

Revolutionizing Educational Assessment Using Bloom's Taxonomy Bot

MINAKSHI ATRE¹, SARTHAK KARANDIKAR¹, KABEER AHMED MERCHANT¹,
ABHIJEET SURYAWANSHI¹, HERAMB PATIL²

¹Department of Artificial Intelligence and Data Science

²Department of Electronics and Telecommunication

Pune Vidyarthi Griha's COET and GKPIM

44, Shiv Darshan Rd, Vidya Nagari, Parvati Paytha, Pune, Maharashtra 411009

INDIA

Abstract: The authors provide a transformative solution to address prevalent challenges in educational assessment in this research work. These challenges include aligning examination papers with syllabi, framing question papers to cover all Bloom's Levels, and upgrading Bloom's Levels by re-framing questions. Bloom's Taxonomy (BT) is a cognitive level-based framework for understanding students' educational progress through the assessment. The educational assessment method is the foundation for our chatbot, Bloomify. It is designed to revolutionize the assessment and help the keen educationists in three key scenarios. In the first scenario, mostly the educationists are not able to reflect the all the course outcomes of the subject. Bloomify will help them to design the question paper to cover all the course outcomes mapped with Bloom's levels (BL). Second, often in a hurry, juggling between students' assessment and institution's accreditation tasks, the educationists fail to balance the question papers with all BLs covered. Bloomify will overcome the human fatigue and exhaustion associated with the manual creation of balanced question papers. In the third scenario, it is observed that many of the universities use of taxonomies and hence fail to design the assessment for the cognitive development of the students. Bloomify helps the universities that lack taxonomy or use alternative taxonomies, and offers a Bloom's approach for assessing student development. Further, authors define a Bloom's Score (BS) to find the average cognitive level of a question paper. BS will help to focus on specific cognitive level of an assessment. For example, the assessment for the graduate students can be defined with BS of 5 or 6 indicating higher Bloom's levels: evaluate and create. Bloomify automates the calculation of Bloom's scores ensuring coverage of all the cognitive levels. This will help the educationists in question paper creation process, saving time and effort. Bloomify includes three features: 1. Question classification 2. Suggestion of BL, and 3. Generation of entire question paper based on specific criteria like the syllabus, marking scheme, and desired average Bloom's score. Bloomify provides a user-friendly interface, ensuring seamless integration with the assessment creation process. Bloomify enhances educational assessment by addressing all cognitive levels, ensuring critical thinking in students. It supports the educationists' efforts to design an assessment focusing on knowledge application, analysis, evaluation, and creation, aligning with the real-world demands. The effectiveness of Bloomify is gauged by allowing different educationists to use this chatbot and calculating the Mean Opinion Score (MoS). Bloomify received scores of 9, 9.5, and 8 for classification, suggestion, and generation features, respectively from the educators of the authors' institute.

Key-Words: Educational Assessment, Cognitive Skills, Large Language Models (LLMs), Bloom's Taxonomy, Chatbots, Cognitive Assessment

Received: August 15, 2024. Revised: May 14, 2025. Accepted: June 11, 2025. Published: July 29, 2025.

1 Introduction

Educational assessments are necessary for understanding student performance. The fundamental purpose of higher education is to enhance the cognitive skills of students to make them employable. Improving educational assessment with AI could improve employability by fostering critical thinking in students. This enhancement of cognitive skills is important. As highlighted in past research, fostering problem-solving abilities, also among students less inclined towards mathematics, is crucial for overall educational improvement [1]. Benjamin Bloom proposed *Bloom's Taxonomy*

(*BT*) in 1956, along with his colleagues. This introduced a framework in hierarchical way to classify educational objectives and learning outcomes based on cognitive complexity. *Bloom's Taxonomy (BT)* makes categories of cognitive skills into six levels, arranged in increasing order of cognitive complexity. In 2001, *Anderson and Krathwohl* published an updated version of Bloom's Taxonomy. The revision was based on the use of action verbs instead of nouns, making it more action-oriented [2]. A comparison of the two taxonomies is depicted in Fig. 1. The revision also prioritized creativity by placing 'Create' at the top level and emphasized

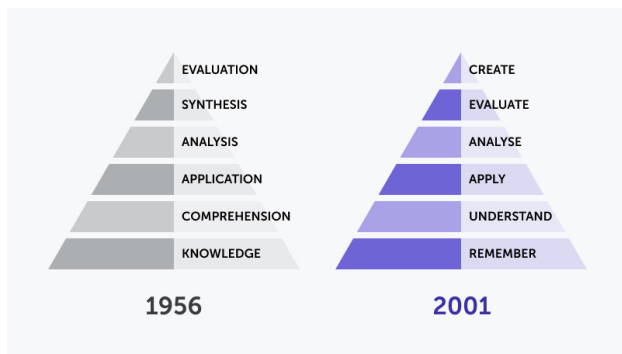


Figure 1: Comparison of the original vs. revised Bloom's Taxonomies, [4]

the interconnectedness of knowledge and skills in learning. The application of Bloom's Taxonomy in educational systems has done significant progress particularly with the incorporation of artificial intelligence (AI). For example, Automatic Question Generation (AQG) systems generate questions based on a topic or some context provided in the form of a paragraph of text or a series of images [3]. These systems use Bloom's Taxonomy to create questions spanning different cognitive levels, from lower levels of recall to higher levels of analysis and creation. This development has provided invaluable support for educators in assessing and promoting *higher-order thinking skills (HOTS)*.

The recent emergence of AI-driven tools, such as ChatGPT, has significantly impacted the educational landscape. The effectiveness of Large Language Models (LLMs) for chatbot interaction, particularly those powered by deep learning, is increasingly recognized in a comprehensive study of ChatGLM,[?]. The debate surrounding the restrictions on ChatGPT usage in academia balances the potential benefits with concerns about academic integrity, [6]. AI models enhance learning by giving personalized feedback and fostering critical thinking. Advances in AI, specially in deep learning, have played a key role in the development of chatbot technology, as explored in the analysis of ChatGLM, [5]. The opportunities and challenges that chatbots present in education are significant and can certainly be leveraged constructively, [7]. They can be integrated with Bloom's Taxonomy to guide students through various cognitive levels, thereby creating a more structured and effective learning experience.

The work of several researchers has been reviewed and analyzed. Researchers have worked in the areas of Automatic Question Generation (AQG), [8], using Large Language Models (LLMs) for setting Bloom's Taxonomy-based psychosomatic medicine exam questions, [9], and assessing students'

performance with BT-based educational quizzes, [10]. In [11], they assess the usefulness of questions generated by LLMs. The classification of academic questions into their respective Bloom's Taxonomy levels using transformer-based approaches in [12], classifier-based approaches with Support Vector Machines (SVM), Naive Bayes (NB) Classifier, and k-Nearest Neighbors (k-NN) in [13], deep learning [14], Term Frequency-Inverse Document Frequency (TF-IDF) in [15], and rule-based approaches that focus on verbs used in questions in [16], have been reviewed for a variety of applications based on LLMs and LLMs based on BT. The process employed for the suggestion and generation features of Bloomify using Quantized Low-Rank Adapters (QLoRA) is novel. Thus, the research team plans to make the classification feature open-source while pursuing process patents for the methods employed in the areas of suggestion and generation.

The work proposed and implemented by the authors showcases the effective use of large language models to gauge the cognitive level of students by designing question papers based on the revised Bloom's Taxonomy (BT). Our solution, Bloomify, uniquely integrates the Mistral 7B large language model with QLoRA for efficient fine-tuning, enabling advanced question transformation and generation capabilities that set it apart from existing AQG systems. The benefits of generative AI, such as ChatGPT, in education also include democratizing access to knowledge and providing continuous learning support, [17]. This promotes both independent learning and creative problem-solving, which are the pillars of the revised Bloom's Taxonomy. The integration of generative AI in education, especially in programming education, is a growing trend. With various directions and approaches being actively explored, the research landscape is increasingly focused on leveraging AI in education, [18]. Ethical considerations are also addressed by sourcing data responsibly from Savitribai Phule Pune University (SPPU) (the authors belong to this university), ensuring data harmlessness and integrity, and considering the environmental impact of AI development. Open-source principles are embraced to promote transparency and collaboration. These enhancements are crucial as the field of generative AI in education evolves, with diverse directions being explored to maximize its impact, as noted in the review by [18].

2 Background

To get a better idea about the present work, it is helpful to first consider the levels of Bloom's Taxonomy (BT). We will discuss the levels here and also mention the older taxonomy terms, as some of them are still

relevant.

1. Remember (Knowledge): This is the fundamental level. It focuses on recalling information such as basic facts and concepts. Learners demonstrate this by recognizing, listing, or describing important elements they have previously encountered.

2. Understand (Comprehension): This level goes above recall. This level focuses on grasping the meaning of the material. Learners must effectively explain ideas, summarize concepts accurately, or interpret information in their own words, hence demonstrating their understanding of the information.

3. Apply (Application): This is where knowledge is put into practice. At the application level, learners apply what they have learned to solve problems, perform procedures, or implement concepts in real (or simulated) contexts.

4. Analyze (Analysis): This is a higher-order level of thinking. Analysis requires breaking down information into its component parts, identifying underlying patterns, identifying relationships between components, or understanding organizational structures within the presented material.

5. Evaluate (Evaluation): This level involves making informed judgments—assessing the value or effectiveness of information, methods, or potential solutions. Learners in this stage must be able to critically examine arguments and justify their viewpoints using defined criteria or standards.

6. Create (Synthesis): Finally, at the apex of the taxonomy, we find "create." This level involves the synthesis of knowledge and skills, combining different elements to develop completely new ideas, products, or solutions. For example, designing experiments, writing original papers, or inventing novel approaches based on the skills gained in the previous levels.

Fig. 2 neatly summarizes the types of action verbs typically associated with each Bloom level. In fact, understanding the levels often depends on interpreting these verbs, especially when they appear in assessment questions or learning objectives. It is this set of action verbs, as we will discuss in more detail in Sections 3 and 4, that formed a crucial part of the training data for our language models.

Bloom's Taxonomy fundamentally offers educators more than just categories; it provides a robust framework. They can use it to formulate clear educational objectives and create assessments that specifically address a full range of cognitive abilities. The framework inherently promotes a

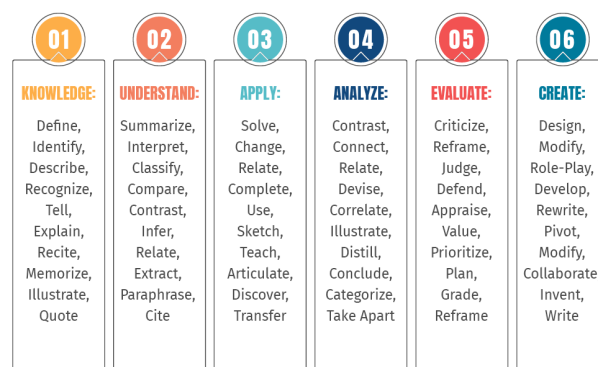


Figure 2: List of Bloom's Taxonomy Verbs (TeachThought)

more holistic view of learning and encourages the development of critical thinking, problem-solving skills, and creativity in students. Given its widespread application in curriculum development, instructional design, and assessment practice at virtually all educational levels and across all disciplines, it seemed a natural fit for our project. This ubiquitous utility is precisely why the authors chose to develop a Large Language Model (LLM) based on BT. The goal? To support educators in structuring their questionnaires to consciously address all cognitive levels required by the taxonomy. We believe this will be a real time saver for educators, potentially allowing them to focus more on innovative teaching and learning methods and their own research, rather than the often tedious task of creating BT-compliant questionnaires.

3 Defining the Problem

The authors have focused on three main problems:

1. **Misalignment between syllabi and question papers:** The first problem arises when universities base their syllabi on Bloom's taxonomy but fail to reflect this in their question papers. This discrepancy often leads to a dominance of rote learning, as students prioritize memorizing information over understanding and applying it. The bot will assist the educator in designing the BT-based test or exam papers. It will effectively map with the syllabus content and marking scheme.

2. **Difficulties in the manual creation of balanced question papers:** Many universities, particularly higher education institutions, adopt Bloom's taxonomy for the cognitive development of students. Professors strive to create balanced question papers that cover all BT levels, with a focus on the application level. The manual process of calculating

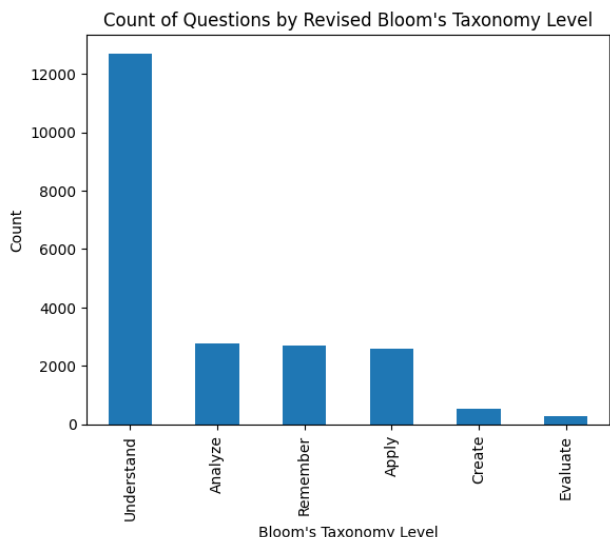


Figure 3: Distribution of Questions by BT Level

Bloom's scores, designing marking schemes, and ensuring a balanced distribution of cognitive levels is time-consuming and challenging. The bot will help rephrase questions and upgrade the BT levels.

3. Naïve approach to BT levels: The third problem pertains to naïve educationists who are unfamiliar with BT verbs and are unable to design questions using the correct BT-level verbs. The 'classify' feature will assist them by allowing them to ask the bot questions and check the corresponding BT level.

The above problems are tackled with three features of the bot named Classify, Suggest, and Generate. The authors believe that educational assessment should focus on the cognitive development of students. If BT-based exams are not promoted, the overemphasis on recall-oriented questions can lead to several drawbacks, such as the dominance of rote learning, inhibition of critical thinking, lack of practical application of theory, diminished problem-solving skills, and limited focus on higher-order skills.

3.1 Data Analysis

The authors analyzed engineering question papers set by Savitribai Phule Pune University (SPPU), identifying that most of the questions are concentrated at lower Bloom's levels (cognitive levels) according to Bloom's taxonomy, [16]. They extracted and examined over **21,000 questions** to assess their taxonomy levels. Fig. 3, Fig. 4, and Fig. 5 illustrate the findings of this analysis.

The observations indicate that there are around 12,000 questions based on a very primitive cognitive level, known as 'Understand'. Very few questions

Proportion of Questions by Bloom's Taxonomy Level

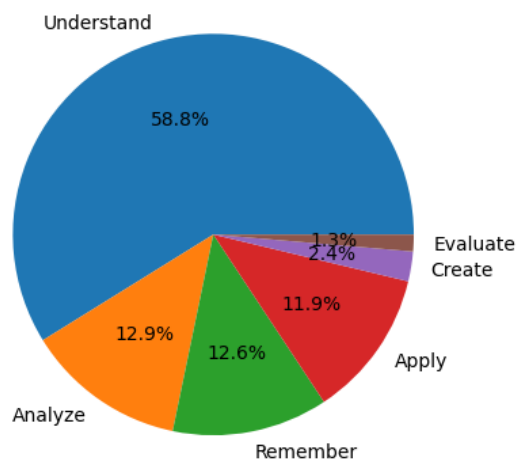


Figure 4: Proportion of Questions by BT Level

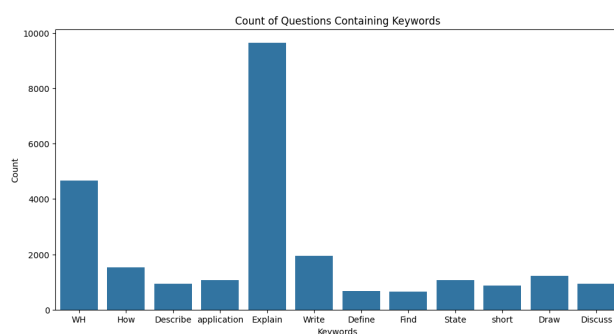


Figure 5: Depiction of the BT Verbs used in SPPU assessments

```

    ### Input:
    Question: Why might a proposed solution for a real
    -world engineering problem be more effective than
    others?

    ### Response:
    Bloom's Taxonomy Level: Evaluate
    
```

Figure 6: Classification Snapshot

```

    ### Input:
    Question: Calculate the bending moment on a beam subjected to a point
    load

    Current Bloom's Taxonomy Level: Apply
    Desired Bloom's Taxonomy Level: Understand

    ### Response:
    Transformed Question: Explain the concept of bending moment on a beam
    subjected to a point load.

    ### Input:
    Question: Explain the concept of bending moment on a beam subjected
    
```

Figure 7: Suggestion Snapshot

```

    ### Input:
    Syllabus: 3 Robot Kinematics - Translation and Rotation representation,
    Coordinate Transformation, Denavit Hartenberg parameters, Forward and
    Inverse Dynamics, Equation of Motion using Euler - Lagrange Formulation
    and Newton - Euler Formulation

    Desired Bloom's Taxonomy Level: Analyze

    ### Response:
    Generated Question: Compare and contrast the advantages and disadvantages
    of using a centralized vs. decentralized approach for robot control.

    ### Instruction
    -----
    
```

Figure 8: Generation Snapshot

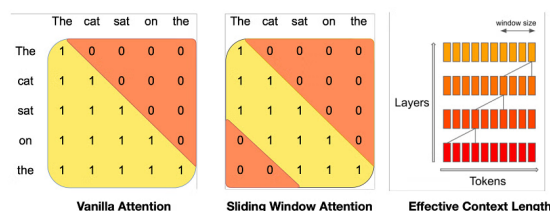


Figure 9: Mistral Architecture Details

tap into the highest cognitive level of students. The bar graph and the pie chart present the reality of the higher education examination and assessment landscape. The time-saving technological solution is presented in the form of the BT-based bot to help educators classify, suggest, and generate questions according to Bloom’s Taxonomy.

4 Presenting the Solution

A large language model (LLM)-based chatbot named “Bloomify” is designed to address the problems outlined in Section 3. Bloomify offers three key features:

1. **Classification:** Accurately classifies academic questions into specific levels of Bloom’s Taxonomy using NLP techniques.
2. **Suggestion:** Suggests diverse ways of representing questions at different cognitive levels, empowering educators to design assessments that encourage higher-order thinking skills.
3. **Generation:** Generates entire question paper based on specific criteria, such as syllabus, marking scheme, and desired average Bloom’s score.

Fig. 6, Fig. 7, and Fig. 8 display the user interface of the bot, providing classification, suggestion, and generation features for the paper setter.

Bloomify leverages large language model-based techniques with Mistral 7B, [19], to achieve its functionalities. Bloom’s Taxonomy of a question can

be identified with specific keywords that serve as important discriminators for the large language model for all three features.

5 LLM-Based Methodology

As mentioned above, the Mistral 7B-based Bloomify bot is designed with three functions from an educator’s perspective: a) classification, b) suggestion, and c) generation. A detailed discussion follows, including the architectures for questionnaire creation and the training process. The results are discussed in terms of LLM performance and the LLM parameters. User feedback is summarized using the Mean Opinion Score (MOS) as shown in Table 1. Fig. 9 illustrates the architecture of the Mistral 7B model (with a simple example sentence): “The cat sat on the mat.” Subsections 5.1, 5.2, and 5.3 provide detailed discussions of the bot’s functions.

To select the baseline model, the authors evaluated several 7B-parameter LLMs as shown in Table 2 (Appendix). Mistral 7B was chosen due to its superior performance in several benchmarks. The fine-tuning process using the Low-Rank Adaptation (LoRA) approach with training loss curves is then explained. Model inference was compared before and after fine-tuning to demonstrate significant improvements.

Table 2 (Appendix) compares the performance of six language models in five benchmarks: logical reasoning, knowledge, reading comprehension, mathematical reasoning, and code generation. Each score represents the performance of models in the respective task, with higher scores indicating

better performance. These scores are derived from standardized evaluation metrics such as accuracy, exact match, F1 score, BLEU score, and pass rate. The benchmarks assess various linguistic and cognitive abilities of the language models:

1. **Logical Reasoning:** Evaluates the model's ability to apply logical reasoning and understand everyday situations.
2. **Knowledge:** Measures the model's ability to retrieve and apply facts from a wide range of domains.
3. **Reading Comprehension:** Tests the model's ability to understand and answer questions based on given text passages.
4. **Mathematical Reasoning:** Assesses the model's competence in solving mathematical problems and understanding mathematical concepts.
5. **Code Generation:** Assesses the model's ability to generate code snippets or programs that fulfill specific tasks or requirements.

5.1 Classification

Previously, researchers have used various NLP techniques and machine learning algorithms, such as random forests and support vector machines with term frequency inverse document frequency (TF-IDF) and word-to-vector embeddings [20]. Deep learning models, including CNNs [21], RNNs and LSTM networks [22], have also been tested for classification of questions [13], [15]. Transformer-based approaches, such as bidirectional encoder representations from transformers (BERT) [23], and large language models such as Mistral 7B, have been investigated by other researchers [12]. These studies highlighted the effectiveness of large language models (LLMs), especially when optimized for classification tasks.

After doing the survey over the performance of these models, the authors adopted a revised yet novel LLM-focused approach. A dataset of 1200 questions paired with their taxonomy levels was used to fine-tune Mistral 7B using Quantized Low-Rank Adaptation (QLoRA), [24]. The classification feature is implemented using the fine-tuned Mistral 7B model.

To enhance classification performance, a LoRA approach was adopted with the following hyperparameters:

- LoRA rank (r): 16
- LoRA alpha (α): 16
- LoRA dropout: 0.05



Figure 10: Training loss curve for the classification feature

These parameters determined based on empirical studies showing that a rank of 16 balances model capacity and efficiency, an alpha of 16 ensures stable adapter scaling, and a dropout of 0.05 mitigates overfitting. The training process utilized a dataset of 600 questions, and the resulting training loss curve (Fig. 10) shows effective convergence.

5.2 Suggestion

The suggestion feature is the second key component of the bot, designed to assist educators in transforming questions across different Bloom's Taxonomy levels. To address memory constraints in fine-tuning large LLMs, authors further employed QLoRA, [24], which uses block-wise k-bit quantization and learnable low-rank adapters. The model is fine-tuned on a 600-question dataset using QLoRA's paged training strategy. Mistral 7B, [25], is selected for its superior performance and adherence to the 7B parameter limit, unlike other models such as Gemma 7B, which exceeded this range.

For fine-tuning, again LoRA is used with the same hyperparameters as in classification:

- LoRA rank (r): 16
- LoRA alpha (α): 16
- LoRA dropout: 0.05

These settings are consistently applied across features for uniformity and proven effectiveness. The training loss curve for the suggestion feature (Fig. 11) indicates successful model fitting.

To highlight the improvement from fine-tuning, the model's output is compared before and after fine-tuning for the suggestion task:

Base Model Response (Without Fine-Tuning):
"Given the provided syllabus on Information Security, explain the concept of Security Policy and its significance in ensuring effective information security."

Fine-Tuned Model Response: *"Transformed Question: Analyze the strengths and weaknesses*



Figure 11: Training loss curve for the suggestion feature

of a proposed solution for a tangible engineering predicament.”

The fine-tuned output demonstrates a more precise transformation aligned with the desired Bloom’s Taxonomy level, showcasing the efficacy of the fine-tuning process.

5.3 Generation

Mistral 7B is also employed for the generation feature, which creates questions based on specific syllabus topics and desired taxonomy levels. This feature leverages the model’s widespread deployment suitability, [8], [11]. To ensure a balanced distribution of questions across Bloom’s Taxonomy levels, the marking scheme is incorporated into the generation process. Advanced language models like GPT-4 or Gemini are used to generate initial questions and then used again to fine-tune Mistral 7B, enhancing its question-generation capabilities. The syllabus is embedded as vectors, and QLoRA is used again for fine-tuning.

The LoRA hyperparameters remained consistent:

- LoRA rank (r): 16
- LoRA alpha (α): 16
- LoRA dropout: 0.05

The training loss curve (Fig. 12) confirms the model’s successful convergence for this feature.

A significant improvement is highlighted by comparing the model’s output before and after fine-tuning for the generation task:

Input: Syllabus: Basic probability: Discrete and continuous random variables, independence, covariance, central limit theorem, Chebyshev inequality, diverse continuous and discrete distributions. Statistics, Parameter Estimation, and Fitting a Distribution: Descriptive statistics, graphical statistics, method of moments, maximum likelihood estimation, Data Modeling Concepts, Understand and model subtypes and supertypes,

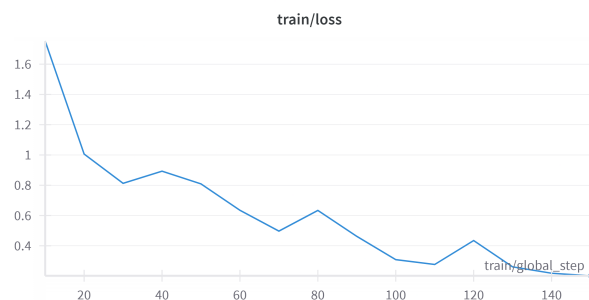


Figure 12: Training loss curve for the generation feature

Understand and model hierarchical data, Understand and model recursive relationships, Understand and model historical data

Base Model Response (Without Fine-Tuning):

“Given the provided syllabus, explain diverse continuous and discrete contributions.”

Fine-Tuned Model Response:

Generated Question: Evaluate the suitability of the Chebyshev inequality for a given data set, considering the distribution of the data and the desired confidence level.

Generated Question: Evaluate the effectiveness of the method of moments for estimating parameters in a given distribution, comparing it to other estimation methods.

Generated Question: Evaluate the appropriateness of using a specific distribution model for a given data set, considering the characteristics of the data and the assumptions of the distribution.

5.4 Workflow

Fig. 13 (Appendix) illustrates the overall Bloomify workflow, integrating the classification, suggestion, and generation features. The fine-tuning process significantly enhanced the model’s performance across all three features.

6 Discussion

This section is focused on dataset description, creation and preparation. The results based on performance parameters and subjective test, known as, mean opinion score are presented in tabular form.

6.1 Dataset Description

Table 3 (Appendix) provides an overview of the datasets used for evaluating Bloomify’s classification, suggestion, and generation capabilities.

6.2 Dataset Creation Process

1. **Synthetic Question Generation:** All questions are generated synthetically using LLMs. By

carefully prompting these models with specific topics, keywords, and desired BT levels, a diverse set of questions is created.

2. Bloom’s Taxonomy Classification: Bloom’s Taxonomy levels are assigned synthetically to the generated questions using specifically designed algorithms and rules. Various linguistic features and patterns associated with different cognitive levels are taken into consideration, ensuring accurate classification.

3. Data Augmentation (Suggestion and Generation Data): The suggestion dataset is created by further synthetically transforming the generated questions into different Bloom’s Taxonomy levels. This involved rephrasing, adjusting the scope and complexity of the questions, and applying predefined transformation rules. The generation dataset is constructed by focusing on specific topics and generating questions tailored to desired Bloom’s Taxonomy levels. This process used LLMs and predefined templates to ensure diversity and relevance.

4. Dataset Splitting: The final datasets are typically divided into training, validation, and test sets to use for the model development and robust evaluation of Bloomify’s performance on unseen data. Specially, for the classification task, the dataset of 1200 questions is split into 600 questions for training and 600 for testing, with an equal distribution of taxonomy levels in each set.

By creating the datasets completely synthetically, several significant benefits are achieved. First, a high degree of control is maintained over the types and distribution of questions within each dataset, enabling controlled experimentation. This control is important for systematically testing specific hypotheses and ensuring the consistency of experimental conditions. Second, synthetic data helped mitigate potential biases that might arise from using real-world data. By designing the datasets themselves, authors avoid inheriting bias present in natural datasets, leading to fairer and more objective evaluations of the models. Lastly, synthetic data offered exceptional scalability, allowing the authors to create large and diverse datasets that support the development and evaluation of robust AI models. This scalability is essential for training models on a wide variety of scenarios, ultimately enhancing their generalization capabilities.

6.3 Mean Opinion Score (MOS)

The Mean Opinion Score (MOS) is a metric used to assess the quality of outputs produced by machine learning models, particularly in tasks where human

judgment plays a key role. It is calculated based on subjective assessments provided by human evaluators, who rate the quality on a predefined scale, typically ranging from 1 (poor) to 5 or 10 (excellent).

In this study, MOS is employed to evaluate the classification, suggestion, and generation capabilities of the bot, Bloomify. A panel of experienced educators provided ratings for a set of outputs from Bloomify, considering factors such as accuracy, clarity, and relevance. Accuracy is assessed based on how well the output aligns with the intended Bloom’s Taxonomy level or the desired question characteristics. Clarity is evaluated by considering the readability, conciseness, and grammatical correctness of the generated questions. Relevance is judged by examining the appropriateness of the questions to the specified topic and their potential educational value. Table 3 presents the MOS results. The high MOS values indicate a strong positive reception from educators. Despite the high MOS scores, educators provided valuable feedback for further improvement. They requested the addition of features for generating mathematical and coding questions, as well as the ability to create multiple-choice question (MCQ) papers.

Feature	Mean Opinion Score (MOS)
Classification	9
Suggestion	9.5
Generation	8

Table 1: Mean Opinion Score (MOS) of three features

6.4 Performance

For the classification feature, a dataset of 600 referential questions (100 per taxonomy level) is used. Mistral 7B with fine-tuning yielded the most promising results. For the suggestion feature, Bloomify transformed questions into different taxonomy levels, and subject matter experts manually labeled the transformed questions to verify their accuracy. For the generation feature, a “Bloom’s Score” is defined as the average taxonomy level of questions within a generated paper, where each taxonomy level is assigned a numerical value from 1 (Remember) to 6 (Create). A score of 3.5 was considered ideal for a balanced mid-term paper. However, the current implementation of Bloomify faces challenges in generating and classifying questions that involve mathematical formulae or coding, common in all STEM subjects. These limitations are planned to be addressed in future work.

7 Conclusion

Educationists often need help to set papers which test the students' cognitive levels and promote their critical thinking. Bloomify's novel integration of Mistral 7B with QLoRA for question transformation distinguishes it from prior AQC systems, offering a unique solution to these challenges. In the world of GenAI, LLMs are helping to automate such time-consuming tasks and enhance the deliverables of educationists. They need help on all three aspects of evaluation: paper setting, BT-based classification and Bloom level transposition. Bloomify presents a transformative solution for educational assessment, addressing key challenges in aligning syllabi and question papers with Bloom's Taxonomy. The chatbot's classification, suggestion, and generation features, powered by advanced NLP techniques, empower educators to create more effective and engaging assessments. Bloomify is further tested to generate mathematical question-papers and succeeded with an MOS score of 9.

8 Future Work

An enhancement of the suggestion feature is underway by integrating Retrieval-Augmented Generation (RAG), [26], to improve the model's ability to navigate the academic syllabus of the university for which assessments are being created without requiring manual inputs from educators. This involves organizing engineering syllabi into individual documents and generating embeddings to capture subject-specific nuances. The bot must evolve to enhance Bloomify's ability to generate and classify numerical questions while addressing the unique challenges posed by numerical data. A plan to leverage stable diffusion models to enrich the learning experience by generating relevant diagrams alongside textual questions is also being discussed. In summary, a versatile Bloomify bot is a priority for the researchers.

Declaration Of Generative AI & AI-assisted Technologies In The Writing Process

During the preparation of this "work" the authors used Grammarly and Gemini in order to avoid grammatical errors, and to improve readability and language. After using these tools/services, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

- [1] Benecke, Klaus. "Comments on School Education in Mathematics and Computer

Science." WSEAS Transactions on Information Science and Applications 22 (2025): 134-145.

- [2] Krathwohl, David R., A revision of Bloom's taxonomy: An overview, *Theory into Practice*, Vol.41, No.4, 2002, pp. 212-218.
- [3] Mulla, N., Gharpure, P., Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications, *Progress in Artificial Intelligence*, Vol.12, No.1, 2023, pp. 1-32.
- [4] Andreev, Ivaan, Comparison image of the original vs. revised Bloom's Taxonomies, *Valamis*, 2022.
- [5] Zeng, Zijian, and Kurunathan Ratnavelu. "Deep Learning-driven Enhancement of Chatbot Interaction: A Comprehensive Study on ChatGLM." WSEAS Transactions on Computer Research 12 (2024): 377-383.
- [6] Yu, Hao, Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching, *Frontiers in Psychology*, Vol.14, No.1, 2023: 1181712.
- [7] Hwang, G. J., Chang, C. Y., A review of opportunities and challenges of chatbots in education, *Interactive Learning Environments*, Vol.31, No.7, 2023, pp. 4099-4112.
- [8] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, Salam Al-Emari, A systematic review of automatic question generation for educational purposes, *International Journal of Artificial Intelligence in Education*, Vol.30, No.1, 2020, pp. 121-204.
- [9] Herrmann-Werner, Anne, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, and Moritz Mahling. "Assessing ChatGPT's Mastery of Bloom's Taxonomy using psychosomatic medicine exam questions: mixed-methods study." *Journal of Medical Internet Research* 26, no. 1 (2024): e52113.
- [10] Elkins, Sabina, Ekaterina Kochmar, Jackie C. K. Cheung, and Iulian Serban. "How Teachers Can Use Large Language Models and Bloom's Taxonomy to Create Educational Quizzes." *arXiv preprint arXiv:2401.05914* (2024).
- [11] Elkins, Sabina, Ekaterina Kochmar, Iulian Serban, and Jackie CK Cheung. "How useful are educational questions generated by large language models?." *International Conference on Artificial Intelligence in Education*. 2023.

- [12] Waheed, Abdul, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. "BloomNet: A robust transformer-based model for Bloom's learning outcome classification." *arXiv preprint arXiv:2108.07249* (2021).
- [13] Abduljabbar, Dhuha Abdulhadi, and Nazlia Omar. "Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination." *Journal of Theoretical and Applied Information Technology* 78, no. 3 (2015): 447-455.
- [14] Laddha, Manjushree D., Varsha T. Lokare, Arvind W. Kiwelekar, and Laxman D. Netak. "Classifications of the summative assessment for revised Bloom's taxonomy by using deep learning." *arXiv preprint arXiv:2104.08819* (2021).
- [15] Mohammed, Manal, and Nazlia Omar. "Question classification based on Bloom's taxonomy using enhanced tf-idf." *International Journal of Advanced Science Engineering Information Technology* 8, no. 4 (2018): 1679-1685.
- [16] Omar, Nazlia, Dhuha Abdulhadi Abduljabbar, and Adel Al-Zubaide. "Automated analysis of exam questions according to Bloom's taxonomy." *Procedia-Social and Behavioral Sciences* 59 (2012): 297-303.
- [17] Baidoo-Anu, Daniel, and Lawrence Owusu Ansah. "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning." *Journal of AI* 7, no. 1 (2023): 52-62.
- [18] DOLINSKY, MICHAEL. "Directions for using Generative Artificial Intelligence in Introductory Programming."
- [19] Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Antoine Bordes, Thibaut Lavril, Devendra Singh Chaplot et al. "Mistral 7B." *arXiv preprint arXiv:2310.06825* (2023).
- [20] Lilleberg, Joseph, Yun Zhu, and Yanqing Zhang. "Support vector machines and Word2Vec for text classification with semantic features." In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, pp. 136-140. IEEE, 2015.
- [21] Jacovi, Alon, Oren Sar Shalom, and Yoav Goldberg. "Understanding convolutional neural networks for text classification." *arXiv preprint arXiv:1809.08037* (2018).
- [22] Nowak, Jakub, Ahmet Taspinar, and Rafal Scherer. "LSTM recurrent neural networks for short text and sentiment classification." In *Artificial Intelligence and Soft Computing: 16th International Conference, ICAISC 2017*, pp. 553-562. Springer, Cham, 2017.
- [23] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [24] Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).
- [25] Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Antoine Bordes, Thibaut Lavril, Devendra Singh Chaplot et al. "Mistral 7B." *arXiv preprint arXiv:2310.06825* (2023).
- [26] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks." *Advances in neural information processing systems* 33 (2020): 9459-9474.

APPENDIX

Benchmark	Mistral 7B	Llama 2 (7B)	Llama 2 (13B)	Code Llama (7B)	XGen-7B
Commonsense Reasoning	82	70	78	74	76
Knowledge	81	68	77	70	74
Reading Comprehension	84	72	80	73	79
Mathematical Reasoning	67	60	65	80	63
Code Generation	78	68	75	82	72

Table 2: Comparison of 7B LLMs

Dataset	Columns	Description
Classification	2 Columns: Question and Bloom’s Taxonomy Level	1200 rows, 200 questions of each level
Suggestion	4 Columns: Question, Current Bloom’s Taxonomy Level, Desired Bloom’s Taxonomy Level, Transformed Question	600 rows, 100 questions of each level
Generation	3 Columns: Topic, Desired Bloom’s Taxonomy Level, Generated Question	600 rows, 100 questions of each level

Table 3: Dataset Descriptions

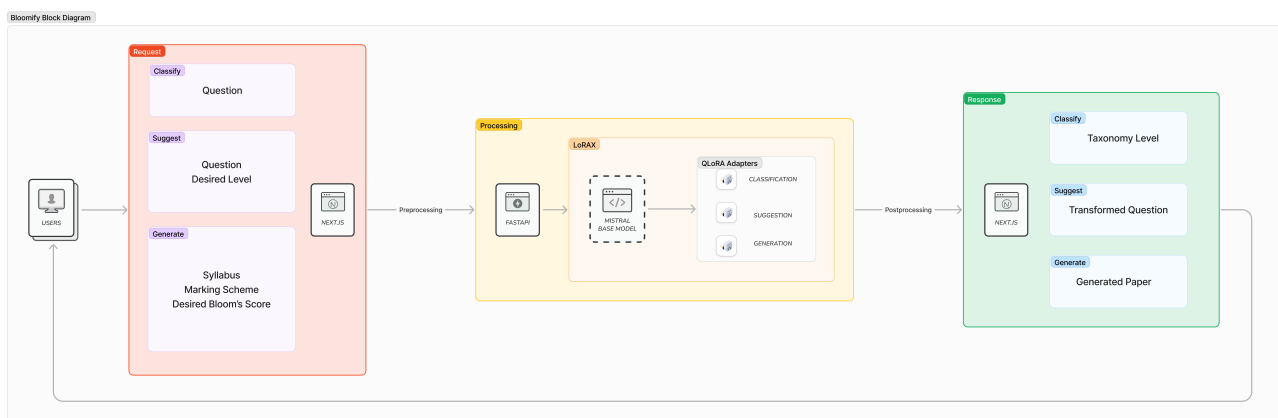


Figure 13: Bloomify Workflow

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The contributions of the authors are as follows: Minakshi Atre provided Resources, Supervision, Writing – review & editing, and Project administration. Sarthak Karandikar contributed to Software, Methodology, and Writing – original draft. Kabeer Ahmed Merchant also contributed to Software, Methodology, and Writing – original draft. Abhijeet Suryawanshi contributed to Data curation, Formal analysis, Investigation, and Visualization. Heramb Patil was responsible for Validation.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.