A Hybrid Multi-Objective AI Framework for Curriculum-Aware Examination Generation

Mohamed Fathy Yehia[1]*, Yehia M. Helmi[2], Mahmoud Mohamed Bahloul[3]
Information Systems Department-Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt[1]
Professors of the Business Information Systems Department-Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt[2]
Business Information Systems Department-Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt[3]

***Abstract*—Automated examination generation has become increasingly important in modern education, where assessments must satisfy multiple pedagogical constraints, including cognitive balance, curriculum alignment, and question diversity. Existing approaches often address these requirements independently, limiting exam coherence and overall quality. To overcome this limitation, this study proposes a curriculum-aware hybrid framework that integrates Curriculum Knowledge Graphs (CKG), NSGA-II, and Proximal Policy Optimization (PPO). The problem is formulated as a constrained multi-objective optimization task that simultaneously maximizes Bloom's taxonomy alignment, difficulty balance, and CLO coverage while minimizing semantic redundancy. The CKG captures relationships among questions, concepts, and CLOs to ensure structured curriculum alignment; NSGA-II generates Pareto-optimal exam candidates, and PPO further refines them through adaptive policy learning. The framework was evaluated on a dataset of 8,000 annotated questions using cross-validation, ablation studies, and statistical significance testing. Results demonstrate strong performance, achieving a Bloom Balance Score of 0.84 and CLO coverage of 87.5%, while reducing semantic redundancy from 10.7% to 3.2% (Δ = 7.5 percentage points; 70 % relative reduction, $p < 0.001$).**

Keywords*—NSGA-II; Reinforcement learning; automated exam generation; Bloom's taxonomy; CLO alignment

I. Introduction

Assessment quality is a fundamental component of higher education, as it directly influences learning outcomes, curriculum effectiveness, and instructional quality. Effective assessments must ensure valid measurement across cognitive levels while maintaining alignment with Course Learning Outcomes (CLOs). However, exam design remains largely manual, time-intensive, and susceptible to inconsistencies in cognitive distribution, difficulty calibration, and CLO alignment, particularly in large-scale and hybrid learning environments where scalability and consistency are essential [1]. Within outcome-based education (OBE), assessment design requires explicit alignment with predefined learning outcomes and constructive integration between instructional activities and evaluation methods [2]. Despite this, conventional practices lack systematic mechanisms for ensuring cognitive balance, redundancy control, and structured CLO mapping, leading to variability in assessment quality across courses and institutions [3–6]. Recent advances in artificial intelligence (AI), including natural language processing (NLP), knowledge representation, and reinforcement learning (RL), have enabled partial automation of assessment-related tasks such as automatic question generation and semantic modeling [7–10]. However, existing approaches predominantly operate at the question level and do not address exam construction as a global optimization problem. Consequently, key pedagogical requirements, such as cognitive balance, difficulty distribution, CLO coverage, and redundancy minimization, are not jointly enforced [11].

From a pedagogical standpoint, Bloom's Taxonomy provides a structured basis for ensuring balanced cognitive coverage across multiple levels of thinking [12]. In parallel, curriculum knowledge graphs (CKGs) enable structured representation of relationships among concepts, questions, and CLOs, supporting curriculum-aware reasoning and traceability [13]. Although these techniques enhance individual components of assessment design, they are typically applied in isolation and lack integration within a unified optimization framework. Moreover, existing optimization-based approaches are often limited by static or pipeline-based designs, restricting their ability to refine candidate solutions after initial generation. The absence of adaptive refinement mechanisms and the limited integration between evolutionary optimization and reinforcement learning constrain the exploration of globally optimal exam configurations [14,15].

To address these limitations, this study investigates two research questions: (RQ1) whether multi-objective optimization can improve exam quality in terms of cognitive balance, CLO coverage, and redundancy reduction; and (RQ2) whether reinforcement learning can enhance Pareto-optimal solutions through adaptive refinement. Accordingly, this study proposes a unified curriculum-aware framework that models exam generation as a constrained multi-objective optimization problem. The framework integrates curriculum knowledge graphs, evolutionary optimization, and reinforcement learning within a single architecture, enabling coordinated optimization across pedagogical objectives. This approach supports the generation of cognitively balanced, curriculum-aligned, and diverse examinations, advancing scalable and consistent assessment design in modern educational systems.

II. Related Work

Recent research on AI-driven assessment has evolved along three main directions: question generation, curriculum-aware

*Corresponding author.

alignment, and automated exam construction. While substantial progress has been achieved within each direction, existing studies remain methodologically fragmented, with limited integration across these dimensions.

A. Question Generation

Neural Question Generation (NQG) has advanced through sequence-to-sequence architectures and attention mechanisms. Hassan et al. (2025) [20] showed that integrating linguistic features such as part-of-speech (POS) and named entity recognition (NER) improves grammatical accuracy and contextual relevance. Mulla et al. (2023) [21] demonstrated that LSTM-based models capture long-range dependencies, enabling more diverse question structures. In multiple-choice settings, Qiu et al. (2020) [22] enhanced distractor quality using attention-based gating mechanisms. More recently, large language models (LLM)-based approaches have improved fluency and semantic coherence. Wang et al. (2025) [23] reported near human-level performance using prompt-engineered LLMs grounded in teacher knowledge bases, while Zahn et al. (2026) [24] combined retrieval-augmented generation (RAG) with human-in-the-loop validation to ensure domain accuracy. Despite these advances, such methods operate at the individual question level and lack mechanisms for coordinating questions within a complete exam. This results in redundancy, inconsistent difficulty distribution, and limited curriculum coverage, indicating that question generation remains a local optimization task.

B. Curriculum and Pedagogical Alignment

Ensuring alignment between generated questions and curriculum objectives remains a key challenge. Yaacoub et al. (2025) [25] incorporated Bloom's Taxonomy into question generation, showing improved cognitive-level classification using transformer-based models. Abdul Wahid et al. (2025) [26] leveraged RAG techniques to enhance factual consistency and curriculum alignment. Nattawuttisit et al. (2024) [27] explored recommendation-based methods for personalized curriculum sequencing, while Ghanimet et al. (2026) [28] proposed the PXF framework integrating pedagogy, explainability, and fairness. Although these approaches improve cognitive labeling and personalization, alignment is typically treated as a post-generation step. Curriculum mapping and cognitive classification are implemented as separate modules, without enforcing consistency across the full assessment. Consequently, they do not ensure balanced CLO coverage or coherent pedagogical structure at the exam level.

C. Optimization-Based Exam Generation

Curriculum: Recent studies have explored AI-driven exam construction using LLMs and hybrid frameworks. Nikolovski et al. (2025) [16] utilized RAG-based agents to enhance semantic coherence, while Papachristou et al. (2025) [17] proposed a multilingual pipeline for exam generation and grading. Mahamad et al. (2025) [18] highlighted the integration of AI techniques within decision-support systems, and Samant et al. (2025) [19] showed that LLM-based grading improves evaluation performance when combined with retrieval mechanisms [30]. Despite improvements in automation and scalability, these approaches model exam generation as sequential pipelines rather than formal optimization problems. Key constraints, including difficulty balancing, redundancy minimization, and CLO coverage, are not jointly optimized, and adaptive refinement mechanisms remain limited. Existing approaches address question generation, curriculum alignment, and exam construction in isolation. There is a lack of unified frameworks that jointly optimize pedagogical constraints, integrate curriculum-aware representations during generation, and support adaptive refinement through reinforcement learning. This gap motivates the proposed framework, which formulates exam generation as a multi-objective, curriculum-aware optimization problem.

D. Research Gap and Novelty of the Proposed Framework

Existing AI-based assessment systems primarily focus on either question generation, curriculum alignment, or optimization-based exam construction as separate tasks. Most existing frameworks rely on sequential pipelines in which generated questions are filtered or evaluated after generation, without jointly optimizing pedagogical objectives at the examination level. The proposed framework differs from previous approaches in three fundamental aspects:

First, exam generation is formulated as a curriculum-aware multi-objective optimization problem in which Bloom's taxonomy balance, difficulty distribution, CLO coverage, and semantic diversity are optimized simultaneously. Unlike existing methods that treat curriculum alignment as a post-processing step, the proposed Curriculum Knowledge Graph (CKG) is embedded directly within the optimization process, enabling curriculum constraints to actively guide solution generation.

Second, the framework introduces a hybrid global-local optimization strategy that combines NSGA-II and PPO reinforcement learning. NSGA-II performs global exploration and generates a diverse Pareto-optimal solution set under multiple competing pedagogical objectives. PPO subsequently performs adaptive local refinement through policy-guided exam modification actions. This enables the framework to overcome local inefficiencies that remain within Pareto-optimal solutions and achieve improved cognitive balance and curriculum coverage.

Third, unlike traditional optimization-based exam generation systems that terminate after evolutionary search, the proposed architecture incorporates a closed-loop refinement mechanism in which reinforcement learning continuously improves candidate examinations based on curriculum-aware reward signals. This transforms exam generation from a static optimization process into an adaptive optimization framework capable of learning improved assessment configurations over time. Therefore, the novelty of this work lies not in the individual components themselves, but in the curriculum-aware integration of knowledge graph reasoning, evolutionary multi-objective optimization, and reinforcement learning refinement within a unified end-to-end framework for automated examination generation.

III. METHODOLOGY

The proposed framework formulates automated exam generation as a constrained multi-objective optimization problem that jointly optimizes cognitive balance, difficulty

distribution, curriculum alignment, and semantic diversity while minimizing redundancy among selected questions. As illustrated in Fig. 1, the framework consists of five interconnected stages: 1) pedagogically annotated dataset preparation, 2) semantic representation and similarity modeling, 3) Curriculum Knowledge Graph (CKG) construction, 4) multi-objective optimization using NSGA-II, and 5) reinforcement learning–based refinement. A curriculum-aware knowledge graph models relationships among questions, concepts, and Course Learning Outcomes (CLOs), providing structured guidance during exam construction. NSGA-II explores the solution space to generate diverse Pareto-optimal exam configurations under competing pedagogical constraints, while a Proximal Policy Optimization (PPO) agent further refines these solutions through adaptive adjustments. By combining curriculum-aware reasoning, evolutionary search, and reinforcement learning refinement within a unified workflow, the framework generates pedagogically balanced and curriculum-aligned examinations.

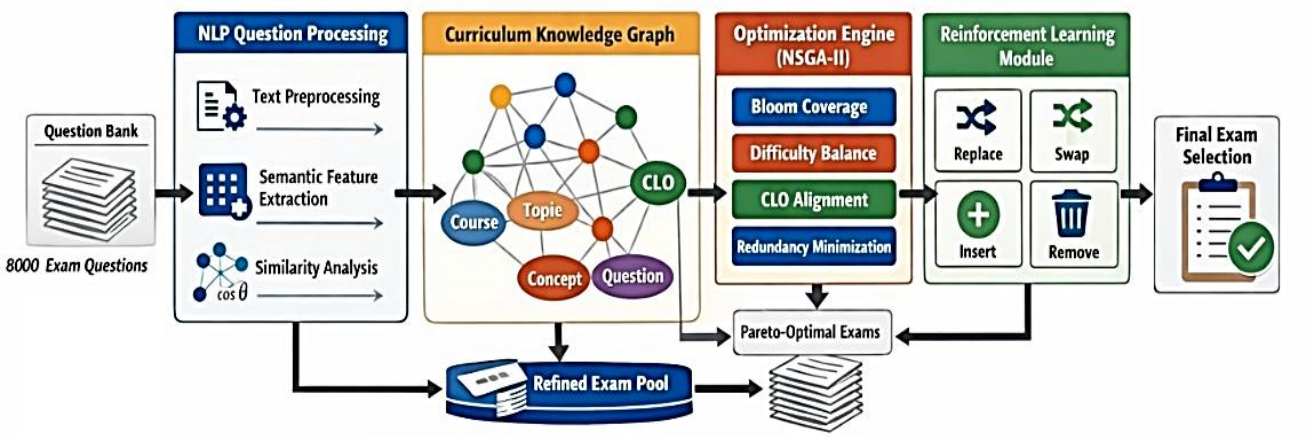

Fig. 1. End-to-end automated exam generation workflow.

The main contributions of this study are:

- Formulating exam generation as a curriculum-constrained multi-objective optimization problem that simultaneously considers Bloom's taxonomy balance, difficulty distribution, CLO coverage, and semantic redundancy.
- Embedding Curriculum Knowledge Graph (CKG) reasoning directly within the optimization process, enabling curriculum alignment during solution generation rather than as a post-processing step.
- Introducing a hybrid global–local optimization strategy in which NSGA-II performs Pareto-based exploration, and PPO provides adaptive policy-driven refinement of candidate examinations.
- Developing a unified end-to-end framework that integrates curriculum modeling, multi-objective optimization, and reinforcement learning for automated exam generation.

Demonstrating, through extensive experiments and ablation analysis, that the proposed hybrid framework improves cognitive alignment, CLO coverage, and question diversity compared with standalone optimization approaches.

A. Proposed Hybrid Curriculum-Aware Framework

1) Question bank preparation and annotation: Question Bank Preparation and Annotation present a curated dataset of 8,000 assessment items drawn from undergraduate Internet Applications and Database Management Systems courses within an Outcome-Based Education framework. Each item is semantically enriched with structured pedagogical metadata, including Bloom's cognitive level, difficulty, CLO alignment, question type, estimated time to solve, marks, and chapter mapping, enabling fine-grained, curriculum-aware optimization. Domain experts independently validated annotations in accordance with standardized OBE guidelines, ensuring reliability and consistency, as assessed using Cohen's Kappa. This multi-dimensional representation supports balanced cognitive distribution, difficulty calibration, and curriculum coverage for assessment generation and evaluation. Table I summarizes the dataset attributes used in this study.

TABLE I. DATASET ATTRIBUTES

Factor	Description
Question ID	Unique identifier for each question (e.g., structured encoding reflecting level and metadata)
Course Code	Institutional alphanumeric code identifying each course
Course Title	Full course name (e.g., DBMS, Internet Applications)
Question Type	Format of the question (MCQ, True/False, Short Answer, Essay, Problem Solving)
Bloom's Level	Cognitive level based on Bloom's Taxonomy
Difficulty Level	Predefined difficulty category (Easy, Medium, Hard)
CLO	Associated Course Learning Outcome
Marks	Score or weight assigned to the question
Estimated Time	Expected time required for solving the question
Chapter	Source chapter or unit within the course syllabus

2) Semantic preprocessing and representation: To support semantic diversity and redundancy-aware exam generation, all

questions were processed through a standard NLP preprocessing pipeline consisting of normalization, tokenization, stop-word removal, and lemmatization. Following preprocessing, questions were transformed into numerical feature representations using TF-IDF vectorization, enabling quantitative estimation of semantic similarity across assessment items. Semantic redundancy between questions $q_i$and $q_j$was computed using cosine similarity:

$$sim(q_i, q_j) = \frac{V(q_i) \cdot V(q_j)}{||V(q_i)|| \, ||V(q_j)||} \quad (1)$$

where, $V(q)$denotes the TF-IDF representation of question q. The resulting similarity scores were incorporated into the optimization process to penalize semantically overlapping questions and improve assessment diversity.

3) Curriculum Knowledge Graph (CKG) construction: A Curriculum Knowledge Graph (CKG) was constructed to represent hierarchical and semantic relationships among curriculum components, including courses, chapters, concepts, CLOs, and assessment items. The graph provides structured linkage between questions and their associated pedagogical elements, enabling curriculum-aligned reasoning during question selection. Relationships such as question–CLO mapping, chapter–concept coverage, and concept- CLO support are explicitly encoded to preserve instructional dependencies. The graph was implemented in Neo4j for efficient relational querying, while Node2Vec embeddings were employed to learn latent structural representations for similarity estimation and optimization. The overall structure is illustrated in Fig. 2, showing integrated alignment across curriculum entities.

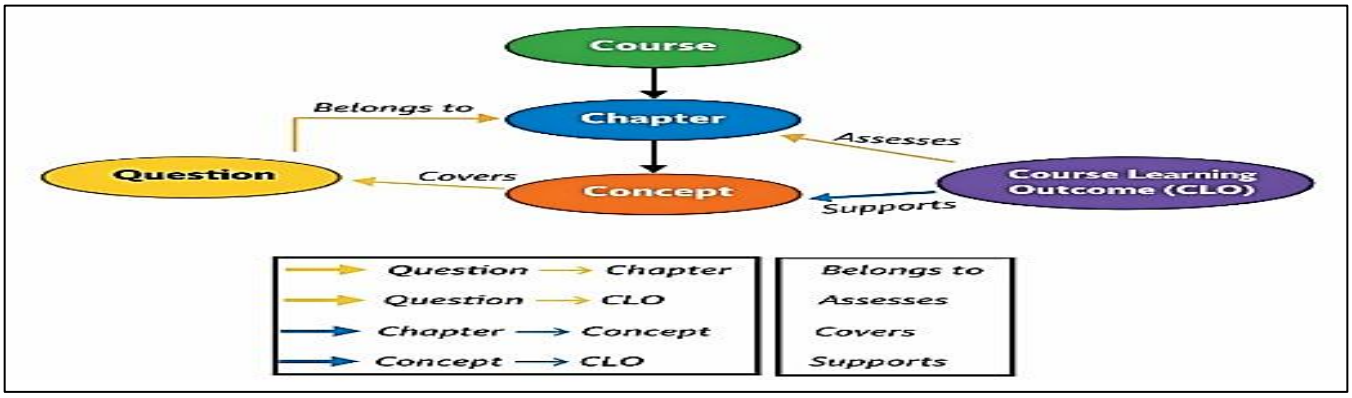

Fig. 2. Curriculum knowledge graph for structured alignment between questions, concepts, and CLO.

4) Multi-objective optimization using NSGA-II: Candidate examinations are encoded as binary vectors, where each element denotes the inclusion or exclusion of a question. Accordingly, exam generation is formulated as a constrained combinatorial multi-objective optimization problem that jointly optimizes four pedagogical criteria: cognitive balance, difficulty distribution, CLO coverage, and redundancy minimization. Given the inherent trade-offs among these objectives, the NSGA-II algorithm is employed to approximate a Pareto-optimal solution set while maintaining population diversity. Specifically, non-dominated sorting is used to organize candidate solutions into hierarchical Pareto fronts, enabling the retention of multiple high-quality exam configurations rather than converging to a single solution. Diversity within each front is preserved using crowding distance, which prioritizes sparsely distributed candidates and mitigates premature convergence.

The optimization process begins with a randomly initialized population to ensure broad exploration of the search space. Each candidate is evaluated using four normalized objective functions: (f_1) cognitive alignment, (f_2) difficulty balance, (f_3) CLO coverage, and (f_4) redundancy minimization, where higher values indicate better pedagogical quality. Evolution proceeds through tournament selection based on Pareto rank and crowding distance, followed by crossover and mutation to generate offspring solutions. At each generation, parent and offspring populations are merged and re-ranked using non-dominated sorting to preserve Pareto optimality. This iterative process yields a diverse set of candidate examinations that capture different trade-offs among competing pedagogical objectives, making the formulation well-suited for curriculum-aware exam generation. The formal definitions of the objective functions are provided in Eq. (2)-(5).

Cognitive Balance (Bloom's Distribution):

$$f_1 = 1 - \frac{1}{2}\sum_{i=1}^{6} | P_i - T_i | \quad (2)$$

where, $P_i$and $T_i$represent the achieved and target Bloom's distributions, respectively.

Difficulty Balance:

$$f_2 = \sum_{d \in \{Easy, Medium, Hard\}} | D_d - T_d | \quad (3)$$

where, D_d and T_d denote the achieved and target distributions across difficulty levels.

CLO Coverage:

$$f_3 = \frac{\text{Covered CLOs}}{\text{Total CLOs}} \quad (4)$$

Redundancy Minimization:

$$f_4 = 1 - R \quad (5)$$

where, R denotes the average pairwise semantic similarity between selected questions.

5) Reinforcement learning-based refinement: Although NSGA-II generates high-quality Pareto-optimal candidate examinations, some solutions may still exhibit local inefficiencies in cognitive alignment or curriculum coverage. To address this limitation, a reinforcement learning refinement stage based on Proximal Policy Optimization (PPO) was introduced. The RL environment state is represented as:

$$s_t = [B_t, D_t, C_t, R_t] \quad (6)$$

where, B_t, D_t, C_t, and R_t denote Bloom's distribution, difficulty distribution, CLO coverage, and redundancy level, respectively. The action space consists of four refinement operators:

$$A = \{insert, remove, replace, swap\} \quad (7)$$

These operators enable controlled modifications of exam structures while preserving feasibility constraints. The PPO agent optimizes the following reward function:

$$R = w_1 f_1 + w_2 f_2 + w_3 f_3 - w_4 f_4 \quad (8)$$

where, f_1, f_2, f_3, and f_4 represent cognitive balance, difficulty alignment, CLO coverage, and redundancy reduction. The framework adopts a hybrid global–local optimization strategy, where NSGA-II explores diverse Pareto solutions while PPO performs iterative policy-based refinement, improving convergence stability and exam quality, as illustrated in Fig. 3.

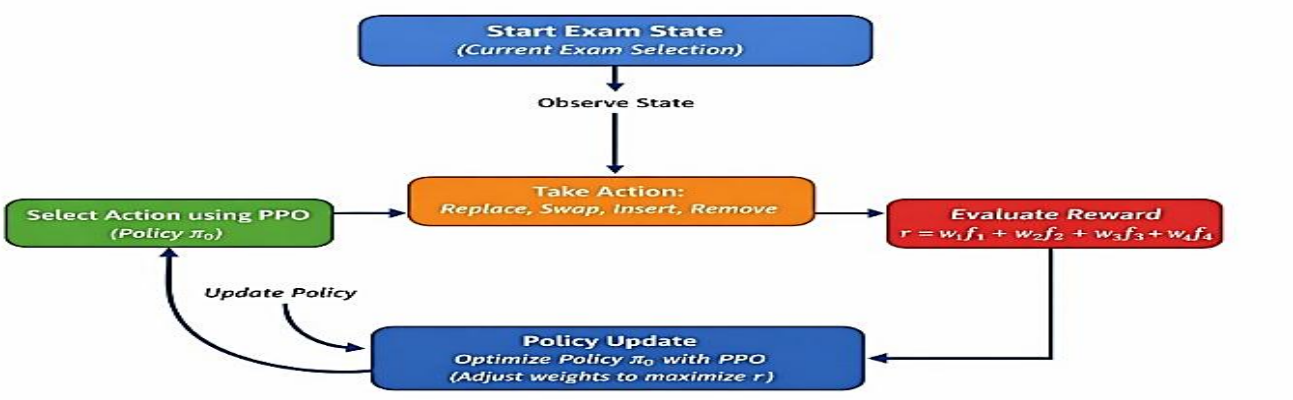

Fig. 3. RL-based optimization loop for iterative refinement of candidate exams.

Reward weights were empirically calibrated using grid search under OBE constraints. Candidate configurations were evaluated through 5-fold cross-validation based on Bloom alignment, CLO coverage, and redundancy minimization. The selected weights ((w_1=0.30), (w_2=0.25), (w_3=0.25), (w_4=0.20)) achieved the best overall performance. A sensitivity analysis was conducted by varying each weight within ±20% and re-normalizing the weight vector (\sum w=1). Across all folds and runs, performance variations remained below 3.5% for Bloom alignment and CLO coverage, indicating strong robustness and low sensitivity to reward-weight settings.

B. Experimental Configuration

The proposed framework was implemented in Python using Scikit-learn for semantic preprocessing, DEAP for NSGA-II optimization, Stable-Baselines3 for PPO training, and Neo4j for graph management. The PPO agent used two hidden layers (128,64) with ReLU activations and was trained for 500 episodes (lr=3e-4, γ=0.99, ε=0.2). Evaluation used 5-fold cross-validation and focused on Bloom Balance Score, CLO coverage, and redundancy rate, as shown in the figure, with statistical significance testing applied at $p < 0.05$.

C. Computational Complexity Analysis

The computational complexity of the proposed framework is governed by the integration of NSGA-II for global optimization and PPO for local refinement. For NSGA-II, the dominant cost arises from non-dominated sorting and fitness evaluation, yielding a complexity of:

$$O(G \cdot M \cdot P^2) \quad (9)$$

This is where G is the number of generations, M the number of objectives, and P the population size. This quadratic dependency on P reflects pairwise dominance comparisons during Pareto sorting. For the PPO refinement stage, the computational cost depends on the number of training episodes E and refinement steps T, expressed as:

$$O(E \cdot T) \quad (10)$$

The overall complexity of the hybrid framework is therefore:

$$O(G \cdot M \cdot P^2 + E \cdot T) \quad (11)$$

With respect to the question bank size N, the scalability of the framework is influenced by the fitness evaluation step, particularly redundancy computation. Using precomputed semantic similarity matrices, redundancy evaluation scales are approximately as:

$$O(k^2) \quad (12)$$

where, k is the number of selected exam questions rather than the full question bank size N, which improves scalability for large repositories. Overall, the hybrid architecture balances computational cost and optimization quality by assigning global exploration to NSGA-II and local refinement to PPO. This

design improves optimization effectiveness while maintaining feasible runtime performance for medium- to large-scale question banks.

IV. Results and Discussion

This section evaluates the proposed curriculum-aware exam generation framework through cross-validation, ablation analysis, baseline comparison, statistical validation, human evaluation, and a representative study. The evaluation focuses on four pedagogical objectives: cognitive balance, curriculum alignment, redundancy reduction, and overall assessment feasibility.

A. Experimental Setup and Dataset

The proposed framework generates structured examinations consisting of 20 questions (100 marks, 90 minutes) under explicit constraints on Bloom's taxonomy distribution, difficulty balance, CLO coverage, and semantic redundancy. Experiments are conducted on a curated dataset of 8,000 annotated questions drawn from Internet Applications (4,200) and Database Management Systems (3,800). Each question is enriched with pedagogical metadata, including Bloom level, difficulty, CLO mapping, question type, estimated solving time, and score weight. This structured representation enables fine-grained pedagogical control during optimization and supports curriculum-aware exam generation.

B. Preprocessing

A structured preprocessing pipeline is applied to ensure robust semantic and pedagogical representation of the dataset. The textual content is first standardized through normalization, tokenization, and lemmatization to reduce lexical variation and ensure linguistic consistency. A TF-IDF representation is then constructed to capture discriminative term-level importance across the corpus. Semantic relationships between questions are modeled using cosine similarity over TF-IDF embeddings, providing a continuous measure of textual overlap. This representation directly supports redundancy-aware optimization by enabling the identification and suppression of semantically similar questions, thereby enhancing diversity, reducing redundancy, and improving overall assessment quality.

C. Cross-Validation Performance

5-fold cross-validation is performed to assess the robustness of the proposed framework across three configurations: NSGA-II, NSGA-II enhanced with Curriculum Knowledge Graph (KG), and the full KG + NSGA-II + RL model. The results, summarized in Table II, show consistent improvements across all evaluation metrics as the model complexity increases.

TABLE II. Cross-Validation Performance of the Framework

Model	Bloom Score (Mean ± SD)	CLO Coverage (%) (Mean ± SD)	Redundancy (%) (Mean ± SD)
NSGA-II	0.72 ± 0.03	75.0 ± 2.1	10.7 ± 1.5
NSGA-II + KG	0.79 ± 0.02	82.3 ± 1.8	6.5 ± 1.2
Full Model (NSGA-II + KG + RL)	**0.84 ± 0.02**	**87.5 ± 1.5**	**3.2 ± 0.9**

The full framework achieves the best performance, with a Bloom score of 0.84, CLO coverage of 87.5%, and redundancy reduced to 3.2%. These results indicate that integrating curriculum-aware knowledge modeling with evolutionary optimization and reinforcement learning significantly enhances both pedagogical alignment and structural diversity of generated exams. Moreover, the low standard deviation across folds demonstrates stable performance and strong generalization capability under different data splits.

D. Comparative Evaluation with State-of-the-Art Methods

Existing methods mainly address question-level classification or generation using transformer-based or LLM-based pipelines. In contrast, the proposed framework formulates exam generation as a constrained multi-objective optimization problem that jointly optimizes Bloom's taxonomy distribution, CLO coverage, and semantic redundancy. Because baseline studies differ in datasets and evaluation protocols, direct numerical comparison should be interpreted cautiously (Table III). Despite this, transformer-based models such as BERT and DistilBERT show strong performance in Bloom-level prediction due to their contextual understanding, but they remain limited to local, question-level optimization without modeling global exam structure. Conversely, the proposed hybrid framework integrates Curriculum Knowledge Graphs, NSGA-II, and PPO reinforcement learning, enabling global exam-level optimization. This unified design leads to superior CLO coverage and reduced redundancy, while maintaining comparable Bloom alignment with classification-based approaches.

TABLE III. Comparative Evaluation of Educational Frameworks

Ref.	Approach	Bloom	κ	Main Outcome
[31]	BERT	0.91	0.891	Cognitive classification
[32]	DistilBERT	0.89	—	Bloom alignment evaluation
[33]	Knowledge Graph	0.84	0.875	Curriculum reasoning
This Work	CKG + NSGA-II + PPO	0.84	0.875	Automated exam generation

E. Ablation Study

This section evaluates the individual and combined contributions of the proposed components, namely Curriculum Knowledge Graph (CKG), NSGA-II, and PPO, to quantify their impact on overall optimization performance.

TABLE IV. Extended Ablation Results

Model	Bloom Score	CLO Coverage (%)	Redundancy (%)
PPO Only	0.67	71.2	12.1
NSGA-II	0.72	75.0	10.7
NSGA-II + PPO	0.81	84.1	5.4
NSGA-II + CKG	0.79	82.3	6.5
Full Model (CKG + NSGA-II + PPO)	0.84	87.5	3.2

As shown in Table IV, standalone NSGA-II achieves moderate performance in terms of cognitive balance (0.72), CLO coverage (75.0%), and redundancy reduction (10.7%),

reflecting its capability for global exploration but limited adaptability in refining solutions. In contrast, PPO alone demonstrates weaker performance compared to evolutionary search, indicating that policy-based refinement without a strong global search mechanism is insufficient for high-quality exam construction.

When NSGA-II is combined with PPO, a clear improvement is observed across all metrics, where Bloom alignment increases to 0.81, CLO coverage to 84.1%, and redundancy decreases significantly to 5.4%. This confirms the complementary nature of global exploration (NSGA-II) and local refinement (PPO). Further improvement is achieved when incorporating the Curriculum Knowledge Graph (CKG), where semantic and pedagogical constraints enhance structural alignment and reduce redundancy. The full hybrid model (CKG + NSGA-II + PPO) achieves the best performance, with a Bloom score of 0.84, CLO coverage of 87.5%, and redundancy reduced to 3.2%. Overall, the results indicate that performance gains are not attributable to any single component but rather emerge from the hierarchical integration of curriculum-aware representation, evolutionary optimization, and reinforcement learning. Specifically, CKG enhances semantic grounding, NSGA-II ensures global Pareto-optimal exploration, and PPO provides adaptive local refinement. An ablation study is conducted to quantify the contribution of each component in the proposed framework. As illustrated in Fig. 4

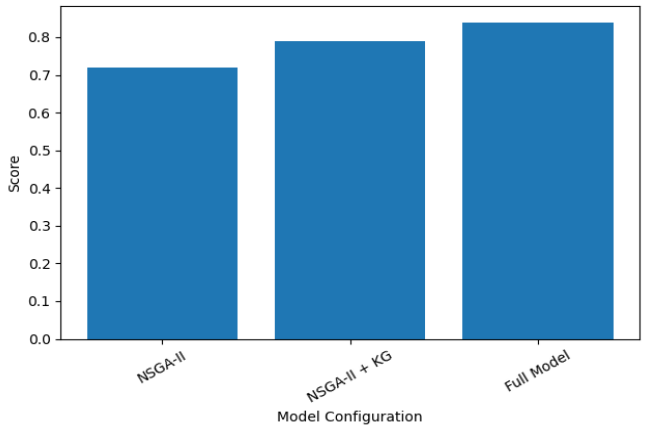

Fig. 4. Ablation study illustrating the contribution of each framework component to overall performance metrics.

F. Cognitive and Difficulty Distribution

The generated examinations exhibit a well-structured distribution across Bloom's taxonomy levels, covering Remember, Understand, Apply, Analyze, Evaluate, and Create. Higher-order cognitive skills are notably represented, with Apply (25%) and Analyze (20%) contributing substantially, while lower-order levels remain sufficiently included. This reflects effective alignment with established assessment design principles [29]. Cognitive coverage is quantified as:

$$\text{Coverage}_i = \frac{|Q_i|}{|Q|} \times 100 \qquad (14)$$

where, $|Q_i|$ denotes the number of questions at Bloom level i, and $|Q|$ is the total number of questions. In parallel, the difficulty distribution demonstrates a balanced allocation across predefined levels (easy, medium, hard), as summarized in Table V.

TABLE V. DIFFICULTY DISTRIBUTION OF GENERATED EXAMS

Difficulty	Percentage
Easy	30%
Medium	45%
Hard	25%

This distribution indicates a reasonable spread across difficulty levels, supporting fair and comprehensive assessment. Difficulty coverage is computed as:

$$\text{Coverage}_d = \frac{N_d}{|Q|} \times 100 \qquad (15)$$

where, N_d represents the number of questions within each difficulty category. Overall, the results indicate that the proposed framework maintains a consistent balance between cognitive complexity and difficulty levels, supporting robust and structured exam design. Fig. 5 further illustrates the proportional distribution of difficulty categories.

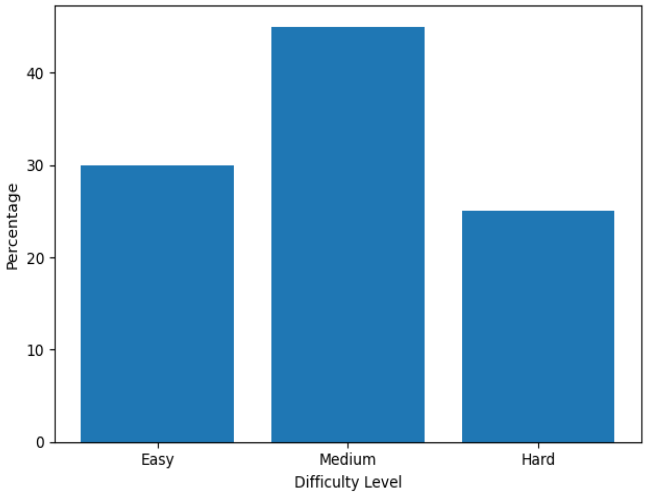

Fig. 5. The proportional distribution of difficulty levels.

G. CLO Coverage and Curriculum Alignment

The proposed framework achieves a CLO coverage of 87.5% (7 out of 8 CLOs), indicating strong alignment with curriculum objectives. CLO coverage is defined as:

$$\text{CLO Coverage} = \frac{\text{Number of covered CLOs}}{\text{Total CLOs}} \times 100 \qquad (16)$$

As illustrated in Fig. 6, all CLOs are represented with non-zero contributions, demonstrating comprehensive curriculum coverage. The distribution remains largely balanced within the 10%–15% range for most CLOs, with CLO 4 receiving the highest emphasis (22%) due to its central role, while CLO 7 and CLO 8 show lower representation (9% and 7%), reflecting their more specialized nature. Overall, the results confirm that integrating Curriculum Knowledge Graph reasoning with reinforcement learning ensures both complete coverage and structured prioritization of learning outcomes, leading to a pedagogically balanced assessment design.

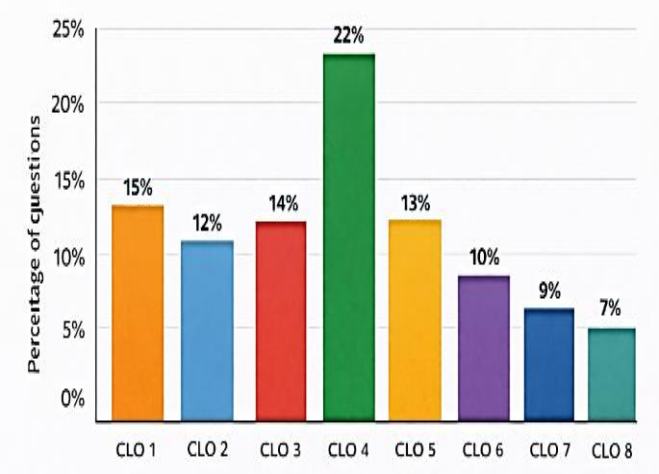

Fig. 6. Presents the distribution of questions across CLOs.

H. Impact of Reinforcement Learning

To further enhance cognitive alignment, a reinforcement learning (RL) refinement stage is applied after the NSGA-II optimization process. Cognitive balance is evaluated using the Bloom Balance Score (BBS) defined in Eq. (2). The RL component iteratively refines candidate exam configurations through actions such as question replacement and swapping, guided by a reward function that promotes closer alignment with the target Bloom's taxonomy distribution. This enables fine-grained adjustments beyond the Pareto-optimal solutions produced by NSGA-II and helps correct residual imbalances remaining after global optimization. As shown in Fig. 7, RL improves the BBS, indicating enhanced cognitive alignment and overall exam coherence.

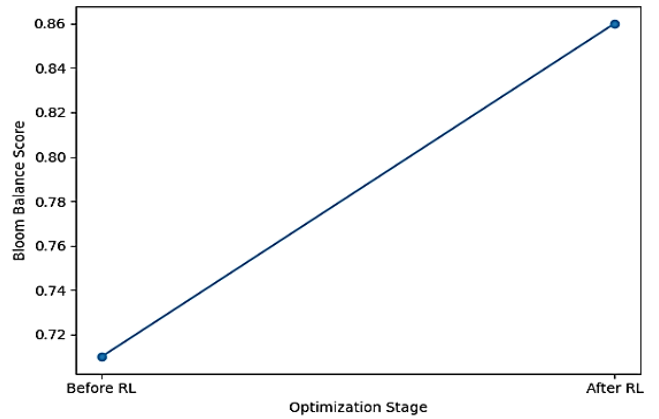

Fig. 7. Impact of reinforcement learning on bloom balance score.

I. Baseline Comparison and Redundancy Analysis

Redundancy measures the degree of semantic overlap among questions in an exam, directly affecting assessment diversity and quality. Let $Q = \{q_1, q_2, \ldots, q_n\}$ denote the set of selected questions, and $\text{sim}(q_i, q_j)$ The semantic similarity between two questions. The redundancy score $R$is defined as:

$$R = \frac{2}{n(n-1)} \sum_{i<j} I(\text{sim}(q_i, q_j) > \tau) \times 100 \quad (17)$$

where, n is the number of questions, τ is a predefined similarity threshold, and $I(\cdot)$is an indicator function. Lower R values indicate higher diversity. The framework's effectiveness is evaluated against baseline methods, with results reported in Table VI.

TABLE VI. COMPARISON WITH BASELINE METHODS

Method	Bloom Score	CLO Coverage (%)
Random Selection	0.52	55%
Rule-Based Generator	0.68	72%
Manual Exam Creation	0.65	70%
Proposed Framework	**0.84**	**87.5%**

The proposed framework achieves the highest performance in both cognitive alignment and CLO coverage. In addition, redundancy is reduced from 10.7% to 3.2% (≈70%), as illustrated in Fig. 8, reflecting effective semantic filtering during multi-objective optimization. Overall, the results indicate improved diversity and stronger alignment with pedagogical objectives compared to baseline approaches.

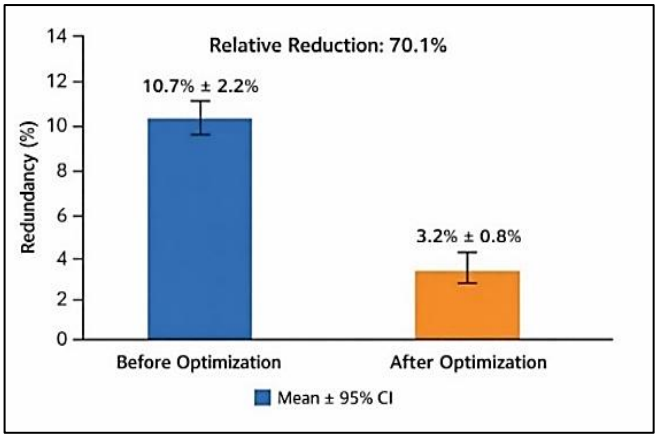

Fig. 8. Redundancy reduction after diversity optimization.

J. Human Evaluation

The human evaluation was conducted using a controlled experimental design involving 5 domain experts in computer science education. Each expert independently evaluated a fixed set of 20 generated exam papers, yielding a total of 100 evaluation instances (5 experts × 20 exams per expert). The evaluation dataset was constructed using stratified random sampling from the full pool of generated exams to ensure balanced representation across Bloom's taxonomy levels, difficulty categories, and CLO coverage distributions. All exam samples were fully anonymized and randomized, and evaluators were blinded to the generation source (NSGA-II, KG-enhanced NSGA-II, and Full Hybrid Model) to eliminate potential bias. The use of a shared evaluation set across all experts enabled the computation of inter-rater reliability (Cronbach's α = 0.87) and ensured methodological consistency in the human assessment process. As shown in Table VII.

TABLE VII. HUMAN EVALUATION RESULTS

Evaluation Criterion	Mean Score	Standard Deviation	Interpretation
Clarity	4.30	0.40	High clarity and readability of generated questions
Difficulty Appropriateness	4.10	0.50	Well-calibrated difficulty levels aligned with student ability
CLO Alignment	4.40	0.30	Strong consistency with intended learning outcomes
Overall Quality Index	**4.27**	**0.41**	High overall pedagogical quality

K. Statistical Validation

Paired t-tests were conducted across 5-fold cross-validation to assess statistical significance. As shown in Table VIII, all metrics demonstrate significant improvements ($p < 0.001$) with large effect sizes. These results confirm the statistical robustness and practical effectiveness of the proposed framework.

TABLE VIII. STATISTICAL ANALYSIS SUMMARY

Metric	Mean ± SD	95% CI	Cohen's d
Bloom Score	0.84 ± 0.05	[0.82, 0.86]	1.50 (Large)
CLO Coverage	87.5 ± 3.4%	[85.9, 89.1]	1.42 (Large)
Redundancy	3.2 ± 1.1%	[2.8, 3.6]	1.58 (Large)

L. Case Study: Generated Exam Analysis

A representative exam from a Database Management Systems (DBMS) course was analyzed to assess practical effectiveness. The overall structure is summarized in Table IX.

TABLE IX. GENERATED EXAM STRUCTURE AND DISTRIBUTION

Difficulty / Bloom Level	Questions	Percentage	Total Score	Time Allocation
Easy/ Remember	2	10%	8	6 min
Easy/Understand	4	20%	16	12 min
Medium / Apply	5	25%	27	23 min
Medium/Analyze	4	20%	27	22 min
Hard / Evaluate	3	15%	22	19 min
Hard / Create	2	10%	22	8 min
Total	**20**	**100%**	**100**	**90 min**

The exam generated demonstrates a coherent and balanced design. Higher-order cognitive levels constitute 70% of the questions, ensuring emphasis on critical thinking, while lower-order levels remain adequately represented. Difficulty levels are consistently aligned with cognitive categories, supporting fair assessment across varying abilities. Additionally, the distribution of marks and time is well-calibrated, resulting in a practical and manageable 90-minute exam. Overall, the case study confirms that the framework produces cognitively balanced, difficulty-aware, and operationally feasible examinations.

V. DISCUSSION

The results demonstrate that integrating curriculum-aware representations with multi-objective optimization significantly enhances the pedagogical quality of automated exam generation. Embedding the Curriculum Knowledge Graph (CKG) within the optimization process enables explicit alignment with Course Learning Outcomes (CLOs), outperforming post-hoc alignment strategies and ensuring structurally coherent assessments. Methodologically, the proposed framework reframes exam generation as a constrained multi-objective optimization problem, allowing simultaneous control over cognitive distribution, difficulty balance, CLO coverage, and semantic redundancy. The evolutionary component (NSGA-II) provides effective global exploration of high-quality solutions, while the PPO-based reinforcement learning stage performs fine-grained local refinement beyond the Pareto front, improving solution optimality and stability. The ablation and comparative analyses confirm that performance gains are not attributable to a single module but emerge from the synergy between curriculum modeling, evolutionary search, and reinforcement learning. In contrast to LLM-based or standalone optimization approaches, the proposed framework offers a unified and scalable optimization-driven architecture for pedagogically consistent exam generation. Despite these strengths, the study is limited by the dataset scope and the size of expert annotations, which may affect generalizability. Future work should extend evaluation across broader academic domains and investigate integration with large language models to enhance semantic understanding while preserving optimization constraints.

VI. CONCLUSION AND FUTURE WORK

This study introduced a hybrid multi-objective framework for curriculum-aware automated examination generation by integrating Curriculum Knowledge Graphs (CKG), NSGA-II evolutionary optimization, and Proximal Policy Optimization (PPO) reinforcement learning. The framework formulates exam generation as a constrained optimization problem that jointly optimizes Bloom's taxonomy distribution, difficulty balance, CLO coverage, and redundancy reduction within a unified architecture. Experimental results demonstrate that the integration of curriculum-aware knowledge representation with evolutionary optimization significantly improves pedagogical alignment and structural quality of generated exams. Furthermore, the reinforcement learning refinement stage enhances solution quality by enabling adaptive local optimization beyond global Pareto-based search, resulting in more balanced and reliable exam configurations. Overall, the findings confirm that combining multi-objective optimization with reinforcement learning provides an effective and scalable mechanism for automated assessment design. This approach reduces manual workload while ensuring curriculum-consistent and cognitively balanced examinations, making it suitable for deployment in Learning Management Systems (LMS). Future work will focus on extending the framework to larger and more diverse datasets across multiple disciplines, improving generalization capability. In addition, integrating large language models with curriculum-aware optimization and exploring real-time adaptive exam generation in intelligent tutoring systems represent promising directions for further advancement.

REFERENCES

[1] Levy-Feldman, I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. Education Sciences, 15(2), 224. https://doi.org/10.3390/educsci15020224

[2] Pageni, S. (2025). Student Assessment in Higher Education: Perspectives and Practices from a Practitioner Inquiry. KMC Journal, 7(2), 225-239.

[3] Monib, W. K., Qazi, A., Apong, R. A., Azizan, M. T., De Silva, L., & Yassin, H. (2024). Generative AI and future education: a review, theoretical validation, and authors' perspective on challenges and solutions. PeerJ Computer Science, 10, e2105.

[4] Sachar, S. (2025). Managing Assessment Challenges in Diverse Classrooms.

[5] Gudoniene, D., Staneviciene, E., Huet, I., Dickel, J., Dieng, D., Degroote, J., ... & Casanova, D. (2025). Hybrid teaching and learning in higher education: A systematic literature review. Sustainability, 17(2), 756.

[6] Saichaie, K. (2020). Blended, flipped, and hybrid learning: Definitions, developments, and directions. New Directions for Teaching and Learning, 2020(164), 95-104.

[7] Jabr, R. B., & Azmi, A. M. (2025). Knowledge-Aware Arabic Question Generation: A Transformer-Based Framework. Mathematics, 13(18), 2975. https://doi.org/10.3390/math13182975

[8] O'Shaughnessy, D. (2026). An Overview of Recent Advances in Natural Language Processing for Information Systems. Applied Sciences, 16(2), 1122. DOI: 10.3390/app16021122

[9] Salian, D. T., Elkhodari, G., Neouchi, R., Brown, S., Babulak, E., & Sbeit, R. (2026). A New Approach to Improving Natural Language Processing Capabilities Using Generative AI: A Systematic Review and Future Perspectives. The Social Impact of Next-Generation Smart Cyber Technology, 139-172. DOI: 10.4018/979-8-3373-5656-3.ch005

[10] Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. arXiv preprint arXiv:2402.01512.

[11] Li, S., Li, S., Yang, Z., Zhang, X., Chen, G., Xia, X., ... & Peng, Z. (2025). Learnalign: Reasoning data selection for reinforcement learning in large language models based on improved gradient alignment. arXiv preprint arXiv:2506.11480.

[12] Li, M., & Rohayati, M. I. (2024). The Relationship between Learning Outcomes and Graduate Competences: The Chain-Mediating Roles of Project-Based Learning and Assessment Strategies. Sustainability, 16(14), 6080. https://doi.org/10.3390/su16146080

[13] Nguyen, H. N. (2025, May). A Knowledge Graph-Based Framework for Personalized Course Recommendations in Higher Education. In 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 853-858). IEEE.

[14] Saleh, A. O. M., Tur, G., & Saygin, Y. (2025). SG-RAG MOT: SubGraph Retrieval Augmented Generation with Merging and Ordering Triplets for Knowledge Graph Multi-Hop Question Answering. Machine Learning and Knowledge Extraction, 7(3), 74. https://doi.org/10.3390/make7030074

[15] Ballah, Anas & Ghanim, Hussein & Hageltoum, Abdallah & Idris, Salwa. (2026). Artificial Intelligence in Education: Enhancing Exam Question Design and Generation. 11. 90-98.

[16] Nikolovski, V., Trajanov, D., & Chorbev, I. (2025). Advancing AI in Higher Education: A Comparative Study of Large Language Model-Based Agents for Exam Question Generation, Improvement, and Evaluation. Algorithms, 18(3), 144. https://doi.org/10.3390/a18030144

[17] Papachristou, I., Dimitroulakos, G., & Vassilakis, C. (2025). Automated Test Generation and Marking Using LLMs. Electronics, 14(14), 2835. https://doi.org/10.3390/electronics14142835

[18] Mahamad, S., Chin, Y. H., Zulmuksah, N. I. N., Haque, M. M., Shaheen, M., & Nisar, K. (2025). Technical review: Architecting an AI-driven decision support system for enhanced online learning and assessment. Future Internet, 17(9), 383.

[19] Samant, T., Bhole, G., & Udmale, S. (2025, December). Automating Exam Evaluation Using Generative AI (GenAI). In Proceedings of the 17th annual meeting of the Forum for Information Retrieval Evaluation (pp. 65-70).

[20] Hassan, A., & Eid, M. (2025). A Systematic Review of Automatic Neural Question Generation. Journal of the ACS Advances in Computer Science, 16(1).

[21] Mulla, N., & Gharpure, P. (2023). Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence, 12(1), 1-32.

[22] Qiu, Z., Wu, X., & Fan, W. (2020). Automatic distractor generation for multiple choice questions in standard tests. arXiv preprint arXiv:2011.13100.

[23] Wang, L., Song, R., Guo, W., & Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. Interactive Learning Environments, 33(3), 2559-2584.

[24] Zahn, Andrew & Overla, Seth & Lowrie, D & Zhou, Christine & Santen, Sally & Zheng, Weibing & Turner, Laurah. (2026). An Artificial Intelligence-Driven Platform for Practice Question Generation. Academic medicine : journal of the Association of American Medical Colleges. 10.1093/acamed/wvaf074.

[25] Yaacoub, Antoun & Da-Rugna, Jérôme & Assaghir, Zainab. (2025). Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment. 10.48550/arXiv.2504.14232.

[26] Abdul Wahid, Rohaizah & Nadim, Muhamad & Sulaiman, Suliana & Shaharudin, Syahmi & Jupikil, Muhammad & Su, Iqqwan. (2025). Automated Generation of Curriculum-Aligned Multiple-Choice Questions for Malaysian Secondary Mathematics Using Generative AI. 10.48550/arXiv.2508.04442.

[27] Nattawuttisit, S. O. O. K. S. A. W. A. D. D. E. E., & Maneerat, P. A. R. A. L. E. E. (2024). AI-driven adaptive curriculum development: Enhancing student learning outcomes aligned with the Thai qualifications framework in higher education. Journal of Theoretical and Applied Information Technology, 102(17), 6512-6520.

[28] Ghanim, H. A., Ballah, A. A., Hageltoum, I. A., & Idris, S. (2026). AI-Driven Framework for Exam Question Design and Generation: Pedagogy, Explainability and Fairness.

[29] Sultan, Sherif & Abdel-Fattah, Manal & Al-Wakeel, Nashat. (2021). A Framework for Automatic Exam Generation based on k-means and Genetic Algorithm. International Journal of Computer Applications. 183. 18-23. 10.5120/ijca2021921576.

[30] Ling, Jintao & Afzaal, Muhammad. (2024). Automatic question-answer pairs generation using pre-trained large language models in higher education. Computers and Education Artificial Intelligence. 6. 100252. 10.1016/j.caeai.2024.100252

[31] Almatrafi, O., & Johri, A. (2025). Leveraging generative AI for course learning outcome categorization using Bloom's taxonomy. Computers and Education: Artificial Intelligence, 8, 100404

[32] Yaacoub, A., Da-Rugna, J., & Assaghir, Z. (2025). Assessing AI-generated questions' alignment with cognitive frameworks in educational assessment. arXiv preprint arXiv:2504.14232

[33] Nguyen, H. N. (2025, May). A Knowledge Graph-Based Framework for Personalized Course Recommendations in Higher Education. In 2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD) (pp. 853-858). IEEE