Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECT OF EMBEDDING MODELS IN RAG SYSTEM FOR PEDAGOGICAL QUESTION GENERATION

JAYASHREE GANESHKUMAR[1] , M. KRISHNAVENI[2]

[1]Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, India

[2]Assistant Professor (SG), Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, India

E-mail: [1]23phcsf002@avinuty.ac.in, [2]krishnaveni_cs@avinuty.ac.in

ABSTRACT

The need for cognitively aligned assessments has highlighted the disadvantages of traditional question generation systems, which often fail in pedagogical structures such as Bloom's taxonomy and suffer from factual hallucinations. This study aims to develop an automated question generation system that integrates a pedagogical framework using a Retrieval-Augmented Generation (RAG) approach. This work uses three open-source embedding models – MiniLM-L6, MiniLM-L12 and msmarco-distilbert to encode educational content and retrieve relevant information using FAISS indices. Retrieved content is then used by T5-based generator, guided by Bloom's taxonomy-specific prompts, to produce pedagogically aligned questions which are then evaluated using standard NLP metrics and validated against manually created ground-truth dataset. This study provides a comparison of embedding models for pedagogically aligned question generation within RAG framework. Findings show that MiniLM-L12 outperforms other embedding models across all the levels of Bloom's taxonomy. The results suggest that educators can use the RAG-based question generation to reduce assessment design workload while ensuring generation of cognitively appropriate questions. Future research will explore larger datasets, fine-tuning strategies, multi-domain scalability to further advance pedagogically aligned automated assessment tools. This work contributes to enhanced learning outcomes and broader educational impact.

Keywords: *Retrieval-Augmented Generation, Bloom's Taxonomy, Educational Technology, Automated Assessment, Question Generation.*

1. INTRODUCTION

Educational assessment is crucial for measuring students' learning and understanding of topics. However, instructors spend a significant amount of time creating assessment questions to evaluate their students across different cognitive levels. One of the major drawbacks is when instructors aim to create assessments that can test various levels of understanding such as basic recall to higher-order thinking. The process of crafting questions for different levels of cognition is a challenge in automated assessment. Improving the assessment creation process could enhance the overall quality of education.

Bloom's taxonomy, developed by Benjamin Bloom in 1956 and is revised by Anderson and Krathwohl in 2001 is a standardized tool. This tool is used to categorize the questions that tests different cognitive complexities that are involved in understanding a concept – Remember, Understand, Apply, Analyze, Evaluate and Create. This tool has been adopted globally due to its approach to measuring the understanding of concepts by learners. Unfortunately, creating questions that align with taxonomical levels is still a time-consuming task and requires expertise. This complexity results in a disparity between intended cognitive level of a question and its actual assessment value. However, there is a need for such tools and methodologies that can facilitate the educators to enhance the quality and effectiveness of educational assessments.

Traditional question generation systems have limitations while manual question creation is labor-intensive and often results in uneven distribution of levels. Many instructors struggle to generate questions that targets higher-order thinking skills such as Analyze, Evaluate and

Create levels. Advances in large language models (LLMs) are expert at generating tasks like this. However, these models are not good at generating domain-specific tasks. They also produce content that are not factually correct and often faces hallucination problems. These hinderances emphasizes the need for careful human oversight and validation when utilizing LLMs for question generation in educational context.

Apart from this, there are critical gaps in developing such systems that can generate questions that are pedagogically aligned. Existing tools often lack in integration of such educational frameworks like Bloom's taxonomy. Further, there is a lack of research in embedding models that are used Retrieval-Augmented Generation (RAG) systems which enhance the LLMs to have domain-specific knowledge. Additionally, implementing rigorous quality control process, including fact-checking and peer review, could help reduce the risk of factual errors and hallucinations in LLM-generated content.

This research aims to addresses these limitations by exploring two main research questions (RQ):

1. How can AI-driven content generation tools be aligned with pedagogical frameworks like Bloom's taxonomy to support educators?
2. What is the effect of different embedding models on Bloom's taxonomy alignment in RAG systems?

This research develops a RAG pipeline that incorporates domain-specific knowledge base to generate questions that are aligned pedagogically. It uses Text-To-Text Transfer Transformer (T5) developed by Google which are good at sentence-to-sentence tasks for generating the questions. The research compares different embedding models such as MiniLM-L6, MiniLM-L12 and msmarco-distilbert and determines the question quality across the cognitive levels. Additionally, this work provides educators with automated question generation capabilities for domain-specific tasks thus reducing their workload. It is also beneficial for self-regulated learners who require unlimited practice questions to master a concept. Thus, contributing to educational technology and Natural Language Processing (NLP) communities. By providing educators with automated question generation capabilities for domain-specific tasks, the system significantly reduces their workload, allowing them to focus on other critical aspects of teaching. Additionally, it benefits the self-regulated learners by allowing them to practice multiple times on certain topics thus enhancing their mastery level of that topic.

Moreover, this study posits that embedding models with higher-dimensional representations will yield more comprehensive content retrieval. This, in turn, is expected to enhance cognitive alignment, which will be assessed using *'BloomAcc'*.

The paper is organized as follows. The paper discusses extensively on existing research to find the gaps. Followed by the methodology section where the procedures adopted are discussed. Further, the results section discusses about the outcomes of the research and briefs about the limitations and future work followed by a brief conclusion.

2. LITERATURE REVIEW

Large Language Models (LLMs) are being used in various educational processes by serving as virtual tutors. These systems provide real-time explanations and feedbacks from assessments which are automated as well [1]. Although LLMs have been developed to produce human-like text across different domains, their usage in the educational sector has serious limitations especially in question generation and assessment systems which are useful for educators.

Models such as ChatGPT are good at generating relevant questions but the quality of it varies based on the prompts users provide[2]. Authors of this paper have conducted an extensive study showing the impact of prompts that produces better quality texts. Hang et al. [3] faced multiple challenges in generating text such as hallucinations and inaccuracies while producing content which is quite common in LLMs. Although the authors experimented with various prompting techniques such as chain-of-thought and self-refine methods, the evaluation metrics reveal authenticity issues especially for domain-specific tasks.

Several research has highlighted major limitations in traditional LLMs in educational setting. Zhang et al. [4] classified questions generated using BERT-based model but the model had difficulty in determining higher order classification in Bloom's Taxonomy. The model had achieved only 59.2% accuracy for first three levels of the taxonomy. Another author showed hallucinations of LLM systems while using it for industry thus proving that the systems are not updated and are not trained on domain specific information. This problem causes factually incorrect and outdated texts being generated from advanced systems [5]. Researchers have acknowledged that LLMs trained on common

datasets results in poor performance when asked about specialized subjects. These limitations are delicate when it comes to educational domain because the models cannot integrate recent developments. This makes them unfit for generating questions on recent topics [6].

Retrieval-Augmented Generation (RAG) is one such promising solution that addresses some of the limitations of traditional LLMs. Veturi et al. [5] demonstrated that RAG could improve the factual accuracy and reduce the hallucination rates when compared to the traditional LLMs. These models have the ability to separate domain- based knowledge storage from generation of text. Apart from that, another research work showcased that these models allow dynamic content uploading without retraining thus allowing new knowledge from articles which are used for generation of questions [3]. Saha et al [6] achieved 82.61% accuracy when switching from traditional models to RAG systems in university information systems while Veturi et al revealed that subjects preferred RAG responses 75% of the time.

RAG systems have already been implemented in various educations applications such as intelligent tutoring systems. These systems use it for Q&A module such as Anatbuddy for medical education and has shown higher accuracy compared to other models [7]. Existing research shows that using RAG systems in Learning Management Systems (LMS) has improved response accuracy and student satisfaction. Laaroussi et al [8] developed BloomTutor that generates questions in real-time but the authors failed to show and compare the quality of questions that are being generated in the system for each Bloom's level. Some research focused in developing RAG models but only used them in generated common MCQ topics without incorporating pedagogical frameworks such as Bloom's taxonomy.

Although RAG has been used in question-answering tasks, most of the existing research focuses on answering queries rather than generating questions. Moreover, the questions that are being generated are not aligned with pedagogical framework making it unfavorable for educators. These frameworks are useful for both educators and learners because it tests the cognitive complexity of the learners. Questions aligned with such frameworks promote cognitive load from recall to higher-order thinking skills such as analysis, creation and evaluation. Elkins et al. [2] incorporated Bloom's taxonomy but only implemented using traditional LLMs. Current RAG systems do not have integration with pedagogical frameworks that could potentially enhance the learning outcomes and make the educators task easier as well.

Although some researchers have shown that these systems have been used for question generation, the exploration of embedding models incorporated in RAG systems are minimal. Additionally, the research on embedding models that generated questions that adhere to Bloom's taxonomy are limited. Pradeesh et al. [7] used FAISS embeddings for the task of generating questions from PDF documents and shown significant improvements compared to traditional text generation methods. Their approach shown that ChatGPT scored better in terms of similarity when compared to manually created questions using BERT, CodeBERT and other such metrics. Yet, the questions that were generated does not follow Bloom' taxonomy. A comprehensive overview of embedding methodologies in RAG systems revealed dense retrieval methods like BERT-based embedding which has dual-encoder frameworks, and hybrid approach like BM25 combined with dense embeddings [9]. This paper showed various embedding models along with their performance, but the paper did not show if it generated educational content which generated domain-specific knowledge that adhere to pedagogical frameworks.

COCOM developed by Rau et al. [10] uses embedding compression techniques that results in the speedup of inference by 5.69 while it maintains the quality of question. However, this method does not explore other embedding models that could possibly affect the meaning that are required for question generation in context-specific field. The work only showed exact match and retrieval accuracy rather than following cognitive level alignment.

Some researchers have implemented course-specific embeddings for CS1 education demonstrating installable small language models with RAG that can utilize multi-layered retrieval dynamic adaptability. But the embedding evaluation is restricted to computational metrics like response time and token generation rate and it doesn't check the effectiveness or the quality of response [11]. The research did not compare different embedding models that can retrieve educational content for different cognitive levels. Casperi et al. [12] carried a similarity analysis of embedding models on different datasets to evaluate the behavior in RAG contexts revealing that these models show high similarity within respective embedding model families and has inter-family

clusters as well. The work showed that models with high representational similarity produces different retrieval results with most used RAG applications being models with low k-values. Although this work emphasizes the complexity of embedding model selection, it lacks in discussion on educational contexts. Alawwad et al. [13] developed a question-answering framework that handles limitations such as lengthy textual contexts by combining RAG with Llama-2 along with supervised fine-tuning and parameter-efficient adaptation techniques. By using the CK12-QA dataset, the work showed some improvements for textual questions.

Salemi & Zamani [14] examined retrieval assessment methods and demonstrated a clear mismatch between traditional information retrieval and RAG systems. A novel method was proposed called eRAG—an evaluation framework that uses LLM performance on retrieved documents as signals that are relevant. This work is useful for researchers seeking efficient evaluation methods for RAG systems. Xu et al. [15] discussed the drawbacks of RAG for customer service by combining knowledge graphs that maintain structural relationships by capturing both intra-issue and inner-issue relationships. The research utilizes hybrid parsing to develop these knowledge graphs and applies embedding-based retrieval for question answering on customer service data from LinkedIn. The study showed improvement over baseline text and BLEU scores emphasizing that domain-specific RAG can be beneficial.

Several researchers attempted to integrate Bloom's taxonomy with AI-powered questions generation systems. MCQ of introductory chemistry and biology courses that aligned with Bloom's taxonomy was generated using GPT-3.5. the work employed zero-shot prompting, automated quality evaluation using Item Writing Flaw(IWF) and RoBERTa-based classifier for taxonomical validation [16]. Although GPT showed strong competency in question generation based on cognitive levels, there were important differences in human and machine quality assessments, highlighting challenges in producing high-quality questions that adheres with pedagogical frameworks. An AI powered chatbot called Bloomify was created to address the limitations such as misalignment between syllabi and examinations. Other drawbacks such as manually creating questions equally for all Bloom's levels and instructors' unfamiliarity with taxonomy implementation [17]. This research work provided an answer that addresses both technical challenges and real-world needs of instructors who wants help in building and executing taxonomy-aligned assessments.

Scaria et al. [18] created concept maps covering 19 units of a subject textbook with extensive information such as units, topics, subtopics and prerequisites using Llama 2.2 70B to generate MCQs across Bloom's taxonomy levels. The work also developed an automatic validation system that evaluates the question correctness. Finally, the authors concluded that the concept map approach had a better success rate compared to RAG and LLM methods. A question generator called SmartTutor using GPT-3.5 Turbo to generate questions that follow Bloom's taxonomy which was developed by Toba et al. [19]. The model extracts nouns from PDF materials and generates questions but only at lower levels of Bloom's taxonomy. They surveyed students and concluded that 77% found that generated questions are easier to understand. However, the system failed to produce high-level questions, leaving the authors to enhance the system for deeper analytical skills that can align with the pedagogical framework.

Text embedding approach proposed by William & Altamimi [20] that combines RAG with LLMs especially with GPT-3 that improves semantic representation of data. It captures explicit knowledge using retrieval methods and undeclared information through generation-based models.

This extensive literature review reveals a clear and certain need for RAG-based systems that can include Bloom's taxonomy thus representing a significant advancement in educational technology. Existing system either focus on traditional LLMS with limited accuracy or RAG systems that does not follow cognitive levels for question generation. This results in a gap that can be addressed by creating a thorough educational question generation system. Current systems like BloomTutor[8] create questions but do not provide quality metrics for each level, while SmartTutor focuses on generating lower-order questions. This study examines embedding models specifically designed for generating questions aligned with Bloom's taxonomy within a RAG pipeline. It also assesses performance across all six standard levels of the taxonomy.

3. METHODOLOGY

This section details the methodology employed in this study, which follows a five-step experimental protocol: (1) data collection and preprocessing, (2) embedding and indexing, (3) retrieval method, (4)

question generation using the T5 model, and (5) evaluation.

This work employs a retrieval-augmented generation (RAG) pipeline that uses open-source embedding models, efficient vector indexing and a T5-based question generator.

The data used to build a knowledge system is first few chapters of "OS concepts" by Abraham Silberschatz and Peter Baer Galvin [21]. This work predominantly uses computer science subjects, utilizing content in PDF format that is usually unstructured with figures and tables.

A manually created dataset called *'questions.json'* was used as the ground-truth dataset. The dataset contains detailed information such as *'context', 'question', 'answer'* and *'bloom_level'* fields. This serves as the standard for evaluating the generated questions by comparing them with traditional metrics and assessing the alignment with pedagogical levels. The performance of the retrieval and generation pipeline was evaluated using this dataset.

The PDFs were preprocessed using *'pypdf'* to extract raw text from each page, followed by splitting the text into paragraphs. This process removes pages where there is no data and joins all text into a large string. The data cleaning process eliminates empty pages, paragraphs, references, headers and footers. Each paragraph becomes one entry in the corpus with paragraph-level chunking serving as the standard RAG implementation for document retrieval.

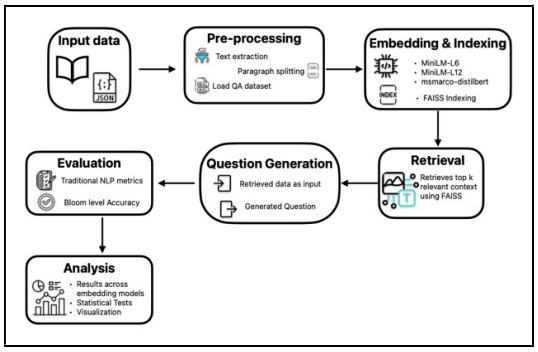

Figure 1: Process of Retrieval-Augmented Question Generation

Research suggests that open-source embedding models such a miniLM-L6, miniLM-L12 and msmarco-distilbert are fine-tuned specifically for multi-domain QA tasks and passage level QA and retrieval tasks [22]. Hence, these models were used for comparison, the size and the dimension of the models are given in the table 1 below. An optimized library for similarity search – FAISS index which enables fast k-nearest neighbor retrieval is used in this research [23]. Each paragraph is embedded and stored in FAISS index. The target text is encoded and used to query the index to retrieve the top-k relevant paragraph. Paragraph-level chunking is standard RAG implementation for document retrieval. The Figure 1 describes the flow of the pipeline in detail for reference.

Table 1: Embedding model card

Models	Size	Dimension
MiniLM-L6	22MB	384
MiniLM-L12	66MB	384
msmarco-distilbert	250MB	768

Finally, T5-small model which are good at sequence-to-sequence generator is employed to the pipeline. The prompt used was *'generate a question based on context: [context] for the level [bloom_level]'*. Below figure 1 represents the flow of the processes this research work has implemented.

This work does not investigate chunking and prompting strategies, alternative vector databases or other open-source models as they are beyond the scope of this study. These limitations may be addressed in the future research work.

3.1 Evaluation Metrics

This research work employs a mix of surface-based n-gram overlap metrics, semantic similarity measures and pedagogical alignment to assess the quality of generated questions.

- Bilingual Evaluation Understudy (BLEU): measures the words being overlapped between generated question and the reference text [24].
- Recall-Oriented Understudy for Gisting Evaluation (ROUGE): which are categorized into three, ROUGE-1 measures word-level overlap whereas ROUGE-L measures the longest common subsequence (LCS) between generated and reference text [25].
- BERTScore: captures the paraphrasing and meaning. It is better suited for natural language generation outputs [26].
- Bloom's taxonomy Accuracy (BloomAcc): a metric introduced in this study to measure of the cognitive level in the generated questions align with levels from the ground-truth dataset.

4. RESULTS AND ANALYSIS

The choice of embedding model substantially influences the quality of texts being generated and their cognitive level alignment leading to improved question generation tasks. This

research work initially compared the three embedding models – MiniLM-L6, MiniLM-L12 and msmarco-distilbert for the retrieval augmented question generation task. These models were evaluated using standard metrics such as BLEU, ROUGE-1, ROUGE-L, BERTScore and Bloom's Taxonomy Accuracy.

The table 2 below shows that MiniLM-L12 has achieved highest scores across most metrics while msmarco-distilbert performs comparably to MiniLM-L12. However, it scored slightly lower on BLEU and BloomAcc. Whereas MiniLM-L6 has scored consistently lower across all the metrics, especially in BloomAcc which suggests that it has weaker alignment with the pedagogical framework

Table 2: Final Metrics Table for Embedding Models

Model	BLEU	ROUGE-1	ROUGE-L	BERTScore	Bloom Acc
MiniLM-L6	0.0227	0.216	0.159	0.851	0.15
MiniLM-L12	0.0230	0.238	0.188	0.852	0.31
Msmarco-distilbert	0.0216	0.237	0.187	0.850	0.30

The figure 2 below is a visualization on the performance of the three models across the traditional metrics. It is evident that MiniLM-L12 consistently has higher score across the metrics indicating balanced performance in both text similarity and alignment with Bloom's taxonomy. Although MiniLM-L6 has scored almost like MiniLM-L12, it is still lower when compared.

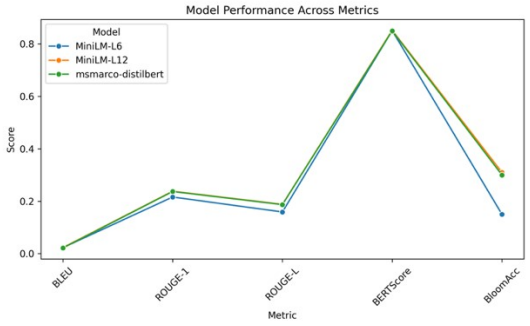

Figure 2: Visualizing the model performance across standard metrics

The paired t-tests indicates that there is no statistically significant difference between these models implying that there is variability across question examples is high and should be interpreted carefully. Despite the lack of statistical significance, the results reveal that embedding model choice influences traditional n-gram metrics and pedagogical relevance thus highlighting the importance of evaluating across multiple dimensions as shown in table 3 below.

Table 3: Paired t-test Results across metrics

Model 1	t-stat	p-value
MiniLM-L6 vs MiniLM-L12	-1.418812	0.228951
MiniLM-L6 vs msmarco-distilbert	-1.393457	0.235918
MiniLM-L12 vs msmarco-distilbert	1.770466	0.151357

Further, this work tested these embedding models on different levels of Bloom's taxonomy using traditional metrics. Across all the levels, MiniLM-L12 demonstrates superior performance on most of the metrics when compared with other models. This model scored higher for *'BloomAcc'* at understand and remember levels while BLEU and ROUGE scores were higher for all levels.

MiniLM-L6 underperforms for higher order levels such as analyze and create whereas msmarco-distilbert has better lexical similarity metrics at remember and create levels. Additionally, paired t-tests were also implemented on all bloom's level resulting in no significant difference. These results indicate that mean performance favors MiniLM-L12 whereas variability across Bloom levels and questions prevents the differences from being statistically significant.

The figure 3 below represents the comparative analysis of all the embedding model performance across Bloom's levels – remember, understand, apply, analyze, evaluate and create. Each line plot represents different evaluation metric. Overall MiniLM-L12 consistently achieves higher scores across most metrics, this indicates balanced performance in textual similarity and pedagogical relevance. The figure highlights the trends across the metrics and blooms levels on each embedding model for educational question generation.

Figure 4 is a heatmap that visualizes the relative ranking of the embedding modes across the pedagogical levels. The heatmap provides a clear, comparative overview of model strengths and weaknesses across the traditional evaluation dimensions. This mode of representation offers insights into each model's suitability across the cognitive levels where 1,2,3 are ranks in the map suggests from best performing model for that specific bloom level to the worst among the three.

These results suggest that MiniLM-L12 as a stable choice for all bloom's level as it shows

Journal of Theoretical and Applied Information Technology
15th April 2026. Vol.104. No.7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

consistently high metrics. This suggests that the model captures both semantic content while it aligns with the pedagogy that can be useful to the educators and learners. The superior model provides strong embedding representation for generating educational questions. It balances semantic meaning of the text while maintaining the cognitive levels as defined by Bloom.

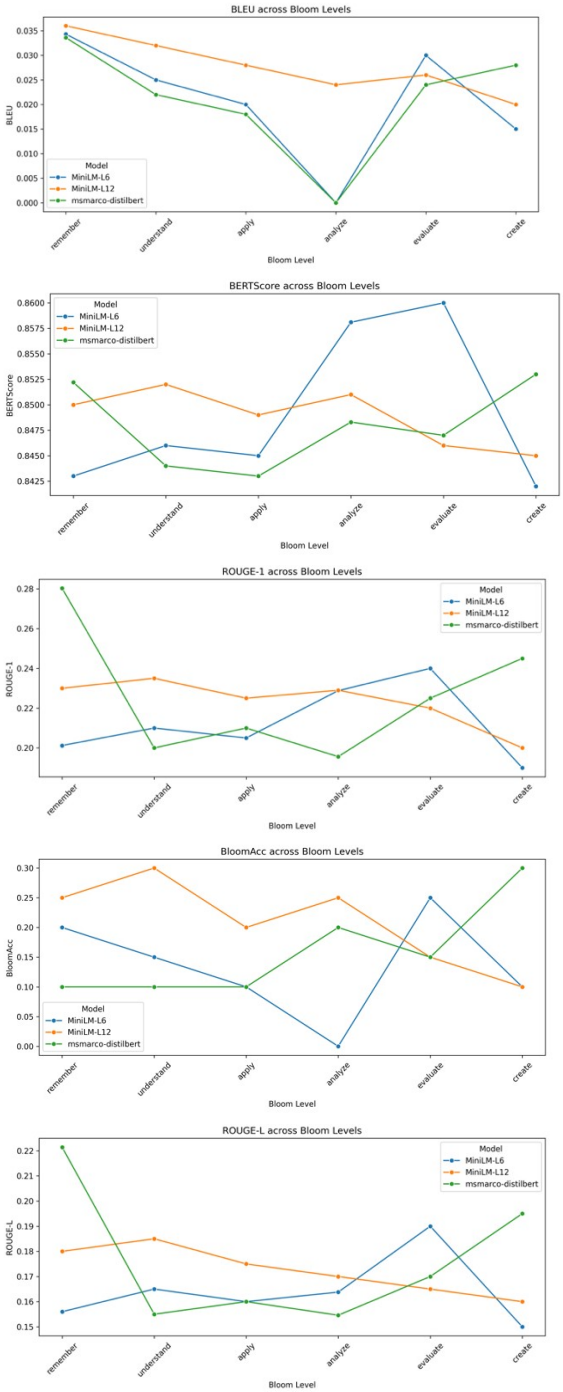

Figure 3: Performance analysis of embedding models across Bloom's levels

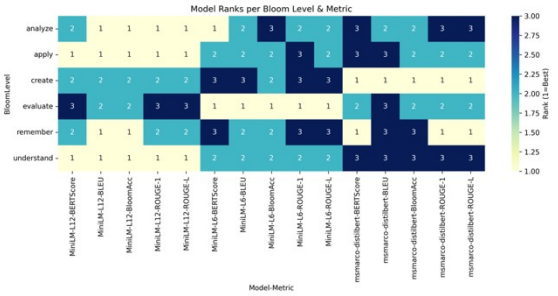

Figure 4: Ranking of embedding models across Bloom's levels

In comparison to the findings of Scaria et al.[18], which indicated improved multiple-choice question (MCQ) success rates through the use of concept maps over RAG, this study demonstrates competitive BloomAcc scores. Notably, this is accomplished without necessitating the creation of structured concept maps. Similarly, while Toba et al.[19]reported a 77% student satisfaction rate, their system was inadequate at addressing higher-order Bloom's levels. Conversely, the results of this research suggest that MiniLM-L12 exhibits superior performance specifically at these higher levels.

The limitations of this research include the use of a single-domain dataset, which restricts the generalizability of the findings. Regardless of retrieval quality, the T5 model may diminish the quality of the generated questions. Furthermore, the standard paragraph-level chunking employed in this study may not represent an optimal strategy, potentially leading to inefficiencies in retrieval and generation quality. Additionally, the 'BloomAcc' metric introduced in this study has not yet undergone external validation.

5. CONCLUSION

This research showcases a significant advancement in educational technological domain by developing a RAG-based model that can generate pedagogically aligned questions. The study emphasizes that embedding model selection plays a crucial role in achieving both semantic accuracy and cognitive alignment. This suggestion offers major advantages for educational practitioners who can leverage this tool to reduce their workload in assessment preparation. The system ensures that these questions maintain pedagogical rigor while generating basic recall levels to higher order thinking levels. Additionally, the RAG based system allows for dynamic content integration that allows educators to integrate current domain-specific knowledge without retraining the entire LLM.

From the student perspective, this system provides opportunities for personalized learning and practice. It supports learning by allowing

students to practice at multiple cognitive levels. This question generation capability guarantees that students can potentially have unlimited practice materials which may reduce the risk of memorization. This framework is particularly useful for self-regulated learners who require such many questions across multiple cognitive levels to learn complex concepts.

Moreover, this work addresses the gap identified through literature review by combining the benefits of RAG with systematic cognitive alignment through Bloom's taxonomy. This is identified as an advancement in creating AI-powered tools that are technically robust and pedagogically aligned. This promises that educators can rely on the tool for assessment creation without cognitive bias or inconsistent question quality.

Future work will focus on expanding the capabilities through larger datasets and fine-tuning approaches to clarify whether the yielded performance trends attain statistical significance. Moreover, developing the system that supports multiple subject domains and integrating real-time feedback will further enhance the educational value. Exploring various fields beyond computer science will further assess the applicability of these results. Employing open-source models other than T5 will help determine if generation constraints hinder the benefits of improved retrieval.

In conclusion, this research demonstrates the incorporation of RAG technology with cognitive complexity levels to create an educational assessment tool that benefit educators and students. This tool can potentially reduce the educator's workload by guaranteeing a solid assessment creation and help the learners with diverse, cognitively appropriate questions. This work not only represents a quality contribution to the development of educational technology but also lays foundation for future development in AI powered question generation in educational settings.

REFERENCES:

[1] E. Prihar, M. Syed, K. Ostrow, S. Shaw, A. Sales, and N. Heffernen, "Exploring Common Trends in Online Educational Experiments," *Proceedings of the 15th International Conference on Educational Data Mining*, Jul. 2022.

[2] S. Elkins, E. Kochmar, J. C. K. Cheung, and I. Serban, "How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes," *AAAI*, vol. 38, no. 21, pp. 23084–23091, Mar. 2024, doi: 10.1609/aaai.v38i21.30353.

[3] C. N. Hang, C. Wei Tan, and P.-D. Yu, "MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning," *IEEE Access*, vol. 12, pp. 102261–102273, 2024, doi: 10.1109/ACCESS.2024.3420709.

[4] J. Zhang, C. Wong, N. Giacaman, and A. Luxton-Reilly, "Automated Classification of Computing Education Questions using Bloom's Taxonomy," in *Proceedings of the 23rd Australasian Computing Education Conference*, in ACE '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 58–65. doi: 10.1145/3441636.3442305.

[5] S. Veturi, S. Vaichal, R. L. Jagadheesh, N. I. Tripto, and N. Yan, "RAG based Question-Answering for Contextual Response Prediction System," Sep. 06, 2024, *arXiv*: arXiv:2409.03708. doi: 10.48550/arXiv.2409.03708.

[6] B. Saha, U. Saha, and M. Zubair Malik, "QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance," *IEEE Access*, vol. 12, pp. 185401–185410, 2024, doi: 10.1109/ACCESS.2024.3513155.

[7] P. N, R. T, M. Thushara, K. A. Krishna, and P. V, "Retrieval-Augmented Generation for Multiple-Choice Questions and Answers Generation," *Procedia Computer Science*, vol. 259, pp. 504–511, 2025, doi: 10.1016/j.procs.2025.03.352.

[8] Y. Laaroussi *et al.*, "BloomTutor: Retrieval Augmentation for Bloom's Taxonomy Question Generation," 2025.

[9] J. Genesis, "Retrieval-Augmented Text Generation: Methods, Challenges, and Applications," Apr. 08, 2025, *Preprints*: 2025040443. doi: 10.20944/preprints202504.0443.v1.

[10] D. Rau, S. Wang, H. Déjean, and S. Clinchant, "Context Embeddings for Efficient Answer Generation in RAG," Oct. 29, 2024, *arXiv*: arXiv:2407.09252. doi: 10.48550/arXiv.2407.09252.

[11] S. Liu, Z. Yu, F. Huang, Y. Bulbulia, A. Bergen, and M. Liut, "Can Small Language Models With Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science?," in *Proceedings of the 2024 on Innovation and*

Technology in Computer Science Education V. 1, Milan Italy: ACM, Jul. 2024, pp. 388–393. doi: 10.1145/3649217.3653554.

[12] L. Caspari, K. G. Dastidar, S. Zerhoudi, J. Mitrovic, and M. Granitzer, "Beyond Benchmarks: Evaluating Embedding Model Similarity for Retrieval Augmented Generation Systems," Jul. 11, 2024, *arXiv*: arXiv:2407.08275. doi: 10.48550/arXiv.2407.08275.

[13] H. A. Alawwad, A. Alhothali, U. Naseem, A. Alkhathlan, and A. Jamal, "Enhancing textual textbook question answering with large language models and retrieval augmented generation," *Pattern Recognition*, vol. 162, p. 111332, Jun. 2025, doi: 10.1016/j.patcog.2024.111332.

[14] A. Salemi and H. Zamani, "Evaluating Retrieval Quality in Retrieval-Augmented Generation," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '24. New York, NY, USA: Association for Computing Machinery, Jul. 2024, pp. 2395–2400. doi: 10.1145/3626772.3657957.

[15] Z. Xu *et al.*, "Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington DC USA: ACM, Jul. 2024, pp. 2905–2909. doi: 10.1145/3626772.3661370.

[16] K. Hwang, S. Challagundla, M. M. Alomair, L. K. Chen, and F.-S. Choa, "Towards AI-Assisted Multiple Choice Question Generation and Quality Evaluation at Scale: Aligning with Bloom's Taxonomy," *Workshop on Generative AI for Education*, Dec. 2023.

[17] M. Atre, S. Karandikar, K. A. Merchant, A. Suryawanshi, and H. Patil, "Revolutionizing Educational Assessment Using Bloom's Taxonomy Bot," *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS*, vol. 22, pp. 593–603, Jul. 2025, doi: 10.37394/23209.2025.22.49.

[18] N. Scaria, S. J. J. Kennedy, D. Seth, A. Thakur, and D. Subramani, "Harnessing Structured Knowledge: A Concept Map-Based Approach for High-Quality Multiple Choice Question Generation with Effective Distractors," May 02, 2025, *arXiv*: arXiv:2505.02850. doi: 10.48550/arXiv.2505.02850.

[19] H. Toba, L. G. O. P. Yudha, O. Karnalim, H. Bunyamin, and T. Tada, "A Large Language Model Question Generator Based on Bloom's Taxonomy Template," in *Proceeding of the 2024 8th International Conference on Education and E-Learning*, Tokyo Japan: ACM, Nov. 2024, pp. 25–31. doi: 10.1145/3719487.3719537.

[20] I. O. William and M. Altamimi, "Text Embedding Implementation Using Retrieval Augmented Generation (RAG) Model Combined With Large Language Model," *International Journal of Advanced Natural Sciences and Engineering Researches*, 2024.

[21] A. Silberschatz, P. B. Galvin, and G. Gagne, "Operating System Concepts," 2018.

[22] R. Sajja, Y. Sermet, and I. Demir, "Domain-Specific Embedding Models for Hydrology and Environmental Sciences: Enhancing Semantic Retrieval and Question Answering in RAG Pipelines," Jul. 2025, Accessed: Sep. 10, 2025. [Online]. Available: https://eartharxiv.org/repository/view/9635/

[23] A. Rezai and S. Hasan, "LLM: Retreival vs. Parametric Memory Tradeoff," 2025.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[25] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Sep. 10, 2025. [Online]. Available: https://aclanthology.org/W04-1013/

[26] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.