International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 22s, 2025
https://theaspd.com/index.php

BT-GEN: A Retrieval-Augmented And BT-Classified Approach For Enhancing Cognitive Assessment Through Automated MCQ Generation And Classification

Disha Sengupta[1*], Sonam Singh[2], Riya Utekar[3], Jyoti Kumari[4], Chaitra Ghule[5], Siddhi Taskar[6]
[1*]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[2]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[3]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[4]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[5]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune
[6]Dr. D. Y. Patil Institute of Technology, Pimpri, Pune

Abstract *– In the era of education 4.0, the teaching and learning methods are inclining more towards multiple choice questions (MCQs). This leads to increasing needs of a system that could automate the process of MCQ generation while aligning them to the learner's needs. To address these needs, we propose a system that combines question classification with question generation and report generation. The system classifies the questions based on Bloom's Taxonomy (BT) levels and hence the user performance in the tests reflects their cognitive strengths and weaknesses. Using the insights from user's test attempt, the system generates questions using RAG framework in which, it uses Gemini, a Large Language Model (LLM) for generation and Chroma DB, a vector database to ensure the question pattern is followed. The feature of report generation then helps the user to reflect on their test attempts and their overall abilities to better plan their approach. This paper proposes a system that will help learner to understand the different levels of questions as per the bloom's taxonomy.*

Index Terms *– Large language models, retrieval-augmented generation, multiple choice questions, bloom's taxonomy, personalized learning.*

1. INTRODUCTION

Modern education system is inspired by tech-driven classroom, personalized and digital learning platforms that provides data-driven insights. Hence, MCQs serve as an adaptable and reliable tool that supports speed, consistency, and analytical assessment tool for learners. The fact that creation of MCQs that address a range of subjects and individual learning styles is complicate, resource-intensive, time-consuming task and heavily reliant on subject-matter expertise. Such challenges highlight the need for smart and efficient system that make MCQ development easy and also captures the interactive, learner-driven spirit of today's hybrid and personalized instructional models. Therefore, automated MCQs generation marks an important landmark in the evolution of modern teaching. As artificial intelligence (AI) is explored, its potential applications in transforming educational sector is becoming clearer. Generative AI serve as a powerful means for MCQ creation. Use of Large language models (LLMs) in NLP and data analytics, and in automating the creation of pedagogically sound assessments is under study.

However, existing systems often struggle to effectively adjust the difficulty level of questions.

This calls for a refined LLM-powered MCQ generator so it better align with pedagogical goals and diverse student needs.

We present an AI-powered framework designed to automatically generate multiple-choice questions (MCQs) using the Gemini large language model (LLM) and Retrieval-Augmented Generation (RAG). By using a vector database, the system ensures that the questions it creates are both relevant and tailored to individual learners. To make the questions more meaningful, each one is categorized according to Bloom's Taxonomy, allowing the system to understand and reflect different levels of cognitive difficulty. The insights from user-interactions are then used to generate a personalized performance report for each user, highlighting areas of strength and suggesting where they can improve. The goal is to create an engaging, adaptive learning experience that not only tests knowledge but also supports deeper understanding.

Overall, the key contributions of this paper are as follows:

- A generative AI pipeline for MCQ creation that introduces an end-to-end framework that seamlessly integrates the Gemini LLM with retrieval-augmented generation (RAG) to automate question creation.
- Question generation that aligns with users' cognitive understanding by leveraging a vector database through RAG.

- Cognitive-level labelling via Bloom's taxonomy enabling targeted assessment of different thinking skills.
- Automated student feedback reports that highlight each learner's strengths and weaknesses.
- Enhanced personalized learning is provided through the combined use of RAG, LLM generation, and cognitive classification, thus our framework delivers MCQ sets that supports deeper understanding, and advance adaptive education.

This paper is divided into sections which determines its structure as follows. Section 2 reviews the background and related studies on automated MCQ generation. In Section 3, we represent our framework combining an LLM with RAG to automate MCQ creation and classification based on Bloom's Taxonomy and performance report creation. Section 4 evaluates the proposed framework and discusses the results. Section 5 highlights the potential limitations and possible improvements in future work. We conclude the paper in the Section 6.

2. RELATED WORK

As a popular and efficient assessment format, MCQs ask learners to choose the most appropriate answer among several given options. Their strength lies in their adaptability for testing various cognitive skills, from remembering to analytical thinking. MCQs simplify the grading process and generate clear, quantifiable performance data, making them ideal for use in large educational environments where fairness and consistency are crucial. They also play a key role in supporting personalized learning by allowing educators to pinpoint specific strengths and weaknesses, providing opportunities for customized feedback. The structured nature of MCQs minimizes ambiguity, leading to more precise evaluations of student understanding.

An MCQ is typically composed of three essential parts: (1) the stem, (2) the key, and (3) the distractors. The option that is correct answer is called key and other options are called distractors. The part of the question, other than options is called and is the core of the question.

Consider the following MCQ as an example:

What is the capital of Maharashtra? - Stem

A. Mumbai - Key
B. Pune – Distractor 1
C. Kolhapur – Distractor 2
D. Satara – Distractor 3

Over years, there has been significant amount of research in the field of automatic MCQ generation leading to its growth over time. The field has evolved from traditional techniques to advanced methods that use deep learning and is still under exploration. The work in [10] uses an automated BERT that does the text summarization and alignment and performs distractor generation using WordNet. In [9], a dependency-based method to learn semantic relationships for automated MCQ generation is presented. Their method achieved high accuracy and was appreciated by users for its clarity, relevance, and effectiveness in e-learning settings. [2] moves further ahead from investigating the progress in personalized education through the automatic MCQ. MCQ Gen is a large language model (LLM) which is integrated with retrieval-augmented generation and refined prompt engineering to create contextually appropriate questions. [7] also automated the process of generating context-aware questions using RAG. [3] randomly selected 2900 sentences from the book named "Operating System Principles" for analysis. The paper went with an approach that was combination of ontology and Machine Learning. Ontology was used for generating WH type questions and SVM was used in ML for Cloze questions. In [5], the authors a system, driven by NLP for automatically generating MCQs for Computer-Based Testing, where keywords are extracted from instructional content to ensure the relevance and effectiveness of generated questions for examination purposes. [8] designed a model to generate educational questions that align with Bloom's taxonomy by extracting key phrases and applying context-free grammar rules. Tested on software engineering course content, the approach demonstrated a systematic and efficient method for automated question creation.

In case of classifying question based on Bloom's Taxonomy, [6] used web scraped software engineering questions and classifying them using SVM classifier, the model achieved accuracy of 98% and 96%-99% precision scores for all classes. The author further created 6 different agents for each level in BT to handle them independently. [4] experimented with LSTM and CNN models for classification. They concluded that CNN had better performance among those with 80% accuracy and 66.67% testing accuracy where LSTM had 44% testing accuracy. The models were compared based on their results on the dataset that consisted of 844 Software engineering questions from different institutions. [1] used 2 datasets that were a combination of 5 datasets. The study was done in 2 phases. The first phase was dedicated to finding the optimal word embedding technique

which concluded that RoBERTa was the best technique. The 2nd phase was about hyperparameter tuning and achieved recall, precision, F1 score of 0.72, 0.85 and 0.78 respectively.

To the best of our knowledge, most of the studies were carried independently for either Automated MCQ generation or for classifying the questions based on Bloom's Taxonomy. This paper is among the fewest studies who implement LLM with RAG for automated generation and at the same time involves a classifier that classifies questions on basis of Bloom's Taxonomy cognitive level. It also involves a report generation module that summarizes the performance of the end-user for their reference. This integrated system leverages the strengths of each component to help students in exam preparation, ideally suited for modern learning styles and educational dynamics.

3. METHODOLOGY

The proposed system is designed to generate multiple-choice questions (MCQs) tailored to the learner's cognitive proficiency based on Bloom's Taxonomy (BT). The workflow describes below as shown in figure 1:

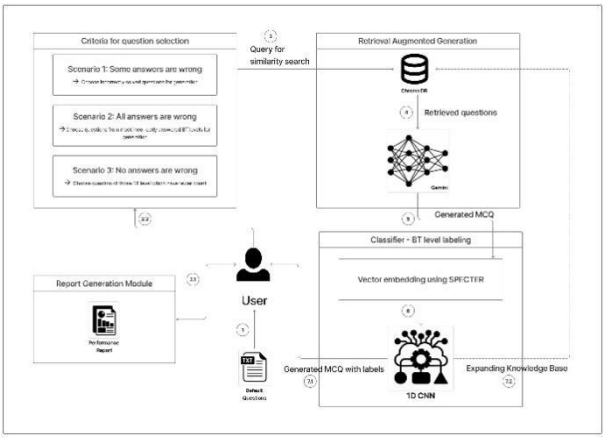

Figure 1 System Architecture

3.0. Research Approach

Our research is structured into three main phases to ensure clarity, repeatability, and academic rigor. First, we prepared and annotated a custom dataset using previous year GATE questions, focusing on SQL and Operating Systems. Second, we developed an end-to-end framework that integrates Retrieval-Augmented Generation (RAG) with Bloom's Taxonomy-based classification using Specter embeddings and 1D CNN. Finally, we evaluated the framework through performance metrics and visual feedback profiling. This phased approach allows us to systematically address the challenges of automated question generation, question classification, and personalized assessment.

3.1. Dataset Description

We created a custom dataset by collecting previous year questions (PYQs) from the GATE examination focused on SQL and Operating Systems (OS) topics. These questions were sourced from publicly available online educational platforms and repositories. The dataset comprises a total of 299 records, each representing a multiple-choice question (MCQ). Each row in the dataset contains the following attributes (Table 1):

Attributes	Description
Question	The text of the question
Option1 to Option4	The four available answer choices
Correct Option	The index of the correct answer
BT	The Bloom's Taxonomy level assigned to the question
Subject	The subject domain of the question (SQL or OS)
Topic	The specific topic within the subject (e.g., Transactions, Paging, etc.)

Table 1 Attribute of dataset and its description

The questions were manually labelled with appropriate Bloom's Taxonomy (BT) levels based on cognitive complexity, following the Revised Bloom's Taxonomy (RBT) framework. This dataset was used both for training and evaluating the classifier model and as input for the question generation pipeline.

3.2. User Interaction and Feedback Capture

The process begins when a user attempts a quiz based on default questions (Step 1). The user's responses are collected and analysed to determine incorrectly answered questions and performance across BT levels (Step 2.1). This analysis forms the basis for personalized content generation.

3.3. Criteria for Question Selection

Based on user's test attempt, one of the following three scenarios is triggered to determine which questions are suitable for further generation (Step 2.2):

- Scenario 1: (Some answers are wrong) If some questions are answered incorrectly, use only those questions for similar question generation.
- Scenario 2: (All answers are wrong) If the user answers all questions incorrectly, the system chooses questions from most incorrectly answered BT levels for generation.
- Scenario 3: (No answers are wrong) If all of the questions are answered correctly the system identifies BT levels that were underrepresented in the test and selects questions from those levels for expansion.

3.4. Question Generation via RAG Framework

To generate contextually similar questions, a retrieval-augmented generation (RAG) pipeline is employed. This process occurs in two stages:

Retrieval Phase: The selected questions are first encoded and used to query a vector database — in this case, ChromaDB. This retrieves semantically similar questions from an existing knowledge base (Step 3).

Generation Phase: The retrieved questions serve as contextual input in a prompt to Gemini, a Large Language Model. Gemini then produces new, analogous questions aligned with the user's understanding (Step 4).

3.5. BT Level Classification

The newly generated questions are classified according to their corresponding BT levels. This classification employs a supervised machine learning model trained on labelled textual data. Given the mathematical nature of the question content, SPECTER is used for vectorization to improve semantic accuracy over standard BERT embeddings. 1D Convolutional Neural Network (1D CNN) that predicts the BT level for each MCQ, enabling further personalization and curriculum alignment.

3.6. Knowledge Base Expansion

Generated MCQs along with their predicted BT levels are added to an expanding knowledge base for future use and training enhancement (Step 6).

3.7. Use Generated MCQs for assessment

The newly generated MCQ which are classified under specific cognitive levels, are sent to the user for further assessment (Step 7.1)

3.8. Report generation

Learner interactions across multiple assessments are continuously monitored to construct a cognitive performance profile (Step 2.1). This profile visualizes the learner's strengths and weaknesses across BT levels and is updated in real time as new questions are answered. The goal is to provide interpretable feedback that reflects not only right or wrong answers but also the underlying cognitive dimensions being developed.

4. RESULTS

In this section we present the outcomes of the question generation and classification components of the proposed system, followed by a summary of the personalized learning visualizations derived from student interaction data.

4.1. Question Classification

To classify questions based on Bloom's Taxonomy, we built the neural network using 1D CNN and comparing MathBERT with Specter for text embedding. The neural network was evaluated using accuracy, recall, precision, and F1-score. Upon comparison, using Specter for text embedding consistently outperformed MathBERT across all metrics. The improved accuracy and balanced classification performance of Specter led us to select it for our final implementation. Below is a summary of the classification results (Table 2, 3 & 4):

Model	Training F1 Score	Testing F1 Score
MathBERT	0.98	0.37

SPECTER	0.85	0.52

Table 2 F1-score comparison

	Precision	Recall	F1-Score	Support
0	0.67	0.15	0.25	13
1	0.17	0.08	0.11	12
2	0.82	0.75	0.78	12
3	0.35	0.69	0.46	13
4	0.38	0.77	0.51	13
5	0.33	0.08	0.13	12
Accuracy			0.43	75
Macro avg	0.45	0.42	0.38	75
Weighted avg	0.45	0.43	0.38	75

Table 3 Evaluation metrics for MathBERT

	Precision	Recall	F1-Score	Support
0	0.53	0.69	0.60	13
1	0.50	0.08	0.14	12
2	0.90	0.75	0.82	12
3	0.31	0.38	0.34	13
4	0.75	0.69	0.72	13
5	0.44	0.67	0.53	12
Accuracy			0.55	75
Macro avg	0.57	0.54	0.53	75
Weighted avg	0.57	0.55	0.53	75

Table 4 Evaluation metrics for SPECTER

The confusion matrices for both models indicate that Specter was better at distinguishing between higher-order Bloom levels such as Analyse and Evaluate, which often exhibit semantic overlap (as shown in figure 2 & 3).The model's stronger performance using Specter (F1 = 0.53) compared to MathBERT (F1 = 0.38) highlights the advantage of using domain-tuned semantic embeddings. Notably, Specter handled higher-order BT levels like 'Analyze' and 'Evaluate' better, likely due to its training on academic corpora. However, confusion between adjacent cognitive levels (e.g., Apply vs Analyze) aligns with the challenges noted in [6] and [8], suggesting a semantic overlap that's hard to disambiguate even with advanced embeddings.

4.2. Question Generation using RAG

For automatic question generation, we implemented a Retrieval-Augmented Generation (RAG) approach. Two state-of-the-art generative models were compared:

- OpenAI GPT-4o-mini
- Google Gemini 1.5 flash

The similarity between generated questions and reference questions was measured using Cosine Similarity. This helped evaluate how semantically aligned the generated content was with expected Bloom-level standards.

Preliminary results show that Gemini achieves a higher average cosine similarity score compared to GPT-4o-mini, indicating greater alignment and contextual relevance in the generated questions (Table 5). Gemini showed a higher cosine similarity score (0.82) compared to GPT-4o-mini (0.78), which means the questions it generated were more closely aligned with the original content and learning goals.

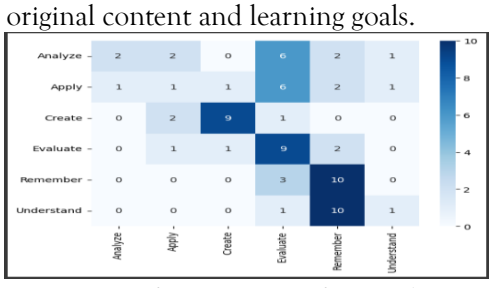

Figure 2 Confusion matrix for MathBERT

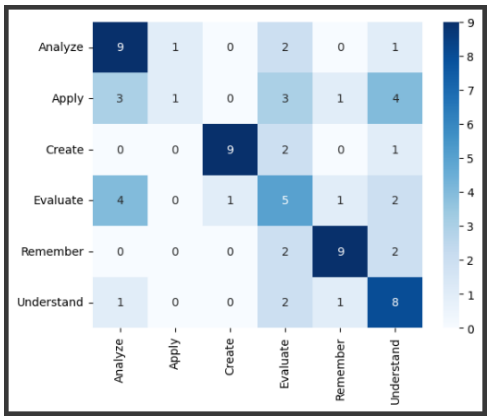

Figure 3 Confusion Matrix for SPECTER

This matches what some earlier studies ([4], [7]) also found that Gemini is better at generating high-quality educational content. However, just relying on similarity scores isn't enough. To truly understand how useful the questions are for learning, we still need feedback from real users like teachers or students, as pointed out in [10].

	Cosine Similarity
gemini-1.5-flash	0.82
gpt-4o-mini	0.78

Table 5 Cosine Similarity comparison

4.3. Visualization and Cognitive Profiling

As part of performance analysis and user insight, we implemented:

- BT-level-wise Performance Distribution: The accuracy of classification per Bloom level was visualized using bar charts to assess the strengths and weaknesses of the model across cognitive categories.

These visual summaries proved valuable for supporting reflective learning and data-driven academic interventions.

Other profiling elements such as time-based performance tracking and improvement suggestions are planned for future iterations but have not yet been implemented.

4.4. Comparison with Existing Works

Unlike prior studies that focus solely on MCQ generation [4] or question classification based on Bloom's Taxonomy [6], our system integrates both components along with a performance feedback module. For instance, [3] used RoBERTa for Bloom-level classification and achieved an F1 score of 0.78 using a combined dataset. In comparison, our system, although trained on a smaller dataset, achieved a competitive F1 score of 0.53 using Specter embeddings and performed notably well on higher-order cognitive levels. Additionally, studies like [7] and [9] implemented RAG for question generation, but they did not link the generation to student performance or cognitive profiling. Our work bridges this gap by creating a personalized feedback loop between the learner's performance, the question difficulty, and the generation mechanism. This comprehensive integration offers a more adaptive and learner-centric assessment solution.

5. LIMITATION & FUTURE WORK

5.1. Limitations

Despite the promising outcomes, the current implementation presents the following limitations:

- Limited Dataset Size and Diversity: The training and testing datasets used for classification were relatively small and domain-specific. This may have limited the generalizability of the model across different subjects or question types.
- No Human-in-the-Loop Evaluation: The effectiveness of the generated questions was evaluated only via cosine similarity scores. The absence of human evaluation (e.g., by teachers or subject experts) limits our understanding of their true educational value and Bloom-level alignment.
- Partial Implementation of Cognitive Profiling: Although BT-level performance visualization was implemented, other components like tracking question-solving time and generating personalized improvement suggestions were not fully realized.

5.2. Future Work

To build on the current system and address its limitations, the following directions are proposed:

- Expand and Annotate Dataset: Curate a larger and more diverse dataset of questions across various domains, manually annotated with Bloom's levels, to improve training effectiveness and model robustness.
- Model Fine-Tuning: Fine-tune transformer models like Specter or Sentence-BERT on BT-specific question datasets to enhance semantic sensitivity and classification accuracy.
- Hybrid Classification Approach: Explore hybrid methods that combine semantic embeddings with rule-based or syntactic features (e.g., question verbs, structure) to reduce misclassification between closely related Bloom levels.
- Human Evaluation of Generated Questions: Incorporate expert reviews and rubric-based assessments of generated questions to validate their cognitive level, clarity, and educational value.
- Full Cognitive Profiling Integration: Implement the remaining components of the cognitive profiling module, including:
  - Tracking student response times
  - Suggesting targeted improvements
  - Analysing learning progress over time
- Custom RAG Tuning: Fine-tune the retrieval and generation stages of the RAG pipeline using Bloom-level aligned corpora to enhance the contextual relevance of generated questions.

6. CONCLUSION

In this research, we presented an intelligent system designed to classify and generate questions based on Bloom's Taxonomy, addressing the growing need for automated cognitive-level tagging and question formation in educational contexts. By employing semantic embeddings through transformer-based models such as Specter and MathBERT, we demonstrated that Specter outperformed MathBERT in accurately classifying questions across the six cognitive levels, achieving improved precision, recall, and F1-scores. The confusion matrix analysis, however, revealed consistent challenges in distinguishing between higher-order thinking categories such as Analyse, Evaluate, and Apply, indicating the semantic overlap inherent in educational texts.

For question generation, the use of Retrieval-Augmented Generation (RAG) models, specifically GPT-4o-mini and Gemini, facilitated the creation of diverse and contextually relevant MCQs aligned with Bloom's cognitive levels. Cosine similarity was used as a metric to evaluate the semantic closeness of the generated questions to the reference questions, demonstrating the effectiveness of the approach.

Furthermore, the system incorporated visualization tools to track question distribution and student performance, laying the groundwork for future cognitive profiling. While promising, this work acknowledges limitations in data scale, model interpretability, and human-in-the-loop evaluation, which open avenues for future enhancements.

In summary, this study contributes a meaningful step toward intelligent educational tools that not only assist educators in creating assessments but also support learners by aligning content with cognitive objectives. With further refinement and integration of explainability and personalization, such systems have the potential to revolutionize digital learning environments.

REFERENCES

[1] Lee, S.hyun. & Kim Mi Na, (2008) "This is my paper", ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.
[2] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.
[3] Mohammed Osman Gani, Ramesh Kumar Ayyasamy, Anbuselvan Sangodiah, and Yong Tien Fui. Blooms taxonomy-based exam question classification: The outcome of cnn and optimal pre-trained word embedding technique. Education and Information Technologies, 28(12):1589315914,2023.
[4] Ching Nam Hang, Chee Wei Tan, and Pei-Duo Yu. Mcqgen: A large language model-driven mcq generator for personalized learning. IEEE Access,2024.
[5] Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy K, Chaitanya, and Kaustav Ghosh. A novel framework for the generation of multiple-choice question stems using semantic and machine-learning techniques. International Journal of Artificial Intelligence in Education, 34(2):332–375, 2024.
[6] Manjushree D Laddha, Varsha T Lokare, Arvind W Kiwelekar, and Laxman D Netak. Classifications of the summative assessment for revised blooms taxonomy by using deep learning. arXiv preprint arXiv:2104.08819,2021.
[7] Chidinma A Nwafor and Ikechukwu E Onyenwe. An automated multiple-choice question generation using natural language processing techniques.arXiv preprint arXiv:2103.14757, 2021.
[8] Rimsha Shahzad, Muhammad Aslam, Shaha Al-Otaibi, Muhammad Saqib Javed, Amjad Rehman Khan, Saeed Ali Bahaj, and Tanzila Saba. Multi-agent system for students' cognitive assessment in e-learning environment.IEEE Access, 2024.

[9] Altaj Virani, Rakesh Yadav, Prachi Sonawane, and Smita Jawale. Automatic question answer generation using t5 and nlp. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), pages 1667–1673. IEEE, 2023.
[10] Bambang Dwi Wijanarko, Yaya Heryadi, Hapnes Toba, and Widodo Budiharto. Question generation model based on key-phrase, context-free grammar, and blooms taxonomy. Education and Information Technologies,26:2207–2223, 2021.
[11] N. Afzal and R. Mitkov, "Automatic generation of multiple-choice Questions using dependency-based semantic relations," Soft Comput., Vol. 18, no. 7, pp. 1269–1281, Jul. 2014.
[12] P. K. Mehta, P. Jain, C. Makwana, and C. Raut, "Automated MCQ Generator using natural language processing," Internation Res. J. Eng. Technol., vol. 8, pp. 2705–2710, May 2021.

Authors

Prof. Disha Sengupta, currently pursuing Phd. at Symbiosis Institute of Technology and Assistant Professor at Dr. D. Y. Patil Institute of Technology, Pimpri Pune. She is having an experience of almost 11 years. She is having expertise in cloud computing, cyber security and machine learning.
Email: disha.24sharma@gmail.com

Prof. Sonam Singh, (Phd. Pursuing) Assistant Professor at Dr. D.Y Patil Institute of Technoly, Pune, having an experience of almost 8 years. She is having expertise in Artificial Intelligence, Deep Learning, Cloud Computing.
Email: sonamchauhan346@gmail.com

Riya Utekar is a final-year undergraduate student pursuing a B.E. in Artificial Intelligence and Data Science at Dr. D. Y. Patil Institute of Technology, Pune. She has gained hands-on experience as a Data Analyst Intern at Opineworks, where she worked on real-world data interpretation. Her research interests lie at the intersection of data science and generative AI, with a strong inclination toward applying these technologies to solve real-world problems in education and policy.Email: riyautekar17@gmail.com

Jyoti Kumari is currently pursuing a Bachelor of Engineering in Artificial Intelligence and Data Science at Dr. D.Y. Patil Institute of Technology, Pune. She is also working as a Data Science intern at a Cognizant. Her research interests include Data Science, Machine Learning, and the application of Artificial Intelligence in Education.
Email: jyoti640771@gmail.com

Chaitra Ghule is a final-year undergraduate student pursuing a B.E. in Artificial Intelligence and Data Science at Dr. D. Y. Patil Institute of Technology, Pune. Her research interests include neural networks and generative AI.
Email: chaitravghule@gmail.com

Siddhi Taskar is a final-year undergraduate student pursuing a B.E. in Artificial Intelligence and Data Science at Dr. D. Y. Patil Institute of Technology, Pune. She completed her diploma in computer technology from K. K. Wagh polytechnic, Nashik in 2022. Her interest lies in data analysis and web development.
Email: siddhitaskar@gmail.com