Adaptive Learning Assessment Enhanced by Language Models

Paul Eyzaguirre Barreda[1][0009−0006−8756−8759] and Carlos Badenes Olmedo[2][0000−0002−2753−9917]

Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid (Spain).
{paul.eyzaguirre@alumnos.upm.es, carlos.badenes@upm.es}

Abstract. We present an educational framework that automates the generation and assessments of questionnaires without domain-specific and language constraints by leveraging Natural Language Processing (NLP) and advanced Language Models (LMs) to internalize Bloom's Taxonomy. Our framework categorizes questions into three distinct difficulty levels, addressing the core challenge of transferring the structured knowledge of cognitive and learning levels into LMs. We hypothesize that these difficulty levels can be effectively represented by grouping Bloom's categories, facilitating the model's understanding and generation of appropriate questions. To test this hypothesis, we generated multiple-choice and open-ended questions, evaluating their syntactic construction and semantic integrity. Our experiments demonstrate a robust alignment between the proposed difficulty levels and the generated questions compared to our baseline. The framework consistently produces questions that are semantically accurate, syntactically precise, and contextually relevant. Additionally, the automatic evaluation model for open-ended questions provides accurate scores and feedback on student responses, further supporting effective self-assessment. These findings underscore the potential of our approach to enhance the learning process by effectively transferring the structured knowledge of Bloom's Taxonomy into LMs. This enables the generation of high-quality, difficulty-leveled questions without domain-specific constraints, thus promoting efficient and adaptive learning experiences.

Keywords: Assessment System · Questionnaire Automation · Adaptive Learning · Bloom's Taxonomy · Language Models (LM) · Semantic Chunking; Automatic Evaluation tool · Multiple Choice questions (MCQ) · Open Ended Questions (OEQ).

1 Introduction

Questionnaires have been a common strategy employed by instructors to test comprehension and actively engage students in the learning process. Question answering offers active learning [8] as demonstrated by decades of educational research, showing its effectiveness in improving comprehension and learning outcomes [4, 7, 18]. Additionally, feedback on quiz questions enables students to self-evaluate their understanding and revisit specific parts of the material for better clarity [17]. However, the manual creation of high-quality questionnaires and providing feedback on responses is a labor-intensive and time-consuming process.

From a test creation perspective, high-quality leveled questions should ask students to highlight important concepts, underline key takeaways, or provide critical thinking about main issues and implications. However, the determination of question quality and difficulty is often subjective and reliant on the evaluator's criteria. In large-scale testing, question difficulty and other measures of quality are established through prior norming [10], where questions are answered by a pool of test-takers before its definitive use with a similar demographic. Difficulty is then determined based on the observed results using Probabilistic Test Theory (PTT) [26]. However, norming is impractical in automated question creation and infeasible in small academic environments. In these contexts, some teachers rely on Bloom's Taxonomy, a classification of knowledge and cognitive dimensions, to manually formulate questions of appropriate difficulty.

Despite the existence of some NLP-based automatic question generation systems, their adoption in classrooms remains low [3] [20]. This is primarily because these models are tailored to specific domains and predominantly restricted to the English language, or they are limited in the types and difficulty levels of the questions they can generate. To address these limitations and meet the need for multilingual support, we introduce a framework meant to assist both students and instructors in creating and assessing high-quality questionnaires in English and Spanish. Our framework adapts to individual needs by offering flexibility in selecting the difficulty level, specifying the number of questions to generate, and choosing the types of questions for the questionnaire. Moreover, it provides feedback and automated grading for open-ended questions for each answer, thus ensuring an adaptive learning experience.

To validate the efficacy of our proposed framework in generating leveled questions, we conducted a comparative analysis against a baseline method across four surveys that included both open-ended (OEQ) and multiple-choice questions (MCQ). Our syntactic evaluation assessed how well Bloom's taxonomy dimensions aligned with and applied to estimating question difficulty in both formats. Our findings suggest that the taxonomy introduced in this work offers an adequate approximation of question difficulty across three levels (easy, intermediate, and difficult). Additionally, we found empirical evidence that the instructional context, specifically the teaching material presented, must be carefully segmented to generate high-quality questions when using language models.

This article comprises four additional sections. Section 2 reviews the related work. A brief theoretical background on PTT and Bloom's Taxonomy is presented in Section 3. Section 4 details the approach for generating questions with varying difficulty levels. Section 5 presents the evaluation, analyzing the difficulty and reliability of the syntactic-generated questions. And finally, we conclude with our findings in Section 6.

2 Related Work

Recent advancements in question generation research have been heavily influenced by the use of Transformer-based large language models (LLMs) [21], [25]. This shift is largely due to the significant performance improvements these models offer over earlier rule-based and other types of supervised systems ([39], [27]). Despite these advancements, there are few documented instances of real-world deployments [16], [21]. According to Wang et al. [41], the primary reason for this gap is the misalignment between research goals and the practical needs of teachers. For example, Van Campenhout et al. [40], developed an NLP system that generates rule-based questions, limited to concept-matching and fill-in-the-blank questions formats. Similarly, Elkins et al. [12] utilized a proprietary language model (GPT-3.5) to generate questions from a given input passage. One of their strategies used Bloom's Taxonomy to reduce redundancy and enhance diversity among generated questions. However, little research has been conducted on implementing question difficulty based on levels.

On the other hand, systems that provide question generation as a service have also been identified, although several limitations were identified. WebExperimenter [13] generates "fill in the blank" style questions for English proficiency tests, while AnswerQuest [34] focuses on reading comprehension, generating only open-ended questions. SQUASH [19] decomposes larger articles into paragraphs, generating a text comprehension question for each. Platforms requiring paid licenses, such as Quizbot.ai [1] and Questgen [32], offer quiz generation from a document with various question types. Only Quizbot.ia offers difficulty-leveled questions categorized broadly as University, High School, or Middle School levels. However, the main drawbacks of these systems are their limited customization for generating leveled questions, language limitations, and, in the case of paid platforms, a lack of transparency in adjusting their behavior.

3 Theoretical Background

To provide a foundation for our framework, this section describes Bloom's Taxonomy, Probabilistic Test Theory, and the Rasch analysis. Bloom's Taxonomy is a hierarchical classification of cognitive skills and learning objectives, which aids in the formulation of educational questions across different levels of complexity, from basic recall to higher-order thinking skills. Probabilistic Test Theory is a statistical approach used to assess the quality and difficulty of test questions by analyzing response patterns from a sample of test-takers, ensuring the reliability and validity of assessments. Additionally, the Rasch analysis, a specific model within item response theory, provides a method for scaling difficulty and evaluating question quality based on the probability of a given response, further enhancing the precision and fairness of educational assessments. These concepts are crucial for understanding the mechanisms behind our leveled-question generation framework.

3.1 Bloom's Taxonomy

Bloom's Taxonomy [6], revised by Anderson and Krathwohl [5], is a well-established framework for creating and interpreting teaching objectives, and writing test questions. The taxonomy comprises two independent dimensions: the Cognitive dimension (CD) and the Knowledge dimension (KD). The CD describes which type of cognitive process is required for answering a question. This dimension categorizes verbs into six levels that describe a progressive development of cognitive skills. The hierarchy starts with the least demanding process, *Remember*, and progresses through *Understand*, *Apply*, *Analyze*, *Evaluate*, and culminates with the most demanding process *Create*.

The KD classifies the type of knowledge required to complete the task. It progresses from the most concrete to the most abstract types: *Factual Knowledge* (facts and terminology), *Conceptual Knowledge* (categories, principles, and models), *Procedural Knowledge* (algorithms, techniques, and criteria), and *Metacognitive Knowledge* (strategic knowledge and self-knowledge). According to Anderson et al. [5] the levels of these dimensions should be inferred from the wording of the question. For example, verbs like "compare" or "generalize" indicate the Understand level, while "identify" or "name" belong to the Remember level. Stanny [38] further identified 433 unique verbs associated with Bloom's Taxonomy, highlighting the potential for automating the process of level identification based on question formulation.

3.2 PTT Difficulty Estimation with the Rasch Analysis

Probabilistic Test Theory (PTT) is composed of statistical models used to analyze the relationship between individuals' responses to test questions (items) and an underlying latent trait or ability [22]. These models assume that a student's ability and the difficulty of a question are not directly observable, but rather depend probabilistically on the manual grades awarded to the student's answers (i.e., after testing). Among the existing models, the best adapted to our work is the Rasch Analysis [33], which is a widely used statistical technique, its use has been reviewed [11] and discussed in multiple texts [37]. Specifically, the Rasch model estimates the difficulty of the question and the ability (ability) of the student given the following relation (where B_n is the ability of a student n and D_i is the difficulty of the question i):

$$P_{ni}(x = 1|B_n, D_i) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \tag{1}$$

Success (x = 1) of a student n on a question i is linked to the difference between the student's ability and the question's difficulty. If the ability is greater than the difficulty, the student is likely to succeed, or if the inverse is true, the student is more likely to fail. Estimates of B and D are made iteratively from the test results. The resulting measures are returned in logits and question difficulty is centered at 0, so easy items have low or negative estimates and difficult items have high and positive estimates.

4 Approach

Our framework is based on *"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 [2], 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 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. The source code is available online [1].

4.1 Semantic Chunking

The chunking strategy proposed, 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 each chunk 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 [14], LlamaIndex [24], Pincone [35], 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 [15]. 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.

1. **Sentence Extraction.** Initially, the text of a document is split into individual sentences.
2. **Embeddings.** For each sentence, we create a combined sentence group by including one sentence before and one sentence after each given sentence. This approach ensures that each group is "anchored" to a central sentence, providing contextual coherence. The number of sentences included before and after the anchor can be adjusted depending on the document size, though we found that including one sentence in each direction yields optimal results for

[1] https://github.com/PaulAndree/SGEC

our purposes. Next, embeddings are generated for each sentence group, and the distances between sequential sentence groups are compared. Sentence clusters are grouped as long as the semantic distance between them remains low, indicating consistent topic or content. Conversely, a higher semantic distance signals a change in topic, thereby delineating distinct chunks of text.

3. **Merging.** To determine the final breakpoints for chunking, a threshold is set at the 80th percentile of the semantic distances, as illustrated in **Figure 1**. By adjusting the percentile limit, the granularity of the divisions can be controlled, ensuring an optimal number of chunks for effective question generation.

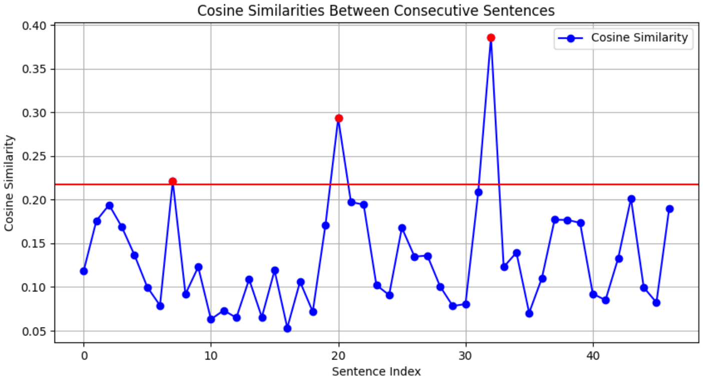

Fig. 1. ***Breakpoints between indices of combined sentences from a 3-page document*** *Note the focus on identifying scattered outliers, which represent significant deviations from the continuity of context or meaning. These outliers, defined as distances above the 80th percentile (indicated by the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

4.2 Bloom's Taxonomy and Difficulty

To effectively categorize question difficulty in three levels, our proposed taxonomy groups the cognitive dimension (CD) and the knowledge dimension (KD) levels of Bloom's Taxonomy. Our focus is on specific levels from each dimension. From the CD, we include the levels of *Remember, Understand, 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 assess creativity or self-reflection effectively. The following taxonomy is proposed:

QUESTION 1	In what year was a study published in the medical journal “The Lancet” that linked the MMR vaccine to autism?
CD	Remember
KD	Factual
Level	Easy
Rasch estimation	-0.71

Table 1. Example of an open-ended question of level 1 (easy), generated from a document passage using our grouped taxonomy. Note the implicit factual requirement (KD - year) and the implicit task of remembering specific information (CD - Remember).

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”*.

Additionally, we incorporated the verbs associated with Bloom’s Taxonomy as identified by Stanny [38]. Previous works [5], 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 with temperature 0.1, which results 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> (see **Figure 2**). **Table 1** provides an example of a generated open question at the easy level, clearly demonstrating the connection with Bloom’s Taxonomy through its syntax. Similarly, **Table 2** showcases a multiple-choice question generated by our system, further illustrating the application of Bloom’s Taxonomy in our approach.

```
Taxonomy description: { “Bloom’s taxonomy is a conceptual structure that
has 2 dimensions ...” }
Few-shot learning: { “What was the method used to evaluate the ...?” }
Instructions: { “Generate a OEQ using the context provided below.” (or
MCQ) “To generate a difficult level question, it must be based on the
“Analyze” or “Evaluate” level of Bloom´s cognitive dimension and ...” }
Contex: { <include the semantic_chunking strategy output> }
```

Fig. 2. Format structure example for question generation at each level (framework method). Due to space limits, instructions and examples are reduced.

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 2. 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.

4.3 Automated grading

Having looked at the difficulty from the point of view of test creation, we now shift our focus to automating the grading of open-ended questions.

Answers to questions formulated at lower levels of Bloom's taxonomy, for example question 1 (**Table 1**), can be effectively graded using shallow features close to the string level, as observed by Pado [30]. However, assessing answers to questions of higher-order levels of Bloom's taxonomy, such as analyzing data, evaluating arguments, or applying solutions, necessitates more sophisticated features derived from syntactic and semantic analysis. For example, grading an answer that involves analyzing the development of a medical solution or evaluating the effectiveness of a connected graph requires understanding the relationships between concepts and the coherence of the reasoning, which cannot be adequately captured by shallow features alone. We explored the relationship between question difficulty and optimal grading metrics, finding that most automatic grading systems compare outputs against reference texts, measuring lexical overlap or semantic similarity [31], [42]. However, these metrics have several limitations. Deviations from reference texts are scored low even if correct, for example, question: *"What is the capital of France?"* answer: *"France's capital is Paris."*. In this case, Paris is the correct answer, but this answer would not receive a perfect score due to how automatic grading systems focusing on lexical or semantic overlap may penalize correct answers that vary in order, expression, or length from the reference text.

On the other hand, studies showed a poor correlation with human judgments [29], [9]. Metrics like BLEU and ROUGE, designed for translation tasks, often do not transfer well to others [23], [28]. Additionally, single similarity scores may overlook important nuances such as creativity, originality, and minor errors like typos.

A more effective approach involves embedding the generation question, instruction, and context to develop a learned metric. Learned metrics tailored to specific tasks emulate human judgments more accurately, as demonstrated by Sellam et al. [36] and Zhao et al. [43]. This approach can be seamlessly adopted by pre-trained language models with instruction tuning. We demonstrate that including zero-shot and few-shot instructions as part of the input enables models to generalize better and perform well in scoring. Our grading scale is based on three levels: 0, 0.5, and 1. A score of 0 corresponds to an incorrect, empty, or very vague answer, 0.5 indicates a partially correct answer or one lacking sufficient detail, and 1 signifies a coherent and valid answer. For the automatic grading process, a LLM was chosen to infer the most appropriate score for a given response within a specific context, similar to the approach used for question generation. Our Llama3-8B-based model predicts a score for each answer using prompts with [8-10] shot learning.

5 Evaluation and results

To validate the efficacy and usage of questions generated by our framework, we developed a baseline method for comparison. The baseline method instructed a language model (Llama3-8B) to generate questions at easy, medium, and difficult levels without employing any taxonomy or specific prompting technique. The format used was: <Instruction> <Context>. In contrast, our framework leveraged Bloom's taxonomy for leveled-question generation.

We conducted four surveys in total, encompassing both open-ended questions (OEQ) and multiple-choice questions (OEQ). Two surveys were based on the baseline method (one with 9 OEQ and one with 9 MCQ). The other two surveys utilized our proposed framework (one with 9 OEQ and one with 9 MCQ). Three university-level texts were randomly selected to feed the model and generate questions of varying difficulty. Each survey comprised three easy, three intermediate, and three difficult questions. Ten university students participated in each survey type without repeating, resulting in a total of 360 responses. Additionally, students rated the perceived difficulty of each question.

5.1 Question Difficulty Evaluation

To ensure a fair evaluation, identical contexts were employed for each method, with a strict focus on assessing question difficulty using both student perception agreement and item response theory, specifically Rasch analysis. Furthermore, we conducted a syntactic analysis of the verbs in the generated questions within

Survey	Metrics	Baseline			Framework		
		L1	L2	L3	L1	L2	L3
MCQ	Agreement	0.71	0.24	0.48	0.75	0.67	0.17
	Rasch estimation	-0.71	0.31	0.69	-0.78	0.1	1.1
OEQ	Agreement	0.71	0.38	0.62	0.95	0.63	0.79
	Rasch estimation	-0.34	-1.45	0.5	-0.86	-0.35	1.38

Table 3. Survey Results Analysis. This table presents the quantified outcomes from the surveys. L1, L2 and L3 indicates the easy, intermediate and difficult levels respectively. The results are for both Multiple Choice Questions (MCQ) and Open-Ended Questions (OEQ) surveys, comparing the baseline and our framework method for question generation.

our framework to gauge its integration and effectiveness.

Open-Ended Questions Results

Our analysis of the open-ended question surveys (OEQ) revealed that the proposed framework taxonomy resulted in better alignment of generated questions with their intended difficulty levels compared to the baseline method. This is evidenced by the higher percentages of student agreement regarding question difficulty perception, as shown in **Table 3**. Specifically, for easy questions, there was a 95% agreement with only 4.7% rated as intermediate. For intermediate questions, there was a 62.5% agreement, with 25% rated as easy, and for difficult questions, there was a 79.12% agreement, with 20.8% rated as intermediate. In contrast, the baseline method resulted in lower agreement percentages across all difficulty levels.

Quantitative measurements using the Rasch model are illustrated in **Figure 3** for the survey utilizing the proposed framework, and are compared to the baseline results in **Table 3**. The box plot (A) highlights more negative values for easy questions (interquartile range from -0.94 to -0.75), values closer to the mean for intermediate questions (-0.45 to -0.01), and higher positive values for difficult questions (interquartile range from 1.18 to 1.57), indicating that the proposed framework provides more significant estimates for differentiating difficulty levels compared to the baseline results.

Multiple-Choice Questions Results

In the context of multiple-choice questions (MCQ), level 3 (difficult) questions received lower agreement on perceived difficulty when using both the baseline method and the proposed framework. This discrepancy can largely be attributed to the plausible options presented in MCQ, which can create an optimistic perception of difficulty influencing the decision of students before knowing the correct answer. Therefore, the Rasch model was employed as a secondary objective metric.

As shown in **Table 3**, our framework yielded a 75% agreement for level 1 (easy) questions, with the remaining 25% rated as intermediate. For level 2 (intermediate) questions, there was a 66.6% agreement, with 29.17% rated as easy and 4.16% as difficult. For level 3 (difficult) questions, there was only a 16.66% agreement, with 83.34% rated as intermediate. Despite the low agreement on difficult-level questions, the Rasch model revealed consistent patterns for both methods: easy questions were rated below the mean, intermediate questions near the mean, and difficult questions positive and above the mean. However, the proposed framework produced more significant Rasch estimates for each difficulty level, as illustrated in the box plot (B) (**Figure 3**), indicating its superior effectiveness in differentiating question difficulty levels compared to the baseline.

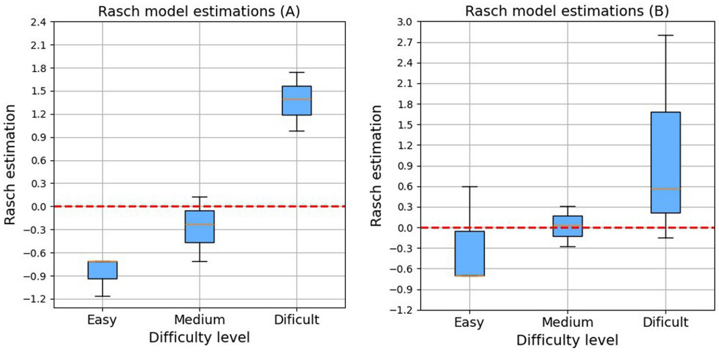

Fig. 3. Results of the Rasch model difficulty estimates on the survey using the Framework method. *Box plot A (left) shows the results estimates for open-ended questions, and Box plot B (right) displays the results estimates for multiple-choice questions for each difficulty level. Note the mean centered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

5.2 Syntactic Analysis

A syntactic analysis for our proposed framework method was performed using the NLP library Stanza [2] in Python. This analysis focused on identifying the frequency of lemmatized verbs corresponding to Bloom's dimensions and interrogative adverbs in the generated questions of the surveys.

[2] https://stanfordnlp.github.io/stanza/

The results revealed that for easy questions the dominant verbs in the questions like “name”, “identify”, and “define” were aligned with Bloom’s *Remember* level. And interrogative adverbs such as “What” and “Which” prompted correct factual responses. For intermediate-level questions, Verbs found like “identify”, “implement”, “follow”, and “develop” fit the *Apply* and *Remember* levels. Adverbs such as “Which” and “What” were appropriate for *Procedural* and *Conceptual* knowledge. Common structures observed in the survey were like: “What are the steps to identify...?”. Finally, for difficult-level questions, verbs like “evaluate”, “consider”, “justify”, etc., matched the *Evaluate* and *Analyze* levels. The interrogative adverb “How” was commonly used, and suitable for eliciting evaluation or critique. Common structures observed were like: “How to critically evaluate...?” and “How can... be evaluated?”.

6 Conclusions

In this paper we introduced a framework designed to automate the generation and assessment of questionnaires, effectively removing domain-specific constraints and enabling implementation across various languages. Our approach leverages our proposed taxonomy that categorizes Bloom’s dimension levels into three difficulty tiers: easy, intermediate, and difficult. By incorporating this taxonomy into Language Models through instruction prompting and few-shot learning, we successfully tackled the challenge of generating leveled questions. Furthermore, we presented a semantic chunking methodology that enhances the quality and granularity of generated questions by considering the document’s semantics. This methodology enables our models to produce contextually relevant and semantically accurate questions without requiring fine-tuning.

The efficacy of our framework was validated through comparative analysis with a baseline method across four surveys employing both open-ended (OEQ) and multiple-choice questions (MCQ). For OEQ, the framework achieved higher student agreement rates across all difficulty levels, with a better differentiation supported by the Rasch model analysis. In MCQ, the framework resulted in more significant Rasch estimates, demonstrating superior effectiveness in generating leveled-question and aligning them accurately compared to the baseline. Additionally, syntactic evaluation confirmed the correct alignment of verbs and interrogative adverbs with Bloom’s Taxonomy cognitive and knowledge dimensions within our framework. Finally, our automated grading method, implemented with learned metrics via instruction tuning, was shown to accurately assess and score answers, providing clear explanations and fair evaluations, thus, reflecting effective self-assessment and adaptive learning.

Future research directions include enhancing the semantic chunking with the ability to process images and tables for better understanding. Additionally, exploring other *in-context learning* techniques for specialized subjects requiring structured reasoning or code writing, such as Mathematics or Programming subjects.

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

References

1. Quizbot.ai. https://quizbot.ai/, accessed: July 1, 2024
2. AI, M.: Llama3 8b model (2024), https://github.com/meta-llama/llama3, accessed: 2024-07-10
3. Alsubait, T., Parsia, B., Sattler, U.: Ontology-based multiple choice question generation. KI-Künstliche Intelligenz **30**(2), 183–188 (2016)
4. Ambrose, S.e.a.: How learning works: Seven research-based principles for smart teaching (2010)
5. Anderson, L.W., Krathwohl, D.A.: A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's. Pearson Education (2014)
6. Bloom, B.S.: The Taxonomy of Educational Objectives, the Classification of Educational Goals, Volume Handbook I: Cognitive Domain (1956)
7. Chi, M.T., Wylie, R.: The icap framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist **49**(4), 219–243 (2014)
8. Clump, M.A., Bauer, H., Bradley, C.: The extent to which psychology students read textbooks: A multiple class analysis of reading across the psychology curriculum. Journal of Instructional Psychology **31**(3), 227–233 (2004)
9. Dhingra, B., Faruqui, M., Parikh, A., et al.: Handling divergent reference texts when evaluating table-to-text generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4884–4895 (2019)
10. Downey, S.: Test development. In: Peterson, P., Baker, E., McGaw, B. (eds.) International Encyclopedia of Education, pp. 159–165. Elsevier Ltd. (2010)
11. Edwards, A.W., Alcock, L.: Using rasch analysis to identify uncharacteristic responses to undergraduate assessments (2010), https://hdl.handle.net/2134/8848
12. Elkins, S., Kochmar, E., Cheung: How teachers can use large language models and bloom's taxonomy to create educational quizzes (2024)
13. Hoshino, A., Nakagawa, H.: Webexperimenter for multiple-choice question generation. pp. 18–19 (2005)
14. Inc. LangChain: Langchain Documentation on Text Splitters (2023), https://js.langchain.com/
15. Kamradt, G.: 5 levels of text splitting. https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/ (2024), accessed: 2024-07-08
16. Kasneci, E., Seßler, Katharina, S.B., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. Learning and Individual Differences **103**, 102274 (2023)
17. Kerr, M.M., Frese, K.M.: Reading to learn or learning to read? engaging college students in course readings. College Teaching **65**(1), 28–31 (2017)

18. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science **36**(5), 757–798 (2012)
19. Krishna, K., Iyyer, M.: Generating question-answer hierarchies. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL '19). pp. 2321–2334 (2019)
20. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education **30**(1), 121–204 (2020)
21. Kurdi, G., Leo, J., Parsia, B., et al.: A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education **30**, 121–204 (2020)
22. van der Linden, W.: Item response theory. In: Peterson, P., Baker, E., McGaw, B. (eds.) International Encyclopedia of Education, pp. 81–88. Elsevier Ltd. (2010)
23. Liu, C.W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. pp. 2122–2132 (2016)
24. Liu, J.: Llamaindex Documentation on Basic Optimization Strategies (2023), https://docs.llamaindex.ai
25. Liu, P., Yuan, W., Fu, J., Jiang, Z., et al.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
26. Lord, F.M., Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA (1968), reprinted by Information Age Publishing, 2008
27. Mulla, N., Gharpure, P.: Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. Progress in Artificial Intelligence **12**(1), 1–32 (2023)
28. Nema, P., Khapra, M.M.: Towards a better metric for evaluating question generation systems. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3950–3959. Association for Computational Linguistics, Brussels, Belgium (2018)
29. Novikova, J., Dušek, O., Cercas Curry, A., Rieser, V.: Why we need new evaluation metrics for nlg. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2241–2252 (2017)
30. Padó, U.: Get semantic with me! the usefulness of different feature types for short-answer grading. In: Proceedings of COLING-2016. Osaka, Japan (2016)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02). pp. 311–318 (2002)
32. Questgen: Questgen: Ai powered question generator. http://questgen.ai/
33. Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen (1960)
34. Roemmele, M., Sidhpura, D., DeNeefe, S., Tsou, L.: Answerquest: A system for generating question-answer items from multi-paragraph documents. pp. 40–52 (2021)
35. Schwaber-Cohen, R.: Chunking Strategies for LLM Applications (2023), https://www.pinecone.io/learn/chunking-strategies/
36. Sellam, T., Das, D., Parikh, A.: BLEURT: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892 (2020)
37. Smith, E., Smith, R. (eds.): Introduction to Rasch Measurement: Theory, Models, and Application. Journal of Applied Measurement Press Books (2004)

38. Stanny, C.: Reevaluating bloom's taxonomy: What measurable verbs can and cannot say about student learning. Educ. Sci. **6**(4), 37 (2016)
39. Steuer, T., Bongard, L., Uhlig, J., Zimmer, G.: 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. pp. 289–294. Springer (2021)
40. Van Campenhout, R., Hubertz, M., Johnson, B.G.: 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. pp. 344–353 (2022)
41. Wang, X., Fan, S., Houghton, J., Wang, L.: Towards process-oriented, modular, and versatile question generation that meets educational needs. arXiv preprint arXiv:2205.00355 (2022)
42. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
43. Zhao, T., Lala, D., Kawahara, T.: Designing precise and robust dialogue response evaluators. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 26–33 (2020)