Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 2 (2025)

Biomentor - Personalized E-Learning Platform for A/L Biology Subject English Medium Students in Sri Lanka.

Dharane S[1], Sajeevan S[2], Sujitha S[3], Gokul Abisheak S[4], K.T.S. Kasthuriarachchi[5], Karthiga Rajendran[6]

[1]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*
[2]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*
[3]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*
[4]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*
[5]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*
[6]*Department of Computer Science and Software Engineering, Sri Lanka Institute of Information Technology Malabe, Sri Lanka*

Abstract: **-** The growing dependence on technology in education has facilitated the creation of customized e-learning solutions that address individual learning requirements. This study introduces BioMentor, a customized e-learning platform designed for Advanced Level (A/L) Biology students in Sri Lanka. The platform incorporates sophisticated artificial intelligence (AI) methodologies, such as Retrieval-Augmented Generation (RAG), transformer-based summarization, adaptive quiz creation, automated answer assessment, and spaced repetition algorithms. The technology improves subject understanding, engagement, and retention via personalized learning pathways, hence enhancing overall academic achievement. Experimental findings demonstrate that BioMentor successfully addresses deficiencies in traditional education by adaptively modifying curriculum according to student performance and improving knowledge retention. The results underscore the capability of AI-driven educational platforms to transform domain-specific learning and provide scalable solutions for individualized education.

Keywords*: Personalized learning, Adaptive Learning, Retrieval-Augmented Generation (RAG), Transformer Models, Spaced Repetition, E-Learning, A/L Biology, Educational Technology.*

1. Introduction

In contemporary education, technology has emerged as an essential instrument for tailored and efficient learning. The Advanced Level (A/L) Biology curriculum presents distinct challenges for students owing to its comprehensive syllabus, sophisticated terminology, and complicated topics. In Sri Lanka, where English medium pupils adhere to a strictly regimented curriculum, conventional teaching and learning methods sometimes neglect specific student requirements. The dependence on rote memorisation, fixed content delivery, and insufficient adaptive educational technologies hinders learners from fully interacting with the material and realising their academic potential.

BioMentor: A tailored e-learning platform for advanced biology students in Sri Lanka, presented in English, serves as a transformative solution to overcome existing deficiencies. The platform offers a comprehensive, personalised learning environment through the integration of adaptive learning technologies. It utilises advanced approaches including spaced repetition algorithms, dynamic summarisation tools, adaptive quiz systems, and automated question-answering and evaluation processes to improve student performance. Each component targets

distinct facets of the learning process—retention, understanding, engagement, and assessment—guaranteeing a comprehensive educational experience.

The platform incorporates optimised language models, Retrieval-Augmented Generation (RAG), and modular architectures to develop a scalable and adaptable solution. By customising education to the requirements of A/L Biology students, BioMentor not only reconciles traditional pedagogical approaches with contemporary educational needs but also establishes a benchmark for specialised e-learning platforms.

2. Literature Review

A. LLM-Based Abstractive Summarization with Voice Output implemented in various Software Architectures.

Recent advancements in transformer-based architectures have significantly enhanced the performance of abstractive text summarization. Models such as BART, T5, and Flan-T5 offer improved coherence, contextual understanding, and fluency compared to traditional extractive methods. Notably, Flan-T5 has demonstrated strong instruction-following capabilities and efficiency, making it particularly well-suited for educational applications [1].

Summarization systems generally adopt one of two primary approaches: long-document summarization or topic-based summarization. The former processes large, structured or scanned files (e.g., PDFs, DOCX) using chunk-based abstraction techniques, often incorporating OCR and tabular data extraction. This multi-stage summarization pipeline has proven effective in managing token limitations while preserving contextual continuity [2]. In contrast, topic-based summarization focuses on generating concise overviews in response to user-defined queries, drawing from multiple relevant sources. BERT-based models have shown effectiveness in this domain by synthesizing targeted, coherent summaries [3].

To improve factual accuracy and contextual relevance, many modern systems integrate Retrieval-Augmented Generation (RAG), which combines semantic retrieval mechanisms with generative models to incorporate external domain knowledge into the summarization process [1].

Complementary features such as notes generation and multilingual translation have been introduced to enhance content accessibility for linguistically diverse users, including Sinhala and Tamil speakers. Furthermore, text-to-speech (TTS) integration enables auditory delivery of content, supporting multimodal learning environments [4].

From an architectural standpoint, monolithic deployments have demonstrated superior performance for real-time applications, offering lower latency and simplified integration compared to microservices. This architecture is particularly advantageous in e-learning contexts requiring seamless interaction between summarization and auxiliary services [5].

B. LLM-Based Answer Generation and Evaluation

Fine-tuning transformer-based models has gained significant attention in Natural Language Processing (NLP) for educational applications. Large Language Models (LLMs), such as BERT and GPT-4, have been extensively utilized for automated question answering (QA) and content generation. Studies indicate that fine-tuning pre-trained models with domain-specific datasets significantly enhances their ability to provide accurate and context-aware responses [6]. Research has demonstrated that employing structured and essay-style training datasets can improve model performance in academic assessments [7].

Recent advancements in automated answer evaluation have explored deep learning and NLP-based methodologies. Transformer-based AI models, including BERT and GPT-3, exhibit strong capabilities in textual comprehension and evaluation [8]. However, these models often require extensive training data and may encounter challenges in handling domain-specific queries. To address these limitations, hybrid models integrating multiple evaluation techniques have been proposed to improve accuracy in educational assessments. These models incorporate structured and unstructured data analysis, yielding enhanced student response evaluation [9]. The integration of semantic similarity measures, lexical similarity techniques, and grammar verification mechanisms further strengthens the reliability of automated evaluation systems, contributing to more effective and insightful feedback in digital learning environments.

C. LLM-Based Adaptive Quiz for Enhancing Biology MCQ Skills

The automatic generation of multiple-choice questions (MCQs) has gained significant attention due to advancements in Natural Language Processing (NLP) and machine learning. Transformer-based models such as T5 and BERT have been widely used for Automatic Question Generation (AQG) because of their ability to generate semantically relevant questions from educational content with minimal human intervention [10][20]. Studies have shown that these models can produce high-quality MCQs that align with learning objectives, making them valuable tools for educational applications.

One of the primary challenges in AQG is ensuring that the generated questions vary in difficulty and are tailored to the learner's proficiency. Adaptive learning systems dynamically adjust quiz difficulty based on student performance, improving engagement and knowledge retention. Research has demonstrated that tracking student responses, including accuracy and response time, allows for better-targeted quizzes that address weak areas and enhance learning outcomes [11], [12].

Recent advancements have focused on domain-specific MCQ generation, particularly in subjects like Biology. Fine-tuning models like LLaMA on specialized datasets ensures that questions align with curriculum standards. Additionally, the incorporation of Retrieval-Augmented Generation (RAG) further enhances question relevance by retrieving contextually appropriate information [13], [14].

The growing research in AQG and adaptive learning highlights the potential of personalized quiz generation. By leveraging transformer-based models and dynamic difficulty adjustments, these systems improve student engagement and performance, forming the foundation for the proposed adaptive MCQ quiz platform for A/L Biology students.

D. Enhancing Vocabulary Memorization through Adaptive Spaced Repetition

Spaced repetition is a learning method that optimizes review intervals to enhance long-term memory retention. The SM-2 algorithm, developed by Piotr Woźniak in 1987 for SuperMemo, is one of the most well-known implementations. It is based on Ebbinghaus's forgetting curve, which shows that memory retention declines exponentially without reinforcement. Spaced repetition mitigates this by strategically reviewing information at increasing intervals. Research by Cepeda et al. [15] confirms that distributed practice improves long-term retention more effectively than cramming, while Roediger and Butler [16] found that repeated testing with intervals enhances memory performance better than passive rereading.

SM-2 adjusts review intervals based on user performance, using an easiness factor (EF) to determine how often an item should be reviewed. This ensures difficult concepts are revisited more frequently, while easier ones have longer gaps. Pavlik and Anderson [17] demonstrated that SM-2 significantly improves efficiency in language learning and medical education. Digital platforms like SuperMemo and Anki implement SM-2, enabling adaptive study schedules. Kang [18] highlighted its role in modern educational technology, while Karpicke and Bauernschmidt [19] found that active retrieval combined with spaced repetition enhances memory better than passive review.

3. Methodology

A. LLM-Based Abstractive Summarization with Voice Output implemented in various Software Architectures.

The summarization component, built using the Flan-T5 Base model, was fine-tuned for structured, syllabus-aligned summaries of Sri Lankan A/L Biology content. Key steps included text extraction, preprocessing, model selection, fine-tuning, RAG framework integration, deployment, and evaluation using ROUGE scores, ensuring accuracy, coherence, and real-time performance optimization.

The system supports two primary modes of summarization: document-based and topic-based summarization. In the document-based approach, users upload extensive content in formats such as PDF, DOCX, PPTX, or scanned documents. A modular extraction pipeline is employed, utilizing PyMuPDF for text-based PDFs, python-docx and python-pptx for Word and PowerPoint files, and Tabula for extracting structured tables from PDF documents. For image-based or scanned documents, the system leverages pdf2image in conjunction with pytesseract to

perform Optical Character Recognition (OCR). The extracted content undergoes comprehensive preprocessing, including regular expression-based cleaning, grammar correction, and spell-checking. The cleaned text is then segmented into token-constrained chunks, which are individually summarized using the fine-tuned Flan-T5 model. A final, coherent summary is generated by contextually merging these chunk-level outputs. This multi-stage abstraction approach aligns with established long-document summarization methodologies [2].

In the topic-based summarization mode, users submit a keyword or query representing a specific subject area. The system first validates the input to filter out inappropriate or irrelevant content. It then employs a FAISS-powered Retrieval-Augmented Generation (RAG) framework to semantically retrieve relevant information from a domain-specific, pre-indexed knowledge base. The retrieved content is aggregated, preprocessed, and subsequently passed through the Flan-T5 model to produce a concise, contextually relevant summary. This methodology reflects best practices in query-driven abstractive summarization, as demonstrated in prior BERT-based summarization frameworks [3].

The training dataset, sourced from government-approved books and guides, ensured syllabus-aligned content for A/L students. A preprocessing pipeline, including text cleaning, segmentation, tokenization, normalization, and stop word removal, enhanced model efficiency and summary quality. The Keywords column was removed as it was irrelevant to abstractive summarization.

Several transformer-based models were evaluated before selecting Flan-T5 Base. BART was considered for its fluency but was found to be computationally expensive for real-time usage. PEGASUS showed promise for long-form summarization but required large training data and infrastructure. Flan-T5 Base was selected for its strong instruction-following capabilities, domain adaptability, and efficient performance [1]. Compared to larger models like Flan-T5 Large and XXL, the Base variant offered a practical balance of speed and accuracy for deployment in resource-constrained environments.

A Retrieval-Augmented Generation (RAG) framework was incorporated to enhance contextual relevance. The RAG pipeline uses FAISS for semantic similarity retrieval, allowing the model to synthesize summaries based not only on user input but also on retrieved supporting material [2], [3].

To evaluate the effectiveness of the summarization model, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores were used. ROUGE is a widely adopted metric in text summarization, measuring the similarity between generated summaries and reference summaries. It evaluates the quality of summaries by comparing word and phrase overlaps between generated and human-written summaries. Three key ROUGE metrics were used in the evaluation:

Table I: ROUGE METRICS

Metric	Description
ROUGE-1	Measures unigram overlap, assessing keyword retention.
ROUGE-2	Evaluates bigram overlap, analyzing phrase continuity and fluency.
ROUGE-L	Assesses longest common subsequence overlap, determining structural similarity.

Deployment was realized in both monolithic and microservices architectures. The monolithic design embedded the summarization module within the main application, reducing latency and simplifying integration. Alternatively, in the microservices architecture, the summarization module was deployed as an independent REST API service. While more modular, this introduced network latency and additional infrastructure complexity. Based on current real-time requirements, the monolithic approach was found more efficient [5].

A text-to-speech (TTS) component, powered by the gTTS library, was added to convert textual summaries into audio. This feature supports auditory learners and enhances accessibility through multimodal delivery [4].

In addition to summarization, the system incorporates notes generation and multilingual translation capabilities. Notes are generated using content retrieved through a Retrieval-Augmented Generation (RAG) framework.

Relevant educational information is semantically retrieved from a pre-indexed knowledge base and formatted into structured notes through grammar and spelling correction routines. Optional translation into Sinhala and Tamil is performed using the deep_translator library. While English notes are made available as downloadable PDFs, translated outputs are returned in plain text format to ensure language compatibility and proper rendering.

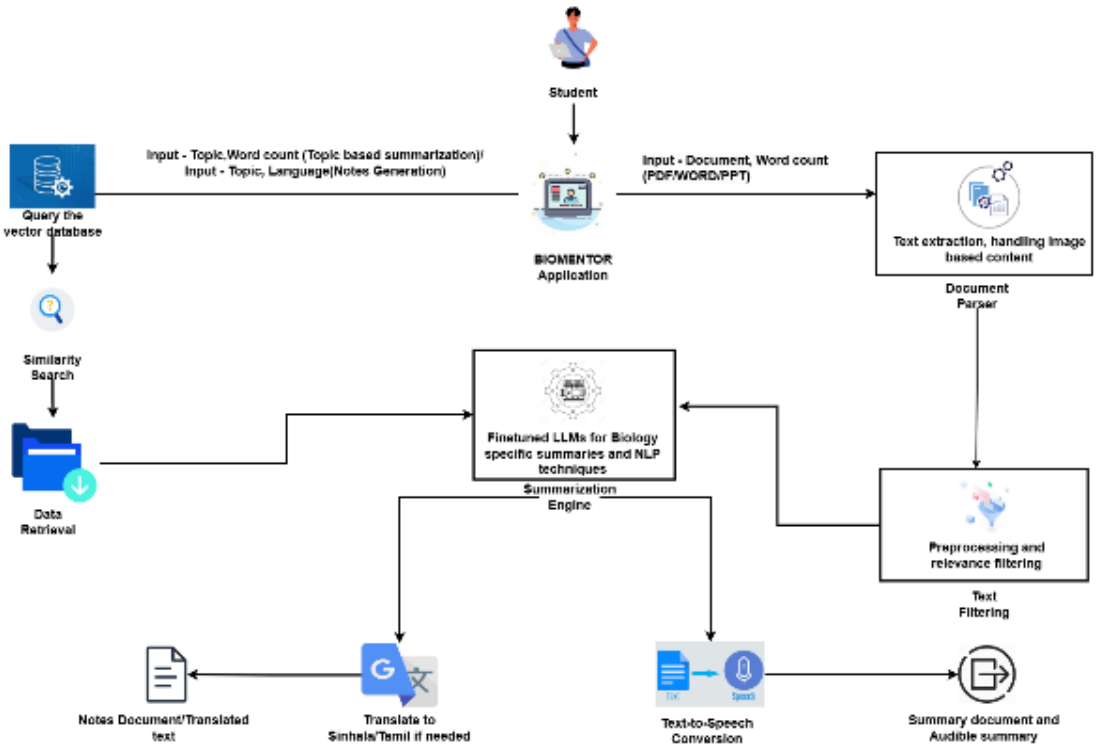

Fig 1: Architecture Diagram of the Summarization System

B. LLM-Based Answer Generation and Evaluation

The proposed methodology enhances students' ability to answer structured and essay-type questions through two key systems: answer generation and evaluation. The answer generation system generates contextually relevant responses, while the evaluation system analyzes student answers for accuracy, relevance, and grammar, providing adaptive feedback for improvement.

The LLaMA 3 Instruct model was fine-tuned using Parameter-Efficient Fine-Tuning (PEFT) techniques, including Low-Rank Adaptation (LoRA) and adapters from Hugging Face's PEFT library, to reduce computational costs. This approach adapted a small subset of parameters while keeping most of the pre-trained model frozen, significantly reducing computational overhead. Additionally, quantized fine-tuning with the BitsAndBytes library enabled 4-bit Normal Float compression for efficient training on consumer GPUs. Supervised fine-tuning optimized structured and essay-type QA using cross-entropy loss.

Instruction-tuning was applied to align the fine-tuned LLaMA-3 8B Instruct model with educational Q&A requirements, ensuring effective generation of structured and essay-formatted responses in an academic setting.

Retrieval-Augmented Generation (RAG) was implemented to enhance contextual understanding. Semantic embeddings from the SentenceTransformer (multi-qa-mpnet-base-dot-v1) model enabled semantic mapping of queries. A FAISS index efficiently stored and retrieved the top-k relevant entries from structured Q&A and notes datasets, ensuring relevant context retrieval for user queries.

The fine-tuned LLaMA-3 model generates responses based on retrieved context, ensuring alignment with educational expectations. It is trained on structured and essay-based questions and deployed using Hugging Face's

transformers pipeline for inference. Before response generation, FAISS retrieves concise Q&A pairs for structured questions, while essay-type questions retrieve detailed context from both Q&A and notes. Safe generation techniques, including logits clamping and softmax normalization, ensure numerical stability and prevent NaN/Infinity errors.

For structured answer generation, retrieved Q&A pairs construct concise 1-2 sentence responses with word limit constraints. Essay answer generation utilizes retrieved Q&A and notes for long-form responses with a minimum word count. Response diversity is controlled using temperature, top-p, and top-k sampling, while a repetition penalty prevents redundancy.

A structured multi-step approach is used to evaluate student answers. A model-generated answer serves as a benchmark, and student responses are assessed using SciBERT-based semantic similarity, TF-IDF cosine similarity, and Jaccard similarity. SciBERT computes semantic closeness, TF-IDF captures lexical overlaps, and Jaccard similarity quantifies keyword commonality. Language Tool ensures grammatical accuracy. A hybrid scoring model assigns predefined weights to similarity metrics and grammar assessment. Evaluated data is stored in MongoDB for performance analytics, trend identification, and personalized learning recommendations. Adaptive feedback highlights missing or extra keywords, grammar errors, and customized exercises to enhance learning. The final score (S) is computed as follows:

$$S_{final} = w_{scibert}S_{scibert} + w_{tfidf}S_{tfidf} + w_{jaccard}S_{jaccard} + w_{grammar}S_{grammar}$$

where w values represent the predefined weights for each metric.

Grammar checking ensures language accuracy, while adaptive feedback provides insights into missing/extra keywords, grammar errors, and personalized exercises. This approach integrates retrieval-augmented generation and answer evaluation techniques, enhancing student engagement and academic performance.

To maintain the appropriateness and academic integrity of user interactions, a multi-stage moderation system is integrated into the proposed framework. This system filters inappropriate, harmful, or low-quality queries before answer generation. The pipeline combines rule-based and deep learning approaches. Named Entity Recognition (NER) using spaCy's en_core_web_sm detects sensitive content, while sentiment analysis with VADER flags inputs with highly negative polarity. Additionally, a BERT-based classifier (unitary/toxic-bert) identifies toxic language, rejecting queries with toxicity scores above 0.85.

A custom moderation model, bert-question-moderator, was fine-tuned on a labeled dataset of acceptable and unacceptable educational queries. Based on the bert-base-uncased architecture, the model was trained using the Hugging Face transformers library with tokenized and preprocessed data. The training utilized binary cross-entropy loss, early stopping, and evaluation metrics such as accuracy and F1-score. The model is deployed via Hugging Face Spaces using a Gradio interface to classify queries in real-time. Only queries labeled as "Acceptable" proceed to the generation pipeline, ensuring safe and relevant educational content delivery.

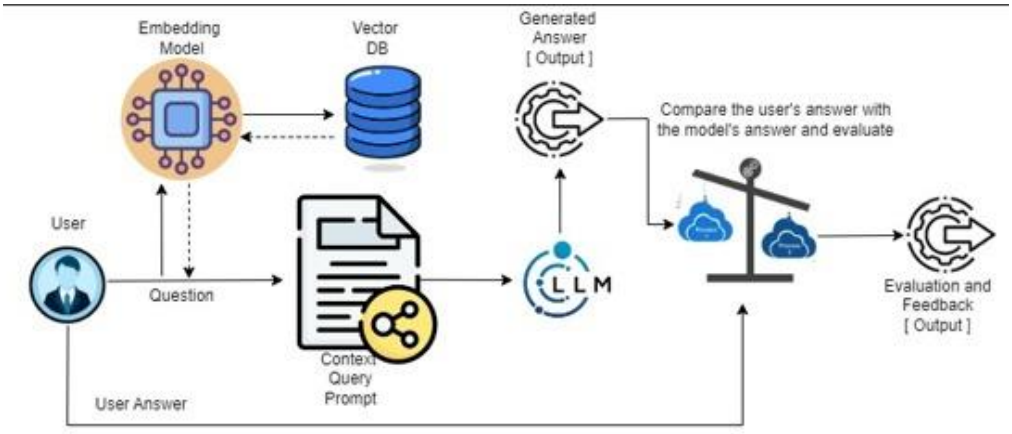

Fig 2: Architecture Diagram of the Question Answering System

C. LLM-Based Adaptive Quiz for Enhancing Biology MCQ Skills

This study presents an adaptive multiple-choice question (MCQ) generation system for A/L Biology students, combining retrieval-augmented generation (RAG), Item Response Theory (IRT), and transformer-based language models to produce contextually relevant and dynamically personalized assessments.

The dataset was compiled from past A/L Biology examinations and school-level assessments. Each question includes a stem, five answer choices labeled A through E, a correct answer, and an expert-validated difficulty label (easy, medium, or hard). Sentence embeddings were generated using the MiniLM-L6-v2 model, followed by KMeans clustering. This allowed questions to be grouped based on contextual similarity, to enhane diversity in generated content. Difficulty annotations were reviewed and refined by subject matter experts to ensure accuracy and alignment with curriculum standards, following practices established in prior MCQ-generation studies [10], [11].

Question generation is powered by a fine-tuned LLaMA-2-7b-chat-hf model, optimized using Quantized Low-Rank Adaptation (QLoRA) for efficient memory usage without compromising output quality. Fine-tuning was carried out using the Hugging Face transformers library, with QLoRA implemented via peft and 4-bit bnb.NF4 quantization supported by bitsandbytes. LoRA adapters were integrated into the attention layers to enable low-resource training while preserving model performance.

To ensure contextual coherence, the system implements a retrieval-augmented generation (RAG) approach. For each question to be generated, the system retrieves semantically similar questions from a FAISS index built on sentence-level embeddings. These retrieved examples are incorporated into the generation prompt, guiding the model to produce questions that are novel yet grounded in syllabus-aligned content [12], [14].

The system adapts question difficulty using principles from Item Response Theory (IRT), specifically the two-parameter logistic model. The probability that a student with latent ability θ will answer a question correctly is given by:

$$P(\theta) = 1 / (1 + e\text{^}(-a(\theta - b)))$$

Here, a denotes the item discrimination parameter and b denotes the difficulty of the question. Both parameters are assigned dynamically. The discrimination parameter a is sampled from a defined range to ensure the question can differentiate between learners of varying ability. The difficulty parameter b is selected in relation to the learner's estimated ability level θ, ensuring alignment between the question's complexity and the student's proficiency. The student's ability θ is estimated based on their recent performance and average response time. The estimation function is defined as:

$$\theta = log(\, accuracy \,/\, (100 - accuracy + 1)\,) - penalty(time)$$

The penalty term adjusts θ based on observed response time, reducing the impact of overly fast or slow responses and ensuring a more behaviourally grounded ability score. This continuous ability estimation ensures that the system delivers appropriately challenging questions as the student progresses.

Each quiz begins with a balanced mix of question difficulties. Over time, the distribution is adapted based on the student's interaction history. Rather than using rigid thresholds, the system uses performance-informed heuristics to adjust the ratio of easy, medium, and hard questions in subsequent quizzes. This strategy enables progressive skill development while reinforcing foundational concepts, consistent with pedagogical theories of adaptive learning [13], [14].

To ensure uniqueness and maintain question novelty, the system applies cosine similarity filtering across generated content. Sentence embeddings of newly generated questions are compared against previous quizzes using FAISS-based retrieval. Questions that exceed a defined similarity threshold are discarded and regenerated. This mechanism prevents repetition and ensures that each quiz contributes new value to the learner's understanding.

Through the integration of RAG, IRT-based difficulty calibration, and transformer-based generation, the system achieves dynamic personalization of assessment content. By continuously adapting to learner performance and

contextualizing generation within a semantically rich corpus, it delivers an intelligent and scalable solution for personalized education in the domain of A/L Biology.

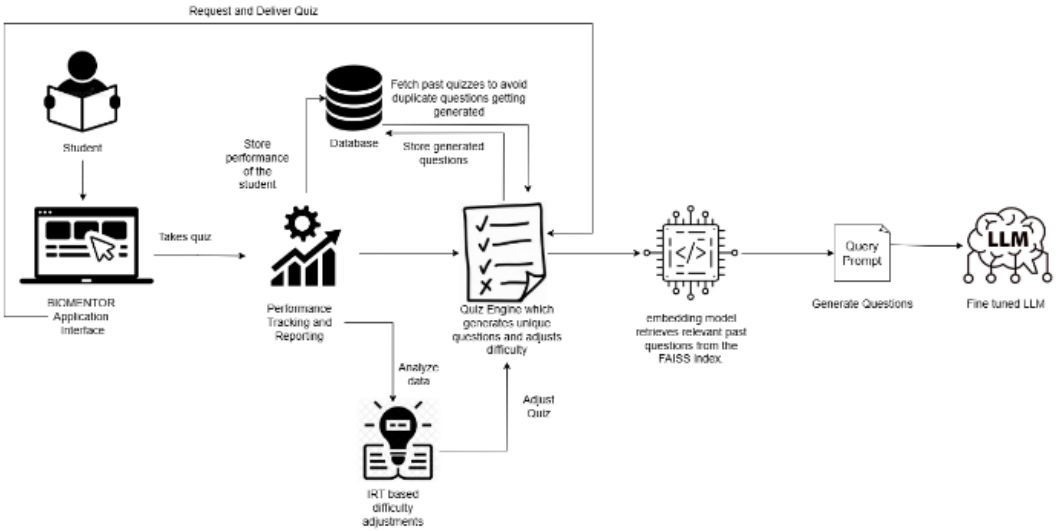

Fig 3: Architecture Diagram of the Adaptive MCQ Quiz System

D. Enhancing Vocabulary Memorization through Adaptive Spaced Repetition

Adaptive Spaced Repetition (ASR) aimed at improving memory retention through the integration of cognitive concepts, and gamification. The solution utilises a FastAPI backend for effective API management and a React frontend for an engaging user experience. The amalgamation of these technologies guarantees seamless data transfer, rapid reaction times, and a user-friendly interface for learners.

The system's foundation is the SM-2 algorithm, an established spaced repetition method that adjusts review intervals based on user performance. The system modifies review schedules based on answer accuracy, retrieval duration, and review attempt frequency. The FastAPI backend analyses user performance data and adjusts scheduling in real-time to guarantee optimal review timeliness. Technology personalises learning strategies and optimises knowledge retention by continuously monitoring user performance and suitably spacing reviews.

$$EF` = EF + (0.1 - (5 - q)\ \times\ (0.08 + (5 - q) \times 0.02))$$

This equation updates the Ease Factor (EF), which determines how quickly the review intervals grow. The term (5−q) represents how difficult the recall was (where q is a rating from 0 to 5). If the recall was poor (q is low), the equation decreases EF more significantly, making future reviews happen sooner. If recall was easy (q is high), the change is minimal, allowing longer intervals. The constants 0.1, 0.08, and 0.02 fine-tune the adjustment, ensuring that difficult items are reviewed more frequently while easy ones are spaced out more.

The backend, constructed with FastAPI, oversees user interactions, retains learning progress, and organises review sessions. Essential elements comprise User Authentication, employing JWT-based mechanisms for secure access; Progress Tracking, utilising a NoSQL database (MongoDB) to archive user progress, review intervals, and performance metrics; and an Adaptive Scheduling API, which ascertains optimal review timelines based on personalised performance data. The backend facilitates seamless transitions between study sessions, enabling users to resume from their previous point with tailored information.

The frontend, developed with React, provides an interactive and captivating user experience. Interactive Flashcards function as the principal educational instrument, enabling users to interact with dynamically organised flashcards while obtaining prompt feedback. These flashcards features annotated illustrations for visual learners

and pronunciation guidelines for biological terms to support auditory learners. This guarantees a comprehensive learning experience by activating various cognitive pathways. The interface incorporates gamification elements, like leaderboards, badges, and streak trackers, to augment incentive and engagement. By employing a straightforward user interface, learners may monitor their progress seamlessly and maintain motivation through engaging challenges.

The system perpetually alters review schedules in real time through User Performance Analysis, modifying difficulty and frequency according to retrieval success. Progress Visualisation is facilitated via dashboards that offer insights into learning trends, strengths, and shortcomings, enabling students to track their progress over time.

Gamification features are effortlessly integrated to promote continuous learning. Streaks and Achievements incentivise users with badges for persistent practice, so encouraging beneficial study habits. Leaderboards establish competitive rankings, motivating students to remain engaged by contrasting their achievement with that of their classmates. Adaptive Challenges progressively escalate in complexity, maintaining an equilibrium between challenge and learning efficacy. These gamification components are meticulously crafted to offer both intrinsic and extrinsic motivation, ensuring sustained user engagement over time.

User engagement metrics and performance analysis will be employed to evaluate the effectiveness of the system in terms of learning outcomes and retention enhancements. Data will be gathered on session frequency, accuracy rates, and time allocated to various modules. Through the analysis of this data, the system can enhance its adaptive scheduling methods to deliver a more customised learning experience. Furthermore, user feedback will be collected to improve the system's usability and efficacy.

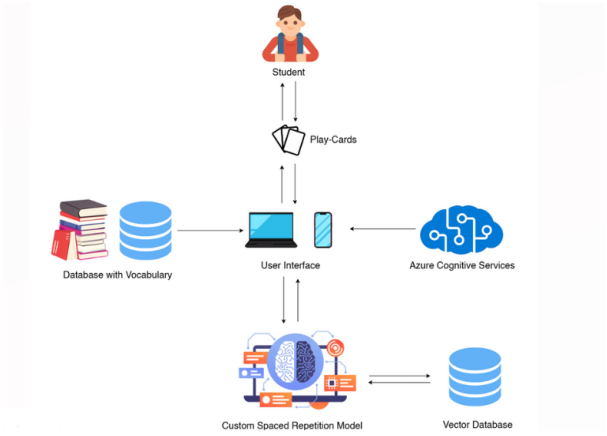

Fig 4: Architecture Diagram of Vocabulary Memorization System

4. Results and Analysis

A. LLM-Based Abstractive Summarization with Voice Output implemented in various Software Architectures.

The summarization component was implemented using monolithic and microservices architectures, evaluated on response time, resource usage, deployment efficiency, and debugging and troubleshooting. The monolithic approach integrated all functionalities within a single application, ensuring faster execution without inter-service communication. In contrast, the microservices architecture separated components into independent services, enhancing scalability and modularity. The table below compares both architectures based on key evaluation metrics.:

Table II: Comparison of Monolithic and Microservices Architectures

Feature	Monolithic Architecture	Microservices Architecture
Response Time	85% faster, no API overhead.	Slower, inter-service delays.

CPU and Memory Usage	Lower (~34% CPU, 28-36% RAM).	Higher (`43-62% CPU, 37-40% RAM).
Deployment Speed	47% faster (37.8 min).	Slower (71.5 min)
Debugging	Easier, centralized logs.	Harder, distributed logs.
Infrastructure	Simple, single container.	Complex, multiple containers.

The performance of the summarization component was evaluated using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, a widely used metric for assessing summarization quality [3]. Unlike traditional accuracy-based evaluation metrics that require exact text matching, ROUGE allows for more flexible evaluation by measuring word and phrase overlap between generated summaries and reference summaries. The Flan-T5 Base model was tested on multiple biology-related educational documents, and the results are summarized in the following table:

Table III: ROUGE Score Evaluation Metrics

Metric	**Score**
ROUGE-1 (Unigram Overlap)	0.74
ROUGE-2 (Bigram Overlap)	0.40
ROUGE-L (Longest Common Subsequence)	0.54

The ROUGE scores indicate that the Flan-T5 Base model successfully retains critical information while ensuring summary coherence and readability. The high ROUGE-1 score (0.74) reflects the strong keyword retention capability of the model, while the ROUGE-L score (0.54) confirms that the model-generated summaries maintain good sentence structure alignment. The ROUGE-2 score (0.40), which evaluates bigram phrase continuity, suggests that the summarization model effectively preserves fluency while paraphrasing complex information .

B. LLM-Based Answer Generation and Evaluation

The model was evaluated using a dataset of structured and essay-type questions from the Sri Lankan Advanced Level Biology curriculum. The training and evaluation pipeline incorporated parameter-efficient fine-tuning, retrieval-augmented generation (RAG) for contextual retrieval, and a hybrid scoring mechanism. However, assessment relied solely on domain experts, primarily Sri Lankan A/L Biology teachers, as no computational parameters exist for objectively evaluating response quality. Given the complexity and subjectivity of Biology answers, only human evaluation is feasible.

C. LLM-Based Adaptive Quiz for Enhancing Biology MCQ Skills

The model was evaluated using a curated dataset of MCQs from Sri Lanka's G.C.E. Advanced Level Biology past papers. While the training pipeline incorporated parameter-efficient fine-tuning and retrieval-augmented generation (RAG), the evaluation relied solely on human judgment due to the subjective nature of biology question quality. A/L Biology teachers served as domain experts, reviewing generated questions for curriculum alignment, conceptual accuracy, linguistic clarity, and appropriate difficulty. Given the limitations of computational metrics in evaluating educational content, expert feedback provided the most reliable measure of the system's effectiveness and informed iterative improvements to the generation pipeline.

D. Enhancing Vocabulary Memorization through Adaptive Spaced Repetition

Adaptive Spaced Repetition (ASR) improves learning efficiency by dynamically modifying review intervals according to individual performance, hence optimising memory retention. The findings demonstrate that ASR markedly enhances recall rates relative to conventional fixed-interval repetition, as it emphasises challenging topics for more frequent review while extending the intervals for easy ones. Research indicates that learners

utilising ASR have enhanced long-term retention, lower cognitive load, and superior learning outcomes across multiple fields, such as language acquisition and medical education. Nonetheless, its efficacy is contingent upon precise difficulty assessment and user involvement. Subsequent research should enhance adaptation algorithms to better individualise learning.

5. Conclusion

This study presents BioMentor, an AI-powered e-learning tool aimed at assisting Sri Lankan A/L Biology students by tackling issues inherent in conventional teaching. BioMentor personalises the learning experience by incorporating advanced AI techniques, like abstractive summarisation, adaptive quizzes, automated answer assessment, and spaced repetition, to improve engagement and retention. The system's use of transformer models and Retrieval-Augmented Generation (RAG) exhibits substantial enhancements in content relevance and contextual comprehension.

The trial findings confirm the platform's efficacy in customising educational resources to meet individual student requirements. The adaptive quiz system modifies difficulty in response to performance, while automatic answer evaluation guarantees precise assessments with tailored feedback. Moreover, the incorporation of spaced repetition enhances long-term retention, solidifying learning results.

Future study may investigate the enhancement of BioMentor's functionalities to encompass new subjects and the incorporation of real-time data to optimise adaptive learning methodologies. The research highlights the revolutionary effect of AI in education, facilitating more accessible and individualised digital learning settings.

Refrences

[1] D. Sudharson et al., “An Abstractive Summarization and Conversation Bot Using T5 and its Variants,” ICAICCIT 2023, IEEE, pp. 431–437.
[2] K. Maurya et al., “NLP-Enhanced Long Document Summarization: A Comprehensive Approach for Information Condensation,” 2024 2nd Int. Conf. on Advancement in Computation & Computer Technologies (InCACCT), pp. 187–192.
[3] M. Ramina et al., “Topic Level Summary Generation Using BERT-Induced Abstractive Summarization Model,” Proc. ICICCS 2020, IEEE, pp. 747–752.
[4] A. Goyal et al., “TalkifyPy: The Pythonic Voice Assistant,” 2024 1st Int. Conf. on Advanced Computing and Emerging Technologies (ACET), IEEE, DOI:10.1109/ACET61898.2024.10730081.
[5] J. Christian et al., “Analyzing Microservices and Monolithic Systems: Key Factors in Architecture, Development, and Operations,” IC2IE 2023, IEEE, pp. 64–69.
[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv preprint arXiv:1810.04805.
[7] Brown, T., Mann, B., Ryder, N., et al. (2020). “Language Models are Few-Shot Learners.” NeurIPS 2020.
[8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.
[9] Dutta, H. S., & Das, B. (2020). "Hybrid AI Models for Educational Applications." Journal of Artificial Intelligence Research, vol. 67, pp. 231-245.
[10] A. S. K. Shukla, D. Arora, and A. K. Sharma, "Automatic Question Answer Generation Using T5 and NLP," IEEE ICCCA, 2019.
[11] P. Kumar, N. Agarwal, and R. Nath, "Generation of Multiple-Choice Questions From Textbook Contents of School-Level Subjects," IEEE ICCCA, 2019.
[12] S. Kumar and M. Gupta, "Automatic Question Generation for Intelligent Tutoring Systems," IEEE ICCCA, 2019.
[13] A. R. Patel, P. K. Jha, and S. Roy, "MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning," IEEE ICCCA, 2019.
[14] R. Sharma and K. Singh, "Generation of Multiple-Choice Questions From Indian Educational Text," IEEE ICET, 2023.

[15] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer, "Distributed practice in verbal recall tasks: A review and quantitative synthesis," Psychological Bulletin, vol. 132, no. 3, pp. 354–380, 2006.
[16] H. L. Roediger and A. C. Butler, "The critical role of retrieval practice in long-term retention," Trends in Cognitive Sciences, vol. 15, no. 1, pp. 20–27, 2011.
[17] P. Pavlik and J. R. Anderson, "Using a model to compute the optimal schedule of practice," Journal of Experimental Psychology: Applied, vol. 14, no. 2, pp. 101–117, 2008.
[18] S. H. Kang, "Spaced repetition promotes efficient and effective learning: Policy implications for instruction," Policy Insights from the Behavioral and Brain Sciences, vol. 3, no. 1, pp. 12–19, 2016.
[19] J. D. Karpicke and A. Bauernschmidt, "Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 37, no. 5, pp. 1250–1257, 2
[20] Y. Tomikawa, A. Suzuki, and M. Uto, "Adaptive Question–Answer Generation With Difficulty Control Using Item Response Theory and Pretrained Transformer Models," IEEE Transactions on Learning Technologies