Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 04, No. 01, 2024 ♦ PP. 01–11

Article DOI: https://doi.org/10.54216/FinTech-I.040101

REVIEW ARTICLE

Supervised Machine Learning Algorithms for Equity Market Regime Classification: A Systematic Literature Review of Comparative Performance, Feature Engineering, and Generalizability (2015–2024)

Suvonkulov Abdulaziz[1,*] **Eugene Q. Castro**[1]

[1] Department of Computer Science, Central Asian University, Tashkent, Uzbekistan

Emails: 220456@centralasian.uz · e.castro@centralasian.uz

Received: January 12, 2024 Revised: March 01, 2024 Accepted: June 28, 2024 ⋆ Corresponding author

ABSTRACT

The application of supervised machine learning (ML) algorithms for equity market regime classification has gained significant attention in recent years. This systematic literature review (SLR) synthesizes findings from 16 peer-reviewed studies published between 2015 and 2024 to address three research questions: (1) How do supervised ML algorithms (XGBoost, Random Forest, SVM, Neural Networks, Ensemble methods) compare in accuracy, robustness, and computational efficiency for market regime classification? (2) What feature engineering approaches are most effective? (3) How generalizable are these models across different equity markets and time periods? Following PRISMA 2020 guidelines, we searched IEEE Xplore, ScienceDirect, and Springer, identifying 2953 records and including 16 studies after screening. Our findings indicate that ensemble methods (particularly Random Forest and XGBoost) and deep learning approaches (LSTM, DNN) consistently outperform traditional classifiers. Technical indicators remain the most common features, though novel approaches including event embeddings, network centrality measures, and signal decomposition show promise. Generalizability remains a challenge, with most studies focusing on developed markets. We identify gaps in cross-market validation and interpretability, providing directions for future research.

Keywords: Systematic literature review ▪ Machine learning ▪ Stock market prediction ▪ Regime classification ▪ XGBoost ▪ Random Forest ▪ LSTM ▪ Deep learning ▪ Feature engineering

1. INTRODUCTION

Assessment is one of the most influential parts of education because it shapes what students practice, what teachers prioritize, and how learning progress is measured. In higher education and online learning environments, assessment is also a major operational burden: instructors must repeatedly design quizzes, produce alternative versions, write rubrics, grade responses, and provide feedback, often under time pressure and with large class sizes. As a result, students may receive fewer opportunities for practice and delayed feedback, even though frequent low-stakes assessment is strongly linked to improved learning outcomes [1].

To address these constraints, digital learning platforms have increasingly adopted automated and semi-automated approaches for quiz creation and evaluation. In recent years, the rise of transformer-based models and large language models (LLMs) has accelerated this trend by enabling systems that can generate questions, produce distractors for multiple-

choice items, create explanations, summarize content into practice tasks, and even score student answers using rubrics or model-based judging [2, 3, 4, 5]. These capabilities are particularly attractive for adaptive learning, where content and difficulty should adjust to the learner's level and where timely feedback is essential to guide improvement [3, 4].

However, despite rapid progress, the research landscape around AI-based quiz and assessment systems remains fragmented. Studies often focus on different parts of the pipeline (e.g., question generation, answer scoring, feedback generation, conversational assessment) and use different datasets, task definitions, and evaluation metrics [2, 6, 7]. Even when two works claim to solve the same problem, they may evaluate quality in incomparable ways, such as using different difficulty labels, different human-rating procedures, or different measures of linguistic quality and educational validity [2, 6]. This diversity makes it difficult to draw reliable conclusions about which methods are consistently effective, which evaluation practices are trustworthy, and which systems are ready for real classroom deployment [7].

In addition, AI-driven assessment introduces risks that are especially important in educational contexts. Generated questions can be ambiguous, factually incorrect, or misaligned with learning objectives; distractors can become unrealistic or reveal the correct answer [4, 2]. Automatic scoring can be sensitive to phrasing, length, or writing style rather than true understanding, and LLM-based judging may suffer from instability across prompts or biased evaluations [5, 2]. Beyond technical issues, practical concerns include fairness, transparency, privacy, and academic integrity, particularly when students interact with AI tools that can also help them answer the very questions being assessed [6, 2]. These limitations highlight the need for careful evaluation standards and responsible design, not only improved model performance [7].

Given these challenges, a structured synthesis of the literature is needed to clarify what has been studied, how quality has been measured, and where the most critical gaps remain [2, 6, 7]. This paper presents a Systematic Literature Review (SLR) of AI-based quiz and assessment systems for adaptive learning. Following a PRISMA-guided review process, we collect and analyze peer-reviewed studies from major academic databases, categorize approaches across the quiz/assessment workflow, and summarize the datasets, evaluation methods, and metrics used in prior work [2, 6, 7]. We also consolidate recurring limitations, ethical and practical risks, and research opportunities that can improve reliability and real-world adoption [7].

The contributions of this paper are as follows:

- We provide a structured overview of AI techniques used for quiz generation, automated assessment, and feedback in adaptive learning settings [3, 4, 8, 5].
- We summarize evaluation practices used in the literature, including datasets, human-judging methods, and commonly reported metrics [2, 6, 7].
- We highlight recurring challenges and open issues that affect robustness, fairness, transparency, and academic integrity [2, 6, 5].
- We identify directions for future research and practical recommendations to support responsible and scalable educational assessment tools [7, 6].

2. METHODOLOGY

This study employs a Systematic Literature Review (SLR) to synthesize recent research on Artificial Intelligence (AI), Natural Language Processing (NLP), and Large Language Model (LLM)-based quiz and assessment systems that support adaptive learning in higher education. The review follows established SLR practices commonly used in educational technology and AI-based question generation research to ensure transparency, rigor, and reproducibility [2, 6, 7].

2.1 Research Questions

The review is guided by the following research questions, which reflect the major technical, evaluative, and practical concerns identified in prior surveys and review studies of AI-driven educational assessment:

- **RQ1:** What AI/NLP/LLM techniques are used to generate quiz and assessment items for adaptive learning in higher education?
- **RQ2:** What learning, scoring, and engagement outcomes are reported for AI-driven quiz/assessment systems in digital learning contexts?
- **RQ3:** What challenges, limitations, and research gaps are reported when applying AI-based quiz/assessment systems in higher education?

2.2 Review Protocol

The SLR protocol consisted of the following stages, aligned with commonly reported workflows in prior systematic reviews of automatic question generation and AI-based educational assessment [2, 7]:

1. **Identification:** Search and collect records from selected scholarly databases and a grey-literature source.
2. **Deduplication:** Combine all retrieved records and remove duplicates (manual check in Excel).
3. **Screening:** Screen titles and abstracts against predefined inclusion/exclusion criteria.
4. **Eligibility:** Retrieve full texts and assess eligibility; exclude non-matching studies with documented reasons.
5. **Data Extraction:** Extract standardized fields (method/model, task focus, evaluation evidence, findings, limitations, etc.) into an Excel sheet (one row per included study).
6. **Synthesis:** Conduct qualitative thematic synthesis by grouping studies into common themes (e.g., generation methods, scoring/feedback, evaluation outcomes, and reported gaps).

Study selection summary (PRISMA counts). A total of 57 records were retrieved (27 from databases and 30 from Google Scholar). After removing 2 duplicates, 55 records

were screened by title and abstract, resulting in the exclusion of 39 records. Sixteen papers were moved to full-text retrieval; 3 full texts were not accessible. Thirteen full-text papers were assessed for eligibility and 4 were excluded with documented reasons, resulting in **9 included studies** for qualitative synthesis. This reporting structure follows PRISMA-style documentation practices commonly adopted in recent SLRs in this domain [6, 7].

2.3 Databases Searched

Three sources were selected to ensure comprehensive coverage of peer-reviewed research in computing, artificial intelligence, and educational technology:

- **IEEE Xplore**
- **ACM Digital Library (ACM DL)**
- **Google Scholar** (treated as a grey-literature source)

These databases are widely used in prior reviews of automatic question generation, AI-based assessment, and educational NLP systems, and together provide access to high-impact journal articles and conference proceedings in this research area [2, 6].

2.4 Search Strategy

The search strategy was designed using a structured concept-block approach, consistent with systematic reviews in AI-based educational assessment, to capture studies addressing (i) AI/NLP/LLM techniques, (ii) quiz and assessment tasks, and (iii) higher education and digital/adaptive learning contexts [6, 7]. Keywords were grouped into three concept blocks and combined using Boolean logic:

- **AI Technique Block:** "large language model", LLM, "generative AI", NLP, "natural language processing"
- **Assessment Task Block:** "automatic question generation", "question generation", "item generation", quiz, assessment, test, examination
- **Education Context Block:** "higher education", university, college, "tertiary education", "e-learning", "online learning", "mobile learning"

For IEEE Xplore and the ACM Digital Library, advanced search interfaces were used to combine the three blocks with the **AND** operator, while synonymous terms within each block were linked using **OR**. Quotation marks were applied to multi-word phrases to preserve semantic meaning and reduce irrelevant matches. Database filters were applied to improve precision and reproducibility, including publication years (IEEE Xplore: 2020–2025; ACM DL: 2024–2025) and document type (peer-reviewed journal articles and conference papers), following common practice in recent SLRs [2, 7].

Google Scholar was used as a complementary source to capture recently published or highly cited review-oriented studies that may not yet be consistently indexed across curated digital libraries. Because Google Scholar result counts vary over time and are not fully reproducible, the search was restricted to publications since 2025 and limited to review articles, sorted by relevance. Only the first set of result pages was screened, consistent with accepted SLR practices in educational technology research [6].

All searches were executed on the same date, and the exact search strings, filters, and initial result counts were recorded to form a transparent search log.

2.5 Search Log

To ensure transparency and reproducibility, a search log was maintained for every database query. The log records: (i) the database/source, (ii) the exact search string used, (iii) applied filters (year range and document type), (iv) the number of results returned at the time of search, and (v) notes about database behavior (e.g., Google Scholar count variability). Maintaining a detailed search log is a standard practice in systematic literature reviews to support auditability and reproducibility of the review process [2, 6, 7]. All searches were executed on **15 November 2025**.

IEEE Xplore (2020–2025; conferences and journals):

```
("large language model" OR LLM OR "generative AI" OR
NLP OR "natural language processing") AND ("automatic
question generation" OR "question generation" OR
"item generation") AND (quiz OR assessment OR test OR
examination) AND ("higher education" OR university OR
college OR "tertiary education") AND ("e-learning" OR
"online learning" OR "mobile learning")
```

ACM Digital Library (2024–2025; Research Article):

```
("large language model" OR llm OR "generative ai" OR
nlp OR "natural language processing") AND ("automatic
question generation" OR "question generation" OR
"item generation") AND (quiz OR assessment OR test OR
examination) AND ("higher education" OR university OR
college) AND ("e-learning" OR "online learning" OR
"mobile learning")
```

Google Scholar (since 2025; Review Articles; sort by relevance):

```
"automatic question generation" ("large language
model" OR LLM) ("higher education" OR university)
(quiz OR assessment)
```

Google Scholar was treated as a complementary grey-literature source to capture recently published review-oriented studies that may not yet be consistently indexed in curated digital libraries. Because Google Scholar result counts vary over time and ranking is not fully reproducible, only the top-ranked results were screened, following accepted SLR practice in educational technology research [6, 7].

Table 1. Summary of Search Log

Source	Filters / Notes	Results
IEEE Xplore	Year + document type filters; peer-reviewed sources; exported and recorded in Excel.	11
ACM Digital Library (ACM DL)	Year + document type filters; peer-reviewed sources; exported and recorded in Excel.	16
Google Scholar	Since 2025; review articles; sorted top-ranked results due to ranking variability.	30
Total	–	**57**

2.6 PRISMA Identification and Screening

The study selection process followed the PRISMA 2020 flow, which is widely adopted for systematic reviews to ensure transparent reporting of identification, screening, eligibility assessment, and inclusion decisions [6, 7].

Identification:

- Records identified from database searching: **57**
  - IEEE Xplore: **11**
  - ACM Digital Library (ACM DL): **16**
  - Google Scholar: **30**

Duplicate removal:

- Duplicates removed: **2**
- Records after duplicates removed: **55**

Screening:

- Records screened (title and abstract): **55**
- Records excluded during screening: **39**

Eligibility:

- Reports sought for retrieval (full-text): **16**
- Reports not retrieved (full-text not accessible): **3**
- Reports assessed for eligibility: **13**
- Reports excluded with reasons: **4**
  - Out of scope (not AI-based quiz/assessment for adaptive learning in higher education): **2**
  - Not peer-reviewed / not an eligible publication type: **1**
  - Duplicate or earlier version of the same study: **1**

Included:

- Studies included in qualitative synthesis: **9**
- Studies included in quantitative synthesis (meta-analysis): **0**

2.7 PRISMA Diagram

Fig. 1 presents the PRISMA 2020 flow diagram summarizing the identification, screening, eligibility assessment, and inclusion steps applied in this SLR. The PRISMA framework is widely used to support transparent reporting of study selection decisions in systematic reviews, particularly in computer science and educational technology research [6, 7].

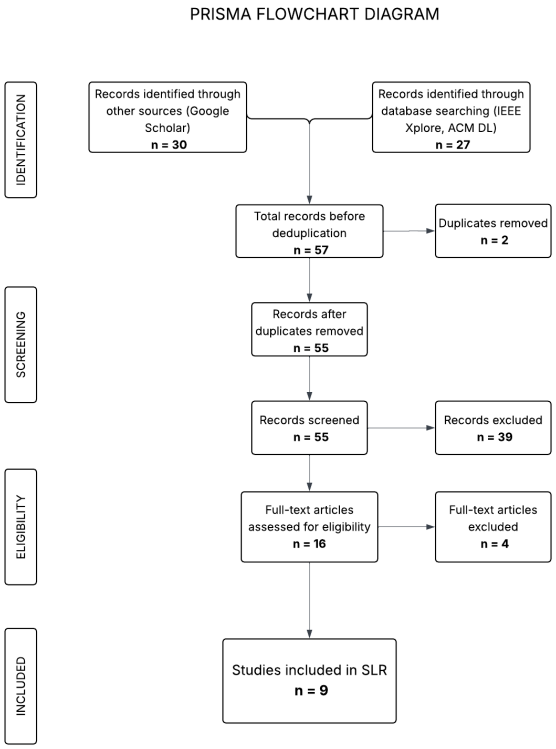

Figure 1. PRISMA 2020 flow diagram illustrating the study selection process.

2.8 Inclusion and Exclusion Criteria

To ensure consistency during screening and eligibility assessment, explicit inclusion and exclusion criteria were defined and applied across all retrieved records. The criteria were formulated based on common practices reported in prior systematic reviews of automatic question generation and AI-based educational assessment systems [2, 7].

Inclusion criteria:

- Peer-reviewed journal articles or conference papers.
- Studies focused on AI/NLP/LLM-based quiz, test, or assessment systems (e.g., question generation, automated scoring, feedback generation, or assessment-oriented conversational systems).
- Studies situated in higher education or clearly applicable to university/tertiary learning contexts.
- Studies reporting an empirical evaluation (e.g., human judgment, automatic metrics, user study, classroom/learning study, or benchmark-based experiments).
- Full text available in English.
- Published within the applied search windows (IEEE Xplore: 2020–2025; ACM DL: 2024–2025; Google Scholar: since 2025 per query constraints).

Exclusion criteria:

- Non-peer-reviewed sources (e.g., blogs, opinion pieces, non-scholarly articles) or publication types not aligned with the review scope.

- Studies not addressing quiz or assessment functionality (e.g., general tutoring/chatbots without assessment, generic learning analytics without quiz/assessment tasks).
- Studies outside higher education or tertiary contexts with no clear transferability to university-level assessment.
- Duplicate records, extended abstracts, or earlier versions of the same study (only the most complete version retained).
- Papers with inaccessible full text at the time of eligibility assessment.
- Out-of-scope studies (e.g., not AI-based, not related to adaptive learning assessment objectives, or unrelated application domains).

2.9 Quality and Relevance Assessment

Each included study (S01–S09) was evaluated using a lightweight quality and relevance appraisal to estimate confidence in the reported evidence and ensure that each study meaningfully supports the objectives of the review. Similar appraisal strategies are commonly employed in systematic reviews of AI-based question generation and educational assessment to balance rigor with feasibility when the number of included studies is limited [2, 6, 7].

A quality score (1–5) was assigned based on the following criteria: (1) methodological clarity, (2) evaluation rigor, (3) transparency of reporting (e.g., datasets, prompts, metrics, and limitations), (4) usefulness for answering RQ1–RQ3, and (5) overall robustness and reproducibility of the reported evidence. In parallel, each study was labeled for relevance (High/Medium/Low) depending on how directly it addressed AI/LLM-based quiz or assessment systems for adaptive learning in higher education. This combined quality–relevance assessment approach aligns with practices reported in prior review and survey studies in this domain [6, 7].

Across the nine included studies, the quality-score distribution was as follows: score 1: 1 study; score 3: 3 studies; score 4: 1 study; score 5: 4 studies (average quality score: 3.78/5). Relevance distribution was: High: 5 studies; Medium: 3 studies; Low: 1 study. Lower-quality or lower-relevance studies were retained but interpreted cautiously, while higher-confidence studies were prioritized during thematic synthesis, consistent with standard SLR interpretation practices [2, 7].

Table 2. Quality and relevance scoring of included studies.

Study ID	Quality (1–5)	Relevance (H/M/L)
S01	1	L
S02	4	H
S03	3	M
S04	5	H
S05	3	M
S06	5	H
S07	5	H
S08	5	H
S09	3	M

2.10 Data Extraction

A standardized data-extraction form (Excel) was used to record one row per included study (S01–S09). The extraction form was designed to directly support answering RQ1–RQ3 by capturing (i) bibliographic metadata, (ii) the AI/LLM technique and task focus, and (iii) the reported evaluation evidence, outcomes, and limitations. The use of structured extraction forms is a common practice in systematic literature reviews to ensure consistency, traceability, and comparability across included studies [2, 6, 7].

For each included study, the following fields were extracted and coded, following data-coding strategies reported in prior SLRs of AI-based question generation and educational assessment systems [6, 7]:

- **Bibliographic information:** study ID, title, venue/source, year.
- **Study type and context:** conference/journal/review; target education context (higher education / transferable setting).
- **Primary task focus:** question/item generation, automated scoring/judging, feedback generation, or conversational assessment.
- **Technique/model category:** e.g., prompt-based LLM generation, taxonomy-guided generation (Bloom's), retrieval-augmented generation (RAG), LLM-as-judge/rubric-based scoring, or survey/SLR (no primary model).
- **Dataset/materials:** course content, lecture notes, repositories, benchmark datasets, or *Not specified*.
- **Evaluation design and metrics:** human ratings (relevance/clarity/difficulty), agreement with human graders, learning outcomes (pre/post tests), engagement metrics, or qualitative evidence.
- **Key findings and limitations:** main outcomes plus reported weaknesses or risks relevant to reliability, fairness, academic integrity, and deployment.
- **Mapping to RQs:** which extracted evidence supports RQ1–RQ3.

To keep the paper concise under IEEE page limits, Table 3 provides a compact three-column overview of the included studies (title, venue/year, and primary technical focus). Table 4 aggregates the main technique and model categories across studies and summarizes dataset and evaluation characteristics. This form of table-level abstraction is commonly used in SLRs to balance completeness with readability [2, 7]. The complete extraction sheet, including all coded fields and per-study details, is provided as supplementary material.

When a study did not explicitly report a required field (e.g., dataset details or evaluation procedure), the value was recorded as *Not specified* to avoid over-interpretation and to maintain an evidence-grounded synthesis, consistent with recommended SLR reporting practices [6].

Table 3. Compact 3-column summary of included studies. Full extraction tables are provided as supplementary material.

ID	Title (Venue, Year)	Tech / Focus
S01	VocQuiz: An LLM-based Vocabulary Quiz System for English … (ACM, 2025)	LLM quiz generation
S02	A Bloom's Taxonomy Aligned Question Generation Template in … (ACM, 2025)	LLM QG + Bloom alignment
S03	Domain-Adaptive Dialogue as Quiz: Towards Conversational … (ACL Anthology, 2025)	Conversational assessment (LLM)
S04	The Who, When, Why? A Survey of ChatGPT and Generative AI … (ACM, 2025)	GenAI in education (survey)
S05	An Attention Head Is All You Need: Large-Scale Analysis of … (ACM, 2024)	Learning analytics + assessment signals
S06	Examining the Effect of Time for Assessment in … (ACM, 2024)	Assessment design + performance impact
S07	A Review on Smart Examination System using Generative AI … (JSR, 2025)	GenAI exam systems (review)
S08	A Systematic Review of Automatic Question Generation for … (IJAIED, 2020)	Question generation (SLR)
S09	Scaling the doer effect: A replication analysis using AI-generated questions (L@S, 2025)	AI-generated formative practice

3. RESULTS

This section presents the results of the systematic literature review based on the final set of **9 included studies** that met all inclusion criteria. The findings are organized by the *primary technical focus* (e.g., LLM-based question generation, conversational assessment, assessment analytics/design, and survey/review evidence) to provide a clear overview of what has been studied and how the evidence is distributed across the AI-based quiz and assessment pipeline.

3.1 Overview of Included Studies

A total of **9 studies** were included in the qualitative synthesis after PRISMA-guided identification, screening, and eligibility assessment. The included studies span the years **2020–2025** and represent a mix of conference and journal publications, as well as secondary evidence in the form of surveys, reviews, and systematic literature reviews [2, 6, 7]. Table 3 summarizes the included studies, while Table 4 consolidates the main techniques/models and evaluation notes reported across the evidence base.

The distribution of primary study focus across the included studies is as follows:

- **LLM-based question/item generation and AI-generated practice: 3 studies** (S01, S02, S09)
- **Conversational assessment (dialogue-based quiz/assessment): 1 study** (S03)
- **Assessment analytics and assessment design/performance impact: 2 studies** (S05, S06)
- **Survey/review/SLR synthesis evidence (secondary studies): 3 studies** (S04, S07, S08)

Overall, the included literature shows a strong emphasis on **LLM-supported question generation and practice content creation**, while a smaller portion of studies directly evaluate **assessment validity, scoring reliability, and learning impact** through empirical or classroom-oriented designs. This imbalance reinforces the need for stronger and more standardized evaluation protocols to support trustworthy real-world adoption.

3.2 Technique Categories Identified (RQ1)

Fig. 2 shows the distribution of primary technique categories across the included studies. The evidence base is dominated by **LLM-based question/item generation and AI-generated practice** (S01, S02, S09) and **secondary synthesis papers** (S04, S07, S08), while fewer studies directly address **conversational assessment** (S03) or **assessment analytics/design and performance impact** (S05, S06).

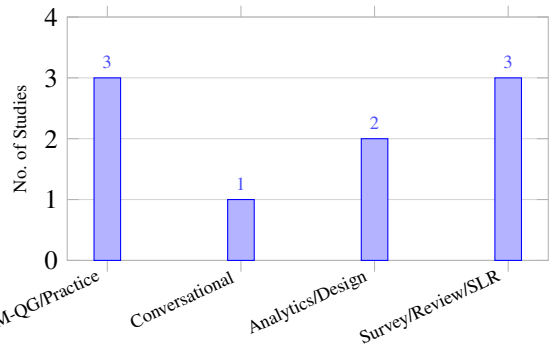

Figure 2. Distribution of primary technique categories across the included studies.

1) LLM-Based Question/Item Generation and AI-Generated Practice: Three studies (S01, S02, S09) focus on generating quiz items (e.g., MCQ or short-answer) and practice content using prompt-based LLM pipelines, sometimes supported by structured templates or retrieval of course content [3, 4, 1]. Commonly used design elements include:

- Prompt-based generation with constraints (format, difficulty, learning objective) [3, 4]
- Taxonomy-guided generation (e.g., Bloom's level control) for educational alignment (S02) [4]
- Retrieval-Augmented Generation (RAG) grounded in lecture notes/slides/course text to reduce hallucination (S01, S09) [3, 1]
- Automatic sanity checks (e.g., duplication/similarity, leakage, surface-form quality) alongside human review [3, 2]

Evaluation in this category is most often based on human ratings of *relevance, clarity, difficulty, and educational usefulness*, with some studies additionally reporting comparisons to baseline (non-LLM or template-only) generation [3, 4, 2, 6].

2) Conversational Assessment (Dialogue-Based Quiz/Assessment): One study (S03) employs an LLM-driven dialogue flow in which assessment occurs through interactive questioning rather than static quiz items [5]. In this approach, learner understanding is probed through a sequence of adaptive questions and follow-up prompts, enabling the system to refine its assessment based on prior responses.

Table 4. Compact summary of techniques/models, study IDs, and datasets/evaluation notes. Full extraction details are provided as supplementary material.

Technique / Model	Used in Study ID(s)	Dataset / Evaluation Notes
Prompt-based LLM question generation (MCQ/short-answer)	S01, S02, S09	Typically evaluated with human ratings (relevance, clarity, difficulty), plus automatic checks (grammar, similarity, leakage).
Taxonomy-guided generation (e.g., Bloom's levels, difficulty control)	S02	Often assessed via label agreement (human vs model), distribution across levels, and educational validity checks.
Retrieval-Augmented Generation (RAG) from course materials	S01, S09	Uses lecture notes/slides/text passages as grounding; evaluation includes factuality/groundedness and coverage of source content.
Conversational assessment (dialogue-based quiz/diagnosis) using LLMs	S03	Evaluated with dialogue quality + assessment validity (rubric adherence, consistency across prompts, robustness).
Automated scoring / judging (rubric-based + LLM-as-judge)	S03, S06	Evaluated via agreement with human graders (e.g., accuracy, correlation, κ), stability across prompts, and bias/fairness checks.
Feedback generation (formative explanations, hints, remediation)	S01, S09	Measured by usefulness/clarity ratings, learning gains in small studies, or expert review.
Survey / SLR synthesis papers (no primary model)	S04, S07, S08	Evidence base is the reviewed literature (N/A for experimental dataset). Summary of scope, databases, and paper counts.
Empirical classroom/user-study evaluation designs (learning outcomes)	S06, S09	Measured via pre/post tests, engagement metrics, completion rates, or performance deltas.

Conversational assessment systems typically include:

- Multi-turn question sequencing and follow-ups to support adaptive probing of learner knowledge [5]
- Rubric-guided judging of student responses using explicit assessment criteria [5]
- Consistency controls, such as prompting constraints, calibration strategies, or repeated evaluation passes, to reduce instability in model judgments [5, 2]

Reported evaluation emphasis in this category centers on *assessment validity* (i.e., whether the dialogue measures the intended skill), *rubric adherence*, and the *stability* of judgments across prompts and response styles, reflecting concerns about reliability and bias in LLM-based assessment [5, 2, 6].

3) Assessment Analytics and Assessment Design/Performance Impact: Two studies (S05, S06) focus on assessment analytics or assessment design choices rather than on question generation alone [9, 1]. These works analyze how assessment structure and delivery influence learner performance and engagement.

Typical approaches include:

- Modeling assessment signals or behavioral traces to identify performance patterns and assessment-related indicators (S05) [9]
- Studying assessment design factors, such as timing constraints, assessment format, or delivery conditions, and their effect on learning outcomes and engagement (S06) [1]

These studies contribute evidence relevant to **RQ2** by linking assessment configuration to measured outcomes (e.g., performance gains or engagement effects). However, they do not necessarily propose a complete quiz-generation pipeline, instead emphasizing evaluation design, analytics, and the conditions under which AI-supported assessment is most effective [9, 1].

4) Survey/Review/SLR Evidence (Secondary Studies): Three studies (S04, S07, S08) synthesize prior work rather than introducing and benchmarking a new model [2, 6, 7]. These secondary studies provide structured overviews of AI-based quiz and assessment research and are useful for:

- Mapping the research landscape, including task formulations, application domains, risks, and opportunities [2, 6]
- Summarizing common evaluation practices and recurring limitations reported across primary studies [2, 7]
- Highlighting gaps such as the lack of shared benchmarks, inconsistent rubrics, and limited comparability across evaluation settings [6, 7]

Because these are secondary studies, their "datasets" consist of the reviewed literature rather than a single experimental benchmark. Consequently, the strength of evidence depends on the scope, selection criteria, and rigor of the synthesis process employed by each review [2, 6].

3.3 Summary of Results

Overall, the results demonstrate that:

- **LLM-based question/item generation is the most common technical focus** in the included primary studies, typically implemented as prompt-based pipelines and, in some cases, grounded with course materials via retrieval mechanisms (S01, S02, S09) [3, 4, 1].
- **Evaluation practices are inconsistent and often qualitative**, with many studies relying on human ratings (e.g., relevance, clarity, difficulty, usefulness) and limited use of standardized benchmarks or directly comparable metrics across papers [2, 6].
- **Direct evidence on scoring reliability and assessment validity is relatively limited** in the included set; only a small subset of studies explicitly emphasize judging consistency, rubric adherence, or agreement with human graders, most notably in conversational assessment and assessment design-focused work [5, 9].
- **Secondary evidence (survey/review/SLR papers) forms a substantial portion of the evidence base**, indicating that the field is still consolidating methods, risks, and evaluation standards rather than converging on shared experimental protocols (S04, S07, S08) [2, 6, 7].
- **Recurring gaps reported across studies include** the reliability of automated judging, lack of shared benchmarks and rubrics, domain transfer challenges, and risks

to fairness and academic integrity, all of which collectively limit confident real-world deployment of AI-based assessment systems [2, 6, 7].

These findings provide a clear foundation for the comparative discussion in the next section, particularly with respect to how technique choice relates to evaluation rigor and the overall strength of evidence supporting adaptive learning outcomes.

4. ANALYSIS AND DISCUSSION

This section interprets the evidence synthesized in Section III and provides analytical answers to the research questions. Unlike the Results section, which reports what was found (e.g., technique distribution in Fig. 2), this discussion examines *why* the literature exhibits these patterns, what strengths and limitations emerge across approaches, and what gaps prevent reliable real-world adoption in adaptive learning contexts [2, 6, 7].

4.1 RQ1: AI/NLP/LLM Techniques Used for AI-Based Quiz and Assessment Systems

The reviewed studies collectively indicate that current AI-based quiz and assessment systems are typically constructed as multi-stage pipelines that combine content generation, pedagogical alignment, grounding in source materials, and evaluation or judging components. This pipeline-oriented design is evident across the included primary studies and is consistent with synthesis findings reported in recent review papers [2, 6]. Based on Fig. 2 and Tables 3–4, four recurring patterns stand out.

1) Prompt-based LLM question/item generation dominates. A substantial portion of the primary evidence focuses on using LLMs to generate quiz items (MCQ or short-answer), explanations, and practice questions [3, 4, 1]. This dominance is expected because LLM-based generation significantly reduces authoring time and enables rapid creation of diverse question variants. However, prompt-only generation is also particularly exposed to ambiguity, hallucination, and uncontrolled difficulty drift, which has motivated many systems to introduce additional constraints or post-generation checks rather than relying on a single prompt [2, 6].

2) Alignment mechanisms are increasingly used to control pedagogical quality. Several approaches incorporate explicit alignment mechanisms, such as structured templates, format and length constraints, and taxonomy-driven controls (e.g., Bloom's cognitive levels), to improve educational validity and alignment with learning objectives [4, 3]. This trend suggests an emerging consensus that raw generation quality is insufficient unless systems can reliably target intended learning outcomes and difficulty levels, a concern repeatedly emphasized in prior review studies [2, 7].

3) Grounding via course materials (e.g., retrieval augmentation) is used to reduce factual errors. When quiz items are generated from lecture notes, slides, or course resources, retrieval-augmented generation (RAG) is commonly employed to improve groundedness and content coverage [3, 1]. This strategy is particularly important in higher education contexts, where domain knowledge is precise and hallucinated or unsupported content can directly undermine assessment validity and learner trust [6, 7].

4) Assessment functionality extends beyond generation into judging and interaction. A smaller but methodologically important subset of studies extends beyond question generation to address conversational assessment via dialogue and automated scoring or judging using rubrics or model-based evaluation [5, 9]. These approaches shift the focus from merely "making questions" to "measuring understanding," which more closely aligns with the core requirements of adaptive assessment. However, prior studies and reviews highlight that these techniques carry higher risk when scoring is unstable, prompt-sensitive, or overly influenced by surface-level linguistic features rather than conceptual understanding [5, 2, 6].

Overall, the techniques used in the included studies suggest that the field is gradually transitioning from isolated generation components toward more integrated, end-to-end assessment pipelines. Nevertheless, maturity across pipeline stages remains uneven: content generation is relatively well explored, while robust, reliable, and transparent scoring and evaluation mechanisms remain less developed and less consistently validated [2, 7].

4.2 RQ2: Reported Learning, Scoring, and Engagement Outcomes

Across the evidence base, reported outcomes rely on heterogeneous evaluation designs, which makes cross-study comparison difficult. This diversity in outcome definitions, metrics, and study designs has been repeatedly highlighted as a challenge in recent reviews of AI-based quiz and assessment systems [2, 6, 7]. The most common outcome types observed in the included studies can be grouped into four categories.

1) Item quality and educational validity outcomes. Many studies evaluate generated questions using human ratings, such as relevance to source content, clarity, difficulty, and perceived educational usefulness, often supplemented by expert review [3, 4, 1]. These outcomes are useful for establishing early-stage feasibility and surface-level quality. However, prior reviews emphasize that such measures alone do not demonstrate whether generated items support valid measurement of learning or fair grading decisions [2, 6].

2) Scoring reliability and agreement outcomes. When automated scoring or LLM-based judging is employed, some studies report agreement with human graders using metrics such as accuracy, correlation, or rubric consistency [5, 9]. Nevertheless, the included evidence suggests that systematic reporting of reliability and robustness is still uncommon. As a result, evidence supporting trustworthy automated grading remains considerably weaker than evidence supporting question or content generation quality, a gap also noted in multiple survey and review papers [2, 7].

3) Learning and performance outcomes. Only a small subset of studies reports direct learning-related outcomes, such as performance differences under alternative assessment designs or the effect of AI-generated practice on learner performance [1, 9]. While these results are promising, they are often context-specific, depending on course design, learner population, and study duration. Replication across different educational settings remains limited, which constrains the generalizability of reported learning gains [6, 7].

4) Usability and engagement outcomes. Some studies in-

clude user-centered outcomes such as usability ratings, perceived helpfulness, or engagement indicators derived from surveys or interaction logs [3, 1]. These measures provide useful signals about feasibility and learner acceptance. However, prior work cautions that positive usability or satisfaction outcomes do not necessarily imply strong measurement validity or reliable assessment outcomes [2, 6].

In summary, the reviewed literature reports many positive signals regarding feasibility, usability, and content generation quality. However, the strongest forms of evidence required for confident real-world adoption in higher education—namely *assessment validity, scoring reliability, and standardized benchmarking*—remain comparatively sparse and inconsistently reported across studies [2, 6, 7].

4.3 RQ3: Challenges, Limitations, and Research Gaps

The included studies consistently report a set of recurring technical and practical limitations that constrain the reliable deployment of AI-based quiz and assessment systems in higher education [2, 6, 7].

1) Validity and reliability of automated judging. A central challenge is ensuring that automated scoring reflects true conceptual understanding rather than surface-level features such as response length, fluency, or phrasing. Several studies note that LLM-as-judge approaches can be sensitive to prompt formulation and may assign inconsistent grades to semantically equivalent answers, which undermines trust in automated assessment [5, 9]. Prior reviews similarly highlight instability and lack of calibration as major barriers to high-stakes adoption [2, 7].

2) Hallucination, ambiguity, and misalignment in generated items. LLM-generated questions may contain factual inaccuracies, underspecified prompts, or ambiguous correct answers, particularly when generation is not grounded in authoritative course materials [3, 4]. Misalignment with intended learning objectives can occur when models optimize for linguistic plausibility rather than pedagogical intent, a limitation repeatedly emphasized in survey and review studies [6, 7].

3) Lack of shared benchmarks and comparable evaluation protocols. Across the literature, studies employ heterogeneous datasets, course materials, rubrics, and rating scales, making meaningful comparison difficult. The absence of shared benchmarks and standardized evaluation protocols limits the ability to distinguish genuine methodological improvements from artifacts of differing experimental setups [2, 6]. This fragmentation has been identified as a key obstacle to cumulative progress in AI-based educational assessment research.

4) Domain transfer and context sensitivity. Approaches that perform well in one subject area often fail to generalize to others due to differences in terminology, acceptable reasoning processes, and assessment formats. Several studies and reviews report that domain adaptation and content grounding remain open challenges, particularly for higher education contexts requiring precise disciplinary knowledge [3, 6, 7].

5) Academic integrity, fairness, and privacy risks. AI-based quiz systems operate in the same environment as AI-based answer assistants, creating new risks such as information leakage, shortcut learning, and proxy completion [1]. Fairness concerns also arise when automated grading systems exhibit sensitivity to writing style, language proficiency, or response structure rather than conceptual correctness. Additionally, privacy risks emerge when student responses or proprietary course materials are processed by external AI services, a concern highlighted in multiple review studies [2, 6].

6) Persistent need for human-in-the-loop workflows. Despite advances in generation and scoring, many systems still require instructor oversight for quality assurance, rubric validation, and question filtering. This indicates that fully automated assessment remains unrealistic for high-stakes educational settings in the near term. Hybrid human–AI workflows are therefore widely viewed as a more credible and responsible deployment model [5, 7].

Overall, the most significant gap revealed by this SLR is the imbalance between rapid progress in content generation and slower progress in rigorous, standardized evaluation of assessment validity and scoring reliability. Future research should prioritize shared benchmarks, transparent and reusable rubrics, calibration procedures for judging stability, and classroom-grounded studies that jointly measure learning outcomes, fairness, and reliability alongside usability [2, 6].

5. CONCLUSION

This paper presented a PRISMA-guided Systematic Literature Review (SLR) on AI-based quiz and assessment systems for adaptive learning in higher education. Searches across IEEE Xplore, ACM Digital Library, and Google Scholar identified 57 records; after deduplication, screening, and full-text eligibility assessment, 9 studies were included for qualitative synthesis. The review addressed three research questions focusing on (i) techniques used for AI-driven quiz and assessment systems, (ii) reported learning, scoring, and engagement outcomes, and (iii) key challenges and research gaps [2, 6, 7].

Overall, the evidence indicates that **LLM-based question and item generation is currently the dominant research focus**, most often implemented through prompt-based pipelines and increasingly supported by structured controls such as templates, taxonomy alignment, and retrieval-augmented grounding using course materials [3, 4, 1]. These techniques are attractive because they reduce instructor authoring effort and enable rapid generation of diverse practice items; however, the reviewed literature consistently shows that generation quality alone is insufficient for trustworthy assessment in higher education contexts [6, 7].

Across the included studies, **evaluation practices remain inconsistent and difficult to compare**. Many works rely primarily on human ratings of relevance, clarity, and difficulty, while fewer provide rigorous evidence on assessment validity, grading reliability, or robustness across prompts and learner response styles [2, 6]. In particular, systems incorporating automated scoring or LLM-as-judge components raise high-stakes concerns, as judgments may vary with prompting, phrasing, or writing style, potentially affecting fairness, transparency, and trust [5, 9]. Evidence on learning gains and engagement is promising in some cases but remains context-

dependent and limited in standardized, reproducible reporting [1].

The review further highlights several persistent gaps that limit confident real-world deployment, including **hallucination and ambiguity in generated items, lack of shared benchmarks and transparent rubrics, limited domain transferability, academic integrity risks, and privacy and fairness concerns** [2, 6, 7]. A recurring practical conclusion across the literature is that **human-in-the-loop workflows remain essential** for quality assurance, rubric validation, and responsible adoption, particularly in higher-stakes assessment scenarios [5, 7].

Future research should prioritize **standardized benchmarks for educational question generation and automated scoring**, transparent and reusable rubrics, systematic stability testing for automated judges, and classroom-grounded evaluations that jointly measure learning outcomes, reliability, fairness, and usability [2, 6]. Progress toward integrated frameworks that combine generation, grounding, validation, and accountable scoring will be critical for moving AI-based quiz and assessment systems from experimental prototypes toward reliable and trustworthy tools for adaptive learning in higher education.

6. LIMITATIONS OF THE REVIEW

This review has several limitations related to database coverage, reproducibility, and study availability. First, the search was restricted to IEEE Xplore, ACM Digital Library, and Google Scholar, which may have resulted in the omission of relevant studies indexed in other databases or educational research venues [6, 7]. In addition, Google Scholar result counts and rankings are known to change over time, which limits full reproducibility despite documented search strings and filters [2].

Second, only English-language full-text papers were considered, and a small number of potentially relevant reports could not be accessed, which may introduce selection bias into the evidence base [6]. Finally, the relatively small number of included studies and the heterogeneity of evaluation designs, datasets, and reported metrics limit strong generalization and prevent quantitative meta-analysis, a limitation commonly noted in prior SLRs on AI-based question generation and assessment [2, 7].

7. FUTURE WORK

Based on the identified gaps and limitations, several directions for future research emerge, many of which are consistently emphasized across recent survey and review studies:

- **Establish shared benchmarks and datasets** for educational question generation, automated scoring, and feedback quality to enable fair and reproducible comparison across methods [2, 6].
- **Conduct multi-institution and classroom-grounded evaluations** to measure real learning impact, reliability, and usability over extended deployments rather than short-term or laboratory-based studies [1].
- **Improve robustness and stability of automated judging**, including systematic prompt sensitivity analysis, calibration strategies, and stronger rubric-based evaluation frameworks [5, 9].
- **Address fairness and academic integrity risks** by evaluating bias across learner groups, monitoring grading consistency, and incorporating anti-cheating and leakage-prevention mechanisms [1, 6].
- **Strengthen privacy-preserving system designs** for handling student responses and proprietary course materials, including secure storage, access control, and minimal data retention policies [2, 7].

Advancing these directions will help move AI-based quiz and assessment systems from promising research prototypes toward trustworthy, scalable, and ethically responsible tools for adaptive learning in higher education.

8. ETHICS CONSIDERATIONS

This review follows established principles of academic integrity, transparency, and responsible research practice for systematic literature reviews [6, 7]. All search queries, database filters, and study-selection criteria were documented to support transparency and reproducibility, in line with PRISMA-guided SLR methodology [2].

The study involved no human participants, personal data collection, or experimental intervention. All analyzed sources were peer-reviewed and publicly available academic publications; therefore, no ethical approval or informed consent was required [7].

AI tools were used only for limited language-level assistance (e.g., grammar and clarity) and were not employed for study identification, screening, data extraction, analysis, or interpretation. This usage aligns with emerging best-practice recommendations for responsible and transparent use of AI tools in academic research [6]. All methodological decisions, analytical judgments, and conclusions were made by the author to ensure accountability, scholarly responsibility, and research integrity.

REFERENCES

[1] M. J. Gierl, H. Lai, and S. R. Turner, "Using automatic item generation to create multiple-choice test items," *Medical Education*, vol. 46, no. 8, pp. 757–765, 2012, doi: 10.1111/j.1365-2923.2012.04289.x.

[2] H. Lu, J. Wang, Q. Wang, and H. Li, "A review of automatic question generation for education," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 8979–9025, 2023.

[3] M. Heilman and N. A. Smith, "Good question! Statistical ranking for question generation," in *Proc. Human Language Technologies: The 2010 Annual Conf. North American Chapter of the Association for Computational Linguistics*, Los Angeles, CA, USA, 2010, pp. 609–617.

[4] B. K. Britton, A. Glynn, M. Meyer, and T. Penland, "Effects of text structure on use of cognitive capacity during reading," *Journal of Educational Psychology*, vol. 74, no. 1, pp. 51–61, 1982.

[5] A. Tack and C. Piech, "The AI teacher test: Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues," in *Proc. Neural Information Processing Systems Datasets and Benchmarks Track*, New Orleans, LA, USA, 2022.

[6] E. Kasneci, K. Sessler, S. Kuechemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Guennemann, E. Huellermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[7] O. Zawacki-Richter, V. I. Marin, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education: Where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, Art. no. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[8] S. Narciss, "Feedback strategies for interactive learning tasks," in *Handbook of Research on Educational Communications and Technology*, 3rd ed., J. M. Spector, M. D. Merrill, J. van Merrienboer, and M. P. Driscoll, Eds. New York, NY, USA: Routledge, 2008, pp. 125–144.

[9] M. D. Shermis and J. Burstein, Eds., *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY, USA: Routledge, 2013.

[10] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hrobjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

[11] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.

[12] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, Keele, U.K., EBSE Tech. Rep. EBSE-2007-01, 2007.