

ReQUESTA: A Hybrid Agentic Framework for Generating Cognitively Diverse Multiple-Choice Questions

Tian, Y., Chakraborty, S., Huynh, L., Christhilf, K., Watanabe, M., & McNamara, D. S.

Cognitive Variety

Hybrid Agentic Framework

Large Language Models

Question Generation

This study presents ReQUESTA, a hybrid agentic framework that integrates LLM-powered and rule-based agents to generate multiple choice questions (MCQs) with distinct cognitive focuses: text-based, inferential, and main-idea. To provide an initial validation of the framework, expert raters evaluated 100 ReQUESTA-generated MCQs using a cognitive classification rubric to assess alignment between intended and perceived cognitive categories. The results indicate a high level of agreement between system-assigned and expert-assigned labels (agreement rate = 0.95), suggesting that ReQUESTA can reliably instantiate targeted cognitive distinctions in question generation. These findings offer preliminary evidence of the framework's capacity to support cognitively diverse and pedagogically meaningful assessment design. ReQUESTA's modular and hybrid design supports scalable, iterative, and evidence-based assessment development. Future work will include psychometric validation and expansion to additional cognitive categories such as application-level questions.

Introduction

Multiple-choice questions (MCQs) remain a widely used format for automated assessment in education due to their efficiency, objective scoring, and scalability across large student cohorts (Simkin & Kuechler, 2005). However, developing high-quality MCQs requires substantial time and expertise from instructors and assessment designers. Consequently, there is a growing need for automated MCQ generation systems that can reduce instructor workload and maintain the pedagogical integrity of assessments while supporting scalable and evidence-based learning evaluation.

Recent developments in transformer-based large language models (LLM) suggest that automatically generating high-quality MCQs has become increasingly feasible. For example, Olney (2023) found that LLM-generated questions drawn from textbook content achieved parity with human-authored items on most metrics of item quality. Likewise, Cheung et al. (2023) showed that ChatGPT generated MCQs that were broadly comparable to those drafted by faculty, exhibiting similar clarity, specificity, discriminative power, and suitability for graduate-level examination.

Nonetheless, key challenges persist. Specifically, alignment of generated items to higher cognitive levels (e.g., inferential or analytical reasoning) remains difficult. Recent evaluations (Scaria et al., 2024), drawing on Bloom's taxonomy (2020), indicate that LLMs perform well for lower-level (recall) items. However, performance and alignment degrade for higher-order items unless supported by additional control mechanisms, such as advanced prompting, verification classifiers, or specialized refinement agents.

Accordingly, we introduce ReQUESTA, a hybrid multi-agent framework that integrates rule-based components with LLM-powered agents to generate cognitively diverse MCQs across three levels: (i) text-based (recall), (ii) inferential, and (iii) main-idea (synthesis). The system decomposes the resource-intensive MCQ authoring process into a sequence of specialized subtasks, each handled by a dedicated agent (e.g., Planner, Question Generators, Evaluator). By combining the generative strengths of LLMs with the precision and control of rule-based logic, the system establishes a scalable, consistent, and pedagogically grounded workflow for automated MCQ generation in educational assessment contexts.

Background and Framework

Agentic systems have emerged as a practical architecture to leverage LLMs on complex reasoning tasks that involve multiple steps (e.g., Christie et al., 2025; Lu et al., 2025). In these systems, agents (e.g., fine-tuned models, external software tools, custom python scripts) may be specialized for distinct responsibilities (e.g., planning, generation, verification, and orchestration), and the frameworks coordinate these roles to decompose difficult tasks into tractable subtasks and to combine deterministic rule-based processing with stochastic LLM outputs.

Recent developments demonstrate the growing success of multi-agent pipelines for automated question generation. Pawar et al. (2024) introduced a LangChain-based multi-step framework using Gemini LLMs to generate and evaluate quizzes from structured instructional text, producing instructor-usable questions with minimal human revision. Similarly, Mucciaccia et al. (2025) presented an LLM-based agentic system combining prompt engineering and automated evaluation, which generated high-quality MCQs while substantially reducing faculty workload. Despite these advances, existing frameworks are not yet able to reliably control the cognitive complexity of generated questions, and alignment with higher-order cognitive levels (e.g., inferential or synthesis questions) continues to be an open challenge.

ReQUESTA Workflow Design

ReQUESTA's design (illustrated in Appendix A) extends beyond conventional automated question generators that produce primarily recall-based items. It strategically analyzes the source text to identify key concepts, inferences, and overarching ideas, enabling the generation of text-based, inferential, and main-idea MCQs based on user specifications. The system is composed of multiple interacting agents, described below.

Core Architecture

1. Preprocessor: A rule-based agent that uses spaCy's sentence segmentation to divide the input text into coherent chunks, preserving contextual integrity while ensuring manageable input size for downstream processing.
2. Planner: An LLM-powered agent prompted with structured, stepwise instructions to (1) summarize each text segment, (2) extract key ideas and inferences, and (3) construct a high-level question generation plan according to user inputs (e.g., desired number of each question type).
3. Controller: A rule-based coordination agent implemented with custom Python scripts. It interprets the Planner's output and assigns generation tasks along with the relevant text segments to appropriate Question Generator agents based on the specified configuration.
4. Question Generators (Text-based, Inferential, Main-idea). Three LLM-powered agents guided by detailed prompts and examples to produce MCQs aligned with the cognitive focus defined in the plan: key concepts for text-based questions, logical reasoning for inferential questions, and overarching understanding for main-idea questions.
5. Evaluator: An LLM-powered agent that applies a pedagogically grounded evaluation checklist (e.g., item clarity, stem-answer alignment, distractor plausibility) to assess the quality of each generated question. Questions meeting the quality criteria are approved for output, while those requiring improvement are automatically returned to the relevant generator for revision.
6. Formatter: A rule-based agent implemented through a set of Python scripts that finalizes the question output by shuffling answer options, standardizing formatting, and applying consistent labeling across all generated items.

Option-Shortening Module

The Option-Shortening Module was developed to ensure uniformity in the length and structure of MCQ options. It detects options that are noticeably longer than others and adjusts them to maintain stylistic consistency without altering meaning. The module consists of four collaborating agents:

1. Syntax Analyzer. An LLM-powered agent that identifies common syntactic patterns across the four options. These patterns guide the generation of revised options to ensure linguistic consistency and structural uniformity.
2. Length Determiner. A rule-based agent implemented with custom Python scripts that establishes acceptable length thresholds and flags options exceeding these limits.
3. Candidate Generator. An LLM-powered agent that produces concise candidate revisions for any overlong option while retaining its original meaning.
4. Candidate Selector. An LLM-powered agent guided by rule-based metrics (e.g., word count, cosine similarity) to evaluate the candidates and select the version that best meets length and stylistic criteria.

Evaluation

The purpose of this evaluation was to provide an initial validation of ReQUESTA's ability to generate MCQs that instantiate intended cognitive focuses (text-based, inferential, and main-idea questions). Specifically, the evaluation examined the degree of alignment between system-assigned cognitive labels and expert judgments, thereby assessing the construct validity of the framework's cognitive differentiation.

Materials and Question Generation

An open-access, college-level textbook, *Introduction to Anthropology* (OpenStax, 2024), was used as the source material for question generation. Ten chapters were randomly selected from the textbook. The chapters averaged 1,835 words in length ($SD = 361$), ranging from 1,309 to 2,480 words.

Each chapter was processed using ReQUESTA with GPT-4o as the major LLM engine. For each chapter, ReQUESTA generated 10 questions, yielding a total of 100 questions. The system was configured to produce four text-based, four inferential, and two main-idea questions per chapter, consistent with the targeted distribution of cognitive focuses.

Expert Evaluation

All 100 generated MCQs were randomized, and their system-assigned cognitive labels were removed before evaluation to prevent potential bias. Two subject-matter experts independently classified each question using the rubric described in Appendix B. Following independent coding, inter-rater agreement was assessed. The two raters achieved a high level of agreement, with an agreement rate of 94% and a Cohen's kappa of 0.91, indicating almost perfect agreement according to the benchmark proposed by Landis and Koch (1977). The raters then met to discuss and resolve disagreements, and the reconciled classifications were used for subsequent analysis.

To examine the extent to which ReQUESTA-generated questions aligned with their intended cognitive focuses, the system-assigned labels were compared with the expert-assigned consensus labels. Standard classification metrics—accuracy, precision, recall, and F1-score—were computed to summarize overall alignment and category-specific patterns. In this context, accuracy reflects the overall proportion of questions whose intended cognitive focus matched expert judgments, while precision and recall characterize alignment within individual cognitive categories. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced summary of category-level alignment.

Results

Example MCQs of different cognitive focuses generated by ReQUESTA are provided in Appendix C. Table 1 reports the alignment between ReQUESTA’s intended cognitive labels and expert-assigned classifications, and Figure 1 presents the corresponding confusion matrix.

Across all categories, ReQUESTA achieved an overall agreement rate of 0.95, indicating that expert raters classified 95% of the generated questions into the same cognitive categories intended by the system. This high level of alignment provides initial evidence that the framework can reliably operationalize distinct cognitive focuses in MCQ generation.

Category-level analyses showed consistently strong alignment across text-based, inferential, and main-idea questions. In particular, main-idea questions exhibited perfect alignment, with all items classified by experts as intended. Minor discrepancies were observed primarily between text-based and inferential categories. As shown in the confusion matrix, four out of 40 questions intended to be text-based were classified by experts as inferential. Qualitative inspection suggests that these items, while grounded in textual content, required additional reasoning or synthesis, thereby blurring the boundary between

Figure. 1. *A confusion matrix comparing ReQUESTA-generated and human-assigned labels.*

recall-oriented and inferential processing.

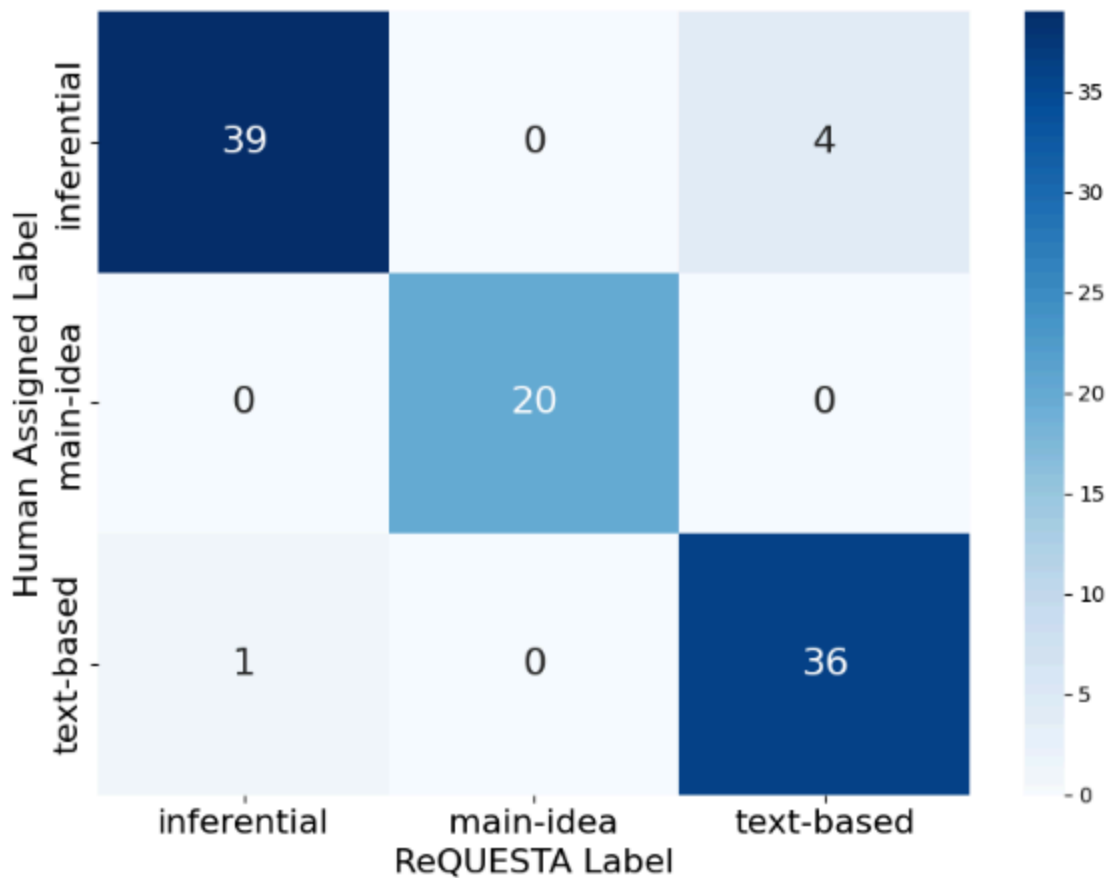


Table 1. Performance Metrics for ReQUESTA Question Type Label against Human Assigned Labels

Label	Precision	Recall	F1 Score
text-based	0.90	0.973	0.935
inferential	0.975	0.907	0.940
main-idea	1.00	1.00	1.00
Overall	0.95	0.95	0.95

Discussion

The evaluation provides initial evidence that ReQUESTA can reliably instantiate distinct cognitive focuses (text-based, inferential, and main-idea) in generated multiple-choice questions, as reflected in the high level of alignment between system-assigned labels and expert judgments. These findings support the construct validity of the framework's cognitive differentiation mechanism. In particular, the complete alignment observed for main-idea questions suggests that ReQUESTA effectively operationalizes global comprehension and synthesis processes that are central to higher-order cognitive assessment. The limited discrepancies between text-based and inferential questions are informative rather than problematic, highlighting the inherently graded boundary between surface-level recall and reasoning-based processing. For example, some questions labeled by the system as text-based required examinees to evaluate the relative relevance or implications of explicitly stated information, prompting expert raters to classify them as inferential. Such cases underscore the continuum nature of cognitive demands in assessment tasks and point to opportunities for further refinement of cognitive category definitions and agent prompting strategies.

The study also demonstrates the effectiveness of using agentic architectures to tackle cognitive complex tasks such as MCQ generation. Specifically, MCQ authoring involves heterogeneous subtasks: content retrieval and segmentation, identification of key concepts, stem composition, distractor generation, stylistic normalization, and pedagogical validation. By assigning these subtasks to dedicated agents, ReQUESTA demonstrates the capacity to move beyond fact-recall items toward more cognitively demanding questions that assess inference and synthesis. This represents an advancement over conventional question generators, which often emphasize surface linguistic quality rather than cognitive depth (Ch & Saha, 2018).

Furthermore, the hybrid and modular nature of ReQUESTA supports scalable, iterative, and evidence-driven processes for designing and refining learning systems, aligning with contemporary characterizations of learning engineering (Baker et al., 2022). The framework facilitates continuous cycles of design, implementation, evaluation, and refinement through the integration of human expert judgment and state-of-the-art intelligent agents. In this sense, ReQUESTA can be conceptualized as a nested learning engineering cycle embedded within a broader assessment design workflow, with the goal of reducing instructor workload while enhancing the cognitive rigor and pedagogical alignment of automated assessments.

Conclusion and Future Directions

This study introduced ReQUESTA, a hybrid agentic framework that integrates rule-based and LLM-powered agents to generate multiple-choice questions with distinct cognitive focuses. Through an initial expert-alignment evaluation, the study provides evidence that ReQUESTA can reliably instantiate intended cognitive categories (text-based, inferential, and main-idea), highlighting the framework's capacity to operationalize cognitively differentiated assessment design. These findings

underscore the potential of hybrid, agentic AI systems to support scalable and pedagogically meaningful assessment generation.

Future work will focus on systematic evaluation of question quality through human judgments and psychometric analyses to assess the validity, reliability, and instructional value of the generated items. Continued development will also aim to extend ReQUESTA's capabilities to produce a broader range of cognitive categories, including application and evaluation questions, thereby enhancing its educational utility and alignment with established learning taxonomies.

Acknowledgments

The research reported here was supported by Arizona State University (ASU), ASU Learning Engineering Institute and the Institute of Education Sciences, U.S. Department of Education, through Grants R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of ASU, the Institute of Education Sciences, or the U.S. Department of Education.

References

- Baker, R. S., Boser, U., & Snow, E. L. (2022). Learning engineering: A view on where the field is at, where it's going, and the research needed. *Technology, Mind, and Behavior*. <https://doi.org/10.1037/tmb0000058>
- Bloom, B. S. (2010). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Ch, D. R., & Saha, S. K. (2018). Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1), 14-25.
- Cheung, B. H. H., Lau, G. K. K., Wong, G. T. C., Lee, E. Y. P., Kulkarni, D., Seow, C. S., ... & Co, M. T. H. (2023). ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong SAR, Singapore, Ireland, and the United Kingdom). *PloS one*, 18(8), e0290691.
- Christie, S. T., Rafferty, A. N., Lee, Z., Cutler, E., Tian, Y., & Almoubayyed, H. (2025, July). An agentic framework for real-time pedagogical plot generation. In *International Conference on Artificial Intelligence in Education* (pp. 309-316). Cham: Springer Nature Switzerland.
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.
- Lu, P., Chen, B., Liu, S., Thapa, R., Boen, J., & Zou, J. (2025). Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*.
- Mucciaccia, S. S., Paixão, T. M., Mutz, F. W., Badue, C. S., de Souza, A. F., & Oliveira-Santos, T. (2025, January). Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 2246-2260).
- Olney, A. M. (2023). *Generating Multiple Choice Questions from a Textbook: LLMs Match Human Performance on Most Metrics*. Grantee Submission.

OpenStax. (2024). Introduction to anthropology. OpenStax, Rice University. <https://openstax.org/details/books/introduction-anthropology>.

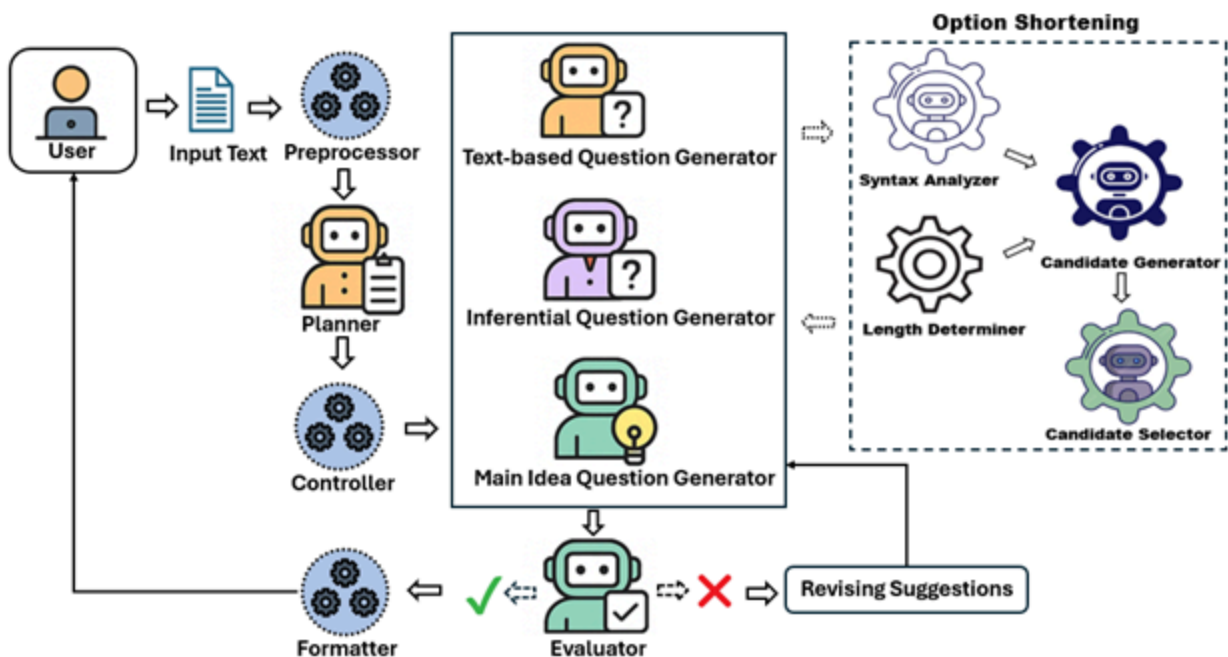
Pawar, P., Dube, R., Joshi, A., Gulhane, Z., & Patil, R. (2024, August). Automated Generation and Evaluation of MultipleChoice Quizzes using Langchain and Gemini LLM. In 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT) (Vol. 1, pp. 1-7). IEEE.

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024, July). Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In International Conference on Artificial Intelligence in Education (pp. 165-179). Cham: Springer Nature Switzerland.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. Decision Sciences Journal of Innovative Education, 3(1), 73-98.

Appendix A

Figure. A1. Overview of the ReQUESTA workflow. The process begins with a user-uploaded file (.pdf, .docx, or .txt). The Preprocessor extracts and segments the text, which is then analyzed by the Planner to identify key concepts and inferences and to construct a question generation plan. This plan is parsed by the Controller, which coordinates and assigns tasks to the three Question Generators responsible for producing text-based, inferential, and main-idea questions concurrently. The Option-Shortening Module standardizes answer options by shortening noticeably longer alternatives to ensure consistent length and structure. The Evaluator assesses each question's quality and determines whether it requires revision by the relevant generator or is ready for output. Finally, the Formatter organizes and delivers the approved questions to the user interface.



Appendix B

Table B1.

Rubrics for Assessing Different Question Types and Their Cognitive Focuses.

Can the exact answer to this question be located from one sentence in the text?

Yes- Text-based

No- Inferential or Main idea

Text-based	Inferential	Main idea
Information is directly stated from the text	Information is not explicitly stated, must be inferred	Overall point, purpose, or central topic of the text
Answer can be found verbatim or paraphrased in one sentence	Answer choice is a combination of ideas from more than one sentences in the text	
	What can be inferred... How would the author feel about...	What is the main idea... Which statement best summarizes...

Appendix C

Example MCQs Generated by ReQUESTA

Text-based

What role do values and beliefs play in shaping society?

- A) They enforce legal standards and judicial processes within a culture.
- B) They suggest what is considered good and bad, beautiful and ugly.
- C) They regulate individuals' emotions and personal feelings.
- D) They determine the distribution of resources among societal groups.

Inferential

How does the Sapir-Whorf hypothesis explain the influence of language differences on cultural perception and behavior?

- A) Language structures can lead to different interpretations of gender roles across cultures.
- B) The adoption of new vocabulary is uniform across cultures, reflecting a shared global culture.
- C) Words have attached meanings that universally influence cognition, regardless of cultural context.

D) Language innovations in technology influence all cultures similarly, despite prior linguistic differences.

Main-idea

What is the main idea of the passage?

A) Anthropology has remained unchanged since the 19th century and still relies on “armchair anthropology”.

B) Anthropology's evolution includes diverse perspectives and the move from indirect to direct research methods.

C) Feminist anthropology was the sole contributor to modern anthropological practices.

D) Modern anthropology focuses only on ethnographic methods without any contributions from other disciplines.

