QuerIA: Contextual Learning-Driven Questionnaire Generation and Assessment based on Large Language Models

Paul Eyzaguirre[1,*], Carlos Badenes-Olmedo[1]

[1]Departamento de Sistemas Informáticos, ETSI, Universidad Politécnica de Madrid, Spain.

Abstract

This paper presents QuerIA, a system based on large language models (LLMs) and contextual learning, to automate the generation and evaluation of educational questionnaires. Central to QuerIA is its integration of Bloom's Taxonomy into the knowledge base of LLMs, enabling the transfer of structured educational objectives to dynamically generate questions that vary in cognitive difficulty. This approach facilitates nuanced customization of assessments that align with individual learning needs and cognitive levels. Using semantic segmentation and in-context learning techniques, QuerIA not only streamlines the creation of questionnaires, but also ensures the relevance and semantic integrity of the generated questions. Both the source code and the online service of QuerIA are publicly available. Our application of the Rasch model to evaluate the system confirms its capability to precisely adapt Bloom's hierarchical framework within the outputs of the LLM, thus achieving adequate control over the difficulty of questions.

Keywords

Assessment System, Questionnaire Automation, Adaptive Learning, Bloom's Taxonomy, Semantic segmentation, Multiple Choice questions (MCQ), Open Ended Questions (OEQ)

1. Introduction

Questionnaires are an essential educational tool for assessing student comprehension and promoting active engagement in the learning process. Decades of research have demonstrated their effectiveness in improving learning outcomes [1, 2, 3]. Feedback from quizzes allows students to gauge their own understanding and revisit unclear content. However, creating high-quality questionnaires and delivering timely feedback is labor-intensive and time-consuming. The quality and difficulty of questions are often subjectively determined, and in automated settings, traditional methods like Bloom's Taxonomy [4] are employed to manually set question difficulty.

Recent advancements in question generation research have predominantly leveraged Transformer-based large language models (LLMs) [5, 6], which have significantly outperformed earlier rule-based and supervised systems [7, 8] . However, real-world applications of these technologies are scarce due to the disconnect between academic research objectives and the practical needs of educators [9]. For example, existing systems such as the rule-based system of Van Campenhout et al. [10] and the GPT-based system of Elkins et al. [11] have focused on basic question formats and utilized empirical strategies to improve question diversity and reduce redundancy.

Despite the existence of automated question generation systems based on natural language processing (NLP), their integration into classrooms has been limited due to domain specificity, language restrictions, and limitations in the types and difficulty levels of the questions generated [5, 12, 9]. Commercial question generation services like WebExperimenter [13] and AnswerQuest [14] offer limited types of questions, often restricted by language and lack of customizable difficulty settings. To address these challenges, we have developed a bilingual framework that not only assists educators and students in

EKAW-24: 24th International Conference on Knowledge Engineering and Knowledge Management

*Corresponding author.

paul.eyzaguirre@alumnos.upm.es (P. Eyzaguirre); carlos.badenes@upm.es (C. Badenes-Olmedo)
https://portalcientifico.upm.es/es/ipublic/researcher/308607 (C. Badenes-Olmedo)
0009000687568759 (P. Eyzaguirre); 0000-0002-2753-9917 (C. Badenes-Olmedo)

creating and assessing high-quality questionnaires in English and Spanish, but also incorporates an approach of transferring the structured knowledge from Bloom's Taxonomy into Large Language Models (LLM) through contextual adjustments. This system offers customizable difficulty levels and question types, along with automated feedback and grading for open-ended questions, ensuring an adaptive learning experience. To validate the effectiveness of our approach to transfer learning, we conducted surveys with 20 Spanish university students, assessing the alignment of Bloom's taxonomy in estimating question difficulty and confirming the importance of precise instructional context segmentation when using language models to generate high-quality questions.

2. Adaptive Learning for Bloom's Taxonomy Alignment

Our framework utilizes *"in-context learning"*, a technique where language models generate outputs based on examples and instructions provided within the input context, allowing for task adaptation without additional fine-tuning. We utilized Llama 3-8B [15], a large language model (LLM) trained on extensive text data, to generate multiple-choice and open-ended questions from a given input document. Our approach employs a semantic chunking strategy, segmenting the document into sequential blocks of text, each serving as the basis for generating a question. To address three different levels of difficulty, we developed a new taxonomy by grouping Bloom's dimension levels [16] into three categories. Question generation and automated grading are achieved through a combination of instructional prompts based on our taxonomy and in-context learning techniques, such as few-shot learning. The following subsections will delve into the specifics of the semantic chunking strategy, our proposed taxonomy, and the automated grading method.

2.1. Semantic Chunking

The proposed chunking strategy, which breaks down long-sequence inputs into manageable parts for a LLM, is a crucial step of the question generation process. These chunks provide the necessary context to the LLM, enabling it to generate relevant and accurate questions. By ensuring that each segment maintains a consistent topic or content, the model can effectively understand the context, leading to the generation of high-quality questions.

Many popular Retrieval-Augmented Generation (RAG) frameworks, such as Langchain [17], LlamaIndex [6], Pincone [18], typically employ empirical or heuristic methods to address this problem. In contrast, our work adopts a semantic chunking approach inspired by methodologies discussed by Greg Kamradt [19]. Here, each chunk serves as the contextual foundation for generating individual questions. The semantic chunking process comprises three key steps: *Sentence Extraction*, *Embeddings*, and *Merging*.

Initially, the text of a document is segmented into individual sentences for the *sentence extraction* phase. In the *embeddings* phase, each sentence is grouped with the preceding and following sentence to form a sentence cluster anchored by the central sentence, providing contextual coherence. The optimal configuration includes one sentence before and after the central sentence, and embeddings are created for these clusters. The semantic distances between sequential sentence groups are then compared, grouping clusters that maintain a low semantic distance, indicating topic consistency, while a higher distance suggests a topic shift, thus delineating distinct text chunks. In the merging phase, the final breakpoints for chunking are determined by setting a threshold at the 80th percentile of the semantic distances, allowing the granularity of the divisions to be adjusted and ensuring an optimal number of chunks for effective question generation.

2.2. Difficulty based on Bloom's Taxonomy

To effectively categorize question difficulty into three levels, our proposed taxonomy groups the cognitive dimension (CD) and knowledge dimension (KD) levels of Bloom's Taxonomy. Our focus is on specific levels for each dimension. From CD, we include the levels of *Remember*, *Understand*,

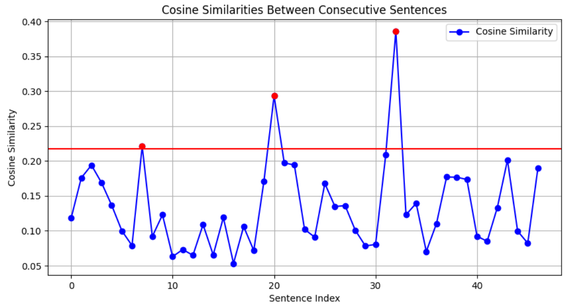


Figure 1: Breakpoints between indices of combined sentences from a 3-page document. Note the focus on identifying scattered outliers, which represent deviations from the continuity of context or meaning. These outliers, defined as distances above the 80th percentile (indicated by the horizontal red line), serve as effective breakpoints for dividing the text into coherent chunks.

Apply, *Evaluate*, and *Analyze*; from the KD, we consider *Factual*, *Procedural*, and *Conceptual* knowledge types. The *Create* level from the CD and the *Metacognitive* level from the KD are excluded due to the nature of multiple-choice questions, which require closed responses and do not provide the flexibility to effectively assess creativity or self-reflection. The following taxonomy is proposed:

1. **Easy level**: Cognitive level "*Remember*" and type of knowledge "*Factual*".
2. **Intermediate level**: Cognitive level "*Understand*" or "*Apply*", and type of knowledge "*Procedural*" or "*Conceptual*".
3. **Difficult level**: Cognitive level "*Analyze*" or "*Evaluate*" and type of knowledge "*Conceptual*".

Furthermore, we incorporated the verbs associated with Bloom's Taxonomy as identified by Stanny [20]. Previous works [16] have shown that the choice of verbs in each category plays a crucial role in determining the cognitive level required to answer a question. The input for question generation is set to temperature 0.1, resulting in more deterministic and focused responses, while a higher temperature would generate more unpredicted and creative outputs. The input is formatted as: <taxonomy_description> <few-shot-learning> <Instructions> <context> Table 1 provides an example of a multiple choice question generated by our system, further illustrating the application of Bloom's Taxonomy in our approach.

2.3. Automated grading

Transitioning from examining question difficulty, the focus shifts to automating the grading of open-ended questions using both basic and complex methodologies. While simpler answers aligned with lower levels of Bloom's taxonomy can be graded on surface-level features [21], responses demanding higher cognitive skills, such as analysis or evaluation, require advanced syntactic and semantic assessments to understand conceptual relationships and reasoning coherence. Traditional automatic grading systems, which predominantly measure lexical or semantic overlaps [22], often fail to accurately score nuanced answers and show poor alignment with human judgment, suggesting limitations in capturing the depth of answers. To address these shortcomings, our approach involves using learned metrics [23] that incorporate the question's context and specific instructions, allowing pre-trained language models to better approximate human evaluations. The system uses a three-tier grading scale and includes instruction to enhance the accuracy of the model's scoring, demonstrating significant improvements in the automated grading of complex answers.

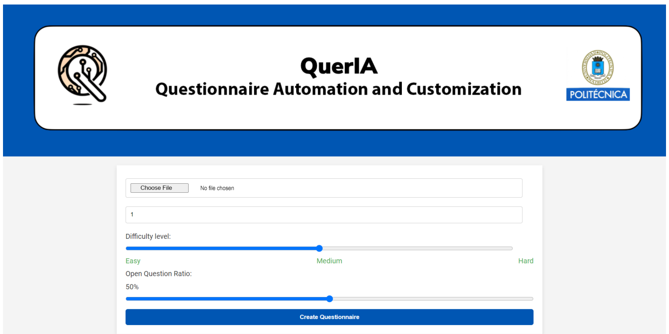
3. QuerIA

QuerIA enables users to upload textual documents such as lecture notes or textbooks for assessment. Users customize their questionnaires by setting the number of questions, choosing between open-

QUESTION 2	What type of information is used to evaluate negative side effects of vaccines and distinguish them from false alarmists?
Option 1	Scientific evidence and statistical data
Option 2	Analysis of the chemical composition of vaccines
Option 3	Opinions of medical experts
Option 4	**Rigorously designed studies published in medical journals**
Evidence	The correct answer refers to the fact that anti-vaccine groups tend to excessively underestimate the complications of infectious diseases that are published in medical articles, while they magnify the side effects of vaccines and offer a very biased view of reality.
CD	Evaluate
KD	Conceptual
Level	Difficult
Rasch estimation	0.91

Table 1
Example of a multiple-choice question of level 3 (difficult), generated from a document passage using our grouped taxonomy. Note the requirement to make a judgment based on evidence (CD - Evaluate) and the implicit task of synthesizing a specific concept (KD - Conceptual). From the context given, the correct answer is option 4.

ended and multiple-choice formats, and selecting the difficulty level. Once uploaded, QuerIA processes the document asynchronously, extracting content to intelligently generate questions and crafting plausible distractors for multiple-choice questions. These questions are displayed in real-time for immediate review to ensure they meet educational standards. Upon questionnaire completion, users can answer directly on the platform, where QuerIA provides instant feedback on open-ended responses, offering corrections, improvements, or confirmations to enhance the learning experience through active engagement. Additionally, the source code is publicly available on GitHub at QuerIA GitHub Repository [1], and there is an online service hosted at QuerIA Online Service [2]. However, the performance of the online service may be slower as it operates on CPUs rather than the more efficient GPUs.


QuerIA
Questionnaire Automation and Customization

No file chosen

1

Difficulty level: Medium

Open Question Ratio:
50%

Figure 2: This screenshot displays the QuerIA user interface, where users can upload educational materials, customize question parameters, and view real-time question generation.

QuerIA evaluates user-submitted answers by analyzing the content extracted from the uploaded educational materials, ensuring that feedback is deeply rooted in the documented evidence. When a user responds to a question, especially in open-ended formats, the system uses advanced NLP techniques to assess the accuracy and relevance of the answer relative to the content of the source material. It then provides a detailed commentary that justifies the answer, highlighting connections to specific information within the document. For instance, if a response is incorrect or partially correct, QuerIA offers constructive feedback that references particular sections or concepts from the document, guiding

[1] https://github.com/cbadenes/queria
[2] https://cbadenes.github.io/queria/

users on how to improve their answers or understand the material more thoroughly. This process not only aids in learning but also reinforces the educational content by linking feedback directly to the text, fostering a more comprehensive understanding and retention of the material.

4. Evaluation and Results

The efficacy of this framework has been empirically validated through surveys involving both open-ended and multiple-choice questions, demonstrating its capability to align generated questions with the intended difficulty levels as confirmed by both perceived difficulty assessments and Rasch analysis. Furthermore, syntactic evaluations have verified the accurate alignment of language used in questions with the cognitive and knowledge dimensions of Bloom's Taxonomy. The innovative automated grading method employed further underscores the framework's utility by providing accurate assessments and feedback, thus enabling effective self-assessment and adaptive learning. Future enhancements will focus on refining the semantic chunker to include image and table processing capabilities and exploring further in-context learning techniques for specialized subjects requiring detailed analytical skills.

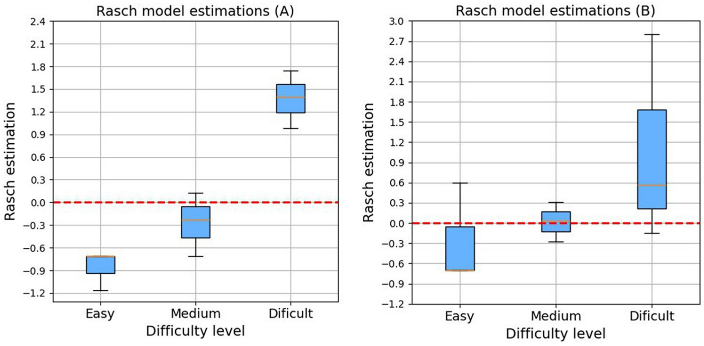


Figure 3: Results of the Rasch model difficulty estimates on the survey using the Framework method. *Box plot A (left) shows the results estimates for open-ended questions, and Box plot B (right) displays the results estimates for multiple-choice questions for each difficulty level. Note the mean centered at 0 (red doted line) and the sign for the interquartile range of each category.*

5. Conclusions and Future Work

In this paper, we introduced a framework that automates the generation and assessment of questionnaires, transcending domain-specific limitations and supporting multilingual implementation. Our method integrates a taxonomy that breaks down Bloom's dimension levels into three difficulty categories: easy, intermediate, and difficult, into language models using instruction prompting and few-shot learning, effectively creating leveled questions. We also introduced a semantic chunking methodology that improves question quality by analyzing document semantics, allowing for the generation of contextually relevant and semantically accurate questions without extensive fine-tuning. The framework's effectiveness was affirmed through surveys evaluating both open-ended and multiple-choice questions, with the results from perceived difficulty assessments and Rasch analysis confirming the accuracy of question difficulty alignment. Additionally, syntactic evaluations upheld the alignment of verbs and interrogative adverbs with Bloom's Taxonomy, and our innovative automated grading method demonstrated accurate response assessments, facilitating effective self-assessment and adaptive learning. Future work will focus on improving the semantic chunker to process visual elements such as images, charts, tables and exploring advanced in-context learning techniques for specialized disciplines that require structured reasoning, such as mathematics or programming.

Acknowledgments. We acknowledge the support of the Educational Innovation Project at Universidad Politécnica de Madrid for facilitating and promoting this research.

References

[1] S. e. a. Ambrose, How learning works: Seven research-based principles for smart teaching, 2010.

[2] M. T. Chi, R. Wylie, The icap framework: Linking cognitive engagement to active learning outcomes, Educational Psychologist 49 (2014) 219–243.

[3] K. R. Koedinger, A. T. Corbett, C. Perfetti, The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning, Cognitive Science 36 (2012) 757–798.

[4] B. S. Bloom, The Taxonomy of Educational Objectives, the Classification of Educational Goals, Volume Handbook I: Cognitive Domain, 1956.

[5] G. Kurdi, J. Leo, B. Parsia, et al., A systematic review of automatic question generation for educational purposes, International Journal of Artificial Intelligence in Education 30 (2020) 121–204.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, et al., Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

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

[8] T. Steuer, L. Bongard, J. Uhlig, G. Zimmer, On the linguistic and pedagogical quality of automatic question generation via neural machine translation, in: Technology-Enhanced Learning for a Free, Safe, and Sustainable World, Springer, 2021, pp. 289–294.

[9] X. Wang, S. Fan, J. Houghton, L. Wang, Towards process-oriented, modular, and versatile question generation that meets educational needs, arXiv preprint arXiv:2205.00355 (2022).

[10] R. Van Campenhout, M. Hubertz, B. G. Johnson, Evaluating ai-generated questions: A mixed-methods analysis using question data and student perceptions, in: Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I, 2022, pp. 344–353.

[11] S. Elkins, E. Kochmar, Cheung, How teachers can use large language models and bloom's taxonomy to create educational quizzes (2024).

[12] E. Kasneci, S. B. Seßler, Katharina, et al., Chatgpt for good? on opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274.

[13] A. Hoshino, H. Nakagawa, Webexperimenter for multiple-choice question generation, 2005, pp. 18–19.

[14] M. Roemmele, D. Sidhpura, S. DeNeefe, L. Tsou, Answerquest: A system for generating question-answer items from multi-paragraph documents, 2021, pp. 40–52.

[15] M. AI, Llama3 8b model, 2024. URL: https://github.com/meta-llama/llama3, accessed: 2024-07-10.

[16] L. W. Anderson, D. A. Krathwohl, A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's, Pearson Education, 2014.

[17] Inc. LangChain, Langchain Documentation on Text Splitters, 2023. URL: https://js.langchain.com/.

[18] R. Schwaber-Cohen, Chunking Strategies for LLM Applications, 2023. URL: https://www.pinecone.io/learn/chunking-strategies/.

[19] G. Kamradt, 5 levels of text splitting, https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/, 2024. Accessed: 2024-07-08.

[20] C. Stanny, Reevaluating bloom's taxonomy: What measurable verbs can and cannot say about student learning, Educ. Sci. 6 (2016) 37.

[21] U. Padó, Get semantic with me! the usefulness of different feature types for short-answer grading, in: Proceedings of COLING-2016, Osaka, Japan, 2016.

[22] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2019.

[23] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892.