NATIONAL ECONOMICS UNIVERSITY
1956-2026

28 – 29 August 2026 · Hanoi

INTERNATIONAL CONFERENCE ON EDUCATIONAL TECHNOLOGY AND ARTIFICIAL INTELLIGENCE

A RAG-LLM-Based Framework for Automatic Multiple-Choice Question Generation in Reading Comprehension and Learning Outcome Analysis

Thanh Nhan Le · Thanh-Hai Le · The Anh Le

FPT University · Viet Nam

Paper 16 · Track 2 · Room B · 14:20 – 14:45, 28 August 2026

Presentation roadmap

1. **Context & problem**
   Why automate MCQ generation, and why is direct LLM generation not reliable enough?

2. **Objectives & scope**
   Two main objectives: reading-comprehension MCQ generation and learning-outcome analysis.

3. **Proposed method**
   Compare Exp1, Exp2, and Exp3; select Exp2 and adapt it to article-only input.

4. **System design**
   Architecture, databases, web examination module, and analytics module.

5. **Results & contributions**
   Automatic evaluation, human evaluation, pilot deployment, limitations, and future work.

Context: the need for scalable MCQ banks

Reading-comprehension MCQs are easy to score automatically, but high-quality item development requires substantial effort.

MCQ
A widely used format for reading assessment

RAG
Grounds question generation in passage evidence

Analytics
Turns test responses into diagnostic feedback

- Teachers need diverse, contextually correct questions with plausible distractors.
- LLMs generate content quickly, but educational questions must be grounded in the passage, not merely fluent.
- Modern assessment systems also need post-exam data on scores, time, question errors, and skills.

This study addresses both automatic question generation and the use of response data after assessment.

Research problem

Direct LLM generation can be fluent, but it does not reliably guarantee grounding, answer preservation, or pedagogical value.

Hallucination
A question or answer may sound plausible without support from the passage.

Answer alignment
The generated correct option may diverge from the target answer or may not be uniquely correct.

Distractor quality
Incorrect options should be plausible, relevant, and semantically comparable.

Evaluation gap
ROUGE measures lexical overlap but not answerability or educational quality.

Deployment gap
Many studies stop at generation and do not connect to exam delivery and analytics.

Research gap

- A deployable pipeline should generate passage-grounded, structurally stable, and controllable MCQs.
- Evaluation should combine format compliance, ROUGE, text informativity, answer alignment, and human judgment.
- Practical feasibility should be demonstrated through teacher review, exam delivery, response logging, and analytics.

Research objectives and scope

The study focuses on English reading comprehension and demonstrates practical system integration.

1 Objective 1: Automated MCQ Generation

Design a passage-to-MCQ pipeline using Qwen2.5-1.5B-Instruct, TF-IDF keywords, salient n-grams, and evidence retrieval.

2 Objective 2: Learning Outcome Analysis

Analyze outcomes at student, question, skill, and exploratory K-means group levels to support teacher decisions.

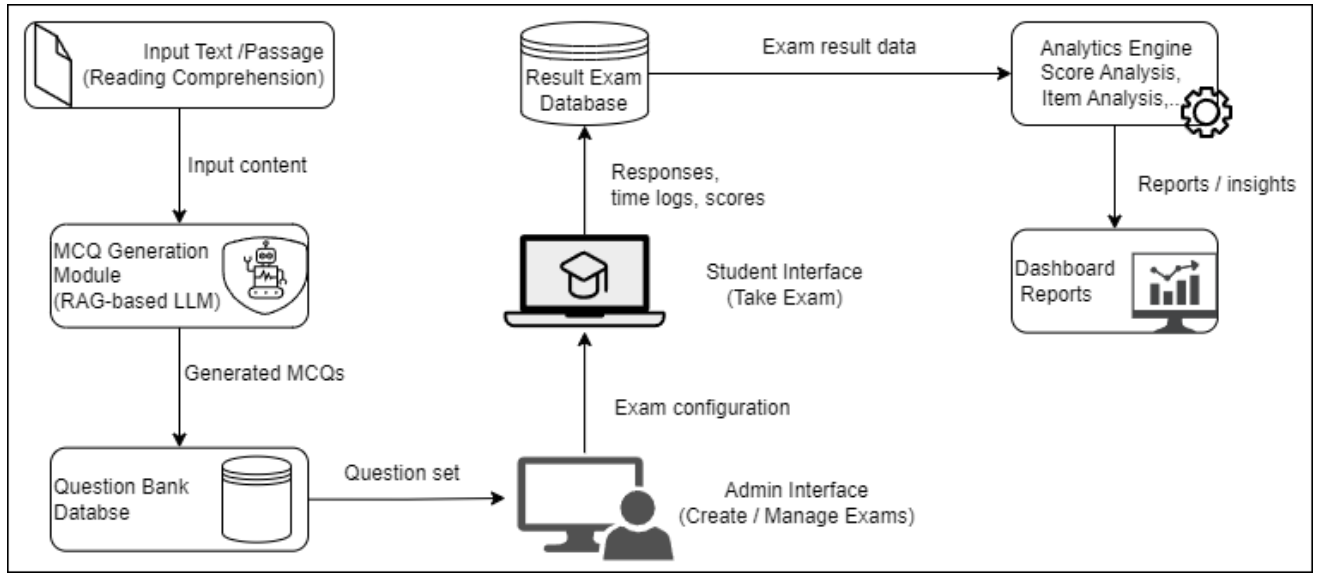
3 Deployment Objective

Adapt the selected benchmark configuration to article-only input and integrate it into a web-based examination system.

Scope: RACE dataset, English reading comprehension.

Overall system architecture

The system connects three layers: question generation, exam delivery, and learning-outcome analysis.



Key distinction: the workflow closes the feedback loop from response logs back to teachers.

1 Input & Generation

Teachers enter a passage or upload CSV data; the pipeline generates an MCQ with four options, a correct answer, and a skill tag.

2 Teacher Review

Teachers inspect, edit, and approve questions before they are stored in the question bank.

3 Exam Delivery

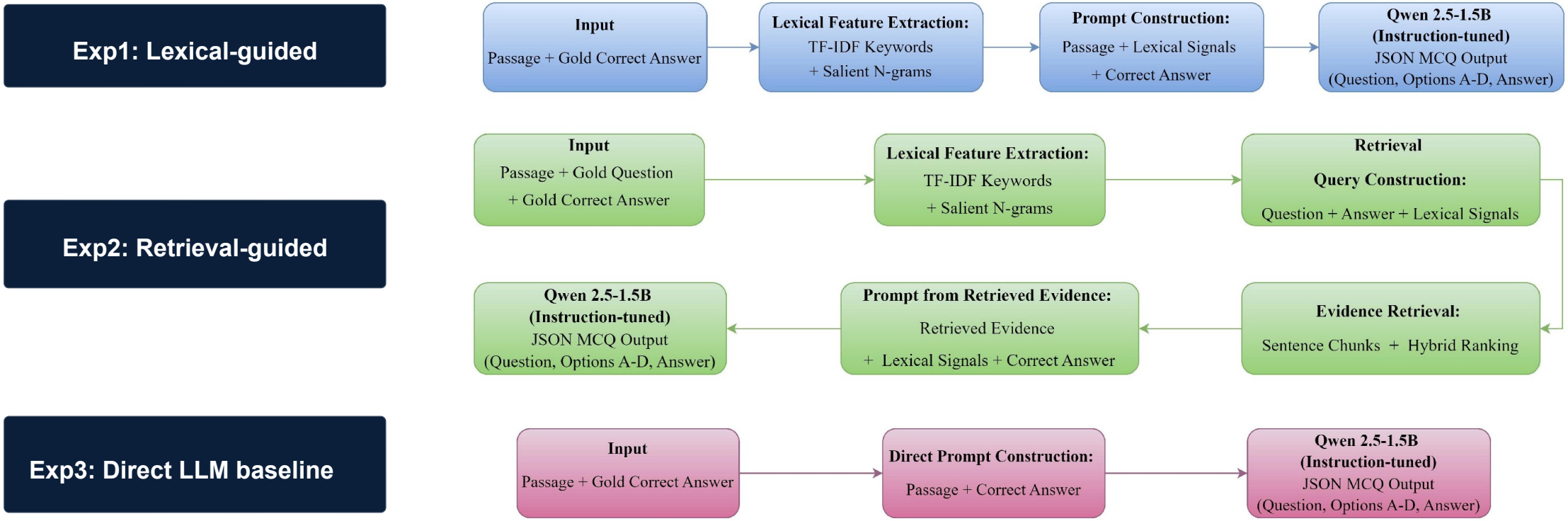
Students take timed web-based exams; the system scores responses and stores attempts and answers.

4 Analytics

The dashboard analyzes scores, question difficulty, skill accuracy, distractor behavior, and exploratory K-means groups.

Three MCQ generation settings are compared

The same base model is used; only lexical guidance and retrieval differ.



- Exp1 tests the contribution of TF-IDF keywords and salient n-grams.
- Exp2 adds evidence retrieval to focus generation on relevant chunks.
- Exp3 provides a minimal direct-prompting reference.

Dataset and experimental protocol

Controlled experiments on 2,000 instances from the RACE validation split.

JSON


```
{
  "answer": "A",
  "article": "\"Schoolgirls have been wearing such short skirts at Paget High School...",
  "example_id": "high132.txt",
  "options": [
    "short skirts give people the impression of sexualisation",
    "short skirts are too expensive for parents to afford",
    "the headmaster doesn't like girls wearing short skirts",
    "the girls wearing short skirts will be at the risk of being laughed at"
  ],
  "question": "The girls at Paget High School are not allowed to wear skirts in that _ ."
}
```


Each RACE sample contains an article, a question, four options A-D, and one answer label, matching the reading-comprehension MCQ task.

1 Base model

Qwen2.5-1.5B-Instruct; the same model is used for Exp1, Exp2, and Exp3.

2 Sampling

2,000 sampled validation instances

3 Generation

Each instance produces one MCQ per setting; output must be JSON with a question, four options, and one answer label.

4 Aggregation

Report mean format compliance, ROUGE-1/L F1, text informativity, and answer alignment for each setting.

Evaluation criteria for generated MCQs

The study combines structural validity, lexical similarity, grounding, and answer preservation.

1 Validity

Proportion of outputs containing one non-empty question, four distinct options A-D, and one valid answer label.

2 ROUGE-1/L F1

Surface similarity to the reference question; useful, but not decisive for deployment.

3 Text Informativity

TF-IDF cosine similarity between the generated question and its generation context; interpreted as a grounding proxy.

4 Answer Support

TF-IDF cosine similarity between the generated correct option and the dataset target answer.

5 Human Evaluation

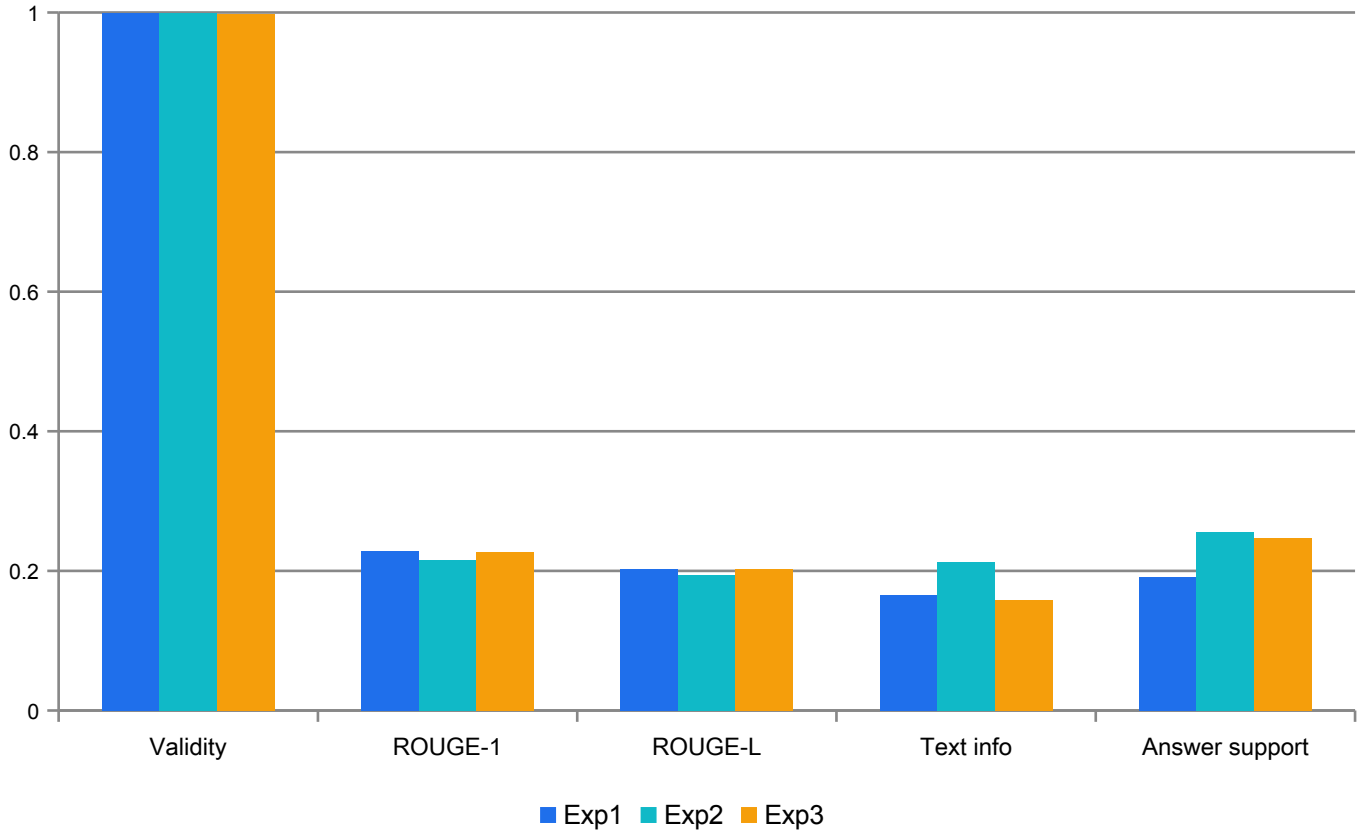
Expert ratings of clarity, correctness, answerability, distractor quality, and pedagogical usefulness.

Deployment quality prioritizes grounding, answerability, and controllability over reference-style wording.

Human evaluation validates the selected Exp2 pipeline; it is not a three-setting comparison.

Automatic results: Exp2 is strongest on grounding and answer alignment

Exp1 and Exp3 are slightly closer to reference wording, while Exp2 offers the better deployment trade-off.



1 Format compliance is nearly identical
Exp1 = 0.999; Exp2 = 0.999; Exp3 = 0.998. Structural differences are negligible.

2 Reference similarity: Exp1 / Exp3
Exp1 and Exp3 obtain slightly higher ROUGE scores, reflecting surface similarity.

3 Highest grounding proxy: Exp2
Text informativity for Exp2 = 0.212016, the highest among the three settings.

4 Highest answer alignment: Exp2
Exp2 = 0.255796, indicating the strongest preservation of the intended answer.

Why select Exp2 for deployment?

The deployment criterion is not reference similarity alone, but whether an item is grounded and usable in a real assessment.

1. Grounded in the passage
Questions should be answerable from explicit evidence, reducing hallucination.

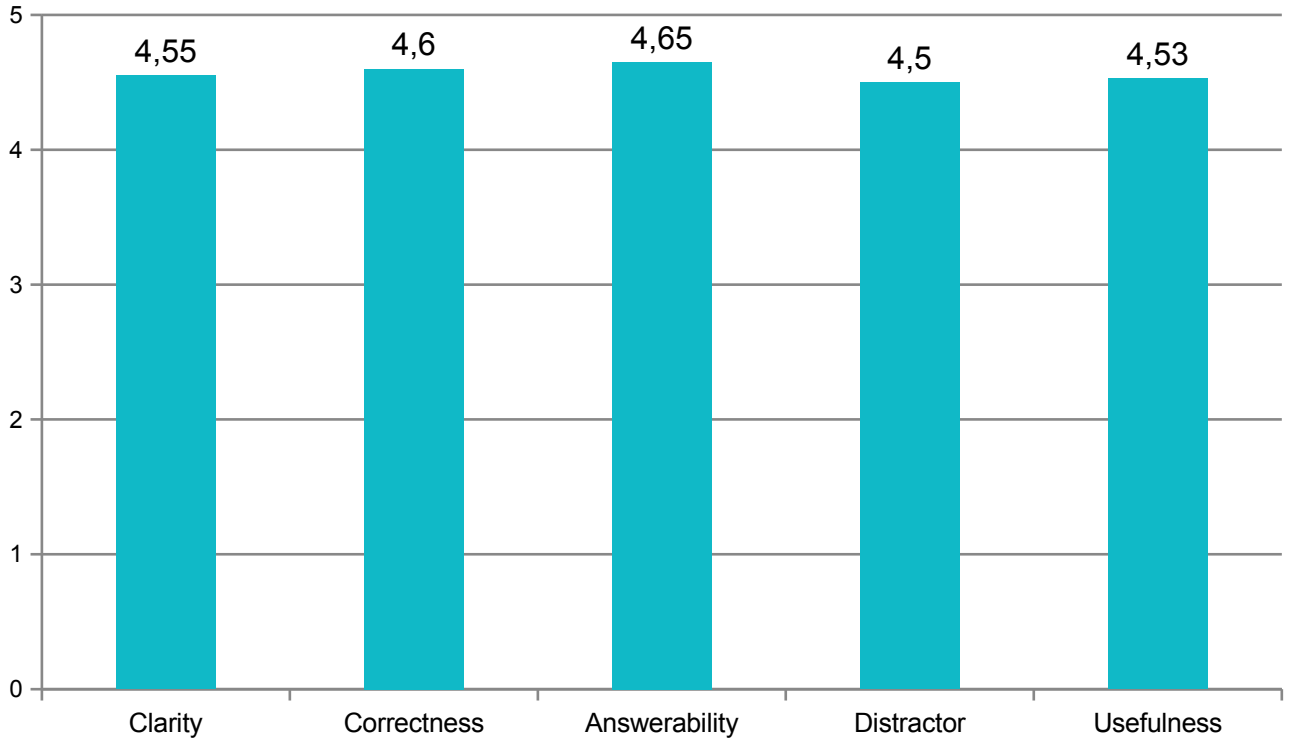
2. Answer aligned
The correct option should preserve the intended answer and remain clearly unique.

3. Controllable & explainable
The workflow exposes evidence chunks, lexical cues, prompt constraints, and validation.

- Exp1 is structurally reliable but weaker on grounding and answer alignment.
- Exp3 is slightly closer to reference wording but lacks retrieval support.
- Exp2 provides the best trade-off among structure, grounding, answerability, and controllability.
- For educational deployment, passage support matters more than lexical imitation.

Expert evaluation of Exp2-generated MCQs

Two educational experts rated 20 valid MCQs on five pedagogical criteria.



4.565/5

Overall mean score

91.3%

Percentage equivalent

1 Evaluator profile

High school teacher + university English lecturer.

2 Highest criterion

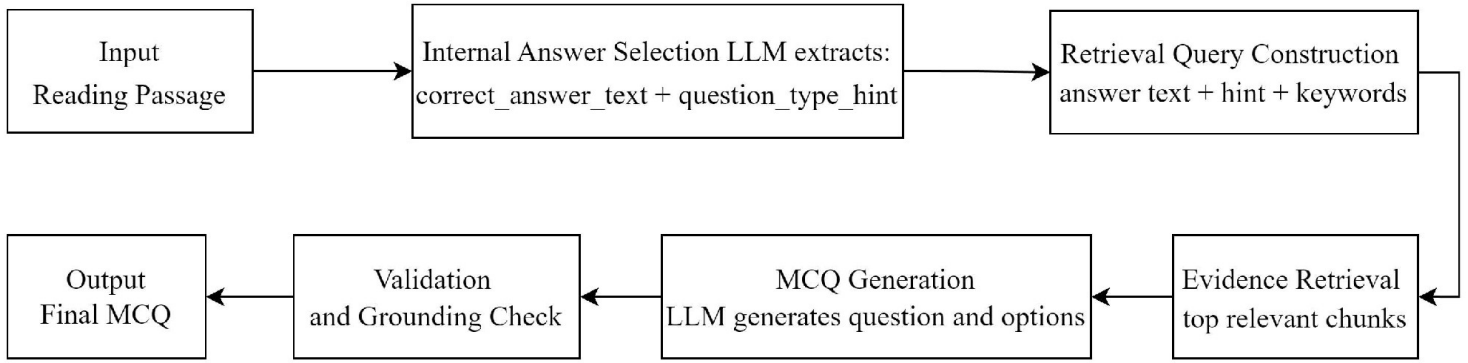
Answerability = 4.65, indicating that questions were generally answerable from the passage evidence.

3 Improvement area

Distractor quality was the lowest criterion, but remained high at 4.50; this is the main improvement area.

Adapting Exp2 to article-only deployment

Benchmark Exp2 uses a reference question and gold answer; real deployment receives only a passage.



- In the benchmark, Exp2 uses the reference question and gold answer to construct the retrieval query.
- In deployment, teachers provide only a reading passage, with no gold annotation.
- The system first infers an answer candidate and a lightweight question-type hint.
- Retrieval-guided generation then preserves the evidence-grounding advantage of Exp2.

1 Input passage

Article-only input provided by a teacher or CSV upload.

2 Infer answer candidate

Extract a concise answer candidate that is explicitly supported by the passage.

3 Infer question-type hint

Examples: detail, inference, main idea, or purpose.

4 Retrieve evidence & generate

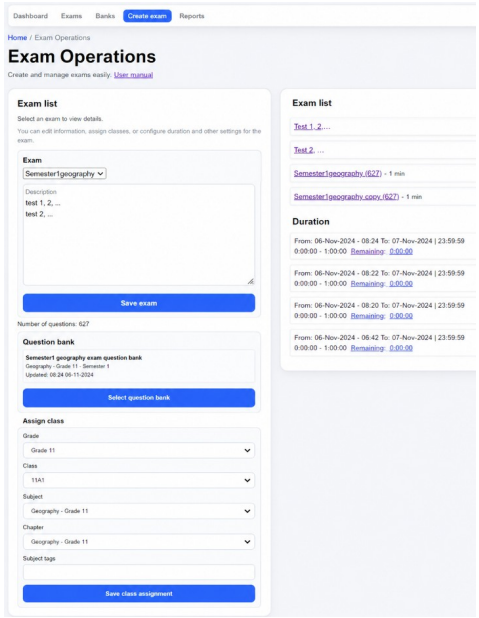
Generate an MCQ using retrieved evidence, lexical cues, and output constraints.

5 Teacher review

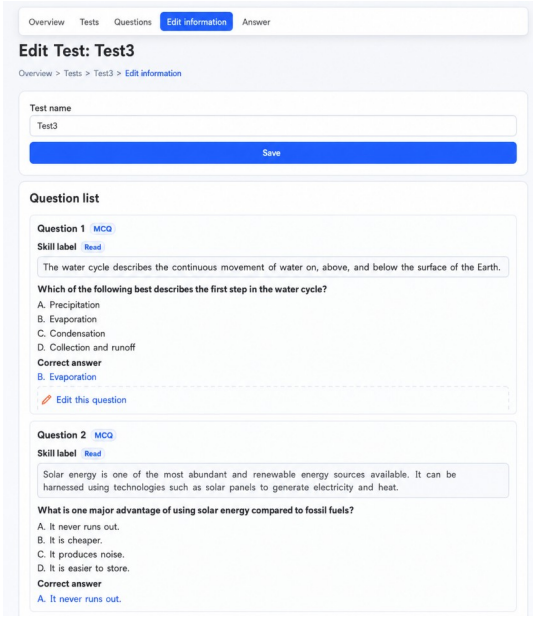
Teachers review and edit items before including them in an exam.

Web-based examination module

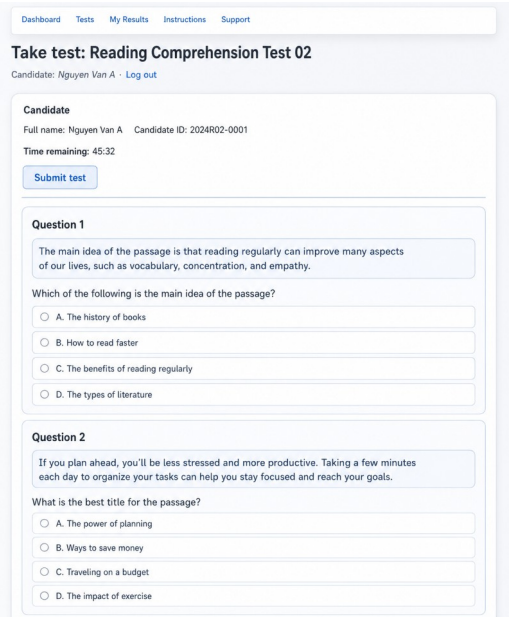
A human-in-the-loop workflow: teachers review first, students take the exam, and analytics follows.



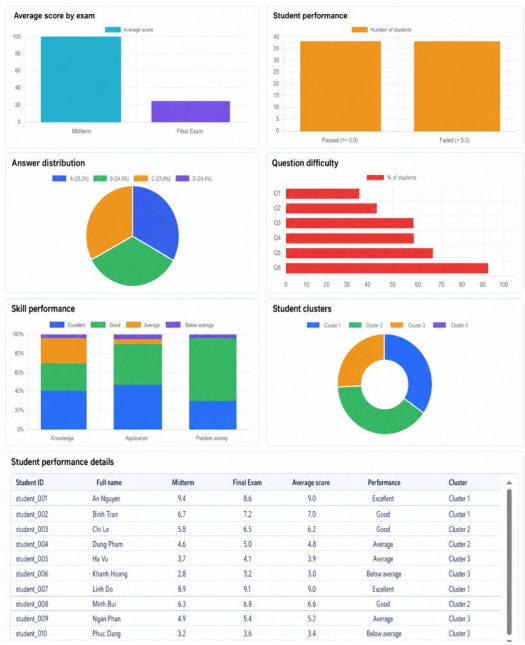
Teacher creates exam



Review generated MCQs



Student test taking



Teacher analytics

- Teacher interface: enter or upload a passage, create an exam, assign a class, and configure timing.
- Review & editing: verify the passage, stem, options, correct answer, and skill label.
- Student interface: take a timed exam; the system stores answers, scores, and completion time.
- Analytics dashboard: summarize results by exam, question, skill, and exploratory student group.

Learning Outcome Analytics Module

Transform raw response logs into indicators that teachers can use.

1 Raw Logs

attempts, attempt_answers, score, time, selected answers

2 Metrics

accuracy, score, difficulty, distractor ratio, skill error ratio

3 Insights

student support, item review, skill remediation, group intervention

---

A Student-level

Accuracy, final score, completion time.

B Question-level

Observed difficulty Pj and distractor selection ratios.

C Skill-level

Detail vs inference accuracy and error ratio.

D K-means grouping

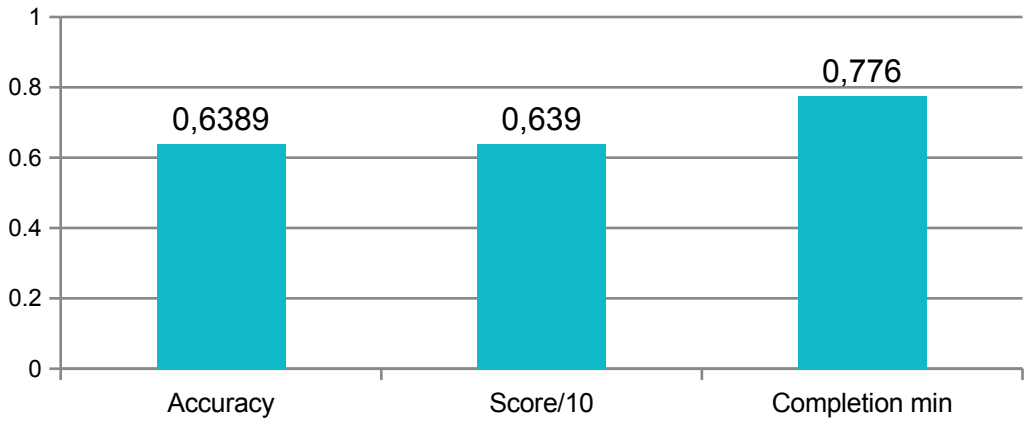
Support, Average, Excellent groups from standardized features.

Goal: help teachers identify students needing support, difficult questions, weak skills, and groups requiring different instructional strategies.

Pilot deployment results

30 students, 1 exam, 6 questions, 180 question-level response records.

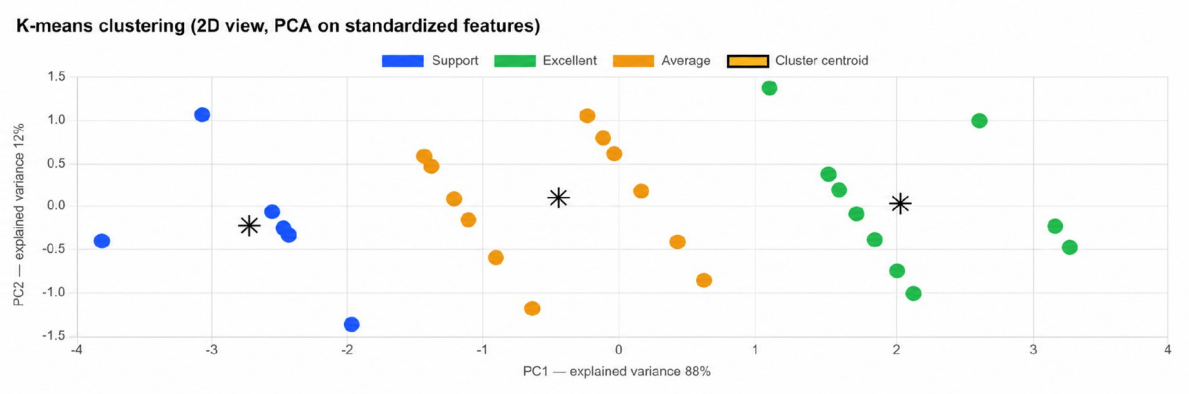
30	180	0.6389	6.39/10	1396s
submitted attempts	question responses	average accuracy	average score	average completion time

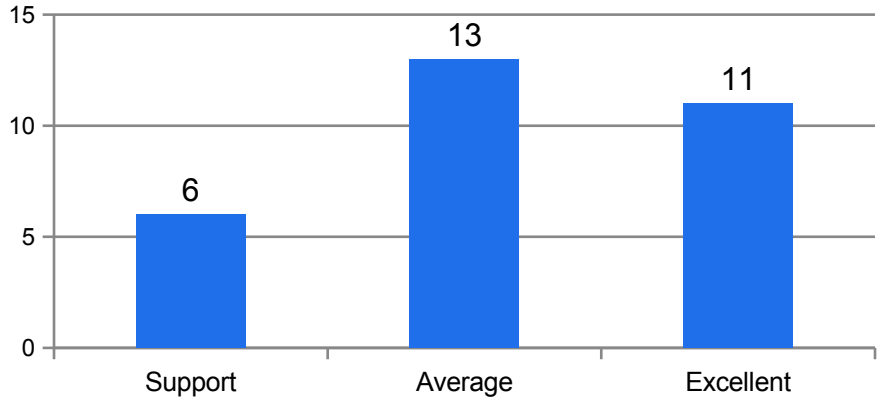


- The average score indicates a usable difficulty level for question- and skill-level analysis.
- Response logs demonstrate that the pipeline operates from exam delivery through analytics reporting.
- The pilot shows operational feasibility, not large-scale learning effectiveness.

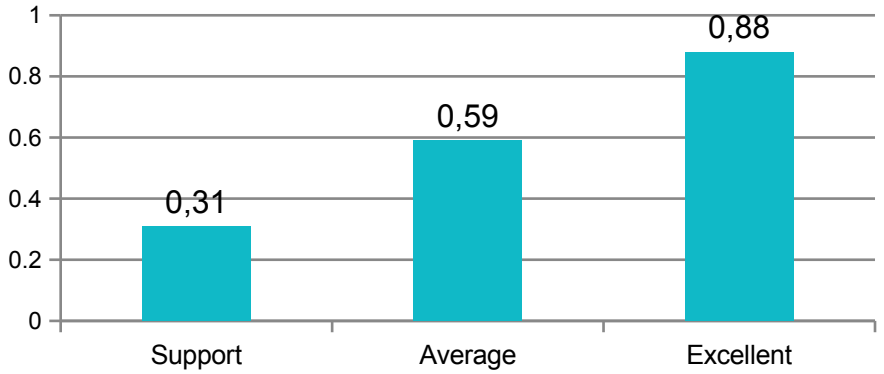
K-means student grouping

Group students by standardized performance features to support differentiated intervention.





Illustrative system output: Silhouette = 0.4386, Davies-Bouldin = 0.7339, Dunn = 0.4387. These values are not reported as primary experimental results in the paper; clustering remains exploratory.



Main contributions

The study does not propose a new model; it develops and evaluates a deployment-oriented workflow.

1 Comparative generation study

Compare lexical-guided, retrieval-guided, and direct LLM generation using the same base model.

2 Adapted article-only pipeline

Adapt Exp2 from benchmark annotations to real-world article-only input.

3 Multi-level evaluation

Combine automatic metrics, expert evaluation, and pilot deployment analytics.

4 Operational web system

Integrate teacher review, exam delivery, response logging, and an analytics dashboard.

Core contribution

A deployment-oriented RAG-LLM framework for reading-comprehension MCQ generation with human review and learning-outcome analytics.

Practical value: reduce item-writing effort, keep teachers in control, and turn responses into diagnostic feedback.

Limitations and future work

The results demonstrate feasibility, but broader evidence is needed for reliability and generalization.

Current limitations

- The scope is limited to English reading comprehension on RACE.
- Human evaluation is small: 20 MCQs rated by two experts.
- The pilot includes 30 students and six questions, which is insufficient for strong psychometric conclusions.
- The deployed article-only generator has not received direct metric-based evaluation.
- The comparison reports means without significance testing or external baselines.

Future work

- Add external baselines and broader model comparisons.
- Evaluate the article-only pipeline directly on larger samples.
- Conduct deeper distractor and psychometric analyses.
- Expand skill and difficulty modeling.
- Measure inference cost, latency, and scalability.

Conclusion

Retrieval-guided MCQ generation can be integrated with online assessment and diagnostic analytics.

1 Research answer

Exp2 is the preferred deployment configuration because it balances grounding, answer alignment, and controllability.

2 System answer

The web system supports question review, exam creation, student test-taking, response logging, and analytics.

3 Educational answer

Analytics summarizes outcomes at student, question, skill, and exploratory cluster levels.

Final message

The study demonstrates the feasibility of a teacher-controlled RAG-LLM workflow that operates in a web assessment environment and produces data-driven diagnostic feedback.

Thank you

Questions welcome

Thanh Nhan Le · nhan24mse43017@fsb.edu.vn

FPT University · Paper 16, ICETAI 2026