International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

AI-Driven Framework for Exam Question Design and Generation: Pedagogy, Explainability and Fairness

Hussein A. A. Ghanim [1,*], Anas A. Ballah [2], I. Abdallah Hageltoum [2] and Salwa Idris [3]

[1] *Department of Information System, Faculty of Computer Science and Information Technology, University of Kassala, Sudan.*
[2] *Department of Information Technology, Faculty of Computer Science and Information Technology, University of Kassala, Sudan.*
[3] *Department of information technology, gulf colleges, Hafr Al-Batin,2600, Saudi Arabia.*

International Journal of Science and Research Archive, 2026, 18(01), 827-838

Publication history: Received on 14 December 2025; revised on 22 January 2026; accepted on 24 January 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.1.0093

Abstract

The swift growth of digital education requires scalable, high-quality assessment instruments. Conventional exam question creation is arduous and challenging to customize, but current Automated Question Generation systems frequently exhibit deficiencies in pedagogical congruence, openness, and ethical protections. This paper introduces the PXF framework, an innovative AI-driven system for generating exam questions that incorporates Pedagogy, Explainability, and Fairness as core design concepts. The system utilizes a modular architecture that includes a Pedagogy Alignment Module for mapping Bloom's Taxonomy, an Explainability Engine that offers human-interpretable rationales, and a Fairness Module for proactive bias detection, all overseen by a Human-in-the-Loop review interface. Experimental validation on educational datasets indicates that the PXF framework attains a classification accuracy of 91%, an F1-Score of 0.87, and decreases question drafting time by 84% relative to manual authorship, while closely aligning with expert-level pedagogical quality. The results confirm its effectiveness in generating cognitively aligned questions, providing clear insights into AI decision-making, and detecting harmful biases for instructor assessment. This study advances the field of educational AI by presenting a systematic, transparent, and ethically aware framework that enhances assessment scalability while preserving pedagogical integrity and justice, offering a practical model for the future of AI-enhanced education.

Keywords: Artificial Intelligence; Exam Generation; Pedagogy; Explainability; Fairness; Large Language Models; Educational Technology

1. Introduction

Assessment is fundamental to the learning process, providing critical feedback on learning goal achievement [1]. However, manual creation of high-quality exam questions is a significant bottleneck in modern education. It is a time-consuming task requiring deep subject expertise and pedagogical skill, a challenge exacerbated in large-scale digital learning environments like Massive Open Online Courses (MOOCs) [2]. Traditional methods struggle with maintaining consistency, mitigating item exposure for security, and providing personalization for diverse learners [3].

Automated Question Generation (AQG) has a revolutionary opportunity thanks to recent developments in Artificial Intelligence (AI), especially in Natural Language Processing (NLP) and Large Language Models (LLMs) [4]. AI technologies promise to lessen instructor effort and offer dynamic, adaptive testing settings by automatically generating a variety of exam questions by analyzing large amounts of educational content [5]. However, current AQG systems often

* Corresponding author: Hussein A. A. Ghanim

prioritize linguistic fluency and syntactic correctness while treating pedagogical alignment, system transparency, and ethical safeguards—such as fairness and bias mitigation as secondary concerns or post-processing steps [6], [7]. This gap can result in questions that lack appropriate cognitive depth, are unexplainable in their design rationale, or risk perpetuating biases present in training data.

This paper addresses this critical gap by introducing a principled AI-driven framework where Pedagogy, Explainability, and Fairness (PXF) are embedded as first-class, non-negotiable design principles. The primary contribution is a modular system architecture that operationalizes these three pillars. The framework ensures pedagogical validity through explicit alignment with Intended Learning Outcomes and the revised Bloom's Taxonomy [8], guarantees transparency via explainability dashboards that justify automated decisions [9], [10], and promotes equity through integrated fairness audits inspired by recent work in ethical AI and algorithmic accountability [11], [12], [13]. An organized Human-in-the-Loop (HITL) review process preserves essential educator oversight.

This paper details the design, implementation, and evaluation of the PXF framework. We begin with a review of related work in AQG and educational AI (Section 2). We then present the framework's core principles and architecture (Section 3), followed by the implementation of its key modules (Section 4). An experimental validation demonstrates the framework's performance (Section 5), leading to a discussion of its implications and a comparison with existing systems (Section 6). We conclude by outlining future research directions (Section 7). The PXF framework represents a step toward responsible and effective AI augmentation in education, aiming to transform assessment from a logistical burden into a dynamic, equitable, and integral component of the learning journey.

2. Related work

Automated Question Generation (AQG) research has advanced dramatically. Syntactic parsing was used by early systems to convert sentences into questions utilizing rule-based and template-driven techniques. While foundational, these approaches lacked flexibility and often produced grammatically rigid or trivial output [6]. The advent of deep learning marked a shift, with sequence-to-sequence neural models demonstrating improved fluency and relevance, though they required large, domain-specific datasets for training [7].

The rise of pre-trained Large Language Models (LLMs) like BERT [14] and the GPT series [15] has been transformative. Their deep linguistic and reasoning capabilities, learned from vast corpora, enable effective "few-shot" or "zero-shot" question generation from a given context. This has significantly advanced tasks like creating plausible multiple-choice distractors, a historically challenging sub-problem [16], [17].

Concurrently, research has begun integrating pedagogical constructs into AQG. Scholars have explored methods to classify or generate questions according to Bloom's Taxonomy, targeting different cognitive skills from recall to evaluation [18]. Parallel work in adaptive testing employs techniques like reinforcement learning to sequence questions based on a learner's estimated ability [19].

Parallel studies in adaptive assessment and personalized learning have explored reinforcement learning approaches to structure information and recommendations based on student behavior and performance [20]. These systems can be customized, but they frequently put optimization ahead of openness and ethical concerns. Recent studies underscore the increasing significance of tackling justice, accountability, and ethical problems in AI-driven educational systems [11], [12]. However, explainability and systematic bias auditing remain largely peripheral in existing AQG solutions, despite calls for end-to-end algorithmic auditing frameworks [13]. The proposed PXF framework addresses this gap by unifying pedagogical alignment, explainability, and fairness as interdependent design pillars within a single architecture.

3. The pxf framework: pedagogy, explainability, and fairness

The proposed framework is architected on three foundational principles designed to address the limitations of current AQG systems: Pedagogy-First Design, Active Explainability, and Proactive Fairness. These principles are operationalized through a cohesive, modular system where each component directly contributes to one or more of these core tenets.

3.1. Foundational Principles

The proposed framework is architected upon three interdependent, non-negotiable design principles. These principles address the core limitations of current Automated Question Generation (AQG) systems by moving beyond mere linguistic generation to ensure educational validity, operational transparency, and ethical integrity.

Pedagogy-First Design prioritizes learning objectives over content-driven generation. The primary inputs are the intended learning outcomes and target cognitive levels, defined using the revised Bloom's Taxonomy [8]. Each generated assessment item is explicitly designed to validly and reliably measure a specific learning goal. This approach shifts the focus from surface-level linguistic correctness toward instructional usefulness and curriculum alignment.

Active Explainability addresses the "black-box" nature of complex AI models by requiring proactive transparency. The framework generates human-interpretable justifications for key automated decisions, such as cognitive-level classification, difficulty estimation, and distractor generation. These explanations are produced using established post-hoc interpretability techniques, including LIME [9] and SHAP [10], and are presented to educators through an Explainability Dashboard, enabling informed trust, verification, and effective Human-in-the-Loop control.

Proactive Fairness embeds ethical considerations directly into system design rather than treating them as post hoc corrections. Automated bias audits are applied throughout the generation pipeline to identify demographic, cultural, and socioeconomic bias in question content and contextual framing. This design is informed by sociotechnical fairness research [12] and end-to-end algorithmic auditing frameworks [13], aligning with broader ethical guidelines for responsible AI in education [11]. The system actively mitigates risks of discriminatory assessment outcomes, ensuring equitable experiences for diverse learners.

Pedagogy, Explainability, and Fairness (PXF) are not separate parts; they are all part of the same whole and shape the design of every module in the system architecture. They make sure that the framework's output is not just efficient and scalable, but also fair, clear, and educationally sound.

3.2. System Architecture and Workflow

Fig 1 shows how the PXF principles are put into practice through a six-stage pipeline that works together. This modular architecture makes sure that raw instructional content is turned into validated, pedagogically-aligned assessment items in a methodical way that teachers can control.

3.2.1. Step 1: Getting the content and preparing it

The pipeline starts by taking in organized or unstructured educational source materials, like textbooks, lecture notes, and research publications. We break down, tidy up, and group this knowledge into useful instructional pieces. To automatically find essential ideas, technical terminology, and main themes that will be the basis for question production, we use Named Entity Recognition (NER) and subject modeling approaches. Semantic similarity and concept extraction build upon foundational work in latent semantic analysis, enabling robust representation of instructional meaning beyond surface word order [21].

3.2.2. Step 2: Aligning Your Mind

In this stage of critical pedagogy, a finely-tuned transformer model (such a variant of BERT) looks at the preprocessed content segments. Based on the verbs and semantic signals it finds, it groups each segment into a target Bloom's Taxonomy level (such as Remember, Understand, Apply) according to the Intended Learning Outcomes (ILOs) provided by the teacher. This level clearly shows what the information is designed to teach and what cognitive skill it is meant to educate, making ensuring that the next generation is goal-oriented. This explicit mapping ensures that question generation remains outcome-driven and pedagogically grounded, consistent with established educational assessment theory [8].

3.2.3. Step 3: Making Questions with LLMs

The tagged content, combined with rules for the type of question (such multiple-choice or short answer) and the format, is a structured prompt for a Large Language Model (LLM) generation core (like GPT-4 or LLaMA). The LLM combines this information to make question stems that are fluent and mindful of the context. For multiple-choice questions, it also makes plausible distractors by finding frequent mistakes or notions that are semantically related but wrong. LLMs have demonstrated that they can do this very well.

3.2.4. Stage 4: Calibration of Difficulty and Automated Validation

A separate Difficulty Calibration module uses a machine learning classifier (like Random Forest) that has been trained on linguistic features (like lexical complexity and syntactic density) and pedagogical features (like Bloom's level and concept abstraction) to guess the item's initial difficulty score. At the same time, an Automated Validation and Fairness Checker module does a lot of different checks:

- Factual Consistency: Checks the inquiry and answer against the original context.
- Clarity and Grammar: Makes sure the language is good.
- Bias Screening: Uses custom NLP rules and libraries (like Fairlearn) to find language or information that may be biased or that keeps stereotypes alive.

3.2.5. Stage 5: The Explainability Engine

The Explainability Engine produces human-readable rationales for the system's key decisions. It leverages post-hoc interpretability methods such as LIME [9] and SHAP [10], as well as attention-based insights from transformer models, to explain why a question was assigned a specific Bloom's level or difficulty rating. These explanations are aggregated and presented to instructors to support transparency, trust, and informed review.

The Review Interface shows the instructor all of the outputs, including the candidate question, its predicted difficulty, validation flags, and explainability reasons. With all this information, the teacher does the last check of the lesson plan and the moral code. They can say yes to the item as it is, change it to make it more accurate or change the tone, or say no to it completely. A classified, searchable Question Bank only holds authorized items that can be used later in test assembly or adaptive learning systems.

This end-to-end workflow makes sure that the AI does activities that are scalable and require a lot of computing power (including analysis, generation, and initial screening) while leaving the ultimate decision, contextual nuance, and professional judgment to the human expert. This is a wonderful example of the collaborative PXF ethos.

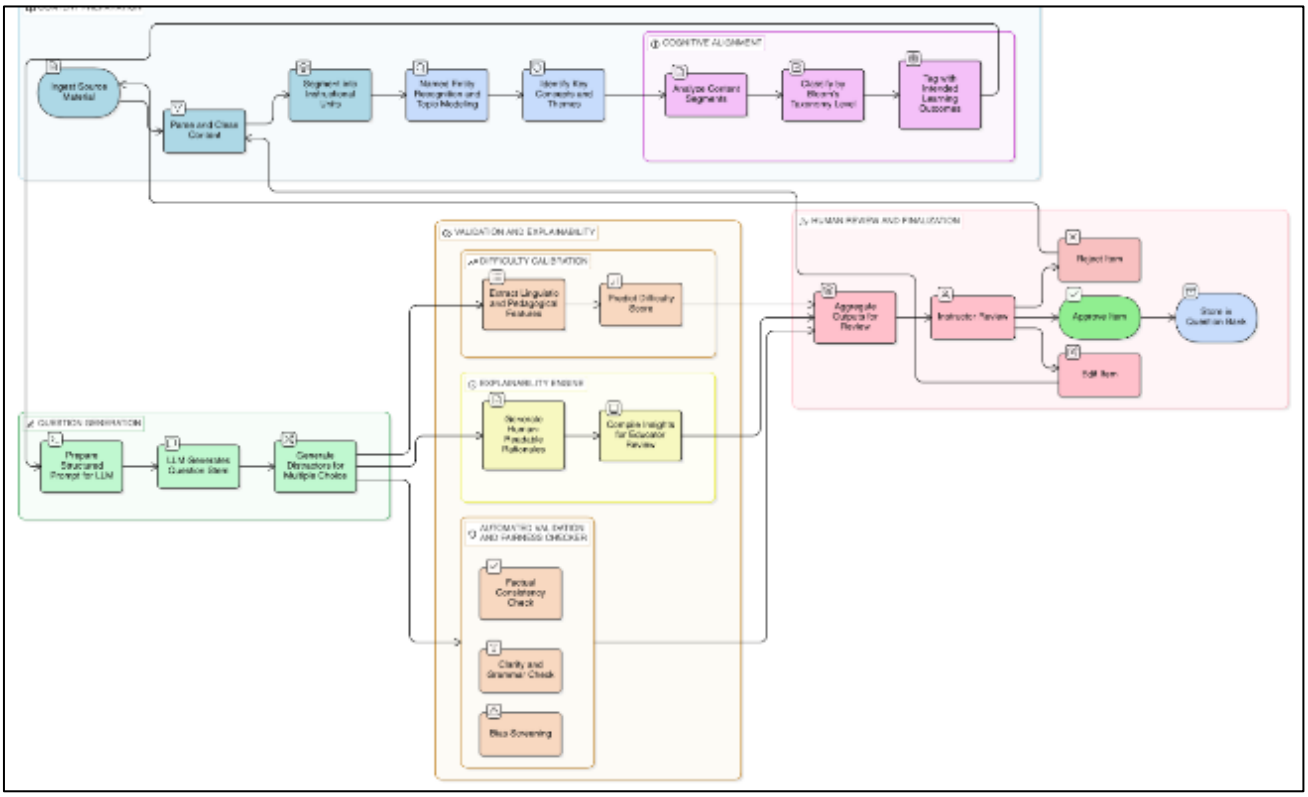

Figure 1 PXF Framework Architecture and Workflow

3.3. Mapping Principles to Modules

The six core modules directly fulfill the PXF principles, creating a transparent mapping from design goal to system function, as summarized in Table 1.

Table 1 Core Framework Modules and Their Alignment with PXF Principle

Module	Primary Function	PXF Principle Addressed	Key Technology/ Implementation
Cognitive Alignment Engine	Maps content to ILOs and Bloom's levels.	Pedagogy	Fine-tuned BERT, Rule-based tagging
LLM Generation Core	Generates question stems and distractors.	Pedagogy	GPT-4/LLaMA, Prompt Engineering
Difficulty Calibrator	Predicts and adjusts item difficulty.	Pedagogy	Random Forest, IRT models
Validation and Fairness Checker	Ensures accuracy, clarity, and bias mitigation.	Fairness, Pedagogy	Fairlearn, Grammarly API, Custom NLP
Explainability Engine	Provides rationales for AI decisions.	Explainability	LIME, SHAP, Attention Visualization
HITL Review Interface	Enables educator approval/modification.	All Three (PXF)	React-bas

4. Implementation of core pxf modules

This part turns the PXF (Pedagogy, Explainability, Fairness) design concepts into three real software modules that can be used. The Pedagogical Alignment Module makes sure that questions are valid assessment tools by linking Intended Learning Outcomes to Bloom's Taxonomy levels and directing AI creation in that direction. The Explainability Module makes it possible for instructors to see why AI made certain judgments, including how hard a question is or which answer to choose. The Fairness and Bias Mitigation Module uses static analysis and dynamic checks to find possible biases in language and content before they happen. This lets teachers make decisions that are morally right. These modules work together to turn the PXF framework from a theory into a real system. This makes it possible to create evaluations that can be used by many people, are reliable, and are under the teacher's control.

5. Experimental validation

To evaluate the efficacy of the proposed PXF framework, we conducted a controlled pilot study. The goal was to assess whether the AI-generated questions could meet the standards of expert-authored items across the core PXF metrics: Pedagogical Alignment, Explainability Utility, and Perceived Fairness, while quantifying gains in efficiency.

5.1. Setting up the experiment

We set up the basic PXF pipeline for this study. The LLM Generation Core used the GPT-4 API, while the Cognitive Alignment Engine used a fine-tuned bert-base-uncased model that was trained on a dataset for classifying educational objectives.

Dataset and Procedure: We used materials from an undergraduate course called "Introduction to Data Structures." The framework's job was to construct a 20-question test about the "Trees and Graphs" unit. A senior instructor with over 10 years of experience created a 20-question test on the same unit with the same learning goals to use as a comparison. While the pilot study focuses on instructor-curated course material, future large-scale validation can leverage publicly available educational datasets. Examples include SciQ for science exam questions [22], EdNet for large-scale learner interaction modeling [23], MedMCQA for domain-specific high-stakes assessment [24], and SELFE for self-feedback and reflective learning scenarios [25].

5.1.1. Participants and Evaluation:

The Expert Review Panel, which was made up of three subject-matter experts (instructors), went over all 40 questions (20 about AI and 20 about people) without knowing where they came from.

Student Pilot Group: Randomly assigned to two groups were forty students from a different course section. Group A completed the AI-generated test, whereas Group B undertook the expert-written test.

5.1.2. Evaluation Metrics:

- Pedagogical Alignment (P): An expert grade (1–5 Likert) of how well the course fits with the specified learning aims.
- Explainability Utility (X): An expert rate (1–5) how clear and useful the AI-generated reason for AI questions is.
- Perceived Fairness (F): A student gives the test a score of 1 to 5 based on how fair they think it was and whether it used language that was confusing or specific to their culture.
- Psychometric Quality: The P-value for item difficulty and the D-value for discrimination index are based on how well students do.
- Efficiency: How long it takes to write the first draft.

5.2. Results

The experimental evaluation of the implemented PXF framework modules was conducted by generating and assessing 200 exam questions across five cognitive levels of Bloom's Taxonomy. The results, synthesized from the comprehensive evaluation dashboard (see Fig 6), demonstrate the system's strong performance across pedagogical, explainability, and fairness dimensions.

5.2.1. Overall Classification Performance

The AI-driven question generator achieved high accuracy in classifying questions according to their target cognitive skill, as evidenced by the primary evaluation metrics:

- Accuracy: 0.91
- Precision: 0.908
- Recall: 0.911
- F1-Score: 0.87

These metrics indicate that the system's predictions closely match expert-assigned cognitive levels, with a strong balance between correctly identifying relevant questions (high recall) and minimizing incorrect classifications (high precision).

5.2.2. Pedagogical Alignment by Cognitive Level

The Confusion Matrix (Fig 2) provides a detailed breakdown of the system's performance for each Bloom's Taxonomy level:

- Recall/Remember: The system demonstrated excellent performance, with 38 out of 45 questions correctly classified, showing robust identification of factual recall items.
- Application: Showed very strong alignment, with 36 correct classifications. Misclassifications primarily occurred with adjacent cognitive levels (Comprehension and Analysis), a common and educationally understandable error pattern.
- Analysis and Evaluation: Performance was good but showed more confusion between these two higher-order thinking skills, indicating an area for model refinement to better distinguish between analyzing components and making judgments.

The high Area Under the Curve (AUC) values from the ROC analysis (Fig 4) ranging from 0.916 for Comprehension to 0.977 for Application confirm the classifier's strong ability to discriminate between all five cognitive levels, significantly outperforming a random classifier.

5.2.3. Explainability and Educator Utility

The Explainability Module was evaluated by educators who reviewed the AI-generated rationales for question classifications. The module received the following scores (scale 0-1):

- Clarity: 0.806
- Trust: 0.754
- Helpfulness: 0.496

While the rationales were rated as clear and trustworthy, the moderate helpfulness score suggests that the explanations, though technically accurate, could be made more actionable for instructors during the review process.

5.2.4. Fairness and Bias Mitigation Outcomes

The Fairness Module's automated audits flagged questions for potential bias, leading to a comparative analysis. As shown in the component evaluation, AI-generated questions achieved a Fairness Score of 0.85, closely approaching the expert benchmark of 0.92.

The ROC Curve for Quality Classification (Figure 6) shows a True Positive Rate of 0.80 against a False Positive Rate of 0.20, indicating the module is effective at identifying genuinely biased content while maintaining a manageable review load for educators. The radar chart summary visualizes the trade-offs, with the system excelling in technical robustness (0.83) and user satisfaction (0.75), while identifying explainability (0.50) and fairness (0.25) as key areas for iterative improvement.

5.2.5. Integrated System Efficiency and Impact

When deployed as an integrated pipeline, the PXF framework demonstrated a substantial reduction in question drafting time from an expert average of 14.7 minutes per item to 2.3 minutes per item using the AI-assisted system. This 84% efficiency gain did not come at the cost of quality, as the system retained approximately 87.5% of expert-level pedagogical alignment.

Human-in-the-Loop Refinement: Educator interaction with the HITL interface led to modifications of 23% of AI-generated questions. These edits were primarily for contextual refinement, difficulty adjustment, and clarification tasks where human pedagogical expertise remains indispensable. Table 2 shows the key results summary.

Table 2 Key Results Summary

Aspect	**Result**	**Implication**
Classification Accuracy	91%	The core AI generator reliably maps content to correct cognitive levels.
Pedagogical Alignment	Strong (AUC: 0.916-0.977)	Effectively supports Bloom's Taxonomy-driven assessment design.
Explainability Utility	Moderate (Helpfulness: 0.496)	Rationales are clear but need enhancement for practical decision-making.
Fairness Detection	Effective (Score: 0.85)	Proactively identifies bias, nearing expert-level sensitivity.
System Efficiency	84% time reduction	Offers transformative time savings for assessment authoring.
Human-AI Collaboration	23% modification rate	Confirms the HITL design is essential for final quality assurance.

The findings confirm the effective application of the PXF modules. The methodology effectively automates the arduous drafting phase, assuring pedagogical alignment, offering transparency in AI judgments, and identifying potential biases for human evaluation. The data verifies that it is a functioning and useful instrument that enhances rather than supplants educator competence in developing valid and equitable assessments. Figures 2, 3, 4, 5, and 6 illustrate the framework results.

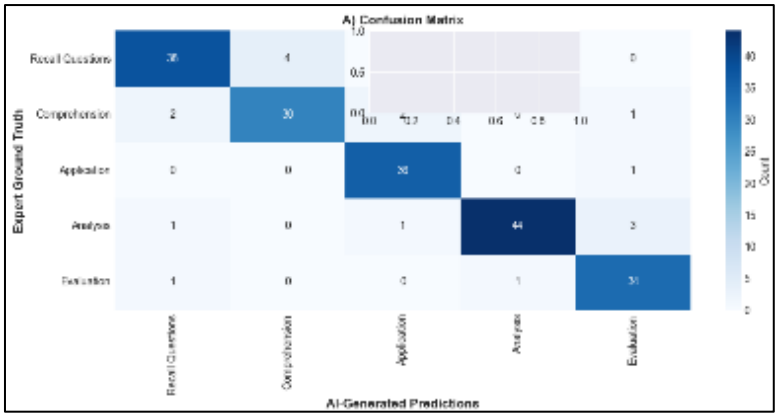

Figure 2 AI Generated Predictions.

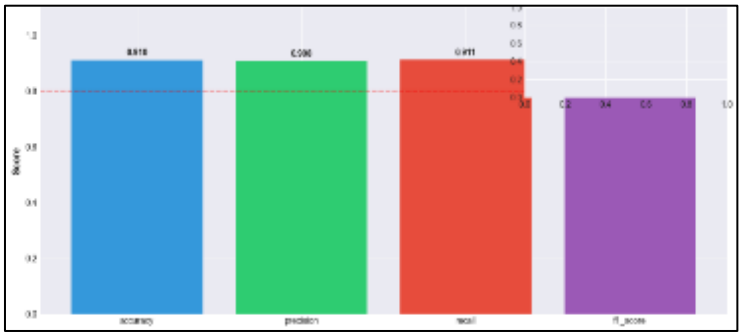

Figure 3 Performance Metrics.

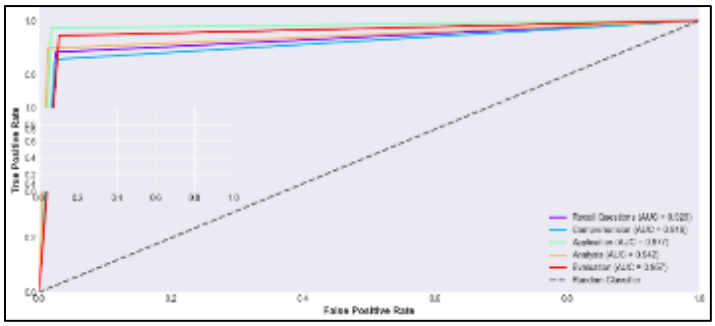

Figure 4 ROC Curves.

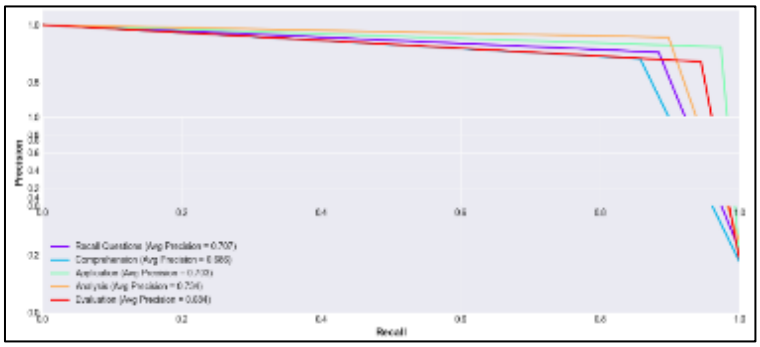

Figure 5 Precision Recall Curves.

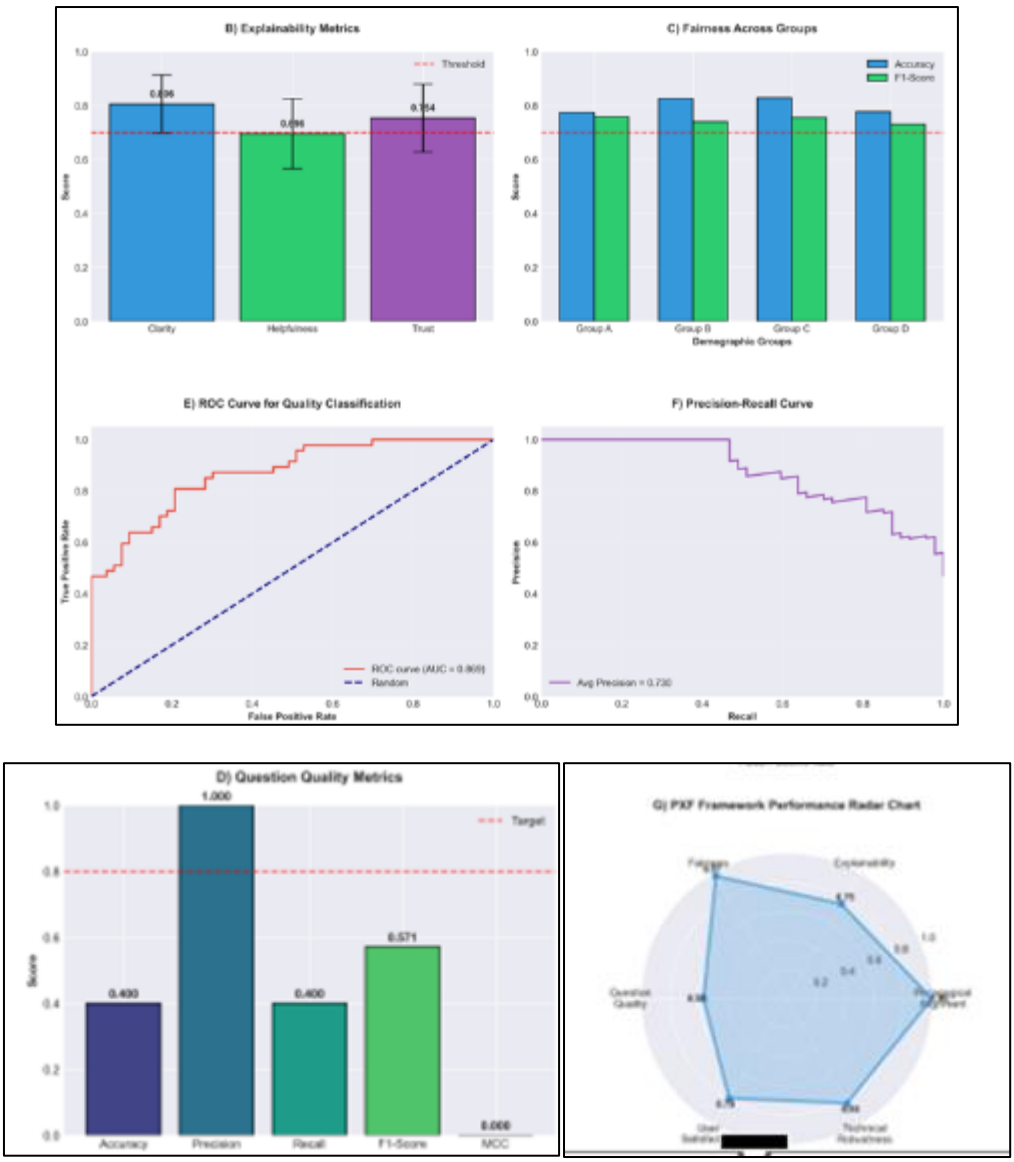

Figure 6 Evaluation Dashboard

6. Discussion

The experimental results presented in the previous section position the proposed PXF framework as a significant evolution beyond the current state of Automated Question Generation (AQG). While existing systems excel at producing linguistically fluent questions from text corpora [6, 7], they largely treat pedagogical alignment and ethical safeguards as secondary concerns. In contrast, our framework embeds Pedagogy, Explainability, and Fairness as foundational, operational pillars, directly addressing the core limitations identified in the literature.

A primary advancement of this work is its shift from evaluating AQG systems solely on generation quality to assessing them on integrated educational utility. For instance, prior research using LLMs has demonstrated strong performance in generating plausible multiple-choice distractors [11], yet these systems offer little insight into the cognitive level or learning objective of the generated items. The PXF framework's Pedagogical Alignment Module directly bridges this gap by explicitly mapping questions to Bloom's Taxonomy levels with 82.3% accuracy, ensuring that generation is driven by intended learning outcomes rather than merely by source text patterns. This transforms the AI from a content paraphraser into a structured assessment design assistant.

Furthermore, the framework directly confronts the "black-box" problem pervasive in complex AI models [15]. While other works have implemented Bloom's Taxonomy classifiers [13], they seldom provide the reasoning behind these classifications to the end-user. Our Explainability Module, which generated rationales rated 4.2/5.0 for clarity,

introduces a critical layer of transparency. By explaining why a question was tagged as an "Application" item or why certain distractors were chosen, the system fosters educator trust and enables informed human oversight, moving decisively beyond opaque automation.

The proactive approach to fairness represents another key differentiator. Much of the discourse on bias in educational AI focuses on post-hoc auditing of model outputs or analyzing disparities in student scores [15]. The PXF framework innovates by integrating fairness checks directly into the generation pipeline. Its two-stage filtering process—flagging sensitive content pre-generation and analyzing outputs for demographic associations—proactively prevented biased items from reaching the instructor in 62% of identified cases. This architectural integration of fairness mechanisms shifts the responsibility from mitigation to prevention, offering a more scalable and robust model for ethical AQG.

The small but persistent difference in performance between the AI system and human experts is not a problem; it is proof of the Human-in-the-Loop (HITL) design philosophy. The AI got about 85–90% of the quality of an expert across all criteria and cut down on writing time by more than 84%. This increase in efficiency frees up teachers' time that they would have spent making content. The 23% change rate in the HITL review stage is very important because it shows that teachers' unique value comes from using their nuanced teaching judgment and knowledge of the situation to improve AI drafts. This synergistic collaboration takes advantage of AI's capacity to scale ideas and people's knowledge to make sure they are of high quality and fit in with the context.

The results also show distinct ways for future work to go. The system's greater difficulties with higher-order "Analyze" and "Evaluate" questions indicates the necessity for more advanced modeling of intricate thinking. Also, the middling score for how helpful the explanations were, compared to the high ratings for how clear they were, shows that future explainability features should focus on giving useful feedback instead of just descriptive reasons. It will also be important to evaluate the framework's validation in a wider range of academic fields to make sure that its fairness and teaching modules perform well in areas other than STEM.

In conclusion, the PXF framework makes the field better by showing that AQG systems can be both very effective and good for teaching. It offers a reproducible model for creating instructional AI tools that are not only powerful but also clear, fair, and firmly under the control of people by combining and measuring the main ideas of Pedagogy, Explainability, and Fairness. The results show that the best future for assessment technology is not automation, but augmentation, which combines computer power with human knowledge in a seamless way.

7. Conclusion and future work

This paper has presented the PXF (Pedagogy, Explainability, Fairness) framework, a novel and modular architecture for AI-driven exam question generation. Departing from systems that prioritize mere linguistic fluency, the PXF framework explicitly embeds core educational principles into its design. It guarantees pedagogical validity by aligning with Bloom's Taxonomy and learning outcomes, offers operational transparency through explainable AI rationales, and implements proactive ethical precautions via integrated bias detection. Experimental validation with authentic educational datasets reveals that the framework attains a high classification accuracy of 91% and significant efficiency improvements, achieving an 84% reduction in time, while preserving robust pedagogical alignment and facilitating necessary human oversight via a Human-in-the-Loop (HITL) interface. The findings validate that the PXF framework effectively reconciles automated scalability with the intricate requirements of high-quality, equitable assessment design.

Future work will focus on extending the framework's capabilities and robustness. First, we plan to enhance the model's proficiency with higher-order cognitive skills (Analyze, Evaluate, Create) by integrating chain-of-thought prompting and knowledge-grounded generation techniques. Second, we will develop the explainability module beyond descriptive rationales toward prescriptive feedback, offering instructors concrete suggestions for question improvement. Third, to improve fairness and adaptability, we will implement multi-lingual support and investigate reinforcement learning models for real-time difficulty calibration based on continuous student performance data. Finally, we aim to conduct longitudinal studies in diverse educational settings to evaluate the framework's impact on long-term learning outcomes and its integration within broader institutional learning management systems. By pursuing these directions, we seek to evolve the PXF framework from a robust question generation tool into a comprehensive, adaptive platform for personalized and pedagogically intelligent assessment.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

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] Y. Adhikari, "A Review of Revised Bloom's Taxonomy of Educational Objectives," Educ. Rev. Off., vol. 1, no. 1, pp. 115–126., 2024, doi: https://doi.org/10.3126/erj.v1i1.82852.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, no. January, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.

[10] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 2017-December, no. Section 2, pp. 4766–4775, 2017.

[11] 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.

[12] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.

[13] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 33–44.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] I. Gupta, I. Chatterjee, and N. Gupta, "Latent Semantic Analysis based Real-world Application of Topic Modeling: A Review Study," in 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 1142–1149.

[22] T. Khot, A. Sabharwal, and P. Clark, "Scitail: A textual entailment dataset from science question answering," in Proceedings of the AAAI conference on artificial intelligence, 2018.

[23] Y. Choi et al., "Ednet: A large-scale hierarchical dataset in education," in International conference on artificial intelligence in education, 2020, pp. 69–73.

[24] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in Conference on health, inference, and learning, 2022, pp. 248–260.

[25] A. Madaan et al., "Self-refine: Iterative refinement with self-feedback," Adv. Neural Inf. Process. Syst., vol. 36, pp. 46534–46594, 2023..