|  |  |
| --- | --- |
| Project Name | *Multimodal evalution of open ended question answer using llm* |
| Background Information | *Evaluating open-ended responses in education has traditionally been a manual, time-consuming process. As educational content diversifies and learning modalities evolve, there's an increasing demand for scalable and intelligent assessment systems. Learners today interact with content not only through text but also through speech, requiring flexible evaluation systems that can adapt to such multimodal inputs.*  *Advances in Large Language Models (LLMs), speech recognition, and context-aware retrieval techniques like Retrieval-Augmented Generation (RAG) have made it possible to automate these processes with a high degree of accuracy. A system that combines voice input, document-driven question generation, and AI-powered evaluation can greatly enhance educational feedback mechanisms and learner engagement.* |
| Literature Review | |  |  |  |  |  | | --- | --- | --- | --- | --- | | ***Paper*** | ***Authors*** | ***Approach Used*** | ***Key Findings*** | ***Relevance to Project*** | | A Systematic Literature Review on LLM-Based Information Retrieval | Diogo Cosme, Fernando Brito E Abreu  2024 | Systematic Literature Review on LLM-based content classification | LLMs significantly improved content classification, but challenges like transparency, computational cost, and hallucinations remain. | Provides foundational understanding of how LLMs enhance content retrieval and classificationâ€”core to generating and evaluating Q&A. | | DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process | Minjun Zhu, Yixuan Weng, Linyi Yang, Yue Zhang | March 2025 (arXiv preprint) | Structured multi-stage LLM review with reasoning steps: novelty verification, multidimensional review, and reliability verification. | DeepReviewer-14B outperforms baseline models in automated peer review, achieving high win rates and robust evaluations. | | Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM | Ruiyang Zhou, Lu Chen, Kai Yu  May 2024 | Evaluation of GPT-3.5 and GPT-4 on review tasks using multiple-choice and scoring metrics | LLMs provide helpful feedback but are error-prone, weak in long context understanding, and critical evaluation. | Highlights LLM limitations in assessment tasksâ€”guides caution when using LLMs to evaluate learner answers. | | The Emergence of Large Language Models (LLM) as a Tool in Literature Reviews | Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, Leslie Lenert | LLM-assisted systematic review of LLM use in literature review automation | GPT-based LLMs dominate review automation, especially in data extraction and synthesis; fewer full-review automation efforts exist. | Shows potential of LLMs in large-scale information synthesis, supporting the use of LLMs in generating Q&A from documents. | |
| Statement of the Problem | *Current learning management systems lack the ability to:*   * *Automatically generate rich, context-specific, open-ended questions.* * *Accept learner input in multiple modalities, particularly voice.* * *Perform automated, unbiased evaluation of subjective answers.* * *Offer educators tools to manage, review, and finalize question sets and evaluation criteria.*   *These limitations hinder personalized feedback and scale, especially in open-ended learning scenarios.* |
| Objectives | *The key objectives of this project are:*   1. ***Authentication and Role-Based Access*** *Implement a login system distinguishing Admin and Learner roles to manage workflows separately.* 2. ***PDF-Based Content Ingestion*** *Enable Admin users to upload documents (e.g., PDFs), which will be processed and stored for question generation.* 3. ***Automated Question-Answer Generation*** *Use an LLM to generate long-form open-ended questions and model answers based on uploaded content.* 4. ***Speech-Based Answering*** *Allow Learners to respond to questions using voice input, which will be transcribed into text.* 5. ***AI-Driven Evaluation*** *Evaluate learner answers against frozen model answers using LLMs with rubric or prompt-based grading.* 6. ***Data Storage and Reporting*** *Maintain records of questions, model answers, learner responses, and evaluation scores in a structured database.* |
| Methodology | *The project will follow a modular approach, covering:*   * ***Login and Session Handling*** *Secure login with hashed credentials and session persistence, enabling role-specific views for Admin and Learner.* * ***Document Upload and Chunking*** *Admin uploads topic-specific documents which are split into meaningful content chunks for Q&A generation.* * ***LLM-Based Q&A Generation*** *Generated Q&A pairs are associated with topics and stored. Admins can review, edit, and freeze answers for consistency.* * ***Voice Input and Transcription*** *Learners select a topic and answer questions via voice, which is transcribed to text using an integrated speech-to-text engine.* * ***Evaluation Engine*** *Learner responses are compared to frozen answers using LLMs, generating a score (e.g., out of 10) based on answer quality and relevance.* * ***Data Persistence*** *All learner interactions, including responses and scores, are stored for reporting and learning analytics.* |
| Proposed Solution/  Expected Results | This project will result in a functional system that:   * Automates the creation and administration of topic-based, open-ended Q&A. * Supports multimodal input, especially voice, making it accessible and engaging. * Enables educators to control the content and maintain consistency through answer freezing. * Delivers AI-powered evaluation that reduces manual grading while ensuring fairness. * Stores data for future analysis and continuous improvement in learning.   Ultimately, the system aims to enhance the scalability, accessibility, and intelligence of educational assessments in environments where open-ended evaluation is crucial. |
| Detailed Scope of Work: |  |
| UseCase | During my discussion with JB sir, we explored the possibility of using the module to evaluate open-ended questions. This aligns closely with a key use case in my organization, where we need to assess user knowledge after each new product release through quizzes. The same module can be used to upload product release documents, automatically generate questions and answers, and evaluate user responses. Similarly, this use case can be extended to assess students' answers for chapters in NCERT textbooks.  The framework will be designed for multiple purposes and can be integrated with both teachers and machines. During the development phase, I will use PDFs related to my own learning. However, once the module is fully developed, it can be used to upload and test documents from my organization or even NCERT textbooks. |
| Environment | 1)Python  2)Deepseek 1.5b  3)OpenWhisper  4)Streamlit |
| References | *1. A Systematic Literature Review on LLM-Based Information Retrieval*   * *Cosme, D., Galvão, A., & Brito E Abreu, F. (2024). A Systematic Literature Review on LLM-Based* * *Information Retrieval: The Issue of Contents Classification. In Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024)* * *Volume 1: KDIR, 135–146. SCITEPRESS. https://doi.org/10.5220/0013062300003838*   *2. DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process*   * *Zhu, M., Weng, Y., Yang, L., & Zhang, Y. (2025). DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process. arXiv preprint arXiv:2503.08569.* [*https://arxiv.org/abs/2503.08569*](https://arxiv.org/abs/2503.08569)   *3. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM*   * *Zhou, R., Chen, L., & Yu, K. (2024). Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Proceedings of LREC-COLING 2024, 9340–9351.* *European Language Resources Association (ELRA).*   *4. The Emergence of Large Language Models (LLM) as a Tool in Literature Reviews*   * *Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., & Lenert, L. A. (2024). The Emergence of Large Language Models (LLM) as a Tool in Literature Reviews: An LLM Automated Systematic Review. Manuscript submitted for publication.* |