud services or funding. Additionally, fine-tuning or custom model training may not be feasible within the scope of the research due to the high computational cost and time required.

VII. Summary

The research addresses the challenge of processing and learning from large volumes of text, a task often hindered by the inefficiency of existing tools. Current solutions, such as static summarization and question-answering systems, fail to adapt to individual user needs or provide interactive, personalized learning experiences. Although advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI), including models like BERT, GPT-4, and PEGASUS, have improved text comprehension and summarization, these systems remain limited in their ability to dynamically retrieve, process, and present contextually relevant information tailored to diverse learning preferences.

This study introduces LearnDoc, a system that leverages Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) to provide interactive document summarization, Q&A, and AI-driven quiz generation. Techniques such as document embeddings, vector search, modular RAG architecture, structured prompting, and dynamic chunking will be applied to ensure efficient document processing and adaptive learning. Unlike existing systems, LearnDoc emphasizes personalized learning by integrating modular RAG for improved retrieval accuracy and generating context-aware quizzes that adjust to user input, introducing a unique focus on adaptability and interactivity in AI-driven learning systems.

The expected outcomes include a fully functional system capable of producing coherent document summaries, accurate answers to user queries, and tailored quizzes that enhance user engagement. LearnDoc's innovative approach bridges the gap between static summarization tools and interactive AI-based learning assistants, contributing to the advancement of adaptive NLP systems. By combining retrieval, generative AI, and personalization, this research advances the field of AI-driven education tools, offering a scalable solution for improving information accessibility and user experience in scientific, technological, and educational domains.

References

[1] Infosys BPM, “The Future of Education: How AI and Adaptive Learning Are Shaping the Classrooms of the Future,” https://www.infosysbpm.com/blogs/education-technology-services/the-future-of-education-how-ai-and-adaptive-learning-are-shaping-the-classrooms-of-the-future.html (available as of February 5, 2025).

[2] Lodovico Molina, Ivo, et al., “Comparison of Large Language Models for Generating Contextually Relevant Questions,” Proceedings of the European Conference on Technology Enhanced Learning, Cham: Springer Nature Switzerland, 2024.

[3] P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020, pp. 9459-9474.

[4] R. Khan and X. Huang, “AI in Education: A Review of Personalized Learning and Adaptive AI Systems,” Journal of Artificial Intelligence in Education, Volume 32, Issue 1, 2023, pp. 45-62.

[5] C. Raffel, N. Shazeer, A. Roberts, et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5),” Journal of Machine Learning Research, Volume 21, Issue 140, 2020, pp. 1-67.

[6] K. Shuster, et al., “Retrieval-Augmented Generation Models for Dialogue,” Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 2022, pp. 1783-1796.

[7] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” arXiv preprint arXiv:1908.08345, 2019.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” arXiv preprint arXiv:1912.08777, 2020.