# A Multi Agentic Framework for Dynamic Objective Test Generation

# Abstract

Assessment plays a pivotal role in the education system by evaluating student knowledge, guiding pedagogy, and enabling effective learning outcomes. Multiple-Choice Questions (MCQs) remain the most widely used assessment format due to their scalability, objectivity, and ease of evaluation. However, manual MCQ creation is time-intensive, requires subject experts, and lacks adaptability to diverse learners. Existing AI-driven approaches, though promising, face limitations in handling multi-format documents, generating context-rich questions, and ensuring originality. This work proposes a multi-agent adaptive MCQ generation framework that leverages Retrieval-Augmented Generation (RAG), Reinforcement Learning (RL), and explainable AI techniques to produce high-quality, personalized, and multimodal assessments. The proposed system addresses challenges such as plagiarism prevention, reasoning-based distractor generation, adaptive learning, and explainability, ensuring reliability and scalability in modern educational environments.

# Introduction

In the context of modern digital education, automated assessments are becoming crucial for scalable, objective evaluation, as manual creation of multiple-choice questions (MCQs) remains resource-intensive and slow. Traditional automated item generation, rooted in psychometrics, was largely template-based creating question variants via structured item models but recent progress in neural networks and LLMs has dramatically improved scalability and content diversity [1].

Adaptive learning systems, which tailor content dynamically to each learner, have gained prominence. Empirical studies show that 86% of adaptive learning implementations yield positive learning outcomes, highlighting the effectiveness of AI-driven educational personalization [2].

More specifically, Large Language Models (LLMs) have been harnessed for MCQ generation. For example, Mucciaccia et al. (2025) built a system using advanced prompt engineering and an evaluation component to generate and validate MCQs derived from university regulations. This approach demonstrates that LLMs can significantly reduce development overhead while maintaining question integrity [3]. In the domain of medical education, a comprehensive review by Al Shuraiqi et al. (2024) explores various ML and NLP methods used to generate case-based MCQs, identifying key strengths such as scalability and new educational opportunities, while also noting gaps in reasoning and domain alignment [4].

Hybrid agent architectures are also emerging. Wang et al. (2025) proposed a multi-agent interactive question generation framework that excels in generating QA pairs for long-form documents (with complex layouts and multilingual text), reducing reliance on costly human annotation and improving long-context understanding [5]. Similarly, Besrour et al. (2025) introduced RAGentA, a trusted multi-agent Retrieval-Augmented Generation (RAG) model emphasizing faithfulness and correctness by iteratively filtering through retrieval, generation, and verification with inline citations for increased grounding [6].

These lines of work spanning adaptive education, automated generation, and multi-agent systems highlight both the strides made and the enduring limitations: limited support for multimodal documents, reasoning-poor distractor generation, insufficient adaptability to learner behavior, plagiarism risk, and hallucination from LLMs. These challenges motivate our proposed multi-agent framework, integrating document ingestion, reasoning-aware MCQ generation, adaptive learning, plagiarism safeguards, and grounded explainability.

# Literature Survey

Early attempts at automated multiple-choice question (MCQ) generation primarily focused on text-to-text transformer models. Li et al. (2021) introduced AGenT Zero, a zero-shot MCQ generation framework that combined logistic regression, T5-based paraphrasing, and FitBERT for distractor creation. While the system enabled efficient generation of a large pool of questions to reduce cheating, its performance declined on unseen datasets due to weak generalization. Building on this, Rodriguez-Torrealba et al. (2022) proposed an end-to-end T5-based pipeline that automated the full process of question generation, distractor selection, and scoring. Their approach leveraged datasets such as DG-RACE and Wikipedia, but distractor quality remained inconsistent and required human review.  
  
To enhance evaluation and adaptivity, Raina and Gales (2022) developed a framework integrating T5 with GPT-3 and ELECTRA models to automate grammar checking, answerability assessment, and difficulty analysis. Although effective in automated scoring, the system suffered from dataset-specific biases and filtering issues. Similarly, Chomphooyod et al. (2023) focused on English grammar MCQs using a T5 model controlled by part-of-speech (POS) tagging. Their approach achieved high accuracy and flexibility in grammar-focused assessments, yet its applicability was limited to linguistic tasks with narrow topic coverage.  
  
Recent research has expanded beyond purely text-based systems by introducing multimodal and domain-specific approaches. For example, Alam and Abuelmaatti (2025) employed Google Gemini 1.5 Pro, a multimodal large language model (MLLM), to generate MCQs aligned with Bloom’s Taxonomy. Their system effectively integrated lecture notes, diagrams, and charts, thereby broadening applicability in engineering education. However, distractor quality and mathematical reasoning accuracy remained significant weaknesses. Similarly, Jiang and Feng (2025) presented UsmleGPT, a multi-agent MCQ generator for the U.S. Medical Licensing Exam (USMLE), which simulated the collaborative workflow of human experts. By assigning roles to different large language models (LLMs), the system improved domain alignment, though its reliance on general-purpose models reduced performance in highly specialized contexts.  
  
Hybrid frameworks have also been explored to integrate semantic knowledge with machine learning. Kumar et al. (2024) proposed a combination of ontology-based techniques (OBT) and machine-learning-based techniques (MBT) to generate Bloom-aligned questions. Their framework ensured semantic and grammatical precision but required extensive expert validation. Liu (2024) further investigated exam paper generation using machine learning techniques such as PCA and Grid Search to balance difficulty levels across test items. Although this method improved overall composition, it lacked semantic control and generative flexibility compared to LLM-based methods.  
  
In summary, existing works demonstrate significant progress in MCQ generation, from zero-shot approaches to multimodal and multi-agent systems. However, key challenges persist:

1. limited handling of diverse document formats (PDFs, images, mathematical content, Tables)
2. insufficient reasoning capabilities in distractor generation
3. weak adaptability to learner performance
4. lack of plagiarism prevention
5. inadequate explainability due to hallucinations in LLMs.

These research gaps motivate the development of a multi-agent adaptive MCQ generation framework that integrates multimodal preprocessing, reasoning-driven generation, and reinforcement learning for adaptive learning.

# References

[1] Automatic item generation. Wikipedia. https://en.wikipedia.org/wiki/Automatic\_item\_generation

[2] Adaptive learning. Wikipedia. https://en.wikipedia.org/wiki/Adaptive\_learning

[3] Mucciaccia, A., et al. (2025). LLM-based MCQ Generation with Prompt Engineering. Proceedings of COLING 2025.

[4] Al Shuraiqi, A., et al. (2024). Automatic question generation and answer assessment: A survey. ResearchGate.

[5] Wang, X., et al. (2025). Multi-Agent Interactive Question Generation for Long-Form Documents. arXiv:2507.20145.

[6] Besrour, M., et al. (2025). RAGentA: A Trusted Multi-Agent Retrieval-Augmented Generation Framework. arXiv:2506.16988.

[7] Li, E., Su, J., Sheng, H., & Wai, L. (2021). AGenT Zero: Automatic MCQ Generation for Skill Assessments.

[8] Jiang, Z., & Feng, S. (2025). UsmleGPT: An AI Application for Developing MCQs via Multi-Agent System. Software Impacts, 23, 100742.

[9] Liu, M. (2024). Study on Exam Paper Generation Methods and Experiments based on Machine Learning. Proceedings of IECT ’24, ACM.

[10] Alam, A., & Abuelmaatti, A. (2025). AI-assisted MCQ Generation with Multimodal Large Language Models in Engineering Higher Education.

[11] Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-End Generation of Multiple-Choice Questions Using Text-to-Text Transfer Transformer Models. Expert Systems with Applications, 208, 118258.

[12] Raina, V., & Gales, M. J. F. (2022). Multiple-Choice Question Generation: Towards an Automated Assessment Framework. arXiv:2209.11830.

[13] Chomphooyod, P., Suchato, A., Tuaycharoen, N., & Punyabukkana, P. (2023). English Grammar Multiple-Choice Question Generation Using T5. Computers and Education: Artificial Intelligence, 5, 100158.

[14] Kumar, A. P., Nayak, A., Shenoy, M. K., & Ghosh, K. (2024). A Novel Framework for the Generation of Multiple Choice Question Stems Using Semantic and ML Techniques. IJAIED, 34(2), 332–375.