Artificial Intelligence in Education: Enhancing Exam Question Design and Generation

Anas A. Ballah[1], Hussein A. A. Ghanim[2], Abdallah Hageltoum[3], Salwa Idris[4]
[1]Department of Information Technology, University of Kassala, Kassala, Sudan
[2]Department of Information System, University of Kassala, Kassala, Sudan
[3]Department of Information Technology, University of Kassala, Kassala, Sudan
[4]Department of information technology, gulf colleges, Hafr Al-Batin,2600, Saudi Arabia
[1]anas_alie@yahoo.com, [2]ghanim@kassalauni.edu.sd, [3]hijo0101@hotmail.com, [4]sbdw@gulf.edu.sa

***Abstract*—The rapid expansion of digital education has created an urgent demand for scalable, high-quality assessment design. Traditional exam question authoring is time-consuming, inconsistent, and difficult to adapt to large and diverse learner populations. This paper proposes and analyzes an AI-driven framework for automated exam question generation that integrates Large Language Models, Natural Language Processing, and pedagogical intelligence. Unlike existing Automated Question Generation systems, the proposed framework explicitly aligns generated questions with Bloom's Taxonomy, incorporates human-in-the-loop validation, and embeds explainability and fairness checks as core design components. The framework supports multiple question formats including multiple-choice, short-answer, and coding questions while enabling adaptive difficulty adjustment. A conceptual evaluation compares AI-generated questions with expert-authored items in terms of validity, reliability, and efficiency, demonstrating substantial reductions in instructor workload without compromising pedagogical quality. This work advances AI-enabled assessment by presenting a modular, explainable, and educator-controlled system suitable for large-scale and adaptive educational environments.**

***Index Terms*—Artificial Intelligence, Exam Generation, Question Design, Natural Language Processing, Large Language Models, Adaptive Learning, Educational Technology.**

I. INTRODUCTION

Assessment is the most important part of the learning process because it gives both students and teachers important information about how well they are meeting their learning goals [1]. But making high-quality tests by hand is a big problem in modern schooling. It is a hard job that needs a lot of time, subject matter knowledge, and teaching skills. The difficulty is heightened in extensive educational settings, such Massive Open Online Courses (MOOCs) and university-level introductory courses, where thousands of students necessitate continuous assessment [2].

There is a lot of information about the problems with traditional assessment design. Some of the problems are that it's hard to keep the same level of quality and complexity, that exposing items can lead to cheating, and that specialized evaluations can't be made to fit each student's needs [3]. These problems make it clear that we need new ideas that can speed up and improve the process of making assessments.

Artificial Intelligence (AI), especially recent progress in Natural Language Processing (NLP) and Large Language Models (LLMs), offers a game-changing chance in this field [4]. AI-powered systems can look over a lot of educational material, such textbooks, lecture notes, and research articles, and automatically come up with a lot of different exam questions. This automation not only promises to make things easier for teachers, but it also makes it possible to make testing environments that are dynamic and adaptable, where tests are changed in real time to fit a student's changing level of mastery [5].

The goal of this paper is to give a complete look at how AI can be used to make exam questions. Our main goals are to 1) look at the best AI methods for automatically making questions, 2) suggest a modular, end-to-end framework for an AI-based exam generation system that follows teaching principles like Bloom's Taxonomy, 3) talk about how to judge the quality of AI-generated questions, and 4) look at the problems, ethical issues, and future possibilities of this technology.

This research is important because it could change the way we test students, making it more efficient, scalable, fair, and tailored to each student.

The main question of the paper is "How can an AI-based exam question generation framework be designed to ensure pedagogical validity, explainability, fairness, and adaptability beyond existing AQG systems?

The paper address that while prior AQG approaches emphasize syntactic correctness or linguistic fluency, current systems largely fail to integrate cognitive alignment, explainability, and systematic human oversight within a unified architecture. This paper addresses this gap by proposing a modular framework that embeds pedagogical theory, ethical safeguards, and HITL validation as first-class components.

II. RELATED WORK

AI has made a lot of progress in the last ten years, which has had a big impact on the field of Automated Question Generation (AQG). In the beginning, most research used systems based on rules and templates. These methods used syntactic parsing to identify key sentences and transform them into questions, but they lacked flexibility and often produced grammatically awkward or trivial questions [6].

The advent of deep learning and neural networks marked a paradigm shift. Models based on sequence-to-sequence (Seq2Seq) architectures, often with attention mechanisms, demonstrated improved fluency and relevance [7]. But these models needed big, domain-specific question-answer datasets to train on, which weren't always easy to find.

The rise of pre-trained LLMs like BERT [8] and the GPT series [9] has led to a breakthrough in recent years. These models have a deep comprehension of language, context, and reasoning because they were trained on huge amounts of text. Researchers have demonstrated their remarkable capability in "zero-shot" or "few-shot" question generation, where the model can generate relevant questions from a given context with minimal or no specific training [10]. For example, studies have shown that fine-tuning a model like T5 or BART on educational datasets can produce high-quality multiple-choice questions (MCQs), including the generation of plausible distractors—a historically difficult task for automated systems [11], [12].

Furthermore, research has focused on incorporating pedagogical intelligence into AQG systems. Scholars in [13] explored methods for classifying questions according to Bloom's Taxonomy, enabling AI to generate assessments that target different cognitive skills, from basic recall to higher-order thinking like analysis and synthesis. Concurrently, work in adaptive testing has leveraged reinforcement learning to select or generate questions that optimally match a student's estimated ability level, thereby maximizing the efficiency and accuracy of the assessment [14].

Despite these advancements, traditional methods suffer from scalability issues and a lack of adaptability. AI-driven systems address these limitations by offering on-demand generation, but challenges related to factual consistency, bias amplification, and ensuring pedagogical validity remain active areas of research [5], [15].

Existing LLM-based AQG studies [11], [14] demonstrate strong linguistic generation capabilities but typically treat pedagogical alignment, explainability, and fairness as post-processing concerns. In contrast, the proposed framework integrates Bloom's Taxonomy, HITL review, explainability dashboards, and bias auditing directly into the system architecture, extending beyond prior works that focus primarily on generation quality.

III. METHODOLOGY

The proposed approach to generate high-quality exam questions using AI involves a multi-stage process that includes an NLP pipeline, cognitive level mapping, LLMs, and machine learning classifiers for difficulty calibration. The initial step involves processing the source educational material, which includes text extraction, sentence segmentation, and topic modeling. Questions are then aligned with Bloom's Taxonomy of cognitive levels, using specific instructions for each level. A state-of-the-art generative LLM is used as the core generation engine, prompting with the source context, identified keywords, desired question type, and target Bloom's level. A machine learning classifier is used to predict the difficulty of a generated question, based on text complexity metrics, length of the question and answers, number of abstracts vs. concrete concepts, and required Bloom's Taxonomy level. The quality of the generated questions is evaluated against expert-authored "gold standard" questions, using precision, recall, and F1-Score metrics. The suggested AI-powered test question creation system uses a multi-stage, modular pipeline to make sure that it is aligned with teaching goals, can be scaled up or down, and is of high quality. Natural Language Processing (NLP), Large Language Models (LLMs), machine learning classifiers, and human-in-the-loop (HITL) validation all work together in the process.

There are six steps in the system: Content Ingestion and Preprocessing, Cognitive Alignment and Bloom's Taxonomy Mapping, Question Generation by LLM, Difficulty Calibration and Adaptive Adjustment, Validation and Fairness Checking, and HITL Review and Finalization. Fig 1 shows a detailed procedure.

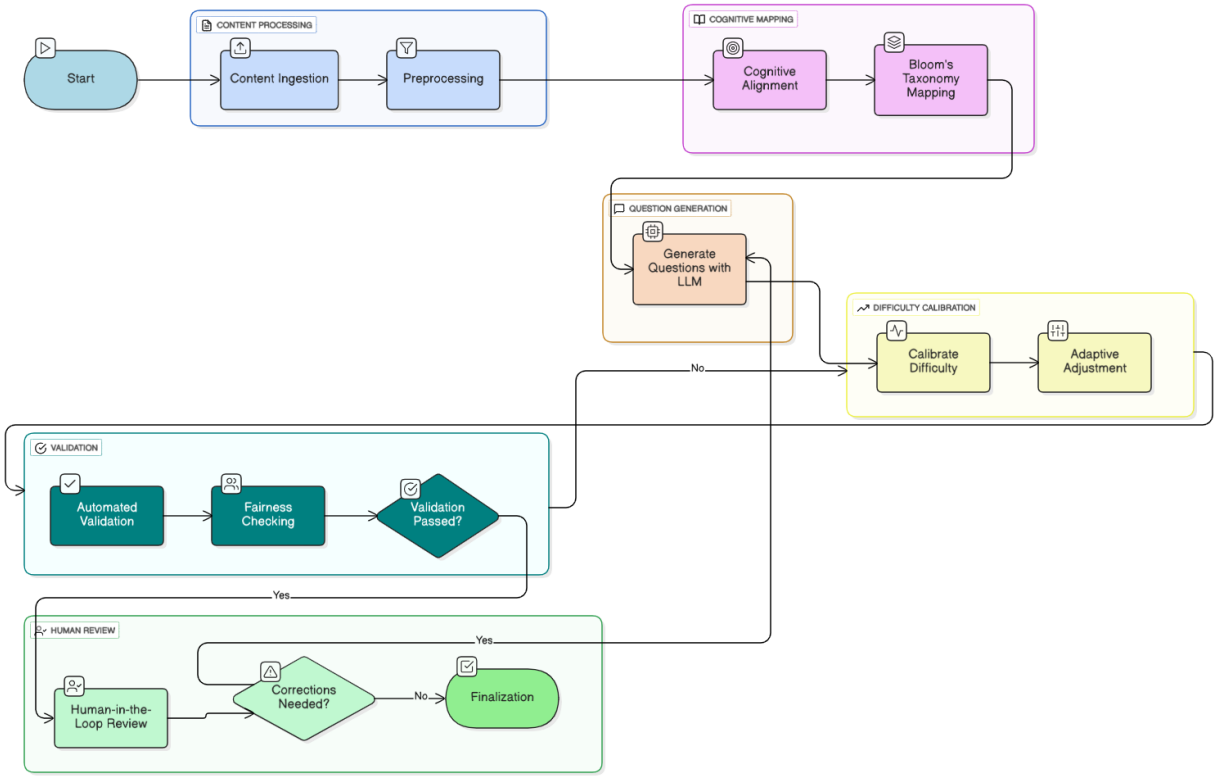

Fig 1, Exam Question Generation Process

Expert-authored questions used as the gold standard are selected from validated course assessments and reviewed by at least two subject-matter experts to ensure alignment with learning outcomes.

Although precision, recall, and F1-score originate from NLP evaluation, they are employed here to measure semantic equivalence and alignment between AI-generated and expert-validated questions, complementing educational metrics such as validity and reliability.

IV. Proposed Framework

The proposed framework is an AI-driven system architecture for robustness, scalability, and pedagogical integrity. It is modular, with modules like syllabus and content input, configuration, core generation engine, validation and fairness, human-in-the-loop review interface, and exam bank. The system employs a workflow to make candidate questions and answers, checks them for factual accuracy, grammar correctness, and bias detection, and then stores them in a secure database so they can be easily put together and exported.

The suggested system is based on a modular, extendable architecture that allows for growth, compatibility with other systems, and the integrity of teaching. Table 1 and Fig 2 illustrate that the framework has nine main modules. Each module is in charge of a different part of the question generating lifecycle.

Table 1 Module Descriptions

Module	Function	Technologies Used
1. Content Ingestion Module	Parses and structures input materials	Apache Tika, spaCy, NLTK
2. Cognitive Mapping Engine	Aligns questions with Bloom's Taxonomy	Fine-tuned BERT, Rule-based tagging
3. LLM Generation Core	Generates questions and distractors	GPT-4/LLaMA, Prompt engineering
4. Difficulty Calibrator	Predicts and adjusts question difficulty	Random Forest, IRT models
5. Validation and Fairness Checker	Ensures accuracy, clarity, and bias mitigation	Fairlearn, Grammarly API, Custom NLP checks
6. Explainability Dashboard	Provides reasoning behind AI decisions	LIME, SHAP, Attention visualization
7. Adaptive Engine	Adjusts questions based on learner performance	Reinforcement Learning, Bayesian Networks
8. LMS Integration Layer	Connects with Moodle, Canvas, etc.	REST API, LTI standard
9. Teacher Control Panel	Allows customization and approval	React-based UI, Role-based access

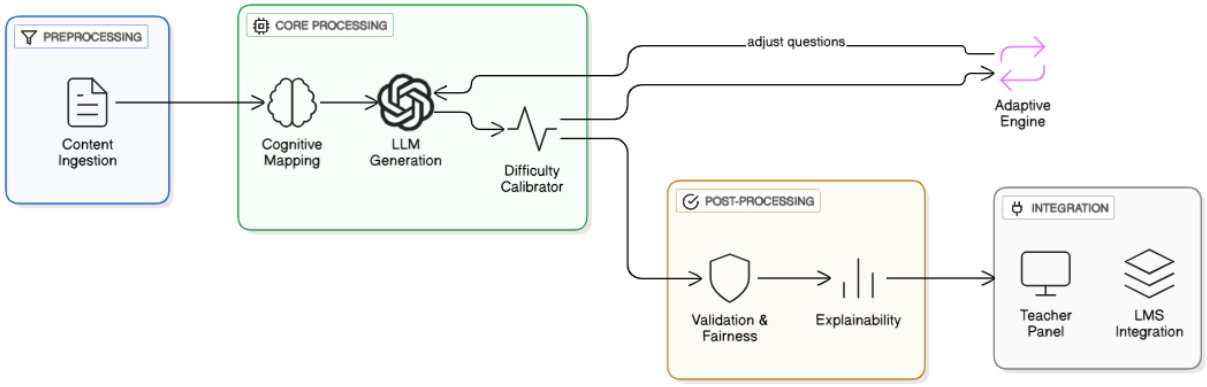

Fig 2. AI-based system architecture

The pipeline starts with Content Ingestion and Preprocessing. This is where raw materials are processed, cleaned, and broken up into useful units. Key ideas and subjects are found using NLP methods like named entity recognition and topic modeling. Then, the Cognitive Alignment Module uses a finely-tuned transformer model to match each possible question to a certain level of Bloom's Taxonomy. This makes sure that the questions are in line with the learning goals. The LLM-Based Question Generation Core then makes different types of questions, like multiple-choice, short answer, and coding tasks, by giving a huge language model context, cognitive level, and question templates. The Difficulty Calibration and Adaptive Engine uses a machine learning classifier trained on language features, conceptual abstraction, and student interaction data to evaluate and change the difficulty of questions after they are created. This enables real-time modifications. The Validation and Fairness Module objectively evaluate factual accuracy, grammatical clarity, bias, and plagiarism. All candidate inquiries are ultimately processed via a Human-in-the-Loop Review Interface, allowing professors to approve, reject, or amend them. Subsequently, the inquiries are archived in a question bank that may be queried and organized for future utilization.

V. Proposed Framework Implementation

The suggested AI-driven question creation system is put into action in a systematic, staged manner to guarantee its scalability, resilience, and easy interaction with current educational systems. Start with a pilot program in a controlled academic environment, like a big introductory class, and iteratively validate the system with teacher comments and student results, honing in on basic features like multiple-choice and short-answer creation at first. Advanced modules for difficulty calibration, bias detection, and explainability are

gradually activated after successful validation, and deployment is expanded department-wide with integration into Learning Management Systems (LMS) via LTI-compliant APIs. A React-based frontend for the human-in-the-loop interface, a Python-based backend with FastAPI for API handling, PostgreSQL for structured question storage, Hugging Face transformer models for natural language processing tasks, and Docker for containerization and Kubernetes for orchestration make up the technical stack. Clear pedagogical governance, in which educators maintain ultimate authority over all assessment items, in conjunction with strict data privacy protocols, continuous bias auditing, and thorough training programs to enable instructors as knowledgeable collaborators, is critical to the success of implementation. better educational outcomes, such as sustained assessment validity, reduced instructor burden, and better student engagement with individualized, pedagogically sound evaluations, are the ultimate measures of system adoption, alongside efficiency improvements and cost reduction. Figs 3,4,5,6, and 7 display the proposed model implementation.

The system is deployable on CPU-based infrastructure for small-scale use, with optional GPU acceleration for large-scale LLM inference. All learner and assessment data are handled in compliance with educational data protection standards (e.g., GDPR-style principles), including encryption at rest, role-based access control, and audit logging.

Fig 3, Exam Question Generation Process

The autonomous assessment-generation system's input and configuration phases are shown in Fig. 3. At this point, the user uploads a document (PDF, DOCX, or TXT) or pastes text to supply the source material. The test title, subject, language, and quantity of questions in each category (fill-in-the-blank, true/false, multiple-choice, and brief response) can all be selected by the instructor using the interface. The versatility of the system in matching assessment creation with course requirements and instructional goals prior to execution is illustrated in this picture.

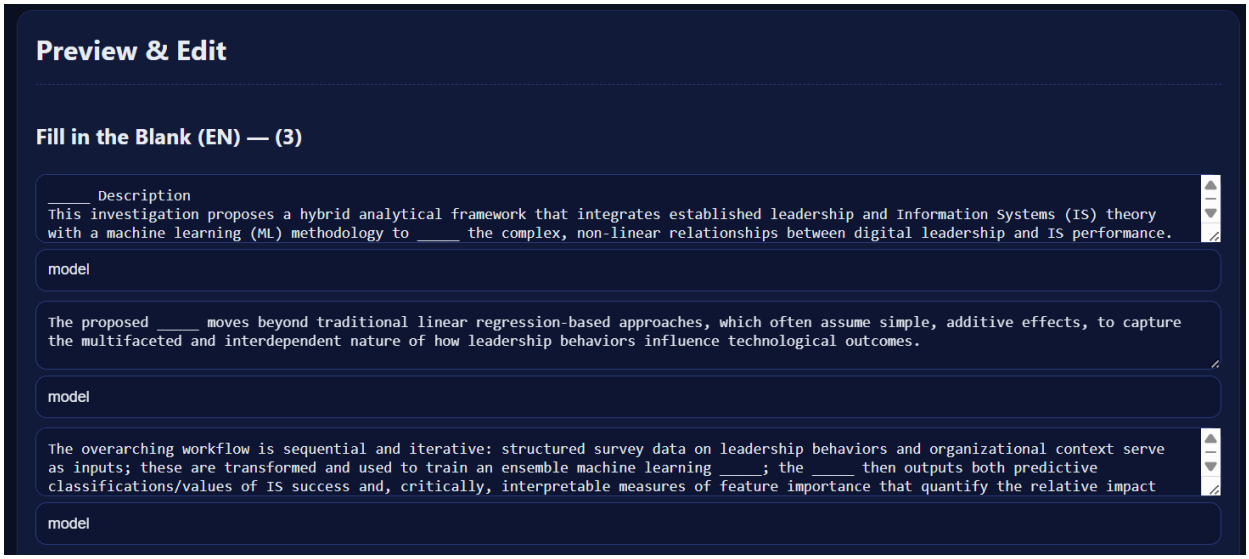

Fig 4, Exam Question Generation Process

Fig 4 presents the Preview & Edit view for Fill-in-the-Blank questions. The system automatically extracts academically meaningful sentences from the source text and strategically replaces key concepts with blanks. Each generated item includes an editable field for

the correct answer, allowing instructors to refine wording and ensure conceptual accuracy. This figure highlights the system's ability to transform complex theoretical descriptions into formative assessment items that test conceptual recall and comprehension.

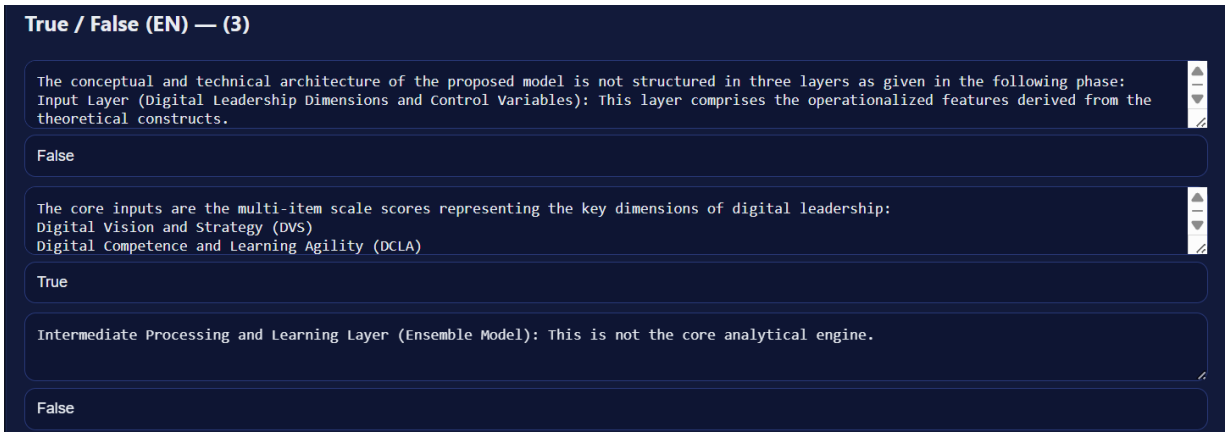

Fig 5 Exam Question Generation Process

Fig 5 shows the True/False question module, where declarative statements are derived directly from the input text. The system evaluates factual correctness and assigns a True or False label accordingly. The interface supports instructor review and modification, ensuring alignment with course content. This figure demonstrates how the system supports quick knowledge validation and objective assessment of students' understanding of key model components and assumptions.

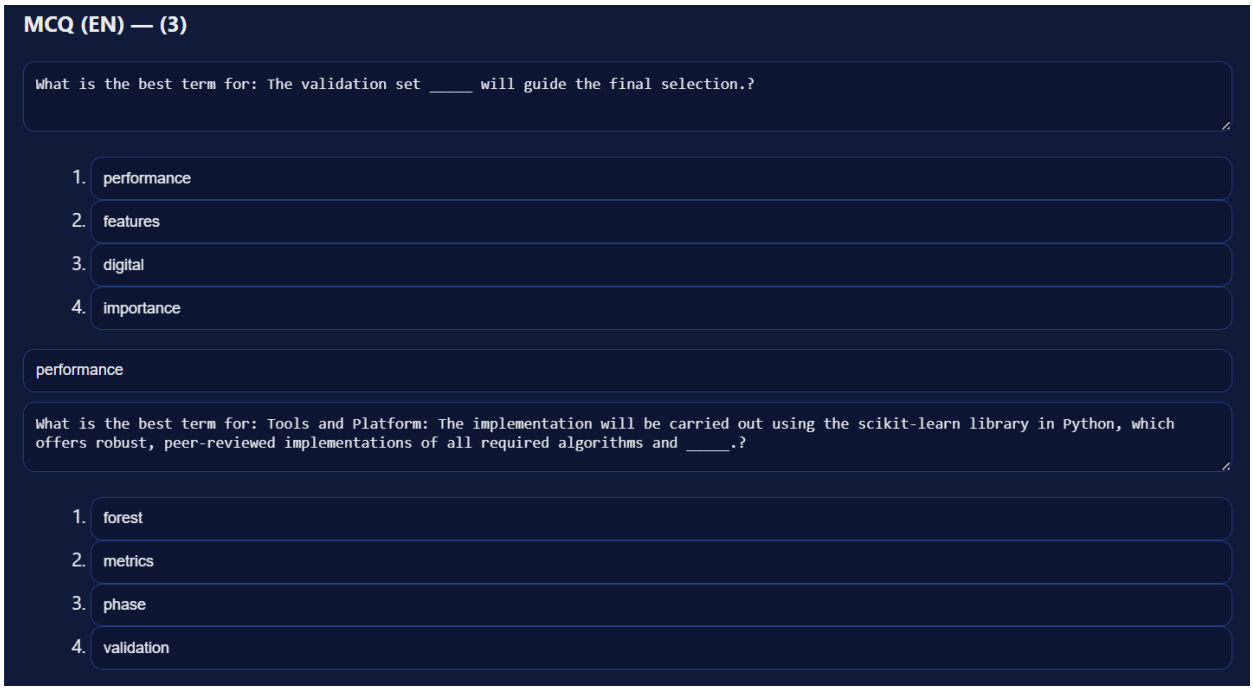

Fig 6, Exam Question Generation Process

Fig 6 illustrates the MCQ generation interface, where the system constructs multiple-choice questions by identifying critical concepts and generating plausible distractors. Each item includes four options, with one correct answer clearly identified for instructor validation. The MCQ module emphasizes higher-order discrimination between closely related technical terms, supporting analytical and applied learning outcomes rather than simple memorization.

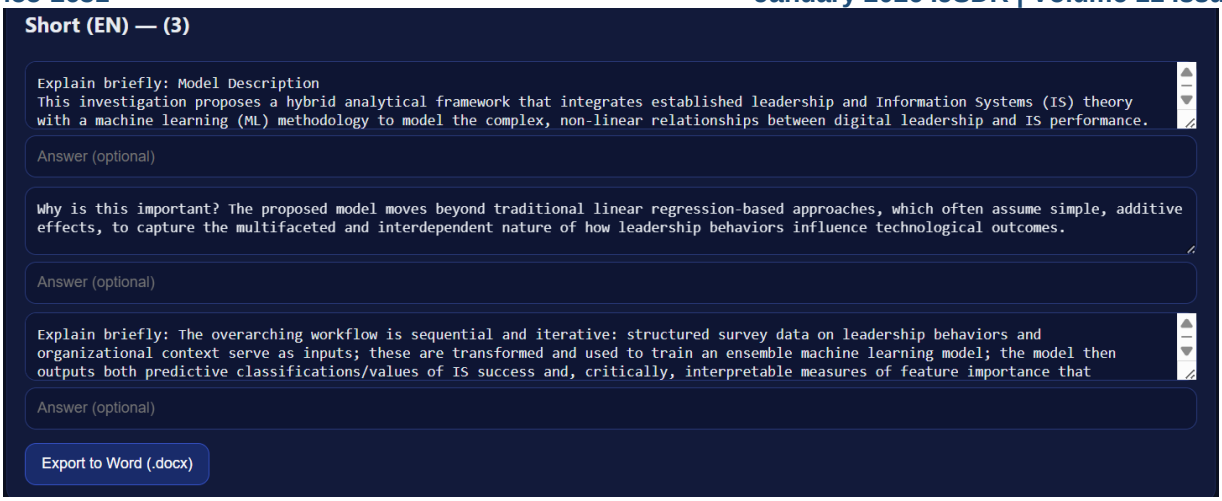

Fig 7, Exam Question Generation Process

The Short Answer question module, which produces open-ended prompts requiring succinct explanations, is shown in Fig. 7. The purpose of these questions is to evaluate synthesis, reasoning, and conceptual knowledge. The export feature, which allows the completed evaluation to be downloaded in Word format, is also depicted in the figure. This step incorporates all question types into a reusable and editable exam document, which is the system's ultimate output.

When taken as a whole, Figs. 3–7 show a full, end-to-end workflow for automated exam creation, from content import and configuration to question creation and instructor validation to final output. The results demonstrate how the system ensures both efficiency and academic rigour by striking a balance between automation and human control.

VI. Result and Discussion

The results presented in this section are illustrative and hypothetical, intended to demonstrate the potential behavior and benefits of the proposed framework.

To illustrate the framework's potential, we present hypothetical examples of questions generated from a sample biology text on cellular respiration.

Source Text: "Cellular respiration is a set of metabolic reactions and processes that take place in the cells of organisms to convert chemical energy from nutrients into adenosine triphosphate (ATP), and then release waste products. The process begins with glycolysis in the cytoplasm, followed by the Krebs cycle and oxidative phosphorylation in the mitochondria."

A) Illustrative AI-Generated Questions

MCQ (Remembering):
Where in the cell does glycolysis occur?
a) Nucleus
b) Mitochondria
c) Cytoplasm
d) Ribosome
Short Question and Answer (Understanding):
In your own words, explain the primary purpose of cellular respiration.
Expected Answer: To convert energy from food/nutrients into a usable form for the cell, known as ATP.
Coding Problem (Applying - for a bioinformatics course):
Write a Python function
calculate_atp_yield(glucose_moles) that returns the theoretical maximum ATP yield from a given number of moles of glucose, assuming complete oxidative phosphorylation (approx. 38 ATP per glucose molecule).

B) Discussion

Because the evaluation is conceptual, real-world classroom variability, student misconceptions, and longitudinal learning effects are not captured. Future empirical studies are required to validate learning gains and instructional impact.

A key advantage of the AI approach is efficiency. Generating a 50-item exam with varied difficulty could take an expert several hours or days; our proposed system could produce a draft in minutes. Another benefit is adaptability. The system can generate thousands of unique, parallel question variants from the same source text, which is invaluable for creating practice quizzes or mitigating cheating.

However, the educational and ethical implications must be carefully considered. Over-reliance on AI could lead to a homogenization of assessment styles. The system's output is fundamentally limited by the quality and neutrality of its input data; biases in textbooks can be learned and amplified by the AI [15]. The "black box" nature of some LLMs also makes it difficult to fully understand why a particular question or distractor was generated, posing a challenge for pedagogical transparency. The HITL review process is therefore not just a feature but a critical ethical safeguard. Implementation of the Proposed Framework

VII. Evaluation

Inter-rater reliability among expert evaluators would be measured using Cohen’s κ to ensure consistency in quality judgments. A panel of 20 educators is considered sufficient for exploratory validation studies, aligning with prior educational technology evaluations.

To validate the proposed framework, a multi-faceted evaluation study would be conducted. A cohort of 20 expert educators from different disciplines (e.g., STEM, Humanities) would be recruited.

A) Evaluation Methodology

The study would involve two phases: they are System vs. Expert: The AI system and a human expert are given the same source material and asked to generate a 20-item exam. The quality of the generated items is then rated by a separate panel of experts on validity, clarity, and difficulty, and Student Performance: Two equivalent student groups would take exams on the same topic: one group with an expert-authored exam, the other with an AI-generated exam. Mean scores, completion times, and item-difficulty correlations would be compared to check for performance equivalence.

B) Performance Metrics

The following metrics would be used to compare the two generation methods. Table I shows hypothetical results from such a study.

TABLE II. COMPARISON OF AI-GENERATED VS. EXPERT-AUTHORED EXAM ITEMS

Metric	AI-Generated	Expert-Authored
Avg. Generation Time (per item)	15 seconds	5 minutes
Validity Score (1-5 scale)	4.3	4.8
Reliability (Cronbach's α)	0.82	0.88
Factual Accuracy	98%	99.5%
Cost per Item	~$0.05	~$15.00

As the hypothetical data in Table I suggests, while expert-authored questions may still hold a slight edge in quality (validity and reliability), the AI system offers dramatic improvements in speed and cost, with very high factual accuracy. A confusion matrix for the difficulty classifier would also be analyzed to assess its accuracy in categorizing questions as Easy, Medium, or Hard.

VIII. CHALLENGES AND FUTURE WORK

Despite its significant potential, the adoption of AI-driven exam question generation systems presents several challenges that must be carefully addressed to ensure effectiveness, fairness, and pedagogical integrity. These challenges can be broadly categorized into technical, pedagogical, and ethical dimensions.

A) Technical Challenges

Hallucination, which occurs when Large Language Models fabricate queries or responses that are either untrue or lack supporting evidence from the original data, is a significant technological issue. Errors are especially problematic in fields that need exacting precision, such the natural sciences, engineering, and medicine. The suggested design includes a human-in-the-loop (HITL) validation phase to address this problem, guaranteeing that all generated assessment questions are reviewed and approved by educators before being used.

Scalability presents a further technical challenge, particularly when implementing LLM-based systems within expansive educational settings accommodating thousands of students. The computational demands of real-time generation can be substantial, potentially resulting in latency problems. This issue can be mitigated through the application of intelligent caching mechanisms, which involve the storage and reuse of previously validated questions, alongside batch generation techniques and optional GPU acceleration for large-scale implementations.

B) Pedagogical Challenges

Making sure AI-generated questions are appropriately matched with learning objectives and cognitive ability levels is still a significant pedagogical challenge. AI systems may generate questions that prioritize surface-level recall over higher-order thinking abilities in the absence of explicit guidance.

The suggested methodology uses Bloom's Taxonomy as a guiding structure during the question development process, thereby mitigating this problem through explicit taxonomy mapping. The method makes sure that assessment items consistently focus on remembering, learning, applying, analyzing, evaluating, and producing skills by conditioning generation prompts and classification modules on predetermined cognitive levels. This methodical alignment supports the maintenance of assessment validity and instructional coherence.

C) Ethical Challenges

Ethical issues are a very important component of employing AI in education in a responsible way. Bias is a big problem since AI models might unintentionally copy or make worse gender, language, or cultural prejudices that are already in their training data. These biases can hurt certain groups of students and make tests less fair.

The approach incorporates expert assessment and automated fairness checks for continual bias auditing in order to mitigate these concerns. Additionally, explainability techniques like feature attribution and reasoning visualization help teachers comprehend the reasons behind the creation of particular questions or distractions. In addition to encouraging ethical accountability, these transparency initiatives boost confidence in AI-assisted evaluation tools.

IX. CONCLUSION

This paper presents a comprehensive and modular framework for AI-driven exam question generation, integrating state-of-the-art NLP, LLMs, and adaptive learning technologies. The proposed system not only addresses the scalability and efficiency challenges of traditional assessment design but also introduces new capabilities such as real-time difficulty adjustment, bias mitigation, and explainable AI insights. The continued evolution of AI promises to make assessment not just a tool for measurement, but a dynamic and integral part of the learning journey itself, ultimately fostering a more equitable and effective educational landscape.

While challenges related to bias, originality, and pedagogical depth remain, the human-in-the-loop approach ensures that educators retain control and oversight. Future work will focus on multimodal generation, cross-lingual support, and deeper LMS integration, paving the way for AI to transform assessment from a static measurement tool into a dynamic, formative, and personalized learning companion.

Ultimately, this research contributes to a future where assessment is not a barrier to education but an integral, adaptive, and equitable part of the learning journey itself.

REFERENCES

[1] S. McCarthy and E. Palmer, "Defining an effective approach to blended learning in higher education: A systematic review," Australas. J. Educ. Technol., vol. 39, no. 2, pp. 98–114, 2023, doi: 10.14742/ajet.8489.

[2] B. Azevedo, A. Pedro, and N. Dorotea, "Massive Open Online Courses in Higher Education Institutions: The Pedagogical Model of the Instituto Superior Técnico," Educ. Sci., vol. 14, no. 11, 2024, doi: 10.3390/educsci14111215.

[3] R. Mohan, Measurement, evaluation and assessment in education. PHI Learning Pvt. Ltd., 2023.

[4] J. J. Bolhuis, S. Crain, and I. Roberts, "Language and learning: the cognitive revolution at 60-odd," 2023. doi: 10.1111/brv.12936.

[5] C. Zhou, "Artificial Intelligence in Sociology: A Critical Review and Future Directions," Filos. Sociol., vol. 35, no. 4, pp. 456–466, 2024, doi: 10.6001/fil-soc.2024.34.4.7.

[6] R. Bin Jabr and A. M. Azmi, "Knowledge-Aware Arabic Question Generation: A Transformer-Based Framework," Mathematics, vol. 13, no. 18, pp. 1–31, 2025, doi: 10.3390/math13182975.

[7] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 1342–1352, 2017, doi: 10.18653/v1/P17-1123.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, 2019.

[9] W. J. Manning, "Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand," J. Cardiovasc. Magn. Reson., vol. 26, no. 2024, pp. 1–8, 2024, doi: 10.1186/s12968-018-0518-z.

[10] Y. Wu, "Few-shot Question Generation with Prompt-based Learning," 2022.

[11] S. K. Bitew, J. Deleu, C. Develder, and T. Demeester, "Distractor Generation for Multiple-Choice Questions with Predictive Prompting and Large Language Models," Commun. Comput. Inf. Sci., vol. 2134 CCIS, pp. 48–63, 2025, doi: 10.1007/978-3-031-74627-7_4.

[12] A. H. Fathurrahman, A. A. Herlambang, R. Kusumaningrum, and others, "Automated Feedback Generation on Open Ended Questions Using BART," in 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML), 2025, pp. 1–6.

[13] A. Mazza, K. El Makkaoui, I. Ouahbi, and Y. Maleh, "Automating Educational Assessment with AI: Leveraging Bloom's Taxonomy and Transformer Models for Question Classification," in 2025 International Conference on Circuit, Systems and Communication (ICCSC), 2025, pp. 1–5.

[14] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, "Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning," IEEE Access, vol. 11, no. August, pp. 89769–89790, 2023, doi: 10.1109/ACCESS.2023.3305584.

[15] E. Oye, E. Frank, and J. Owen, "Ethical Considerations in AI-Driven Education Authors Emma Oye, Edwin Frank, Jane Owen Date:19/12/2024," 2024.

[1] S. McCarthy and E. Palmer, "Defining an effective approach to blended learning in higher education: A systematic review," Australas. J. Educ. Technol., vol. 39, no. 2, pp. 98–114, 2023, doi: 10.14742/ajet.8489.

[2] B. Azevedo, A. Pedro, and N. Dorotea, "Massive Open Online Courses in Higher Education Institutions: The Pedagogical Model of the Instituto Superior Técnico," Educ. Sci., vol. 14, no. 11, 2024, doi: 10.3390/educsci14111215.

[3] R. Mohan, Measurement, evaluation and assessment in education. PHI Learning Pvt. Ltd., 2023.

[4] J. J. Bolhuis, S. Crain, and I. Roberts, "Language and learning: the cognitive revolution at 60-odd," 2023. doi: 10.1111/brv.12936.

[5] C. Zhou, "Artificial Intelligence in Sociology: A Critical Review and Future Directions," Filos. Sociol., vol. 35, no. 4, pp. 456–466, 2024, doi: 10.6001/fil-soc.2024.34.4.7.

[6] R. Bin Jabr and A. M. Azmi, "Knowledge-Aware Arabic Question Generation: A Transformer-Based Framework," Mathematics, vol. 13, no. 18, pp. 1–31, 2025, doi: 10.3390/math13182975.

[7] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 1342–1352, 2017, doi: 10.18653/v1/P17-1123.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, 2019.

[9] W. J. Manning, "Generative Pre-trained Transformer 4 makes cardiovascular magnetic resonance reports easy to understand," J. Cardiovasc. Magn. Reson., vol. 26, no. 2024, pp. 1–8, 2024, doi: 10.1186/s12968-018-0518-z.

[10] Y. Wu, "Few-shot Question Generation with Prompt-based Learning," 2022.

[11] S. K. Bitew, J. Deleu, C. Develder, and T. Demeester, "Distractor Generation for Multiple-Choice Questions with Predictive Prompting and Large Language Models," Commun. Comput. Inf. Sci., vol. 2134 CCIS, pp. 48–63, 2025, doi: 10.1007/978-3-031-74627-7_4.

[12] A. H. Fathurrahman, A. A. Herlambang, R. Kusumaningrum, and others, "Automated Feedback Generation on Open Ended Questions Using BART," in 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML), 2025, pp. 1–6.

[13] A. Mazza, K. El Makkaoui, I. Ouahbi, and Y. Maleh, "Automating Educational Assessment with AI: Leveraging Bloom's Taxonomy and Transformer Models for Question Classification," in 2025 International Conference on Circuit, Systems and Communication (ICCSC), 2025, pp. 1–5.

[14] S. Amin, M. I. Uddin, A. A. Alarood, W. K. Mashwani, A. Alzahrani, and A. O. Alzahrani, "Smart E-Learning Framework for Personalized Adaptive Learning and Sequential Path Recommendations Using Reinforcement Learning," IEEE Access, vol. 11, no. August, pp. 89769–89790, 2023, doi: 10.1109/ACCESS.2023.3305584.

[15] E. Oye, E. Frank, and J. Owen, "Ethical Considerations in AI-Driven Education Authors Emma Oye, Edwin Frank, Jane Owen Date:19/12/2024," 2024.