Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.11 n.54, p. 188-197. July/August., 2025.
DOI: https://doi.org/10.5935/jetia. v11i54.1743

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

QUIZ-TUBE: ENHANCING VIDEO-BASED LEARNING WITH AUTOMATED AI QUIZ GENERATION

Akash Thakur[1], Rashmi Bhat[2], Ayush Shukla[3], Tushar Ram[4] and Vivek Singh[5]

[1,2,3,4,5] St. John College of Engineering and Management, Palghar, Maharashtra.

[1]https://orcid.org/0009-0006-4704-6374, [2]https://orcid.org/0009-0003-3064-1247, [3]https://orcid.org/0009-0005-9677-5509
[4]https://orcid.org/0009-0006-2523-418X, [5]https://orcid.org/0009-0005-5429-8265

Email: akashthakur4553@gmail.com, rashmib@sjcem.edu.in , ayushshukla7291@gmail.com, tusharram085@gmail.com, viveksingh92317@gmail.com

ARTICLE INFO

Article History
Received: March 22, 2025
Revised: April 20, 2025
Accepted: June 15, 2025
Published: August 31, 2025

Keywords:
Large Language Model,
Natural Language processing,
Video-based learning,
model fine-tuning,
Question generation

ABSTRACT

In the digital age, video-based learning has become a dominant educational medium, with platforms like YouTube offering a vast repository of instructional content. However, passive video consumption often leads to suboptimal knowledge retention and limited engagement. To address this challenge, we introduce Quiz-Tube, an AI-powered web platform that converts educational YouTube videos into interactive quizzes. By leveraging a fine-tuned Gemma-9B language model, Quiz-Tube generates contextually relevant multiple-choice questions (MCQs) tailored to users' preferences. The system integrates transcript extraction via the YouTube Transcript API, followed by data preprocessing, question generation, and quiz customization. Users can define quiz parameters such as difficulty level and question count, enhancing personalized learning. Our evaluation demonstrates significant performance improvements, including a BLEU score increase from 45.3 to 68.7 and an accuracy boost from 64.5% to 81.9% post-model fine-tuning. Future enhancements include multilingual support and AI-driven adaptive learning.

I. INTRODUCTION

The expansion of online education has revolutionized the way people access and interact with learning materials. With the rapid advancement of digital platforms, particularly video-based platforms such as YouTube, students and professionals now have unprecedented access to vast repositories of educational content spanning diverse subjects. The rise of video-based learning has played a transformative role in modern education, offering flexibility and accessibility that traditional learning methods often lack. Learners can explore complex topics at their own pace, revisit content as needed, and gain insights from experts worldwide. However, despite these advantages, video-based learning presents inherent challenges. Passive video consumption does not actively engage learners, which can lead to poor knowledge retention, decreased motivation, and limited comprehension.

Research in educational psychology and cognitive science suggests that active learning techniques significantly enhance comprehension and retention [1]. Active learning involves engaging learners through self-assessment, reflection, and interactive exercises, which are proven to strengthen understanding and long-term recall. Among these techniques, quizzes are particularly effective in reinforcing concepts, testing understanding, and ensuring active participation. By prompting learners to recall information, quizzes help consolidate knowledge and bridge the gap between passive consumption and active engagement. However, the manual creation of quizzes from video content is a labor-intensive and time-consuming task, making it impractical for both educators and learners who consume vast amounts of online educational material daily.

As the demand for personalized and interactive learning experiences continues to grow, there is a pressing need for an automated solution capable of generating relevant and customized quizzes from video content. Addressing this challenge, Quiz-Tube introduces an AI-driven approach that leverages large language models (LLMs) [2] and natural language processing (NLP) to convert YouTube video

transcripts into interactive quizzes. By utilizing the fine-tuned Gemma-9B model [3], Quiz-Tube ensures the generation of contextually accurate and personalized multiple-choice questions (MCQs) tailored to individual learning preferences and proficiency levels.
Quiz-Tube enhances user engagement by allowing learners to define quiz parameters, such as difficulty levels, the number of questions, and topic focus. This customization empowers users to tailor their learning experience, making it more adaptive and suited to their specific educational needs. Additionally, by integrating natural language understanding and deep learning techniques, Quiz-Tube can analyze video transcripts, extract key concepts, and generate well-structured assessment items that align with the core ideas presented in the video. This automation streamlines the quiz creation process, reducing the workload on educators while enabling self-learners to engage more effectively with educational content.

This paper details the system architecture, methodology, and performance evaluation of Quiz-Tube, demonstrating its effectiveness in enhancing active learning engagement. We explore the technical aspects of AI model training, quiz generation processes, and user experience design. Additionally, we discuss future improvements that will further expand the platform's impact on digital education. These enhancements include multilingual support, which will enable non-English speakers to benefit from the platform, AI-driven personalization to further tailor quizzes based on individual learning patterns, and mobile integration to increase accessibility for users who prefer learning on-the-go. By addressing the challenges of passive video-based learning and bridging the gap between content consumption and active engagement, Quiz-Tube represents a significant advancement in the field of digital education. This research aims to contribute valuable insights into the intersection of artificial intelligence and online learning, paving the way for more interactive, personalized, and effective educational experiences. Through continued development and refinement, Quiz-Tube has the potential to revolutionize how learners engage with video-based educational content, fostering a more dynamic and enriching learning ecosystem.

II. RELATED REVIEWS

A variety of approaches have been explored in the field of large language models (LLMs) and automated question generation, with applications spanning financial decision-making, education, and efficient AI-driven assessments. Fine-tuning domain-specific LLMs [4] has shown significant promise in enhancing decision-making and automation tasks, particularly when domain vocabulary creation and regulatory compliance are carefully managed. While financial applications have been a key focus, the potential of fine-tuned models extends to other domains, such as education, where models can dynamically generate quizzes from video transcripts and tailor content to user preferences.

Automated multiple-choice question generation (MCQG) has been extensively studied using NLP techniques. Early methodologies relied on static text materials, employing techniques such as TF-IDF and N-grams for keyword extraction and question formulation. While effective, these methods lack integration with multimedia content and fail to incorporate user-driven customization. Advancements in LLMs have enabled more dynamic approaches, allowing quiz generation from video transcripts and enhancing contextual relevance through fine-tuned models [5]. Another approach involves leveraging advanced prompt engineering and retrieval-augmented generation (RAG) techniques to improve question quality. By incorporating chain-of-thought and self-refine prompting, more contextually relevant and challenging questions can be produced. A key distinction in modern systems is the reliance on educational video transcripts, allowing for the extraction of meaningful information and the creation of tailored datasets. This approach not only improves question quality but also integrates user quiz histories to refine personalization and adaptability [6].

Fine-tuning pre-trained language models (PLMs) for educational question generation has also gained traction. Some models focus on additional pre-training using domain-specific datasets, such as those containing scientific educational questions. The emphasis on linguistic fluency and contextual relevance aligns with the goal of generating high-quality quizzes from educational video transcripts. While some approaches rely primarily on textual corpora, integrating transcript APIs broadens real-world applications, making AI-driven quiz generation accessible across diverse subjects and formats [7]. Efforts to optimize memory and computational efficiency in LLMs have been explored through post-training quantization techniques. These methods aim to reduce bit-width while maintaining performance across tasks such as dialogue processing and long-context understanding. While not directly implemented in all quiz generation systems, such techniques hold promise for enhancing scalability, particularly when processing large volumes of educational transcripts or generating multiple-choice quizzes in real time [8].

Further research into quantization has investigated strategies to mitigate information loss through fine-tuning techniques. Methods incorporating information calibration quantization and elastic connection strategies have been proposed to retain model accuracy while reducing resource consumption. These approaches align with the broader goal of efficiency and customization in LLM-based applications, demonstrating the potential for optimizing models to generate adaptive quizzes based on user-selected difficulty levels [9]. Improvements in retrieval-augmented generation have also contributed to enhancing quiz generation processes. By employing long-context LLMs, retrieval methods have been optimized to minimize contextual loss, leading to higher retrieval accuracy and improved content generation. This shift from short retrieval units to longer retrieval mechanisms reduces the burden on information retrieval systems while ensuring the accuracy of generated questions. These advancements further enable large-scale educational applications, allowing for the seamless integration of AI-driven learning tools into modern e-learning platforms.[10]

The automation of test creation using AI has been explored as a means to streamline educational assessments. AI-based systems can generate test questions based on predefined parameters such as difficulty level, question type, and cognitive skills classification. By leveraging cloud-based platforms and API-driven integrations, these systems enhance flexibility, adaptability, and efficiency in educational settings. Automated test generation offers a scalable solution for educators seeking to implement personalized assessments without extensive manual effort [11]. Beyond multiple-choice question generation, research has also explored the use of large language models for subjective question formulation. Some systems have been developed to generate opinion-based questions derived from news articles and trending topics, leveraging sequence-to-sequence generation techniques. This research highlights the adaptability of LLMs for diverse question-generation applications beyond factual recall, extending their utility to discussion-based and critical thinking exercises [12].

Self-play fine-tuning has emerged as a novel technique for enhancing LLM performance without requiring extensive human-annotated data. Unlike traditional fine-tuning approaches, self-play mechanisms enable models to iteratively train themselves by generating and evaluating their own data. This method improves response quality by aligning model-generated outputs with human-like distributions. Studies have demonstrated that this technique can surpass traditional optimization strategies, paving the way for improved performance in educational AI applications. Reinforced self-training has also been introduced as an alternative method for fine-tuning language models. By integrating reinforcement learning techniques, models can refine their outputs iteratively using reward-based mechanisms. This approach enhances generalization capabilities, particularly in domains requiring logical reasoning and complex problem-solving. By balancing exploration and exploitation, reinforcement-based self-training minimizes training instability while optimizing model performance for educational tasks [13].

The interaction between parametric knowledge stored within LLMs and externally retrieved information has also been a subject of study. Research has shown that retrieval-augmented generation methods improve factual accuracy by prioritizing external context over pre-trained memory. Understanding these interactions is crucial for optimizing AI-driven quiz generation systems, as it helps ensure that generated questions remain accurate and contextually relevant. Such insights contribute to refining LLM-based educational tools and optimizing their ability to process diverse input sources effectively. Collectively, these advancements in LLM fine-tuning, retrieval-augmented generation, quantization, and self-training methodologies highlight the transformative potential of AI in educational applications. The ability to generate dynamic quizzes, personalize learning experiences, and enhance efficiency through optimized AI models paves the way for a more interactive and effective digital learning ecosystem. As research continues to evolve, further integration of scalable AI-driven solutions in education will contribute to more adaptive and accessible learning methodologies, ultimately improving knowledge retention and learner engagement.

III. METHODOLOGY

III.I SYSTEM DESIGN

The QuizTube system is meticulously engineered to automate the generation of quizzes from YouTube educational videos by leveraging advanced Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques. The architecture is designed as a structured pipeline, enabling users to input a video URL and receive an interactive quiz, complete with evaluation results and certification. This comprehensive system comprises several interconnected modules that collaboratively process video content, generate quizzes, and assess user performance.

1. User Interaction and Dashboard

The user journey begins with an intuitive dashboard that allows learners to input a YouTube video URL. This central interface not only facilitates quiz generation but also enables users to manage their quiz preferences, track past results, and obtain certifications. Designed for accessibility across various educational levels, the dashboard provides access to historical quiz performances, allowing learners to monitor their progress over time. Features such as progress bars, feedback visualization, and adaptive quiz difficulty based on past performance enhance user engagement and learning outcomes.

2. Video Processing Module

Upon submission of a YouTube URL, the system initiates a multi-stage process to extract and process the video content. Recognizing that YouTube videos contain both visual and auditory information, the system first isolates the audio stream to capture spoken content accurately. The extracted audio is then transcribed into text using speech recognition algorithms, transforming raw spoken content into a structured text format suitable for further processing by the Large Language Model (LLM). This transcription ensures that the entire lecture or discussion is accurately captured, preserving all relevant details necessary for effective quiz generation.

3. AI Processing and Quiz Generation

The transcribed text undergoes sophisticated NLP processing using a fine-tuned Gemma-9B language model. This AI-powered module is responsible for extracting key concepts and main topics from the transcribed text, ensuring the retention of the most pertinent information. The model generates quiz questions of varying difficulty levels, allowing users to test their knowledge across different cognitive domains. The generated questions maintain contextual relevance and coherence with the original lecture, ensuring that the quiz accurately reflects the video content. The use of Retrieval-Augmented Generation (RAG) techniques ensures that the generated questions remain accurate and contextually aligned with the video content. Additionally, the AI-powered question generation system ensures grammatical correctness and diversity in the types of questions generated.

4. Quiz Evaluation and Results Analysis

Once the quiz is generated, users can attempt it in real-time through the QuizTube platform. The system automatically scores multiple-choice, providing instant feedback on correct and incorrect answers. It assesses learning progress by tracking historical quiz performance, helping learners identify their strengths and weaknesses. This interactive feedback mechanism enhances the learning process by offering insights into areas needing improvement. The results analysis feature contributes to a personalized learning experience by adjusting quiz difficulty based on the user's previous attempts.

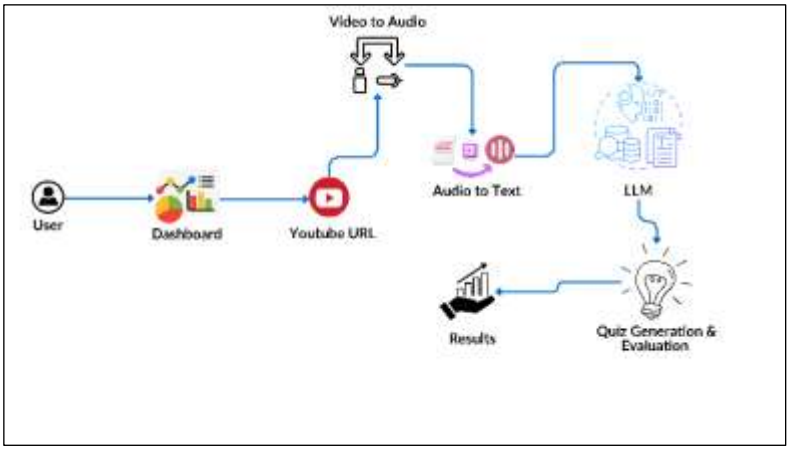

Figure 1: Flowchart of the system.w.
Source: Authors, (2025).

III.1.2 Algorithm:

The algorithmic framework for QuizTube encompasses a structured approach for transforming YouTube educational videos into well-defined, interactive quizzes. The following sections detail the step-by-step process involved in the algorithm's design and implementation:

1. Input Processing and Transcript Extraction

The algorithm begins with acquiring user input, primarily in the form of a YouTube video URL. Upon receiving this URL, the system verifies its validity and attempts to extract the transcript using the YouTube Transcript API.

- **Transcript Retrieval:** The system requests the transcript associated with the video. If the transcript is not available (e.g., due to copyright restrictions or non-English content), the system returns an error message indicating the unavailability of transcription.
- **Audio-to-Text Conversion (if needed):**For videos without transcripts, the audio stream is isolated and converted into text using advanced speech recognition techniques. The algorithm ensures accurate transcription by leveraging pre-trained speech-to-text models optimized for educational content**.**

2. Preprocessing and Text Segmentation

Once the transcript is successfully retrieved, the text undergoes extensive preprocessing to enhance the quality of subsequent processing stages.

- **Noise Removal**: Timestamps, special characters, and irrelevant data (such as advertisements or unrelated dialogue) are removed.
- **Text Normalization:** The text is converted to lowercase, and unnecessary punctuation marks are discarded.
- **Tokenization:** The cleaned text is divided into individual sentences and paragraphs to facilitate easier segmentation.
- **Keyword Extraction:** The Term Frequency-Inverse Document Frequency (TF-IDF) technique is applied to identify the most significant keywords from the transcript. Named Entity Recognition (NER) techniques are also utilized to detect important concepts and topics.

3. Knowledge Graph Construction

A knowledge graph is constructed to map relationships between identified entities and concepts. This step improves the contextual understanding of the text and aids in generating meaningful questions. Techniques such as entity linking and relationship extraction are employed, often supported by knowledge bases like ConceptNet or Wikidata.

4. AI-Based Question Generation

The core of the QuizTube algorithm revolves around generating high-quality quiz questions using the fine-tuned Gemma-9B model.

- **Question Formulation:** The pre-processed text is fed into the Gemma-9B model, which utilizes context-aware filtering and retrieval-augmented generation (RAG) to produce relevant questions.
- **Question Categorization:** The algorithm categorizes questions into three difficulty levels:
  - **Easy:** Direct, fact-based questions aimed at recalling specific information.
  - **Medium:** Conceptual questions requiring comprehension and contextual application.
  - **Hard:** Analytical questions demanding multi-step reasoning or evaluation.

5. Quiz Customization and Formatting

To provide a personalized learning experience, the algorithm allows users to customize quizzes based on their preferences.

- **Customization Options:** Users can select the number of questions (e.g., 5, 10, 15) and choose difficulty levels based on their learning objectives.

- **Content Filtering:** The system filters out questions that are redundant, grammatically incorrect, or poorly structured. This process ensures the coherence and contextual alignment of generated quizzes.

6. User Interaction and Feedback Analysis

The algorithm integrates an interactive quiz interface where users can attempt generated quizzes.

- **Real-Time Evaluation:** Users receive immediate feedback for each question, indicating whether their answer was correct or incorrect.
- **Score: Calculation:** A percentage-based grading system is used to compute the user's overall performance.
- **Performance: Analysis:** Detailed reports are generated, including accuracy rates, topic-wise strengths and weaknesses, and time spent per question.

7. Adaptive Learning Mechanism

Based on the user's quiz performance, the system adapts future quizzes to provide a tailored learning experience.

- **Difficulty Adjustment:** If a user consistently excels, the system gradually increases the difficulty level to maintain engagement and enhance knowledge retention.
- **Topic Reinforcement:** For areas where users demonstrate weakness, the system provides additional quizzes targeting those specific topics.

8. Model Fine-Tuning and Optimization

The fine-tuning of the Gemma-9B model is a crucial aspect of the QuizTube framework, ensuring high-quality question generation.

- **Data Augmentation:** Synthetic question-answer pairs are generated using Gemini-1.5-flash to expand the training dataset, enhancing the model's ability to generalize across diverse educational content.
- **Transfer Learning:** The Gemma-9B model is pretrained on large-scale text corpora and fine-tuned specifically for educational content extraction and quiz generation.
- **Regularization Techniques:** Dropout, gradient clipping, and early stopping mechanisms are employed to prevent overfitting during training.
- **Evaluation Metrics:** Performance is measured using metrics like BLEU Score, F1 Score, and Perplexity.

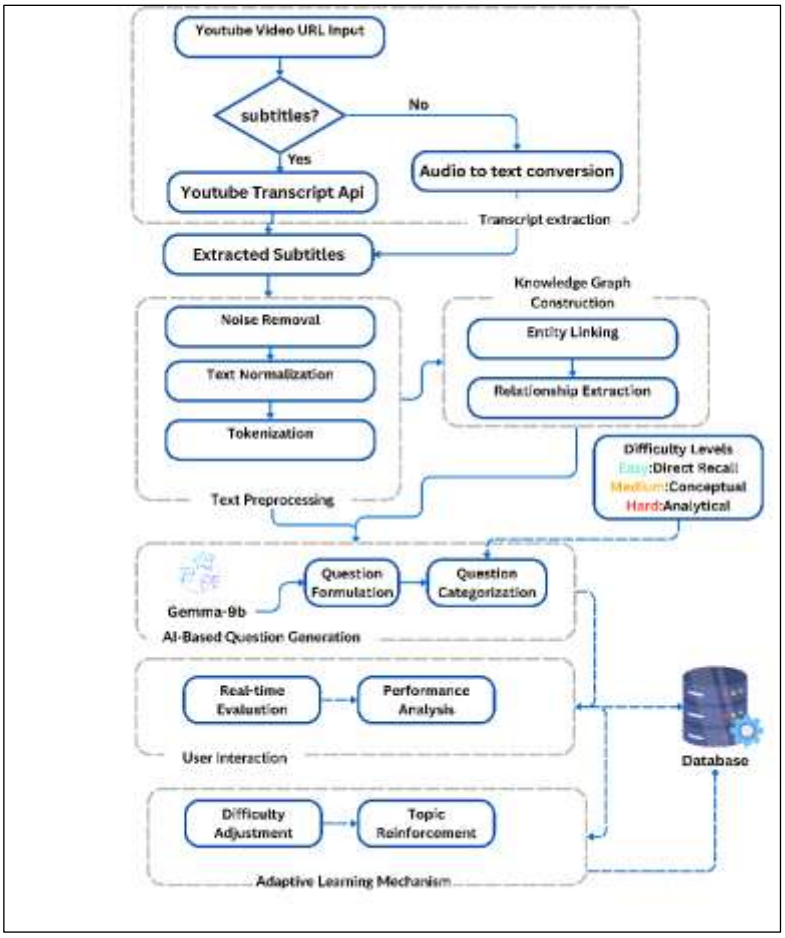

Figure 2: QuizTube Algorithmic Framework
Source: Authors, (2025).

III.II GEMMA-9B MODEL

The Gemma-9B model is an advanced large language model designed for high-accuracy natural language processing (NLP) tasks. It is a 9-billion-parameter model built on Google's Griffin architecture, which integrates gated linear recurrences with local sliding window attention mechanisms. This hybrid approach enhances computational efficiency and is particularly adept at handling long-context prompts. It features 9 billion parameters, making it highly capable of understanding and generating human-like text. The architecture leverages a Transformer-based framework with multiple attention heads and deep feed-forward networks, allowing it to capture complex linguistic patterns and contextual nuances. This model has been optimized for tasks such as text generation, summarization, question answering, and language translation.

III.2.1 Griffin Architecture

The Griffin architecture combines the strengths of recurrent neural networks (RNNs) and attention mechanisms:

- **Gated Linear Recurrences:** These components allow the model to process sequences with a form of memory, capturing temporal dependencies effectively.
- **Local Sliding Window Attention:** This mechanism enables the model to focus on specific parts of the input sequence, improving context management without the extensive computational load associated with global attention.

This architecture ensures that Gemma-9B can manage long sequences efficiently, making it suitable for tasks requiring extensive context understanding.

III.2.2 Model Specifications

Gemma-9B's architecture includes:
1]Multi-Head Self-Attention: Facilitates parallel processing of input sequences to extract contextual representations.
2]Layer Normalization: Enhances stability during training and fine-tuning.
3]Feed-Forward Networks (FFN): Ensures the model can handle non-linear transformations efficiently.
4]Positional Encoding: Maintains word order and contextual relevance within the input text.
5]Residual Connections: Prevents gradient vanishing issues by maintaining information flow between layers.
Collectively, these specifications empower Gemma-9B to perform a wide range of NLP tasks with high proficiency. All Gemma variants operate with a substantial context window of 8,192 tokens, enabling the processing of approximately 6,144 words simultaneously. This extended context capacity allows the model to maintain coherence across longer text sequences and better understand complex, multi-part prompts. However, practitioners should note that effective input capacity fluctuates during interactive use, as generated text consumes portions of the available context window, thereby constraining the space available for new inputs in continuing interactions.

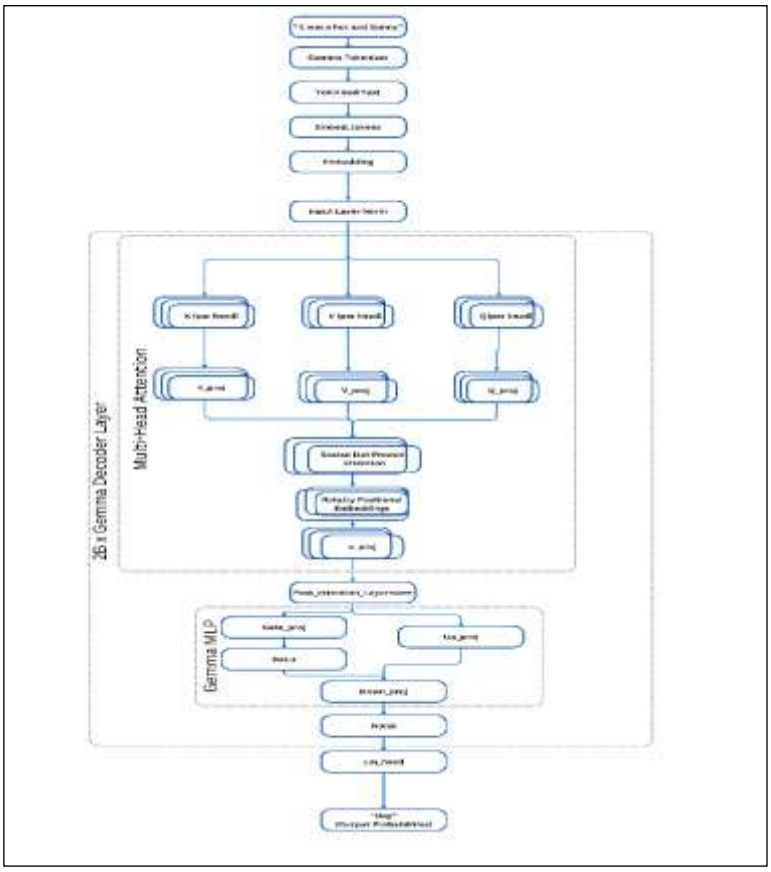

Figure 3: Gemma Architecture.
Source: Authors, (2025).

III.2.3 Key Architectural Parameters and Model Variations

The Gemma architecture is defined by several critical parameters that collectively determine its computational capabilities and performance characteristics:

1] Embedding Dimensions (d_model): The embedding dimension establishes the vector size used to represent individual tokens and significantly influences the model's capacity to capture linguistic nuances and semantic relationships. The family offers progressive embedding dimensionality across its scale variants:

- The 2B model utilizes a 2,048-dimensional embedding space
- The 7B variant employs a more expansive 3,072-dimensional representation
- The 9B model (Gemma 2) further increases dimensional capacity to 3,584, providing enhanced representational capabilities

These dimensional increases correlate with improved performance on complex language tasks, though at the cost of greater computational requirements for both training and inference.

2] Network Depth and Layer Organization: The architectural depth varies significantly across model sizes:

- The 2B configuration implements 18 transformer layers
- The 7B version incorporates 28 layers
- The 9B model features 32 layers

This progressive depth scaling provides increasingly sophisticated pattern recognition capabilities through deeper transformations of the input representations. The additional layers allow higher-parameter models to capture more intricate linguistic patterns, long-range dependencies, and abstract semantic relationships. However, this depth also increases the risk of overfitting and requires substantially more computational resources and training data to achieve proper generalization. The internal organization of each layer follows a consistent pattern across all model sizes, incorporating self-attention mechanisms followed by feed-forward networks, with normalization layers strategically positioned to stabilize training dynamics.

3] Feed-Forward Networks and Activation Functions: Each transformer layer contains a feed-forward network that significantly expands the representational capacity through dimensional projection. The hidden dimensions of these networks scale proportionally with model size:

- 32,768 dimensions in the 2B model
- 49,152 dimensions in the 7B variant
- 57,344 dimensions in the 9B implementation

These expansive hidden representations allow the model to perform complex non-linear transformations on the token embeddings. Notably, Gemma replaces the traditional ReLU activation function with the more sophisticated GeGLU (Gated Linear Unit with GELU activation) function. This advanced activation approach divides processing into parallel sigmoidal and linear projection pathways, with the element-wise multiplication of these pathways producing a more expressive non-linear transformation. The gating mechanism enables more selective information flow, allowing the model to dynamically emphasize relevant features while suppressing irrelevant information.

4] Attention Mechanisms and Efficiency Optimizations: The attention architecture represents a key differentiator between Gemma variants, reflecting sophisticated performance/efficiency trade-offs. The 7B and 9B models implement full Multi-Head Attention (MHA) with 16 attention heads (9B model extends this to 24 heads), enabling the model to simultaneously attend to different aspects of the input sequence from various representational perspectives. This approach maximizes representational capacity but incurs significant computational costs. In contrast, the 2B version employs Multi-Query Attention (MQA) with 8 query heads but only a single key-value head. This architectural optimization dramatically reduces parameter count and memory bandwidth requirements by sharing key and value projections across all query heads. The efficiency gains come with minimal performance degradation on many tasks, making it particularly suitable for resource-constrained environments.
All variants maintain a consistent head size of 256 dimensions, calculated by dividing the embedding dimension by the number of attention heads. This dimensionality provides sufficient representational capacity within each attention head while keeping computational requirements manageable. The models employ rotary positional embeddings (RoPE), which encode positional information directly within the key and query vectors through rotation in the complex plane. This approach provides several advantages over traditional positional encodings, including better extrapolation to sequence lengths beyond those seen during training.

5] Vocabulary and Tokenization Approach: All Gemma models feature an extensive vocabulary of 256,128 tokens derived using the SentencePiece tokenization methodology. This expansive token space enables efficient encoding of diverse linguistic patterns across multiple languages and domains. The SentencePiece approach dynamically learns optimal sub-word segmentation patterns based on training data distribution, allowing the model to effectively handle out-of-vocabulary words through sub-word decomposition. The large vocabulary particularly enhances the model's versatility for multilingual applications and specialized technical domains that may feature unique terminology. This capacity for nuanced tokenization contributes significantly to Gemma's performance on language understanding tasks across diverse domains.

III.III FINE-TUNING THE GEMMA MODEL

Fine-tuning large language models (LLMs) like Gemma-9B necessitates sophisticated strategies to adapt them effectively to specific domains. This section explores advanced fine-tuning methodologies, including Parameter-Efficient Fine-Tuning (PEFT) and Representation Fine-Tuning (ReFT), to optimize the model using a dataset of 1,000 educational transcripts.

III.3.1 Fine-Tuning Techniques

1] Parameter-Efficient Fine-Tuning (PEFT):

Traditional fine-tuning methods involve adjusting all parameters of a pre-trained model to tailor it to a downstream task. While effective, this approach becomes impractical for LLMs due to their massive parameter counts, leading to increased computational costs and memory requirements. PEFT addresses this issue by modifying only a small subset of the model's parameters, thereby preserving the majority of the pre-trained weights. This strategy not only reduces the computational burden but also mitigates the risk of overfitting, especially in scenarios with limited task-specific data.

Several methodologies have been developed under the umbrella of PEFT, each offering unique mechanisms to achieve parameter efficiency. Adapters, for instance, are small neural network modules inserted between the layers of the pre-trained model. They adjust the output of each layer to better fit the target task without altering the original weights, enabling task-specific customization with minimal parameter updates and facilitating efficient multi-task learning. Low-Rank Adaptation (LoRA) introduces trainable low-rank matrices into each layer of the transformer architecture. By decomposing weight updates into low-rank matrices, LoRA reduces the number of trainable parameters, maintaining model performance while significantly lowering the computational burden. This makes it feasible to fine-tune large models on domain-specific data without extensive resource requirements. Prefix Tuning focuses on optimizing continuous task-specific vectors, known as prefixes, that guide the model during generation. By prepending these learnable vectors to the input, prefix tuning steers the model's behaviour without modifying its core parameters, allowing for rapid adaptation to new tasks with a minimal computational footprint.

The primary advantage of PEFT lies in its ability to adapt large models to specific tasks without the need for extensive computational resources. By fine-tuning only a small subset of parameters, PEFT methods significantly reduce memory usage and training time, making them accessible for applications with limited resources. Additionally, by preserving the majority of the pre-trained weights, PEFT minimizes the risk of catastrophic forgetting, ensuring that the model retains its general language understanding capabilities while adapting to new tasks.

PEFT has been successfully applied across various NLP tasks, including sentiment analysis, named entity recognition (NER), and question answering. These applications demonstrate PEFT's versatility in enabling LLMs to perform effectively across diverse tasks without the need for full-scale model retraining. Despite its advantages, PEFT is not without challenges. Determining the optimal subset of parameters to fine-tune requires careful consideration, and there is an ongoing need to balance parameter efficiency with model performance. Future research in PEFT aims to develop more sophisticated methods for selecting trainable parameters and to explore the integration of PEFT with other model optimization techniques to further enhance the adaptability and efficiency of LLMs [14].

2] Representation Fine-Tuning (ReFT):

Representation Fine-Tuning (ReFT) is an emerging paradigm in the field of natural language processing (NLP) that focuses on adapting large language models (LLMs) by intervening on their internal representations rather than their weights. This approach leverages the rich semantic information encoded within the model's hidden states to achieve task-specific adaptation with enhanced parameter efficiency.

Traditional fine-tuning methods adjust the weights of a pre-trained model to tailor it to a specific task. While effective, this approach can be computationally intensive, especially for LLMs with billions of parameters. ReFT offers an alternative by modifying the hidden representations of the model, allowing for task adaptation without altering the model's parameters. This method can be more parameter-efficient than traditional fine-tuning, as it leverages the rich semantic information encoded in the representations.

A notable implementation of ReFT is Low-Rank Linear Subspace ReFT (LoReFT). LoReFT operates by projecting the model's representations into a low-dimensional subspace and learning task-specific interventions within this space. This technique achieves significant parameter efficiency, often requiring updates to less than 1% of the model's parameters while maintaining or improving performance. LoReFT has demonstrated effectiveness across various NLP tasks, including commonsense reasoning, arithmetic reasoning, and instruction-tuning, delivering a favourable balance of efficiency and performance.

The advantages of ReFT are manifold. By focusing on the model's internal representations, ReFT methods can achieve high levels of parameter efficiency, reducing the computational resources required for fine-tuning. This approach also allows for rapid adaptation to new tasks, as the interventions on representations can be learned quickly without the need for extensive parameter updates. Furthermore, by preserving the original weights of the model, ReFT minimizes the risk of catastrophic forgetting, ensuring that the model retains its general language understanding capabilities while adapting to new tasks.

However, ReFT is not without challenges. Determining the optimal interventions on the model's representations requires careful consideration, and there is an ongoing need to balance parameter efficiency with model performance. Future research in ReFT aims to develop more sophisticated methods for learning interventions on representations and to explore the integration of ReFT with other model optimization techniques to further enhance the adaptability and efficiency of LLMs.

In conclusion, Representation Fine-Tuning represents a promising direction in the adaptation of large language models. By leveraging the rich semantic information encoded within the model's representations, ReFT offers a pathway to achieve task-specific adaptation with enhanced parameter efficiency, paving the way for more efficient and effective utilization of LLMs in various NLP applications [15].

Table No. 1: Comparative Analysis of Fine-Tuning Techniques.

Technique	Efficient	Accuracy	Generalization	complex
PEFT	High	High	Medium	Low
REFT	Medium	Medium	High	Medium

Source: [16].

III.3.2 Implementing the fine-tuning process.

The implementation of the fine-tuning process for the Gemma-9B model involves a series of carefully structured steps aimed at enhancing the model's performance for the QuizTube application. The process is divided into three main phases: data preparation, fine-tuning, and evaluation.

In the Data Preparation phase, the transcripts gathered from educational YouTube videos are thoroughly cleaned and preprocessed to ensure consistency and accuracy. This cleaning process involves removing irrelevant symbols, correcting transcription errors, and normalizing text for uniformity. Once the text is cleaned, it undergoes tokenization, where the text is divided into smaller units or tokens that are compatible with Gemma-9B's tokenizer. Tokenization ensures that the model can accurately process and understand the input data, preparing it for the subsequent fine-tuning phase.

The Fine-Tuning Process begins with a baseline evaluation to assess the original performance of the pre-trained Gemma-9B model on the target task. This initial evaluation provides a reference point against which improvements can be measured. Once the baseline performance is established, the selected fine-tuning technique, that is Low-Rank Adaptation (LoRA), is applied to the model. This technique aims to enhance the model's performance while maintaining computational efficiency by only updating a subset of parameters or manipulating internal representations. During the training process, the model is fine-tuned on the prepared dataset, with performance metrics monitored continuously to ensure optimal results. The training process involves multiple iterations over the dataset, allowing the model to learn task-specific knowledge effectively.

The final phase, Evaluation, involves assessing the fine-tuned model's performance using a variety of relevant metrics. Commonly used metrics include accuracy, BLEU score and ROUGE-L, which provide a comprehensive understanding of the model's ability to generate accurate and contextually relevant quiz questions. To determine the effectiveness of the fine-tuning process, the results of the fine-tuned model are compared against the baseline performance established earlier. A significant improvement in these metrics would indicate the success of the fine-tuning process, while minimal or negative changes would prompt further adjustments to the fine-tuning methodology.

IV. RESULTS AND DISCUSSIONS

The results of the Quiz-Tube system demonstrate its effectiveness in transforming YouTube educational content into an interactive learning experience. By leveraging AI-driven Natural Language Processing (NLP) techniques, the system successfully generates quizzes that align with the key concepts of the video content. The automated evaluation process provides instant feedback, enabling users to track their performance and improve their understanding of the subject matter. The adaptive learning mechanism personalizes quiz difficulty based on individual performance, ensuring a tailored learning experience.

Table No. 2: Performance improvements.

Metric	Pre-Finetuning	Post-Finetuning
BLEU score	45.3	68.7
ROUGE-L	57.2	74.8
Accuracy (%)	64.5	81.9

Source: Authors, (2025).

The given Table presents a comparative analysis of the Quiz-Tube system's performance before and after fine-tuning using three key evaluation metrics: BLEU score, ROUGE-L, and Accuracy (%).

1]BLEU Score: Measures the overlap between generated and reference text. The model shows a significant improvement from 45.3 (Pre-Finetuning) to 68.7 (Post-Finetuning), indicating enhanced fluency and alignment with expected outputs.

2] ROUGE-L: Evaluates recall-oriented similarity by considering the longest common subsequence between generated and reference text. The model's ROUGE-L score increases from 57.2 to 74.8, reflecting improved text relevance and coherence.

3]Accuracy (%): Represents the proportion of correctly generated quiz questions or answers. The system achieves a substantial accuracy boost from 64.5% to 81.9%, demonstrating better precision in content generation.

V. CONCLUSIONS

QuizTube is a groundbreaking AI-powered platform designed to revolutionize the educational landscape by converting educational YouTube videos into interactive quizzes. Through the integration of cutting-edge AI techniques, particularly the fine-tuned Gemma-9B model, QuizTube effectively generates quizzes that promote active learning and enhance user engagement. By utilizing the YouTube Transcript API, QuizTube efficiently processes and converts spoken content from videos into structured text, enabling the AI model to extract meaningful information and generate contextually accurate questions. This approach ensures that users are presented with high-quality quizzes that reflect the core concepts and knowledge imparted by the original video content.

The fine-tuning of the Gemma-9B model plays a critical role in the platform's success, with Parameter-Efficient Fine-Tuning (PEFT) being employed to enhance the model's adaptability and performance. By training the model on a dataset of 1,000 educational transcripts, the system demonstrates significant improvements in performance metrics. Notably, BLEU scores have increased from 45.3 to 68.7, and ROUGE-L scores have risen from 57.2 to 74.8, indicating a substantial enhancement in the model's ability to generate coherent and contextually relevant questions. The utilization of these fine-tuning methods ensures that QuizTube remains robust and effective even when processing complex and domain-specific educational materials.

QuizTube's architecture is designed to be highly flexible and user-centric. Through its intuitive dashboard, users can seamlessly input YouTube video URLs, customize quiz preferences, and receive immediate feedback on their performance. The platform's ability to adjust quiz difficulty based on the user's historical performance ensures a personalized and adaptive learning experience. This adaptability is particularly valuable for learners with varying levels of expertise, enabling them to progress at their own pace and focus on areas where improvement is needed. Additionally, QuizTube's integration of real-time quiz generation capabilities distinguishes it from traditional e-learning platforms, making it a valuable tool for both educators and independent learners.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Author One, Author Two and Author Three.
Methodology: Author Two.
Investigation: Author Three and Author Four.
Discussion of results: Author Five, Author Four and Author Three.
Writing – Original Draft: Author One and Author Two.
Writing – Review and Editing: Author Three and Author Five.
Resources: Author Two.
Supervision: Author One.
Approval of the final text: Author One, Author Two and Author Three.

VII. REFERENCES

[1] Maceiras, Rocio, et al. "Effectiveness of Active Learning Techniques in Knowledge Retention among Engineering Students." *Education for Chemical Engineers*, vol. 51, Apr. 2025, pp. 1–8.

[2] Naveed, Humza, et al. "A Comprehensive Overview of Large Language Models." Elsevier Preprint, 18 Oct. 2024.

[3] Gemma Team, Google DeepMind. "Gemma: Open Models Based on Gemini Research and Technology." arXiv preprint, 16 https://arxiv.org/abs/2403.08295. Apr. 2024.

[4] Jeong, Cheonsu. Fine-tuning and Utilization Methods of Domain-specific LLMs. Samsung SDS, 2023.

[5] Nwafor, Chidinma A., and Ikechukwu E. Onyenwe. An Automated Multiple-Choice Question Generation Using Natural Language Processing Techniques. Nnamdi Azikiwe University Awka, 2023.

[6] Hang, Ching Nam, Chee Wei Tan, and Pei-Duo Yu. "MCQGen: [102273, https://dx.doi.org/10.1109/ACCESS.2024.3420709.

[7] Bulathwela, Sahan, Hamze Muse, and Emine Yilmaz. "Scalable Educational Question Generation with Pre-trained Language Models." Centre for Artificial Intelligence, University College London, 2023.

[8] Li, Shiyao, et al. "Evaluating Quantized Large Language Models." Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 2024.

[9] Qin, Haotong, et al. "Accurate LoRA-Finetuning Quantization of LLMs via Information Retention." Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 2024.

[10] Jiang, Ziyan, Xueguang Ma, and Wenhu Chen. "LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs." University of Waterloo, 2024.

[11] Hadzhikoleva, Stanka, et al. "Automated Test Creation Using Large Language Models: A Practical Application." Applied Sciences, vol. 14, no.9125.https://doi.org/10.3390/app14199125.

[12] Babakhani, Pedram, et al. "Opinerium: Subjective Question Generation Using Large Language Models." IEEE Access, vol. 12, 2024,pp.66085–66092.

[13] Chen, Zixiang, et al. "Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models." Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 2024, Vienna, Austria. arXiv:2401.01335.

[14] Xu, Lingling, et al. "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment." arXiv, 20 Dec. 2023, https://arxiv.org/abs/2312.12148.

[15] Luong, Trung Quoc, et al. "ReFT: Reasoning with Reinforced Fine-Tuning." *arXiv*, 2024, https://arxiv.org/abs/2401.08967.

[16] Srinivasan, Krishna Prasad Varadarajan, et al. "Comparative Analysis of Different Efficient Fine-Tuning Methods of Large Language Models (LLMs) in Low-Resource Setting." *arXiv*, 21 May 2024, https://arxiv.org/abs/2405.13181.