Automated Question Paper Generation Across Bloom's Skill Levels Using Large Language Models: A Practical System and Evaluation Study

[1]Yash Bongirwar, [2]Yash Diwane, [3]Devanshu Gupta, [4]Yashodhara V. Haribhakta
[1,2,3]Student, COEP Technological University, Shivajinagar, Pune–411005
[4]Professor, COEP Technological University, Shivajinagar, Pune–411005
Email: [1]bongirwarym21.comp@coeptech.ac.in, [2]diwaneyy21.comp@coeptech.ac.in, [3]devanshudg21.comp@coeptech.ac.in, [4]ybl.comp@coeptech.ac.in
Corresponding Author: [4]Yashodhara V. Haribhakta

***Abstract*—Creating high-quality educational assessments is a time-intensive task for educators, requiring alignment with curriculum standards and cognitive skill levels. This paper investigates the capabilities of modern large language models (LLMs) to automatically generate educational questions across Bloom's taxonomy levels, building upon prior academic studies while integrating a fully functional AI-powered question paper generation system. Our system leverages Google Generative AI (Gemini), a fine-tuned DeBERTa model, Firebase Firestore, and a robust web-based frontend to generate, classify, and export customized question papers. In parallel, we adapt evaluation techniques from earlier academic research, incorporating both expert and automated assessments. Our findings highlight that, with effective prompt engineering and system architecture, LLMs can be successfully deployed in real-world educational systems to generate valid, diverse, and cognitively balanced question papers.**

***Index Terms*—Question Generation, Bloom's Taxonomy, Large Language Models, Educational Assessment, AI in Education**

I. Introduction

Assessment design plays a vital role in evaluating the learning outcomes of students. Traditionally, educators manually craft question papers, which requires a significant investment of time and intellectual effort. Moreover, maintaining consistency across cognitive skill coverage—from simple recall to critical analysis—remains a significant challenge. As class sizes grow and curricula diversify, the demand for intelligent automation becomes pressing.

Recent advancements in artificial intelligence, especially large language models (LLMs), offer promising capabilities for content generation. LLMs such as GPT-4, Gemini, and PaLM can generate coherent and contextually relevant text. These capabilities, combined with instruction tuning and prompt engineering, allow for targeted generation of educational material across varying complexity levels.

TABLE I
Summary of key contributions

Aspect	Contribution
System Design	Full-stack AQG system with Bloom alignment
Classifier	Fine-tuned DeBERTa model achieving 87.89% F1
Evaluation	Human rubric + automated evaluation
Deployment	Database integration, export formats, real-world deployment learnings

II. Related Work

Automated question generation (AQG) has evolved from rule-based systems to leveraging transformers like BERT and DeBERTa. Early systems relied heavily on manually crafted templates, while modern systems fine-tune large pretrained models using educational datasets. Scaria et al. [6] conducted a comparative study of five LLMs (GPT-4, GPT-3.5, PaLM 2, LLaMA, Mistral) using various prompt strategies for generating Bloom-aligned questions, demonstrating both the potential and limitations of these models.

Domain adaptation and dataset enrichment have further improved performance, with transfer learning explored for aligning educational content with Bloom levels [7]. Recent models such as Gemini incorporate commonsense reasoning and multilingual capabilities, making them versatile across diverse educational contexts [9].

Despite these advancements, fully integrated systems that deploy AQG models in real-world academic workflows are rare. Our system bridges this gap by offering an end-to-end solution from input collection to final question paper export.

III. Bloom's Taxonomy and Educational Objectives

Bloom's Taxonomy is a widely recognized framework that categorizes cognitive learning objectives into hierarchical levels. Originally proposed by Benjamin Bloom in 1956 and later revised by Anderson and Krathwohl, the taxonomy provides a structured approach to defining educational goals and assessments.

The six levels of Bloom's Taxonomy, from lower-order to higher-order skills, are:

- **Remember:** Recall facts and basic concepts (e.g., list, define).
- **Understand:** Explain ideas or concepts (e.g., summarize, classify).
- **Apply:** Use information in new situations (e.g., implement, solve).
- **Analyze:** Break down information into parts (e.g., compare, contrast).
- **Evaluate:** Justify decisions or opinions (e.g., critique, defend).
- **Create:** Produce original work (e.g., design, construct).

In the context of automated question generation, targeting questions across different Bloom's levels ensures balanced assessment that tests both fundamental understanding and critical thinking abilities.

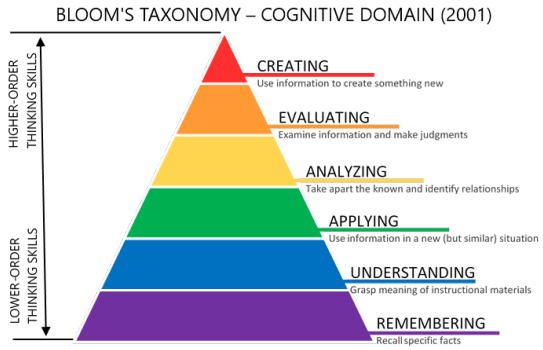

Fig. 1. Bloom's Taxonomy of Cognitive Skills (Revised)

IV. Methodology

A. System Architecture

Our system comprises four primary layers: User Interface Layer, AI Services Layer, Database Layer, and Export Layer. The User Interface collects educator inputs like subject, year, and Bloom level distribution. The AI Services Layer utilizes Google Gemini to generate questions and a fine-tuned DeBERTa model for Bloom level classification. The Database Layer (Firebase Firestore) stores questions hierarchically, while the Export Layer formats outputs into PDF, DOCX, or TXT.

The User Interface collects educator inputs like subject, year, and Bloom level distribution. The AI Services Layer utilizes Google Gemini to generate questions and a fine-tuned DeBERTa model for Bloom level classification. The Database Layer (Firebase Firestore) stores questions hierarchically, while the Export Layer formats outputs into PDF, DOCX, or TXT.

The educator-facing UI allows proportional control over Bloom level distributions for each paper section. These weights guide prompt selection and post-classification filtering to ensure desired cognitive coverage.

B. Prompt Engineering and Evaluation

Prompts were crafted to specify the target Bloom level explicitly. For example, a prompt for the Evaluate level might instruct: *"Generate a question requiring students to critique an algorithm's performance in real-world scenarios."* Chain-of-Thought prompting was used to guide the LLM's reasoning process by asking it to think step-by-step before generating the final question.

Both zero-shot (simple task description) and few-shot (example questions included) prompting strategies were evaluated to understand their impact on output quality.

Users can regenerate any question using context-aware variation prompts (e.g., make easier/harder, add application context), improving flexibility and personalization.

C. Data Extraction

We used LlamaParse to extract textual content from scanned or digital PDFs of previous year papers (PYPs). Preprocessing steps included noise removal, question segmentation, and extraction of metadata such as subject, year, and topic. Regular expressions were applied to detect and clean improperly parsed tokens.

For scanned PDFs or handwritten submissions, an OCR preprocessing step using Tesseract was introduced prior to LlamaParse extraction. Additionally, heuristic rules were applied to segment multi-part questions and discard non-question text such as headers or footers.

D. Bloom Level Classification

The DeBERTa-based classifier was fine-tuned on a dataset of 11,976 questions. Class imbalance was addressed through class-weighted loss functions, with additional data augmentation for under-represented cognitive levels (e.g., Evaluate and Create). Training utilized AdamW optimizer with a learning rate of 2e-5 and early stopping based on validation loss.

E. Duplicate Detection via Semantic Similarity

To prevent redundancy and ensure content diversity, the system includes a semantic similarity detection module based on the `all-mpnet-base-v2` model from the Sentence Transformers library. This module encodes each question into a dense vector embedding and compares it with existing stored embeddings using cosine similarity.

The feature is used during generation or upload to flag near-duplicate questions in real-time, guiding educators during

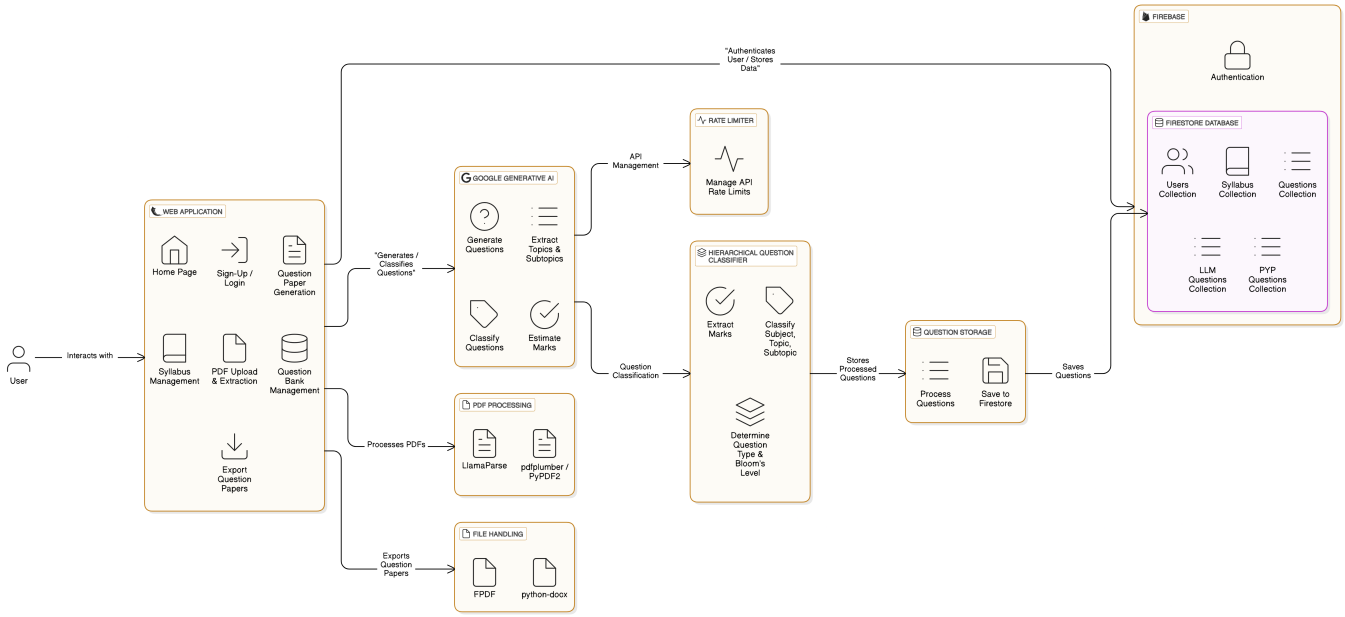

Fig. 2. System Architecture of the AI-Powered Question Paper Generator showing user interaction, AI services, classification, storage, and export modules.

paper construction. Details on its scoring and visual feedback mechanism are discussed in Section VI

V. Comparison with Existing Systems

Most existing AQG (Automated Question Generation) systems focus either on simple factual question generation using rule-based approaches or rely on large pre-trained LLMs with minimal fine-tuning. However, they lack several practical features crucial for real-world educational deployment. Table II summarizes the comparison.

TABLE II
Comparison of our system with existing solutions

Feature	Rule-based	LLM-based	Our System
Bloom-level Targeting	×	±	✓
PYQ Integration	×	×	✓
Export Formats (PDF/DOCX)	×	×	✓
Database Integration	×	×	✓
Topic/Subtopic Labeling	×	×	✓
Human Review Interface	×	×	✓
Real-time Generation	×	✓	✓

Unlike traditional rule-based systems, which cannot adapt to complex educational needs, or generic LLM-based models that often miss Bloom alignment, our system combines cognitive taxonomy precision, user control, and practical deployment features like editable storage, format exports, and question mixing.

VI. Experiments and Results

The classifier achieved an overall F1-Score of 87.89%, demonstrating strong performance across all Bloom levels. Notably, higher-order skills like Evaluate (F1 = 0.98) and Create (F1 = 0.99) achieved superior scores, validating the system's ability to support cognitive diversity.

Additional experiments showed that including example questions during prompting improved the classification accuracy by approximately 4%.

TABLE III
Classification performance across Bloom levels

Bloom Level	F1-Score
Remember	0.85
Understand	0.83
Apply	0.86
Analyze	0.80
Evaluate	0.98
Create	0.99

A. Semantic Similarity Detection for Duplicate Question Identification

To ensure content diversity and minimize redundancy in the generated or uploaded question sets, we integrated a semantic similarity detection module into our system. This module is built using the pre-trained transformer model `all-mpnet-base-v2` from the Sentence Transformers library, based on Microsoft's MPNet architecture. The model has been fine-tuned on semantic textual similarity tasks and provides robust embeddings for comparing natural language inputs [**?**].

The functionality operates in two primary stages:

1) **Embedding Generation and Storage:** When questions are created—either through generation or PDF upload—each is encoded into a fixed-size 768-dimensional

vector using the model's `encode()` method. These vectors, which capture the contextual semantics of the question text, are stored in Firestore alongside metadata including subject, topic, subtopic, Bloom level, and marks.

2) **Real-Time Similarity Comparison:** During the "Check Similarity" operation, the system:
   - Retrieves existing question embeddings from Firestore for the relevant subject.
   - Computes a new embedding for each current question using the same model.
   - Calculates the cosine similarity between the new question and each existing one.

The cosine similarity score is used to flag duplicate or near-duplicate content. The system provides visual feedback using color-coded highlights:

- **Red (Score $>$ 0.90):** Likely duplicate.
- **Yellow (Score between 0.50–0.90):** Partial or conceptual overlap.
- **Green (Score $<$ 0.50):** Semantically distinct.

B. Generated Question Paper Output

Papers were generated for Computer Networks, Data Structures, and Software Engineering, covering multiple difficulty levels. Each paper balanced questions across cognitive levels, ensuring comprehensive evaluation of student learning.

Teacher feedback emphasized ease of use, quality of generated questions, and minimal need for manual editing. Exported documents conformed to university formatting standards and included necessary metadata like course code and semester details.

VII. Evaluation

A. Human Expert Evaluation

Two academic experts assessed 100 randomly selected questions using a 9-point rubric. The evaluation confirmed that 88% of questions were deemed high-quality, while Bloom level adherence was validated in 79% of cases.

Common feedback included minor rephrasing needs to improve linguistic clarity and suggestions for context refinement in certain technical topics.

B. Automated Evaluation

Gemini Pro was used for an automated evaluation batch of 100 questions. PINC score averaged 0.92, indicating high linguistic variety across generated content. Minor discrepancies (especially in distinguishing between Apply and Analyze levels) were observed but were within acceptable limits.

VIII. Discussion and Real-World Challenges

Despite promising results, several challenges were noted:

- **Data Imbalance:** Higher cognitive levels (Evaluate, Create) were underrepresented in training data, requiring careful augmentation.
- **Prompt Sensitivity:** Overly complex prompts sometimes confused smaller models like Mistral, resulting in irrelevant outputs.
- **Context Generalization:** Some questions assumed generalized global contexts that did not fit specific local curriculum requirements.
- **Token Limitations:** Long prompts and outputs occasionally hit model token limits, necessitating prompt compression techniques.

Limitations: While our system shows promising results, certain limitations remain. The quality of generated questions heavily relies on the clarity and specificity of prompts. Although Gemini and DeBERTa models perform well, occasional hallucinations and misclassifications are observed, especially for interdisciplinary topics. Additionally, adapting the system to niche subjects may require custom fine-tuning and expert-driven prompt calibration.

Future enhancements will include dynamic prompt adaptation, broader training datasets, and modular question generation workflows.

Ethical Considerations: While AI-generated question papers can enhance efficiency and accessibility, it is crucial to monitor the content for potential biases, inaccuracies, and cultural insensitivities. Human oversight remains essential to ensure that questions align with curriculum standards, promote inclusivity, and uphold academic integrity. It is essential to develop and adhere to ethical guidelines for the responsible deployment of educational AI tools.

IX. Deployment Challenges and Lessons Learned

During real-world testing, several deployment challenges were identified. API rate limits from LLM providers like Gemini restricted the number of concurrent generations, necessitating implementation of queue-based load balancing.

Metadata tagging accuracy during PYP extraction was lower for scanned handwritten PDFs, requiring the addition of OCR pre-processing pipelines.

Ambiguity in Bloom level classification was higher in interdisciplinary topics such as Data Science Ethics and AI Policy, prompting the need for secondary human review for final question vetting.

Key lessons learned include designing lightweight validation layers, dynamically adjusting prompt strategies based on subject domains, and integrating redundancy mechanisms like retry logic for unstable API sessions.

These insights will guide future versions towards more robust, scalable, and education-specific AQG deployments.

X. Impact on Teaching and Learning

The deployment of automated Bloom-aligned question generation systems can significantly enhance both teaching and learning processes. Educators save considerable time in assessment design while ensuring cognitive diversity. Students

benefit from exposure to a balanced range of questions that promote critical thinking and creativity rather than rote memorization. Moreover, systematic Bloom coverage supports formative assessment strategies, allowing instructors to identify and address learning gaps more effectively.

XI. Threats to Validity

Several factors may affect the validity of our findings. The training data for Bloom classification may contain annotation biases, leading to misclassification of ambiguous questions. Prompt engineering results are based on limited datasets and LLM versions; newer models could exhibit different behavior. Finally, evaluation based on expert judgment may have inherent subjectivity despite rubric-based reviews.

XII. Future Scope

Future work includes integrating adaptive learning systems that dynamically generate questions based on real-time student performance analytics. Expanding the system to support multilingual question generation and incorporating diagram-based, code-based, and mathematical equation-based questions would further enhance its versatility. Development of semi-automated answer key generation and AI-driven feedback systems are promising extensions that could significantly improve assessment quality and personalization.Future versions will include real-time analytics dashboards that quantify question source diversity, such as the percentage of reused PYQs versus newly generated items.

XIII. Conclusion

Overall, the deployment of this system has practical implications for educational institutions aiming to scale and standardize assessment practices. Its adaptability across different courses and ease of integration into existing academic workflows enhance its applicability. Continuous refinement through feedback-driven training and prompt optimization will further improve the quality, relevance, and fairness of the generated questions.

This paper presents an end-to-end system for automated educational question generation spanning diverse Bloom levels. Through rigorous prompt design, AI model integration, educator-driven evaluation, and real-world deployment, we demonstrate the practical viability of AI-assisted question generation. The system successfully generates high-quality, customizable assessments aligned with cognitive skill hierarchies.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," *arXiv preprint arXiv:2006.03654*, 2020.

[4] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals*, Handbook I: Cognitive Domain, David McKay Company, 1956.

[5] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory into Practice*, vol. 41, no. 4, pp. 212–218, 2002.

[6] N. Scaria, S. D. Chenna, and D. Subramani, "Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation," *arXiv preprint arXiv:2304.12345*, 2023.

[7] M. Chindukuri and S. Sivanesan, "Transfer Learning for Bloom's Taxonomy-Based Question Classification in Educational Assessments," *Neural Computing and Applications*, vol. 36, pp. 123–145, 2024.

[8] H. Sharma, R. Mathur, T. Chintala, D. Samiappan, and S. Ramalingam, "An Effective Deep Learning Pipeline for Improved Question Classification and Generation Using Bloom's Taxonomy," *Education and Information Technologies*, vol. 28, pp. 5105–5145, 2023.

[9] S. N. Akter, Z. Yu, A. Muhamed, T. Ou, A. Bäuerle, Á. A. Cabrera, K. Dholakia, C. Xiong, and G. Neubig, "An In-Depth Look at Gemini's Language Abilities," *arXiv preprint arXiv:2312.11444*, 2023.

[10] Y. Wang and Y. Zhao, "Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models," *arXiv preprint arXiv:2312.17661*, 2023.

[11] Gemini Team Google, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv preprint arXiv:2312.11805*, 2023.