© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716734

AI-Driven Automated Question Paper Generator Using Past Exam Datasets

PRAVEEN H S[1], PRAVEEN KHOT[2], OMKAR SHINDE[3]
[1,2,3]*Dept. of Computer Science& Engineering, Dayananda Sagar University Bangalore, Karnataka, India*

Abstract- ***Automatic Question Paper Generation is a crucial problem in current educational technology. The manual process of generating test papers is cumbersome, prone to bias in the distribution of difficulty, and rarely achieves consistency year after year. This article proposes a complete reviews existing works on automated question paper generation through AI and presents a new combined system that uses NLP, deep learning, and difficulty classification based on Bloom's Taxonomy to develop balanced curriculum-compliant question papers using past exams. Five recent publications have been examined regarding rule-based systems, information retrieval, sequence-to-sequence deep learning, transformers, and LLMs. From the above literature review, we suggest a two-step process consisting of a (1) question bank generator using BERT to extract and de-duplicate questions and assign tags, and (2) paper assembler using GPT to assemble questions based on difficulty and topics. Experiments demonstrate our proposed system can achieve 91% question accuracy and takes less than 2 minutes compared to 1 hour previously, which is a 9-30% improvement.***

Keywords—Auto QG, NLP, BERT, GPT, Bloom's Taxonomy, EdTech, deep learning, difficult level classification, dataset for exams, machine learning.

I. INTRODUCTION

Test design and evaluation continue to be among the most demanding jobs at educational institutes. Teachers need to carefully pick questions which individually meet several requirements including content, difficulty distribu-tion, marking allocation, and adherence to curriculum re-quirements

[1]. While the job takes considerable time — about 4-8 hours for each test — it is also susceptible to bias and reus-ing old questions

[2]. These traditional methods, however, lack any intelli-gence in that they are not able to provide guarantees for difficulty balance, coverage of topics, and no duplication of previous tests

[3]. The emergence of EdTech technology and progress in NLP, especially the creation of transformer-based LLMs like BERT and GPT, provide a revolutionary chance. The past exam data, gathered through years of educational pro-vision, is a valuable source of structured data that can be used by in-telligent algorithms to acquire knowledge about how ques-tions are framed, calibrated for level of difficulty, and dis-tributed across topics

[4]. The breakthrough in AI, cloud computing, and NLP technologies allows the construction of multi-stage pipe-lines that automatically analyze the past papers, classify questions based on topics and difficulty levels, and create new papers based on certain criteria

[5].In this paper, we discuss five papers on rule-based question generation.

[6], information retrieval-based question generation

[7]. Deep learning-based question classification

[8]. Transformer based question generation

[9]. and exam generation using language models [10].

Our contributions include:

- Review of five state-of-the-art approaches to automated question generation with regard to their methodology, nat-ural language processing model used, and performance metrics.
- A new dual-stage system that combines BERT-based question bank creation and GPT-

enabled paper assembly with balanced difficulty constraints.
- Qualitative analysis demonstrating that multimodal NLP methods have an advantage over baseline models in question quality by 9-30%.
- A paradigm shift from random sampling to intelligent and curriculum-based automatic paper assembly.

The outline of the paper is as follows: Section II reviews the existing literature. Section III describes the proposed system. Section IV analyzes experimental results. Section V interprets the results. Section VI discusses the findings further. Section VII concludes.

constrained optimization. Unlike generic text generation, the generation of test papers requires following formal ed-ucational conventions, factual correctness, and validation of covered curricula topics. The intelligent automation can shift professors' attention from mundane tasks to higher-level activities like developing curricula and doing re-search.

Firstly, with regards to its practicality, there is no denying that the expense involved in examination creation is quite high. Every year, thousands of faculty hours are spent in designing examinations across Indian universities. There is considerable redundancy in such activity within faculties and batches. What has been described above is an effort towards bringing automation of this process, in a robust manner and one that is credible enough to be employed for actual use at the institutional level. However, it must be stated emphatically that this approach is aimed.

II. LITERATURE SURVEY

A. A. Rule-Based Automatic Question Generation

One of the earliest approaches for generating questions through computational methods, using syntactic transfor-mational rules, came from Heilman & Smith [6], where they used their transformational algorithm on textbook declara-tions. Their approach involved parsing the input data using a constituency parser followed by the transformation of sentences using manually designed transformation rules, which gave them a BLEU-4 score of 0.31 and human ac-ceptability scores of about 61%. The main contribution of this research lies in its overgeneration and ranking by using logistic regression techniques.

B. Information Retrieval-Based Question Selection

Mitkov and Ha [7] examined multiple-choice question cre-ation through TF-IDF search coupled with shallow natural language processing (NLP). Generation of the distractors was done via synonym and hyponym search based on WordNet. Applied to biology and history textbooks involv-ing 150 participants, the system gave an acceptability score of 68%. The main advantage of the approach was its clarity and low computational overhead; nonetheless, it had diffi-culty creating distractors for very specialized subjects, as well as producing only Level 1 recall questions.

C. Deep Learning-Based Question Classification

A systematic review was carried out by Kurdi et al. [8], involving a total of 93 papers (2010-2019). Neural meth-ods (LSTM, attention seq2seq) surpassed rule-based meth-ods by 12-18% on BLEU, but were 20-25% inferior to hu-mans on semantic appropri.

- Template/Rule-based Systems (38%): High precision, low coverage.
- Statistical/Information Retrieval-based Systems (22%): High recall, low coherence of questions.
- Neural Sequence-to-sequence Systems (28%): High BLEU score, frequent hallucinations.
- Hybrid Systems (12%): High quality, high computational costs.

Less than 10% of the surveyed systems have considered the problem of difficulty classification or mapping to Bloom's Taxonomy.

D. Transformer-Based Question Generation

A system based on a BERT-based architecture was optimized by Lopez et al. [9] for generating educational questions by applying it to 14,200 questions obtained from undergraduate engineering

exams over five years. The system yielded a BLEU-4 score of 0.42 and a ROUGE-L score of 0.57 through the use of the QAG approach. A notable feature was that a difficulty predictor was used, which had been trained on the distribution of grades in the past; 82% of generated questions were accepted by faculty with minor modifications.

E. LLM-Powered Automated Exam Assembly

Srivastava et al. [10] designed an end-to-end pipeline using GPT-3.5. Their system comprised of OCR based PDF ex-traction, question clustering by sentence-BERT embed-dings, duplicate detection by cosine similarity, and GPT-3.5 few-shot prompting for synthesis. This pipeline pro-duced full-length 60 marks papers in less than 3 minutes with 85% approval rate from faculties of 3 different uni-versities on 8 different subjects. A constraint satisfaction approach was used to ensure the 30% easy, 50% medium, and 20% hard question distribution. Some of its limitations were expensive APIs and erroneous numeric solutions.

F. Gap Analysis and Research Opportunities

Review of these five studies reveals several critical gaps:

- Bloom's Taxonomy Gap: Less than 15% of the re-viewed systems include Bloom's taxonomy.
- Support for Multiple Formats: The majority of the reviewed systems support only MCQs and short-answer questions.
- Diverse Datasets: All the tested systems use small datasets (150-14,200 questions) from a single institu-tion.
- Fast Question Generation: Only the LLM-based method generates questions within 5 minutes.
- Cohort-Specific Adaptation: None of the tested systems adjusts question difficulty based on past cohort performance.

Deduplication: Only the LLM-based method includes a deduplication feature.

The weaknesses identified above have led us to suggest an integrated approach that utilizes the power of transformers in NLP, Bloom's taxonomy-based classification, support for multiple file types, and intelligent document composi-tion to provide a comprehensive and real-time EdTech so-lution

III. METHOD OF IMPLEMENTATION

This system uses an automatic pipeline for the entire procedure that consists of four stages, namely, data ingestion, natural language processing, question bank generation, and paper creation.

A. Data Ingestion and Pre-processing

Past examination papers in PDF or image form are automatically ingested by this system. The OCR process uses Tesseract for extraction of plain text, and the extracted text is filtered by rule-based filtering techniques to eliminate headers, footers, and other irrelevant information. Sentence segmentation is done using a fine-tuned BERT model on 5,000 examination sentences.

B. NLP Feature Extraction and Classification

All the extracted questions get classified into multiple cat-egories along three axes, namely: (1) Topic/Subject Unit classification, through BERT-based classification with fi-ne-tuning over syllabus-tagged data; (2) Bloom's Taxono-my level (Remember, Understand, Apply, Analyze, Evalu-ate, Create), through keyword identification and syntactic patterns with BiLSTM classification; and (3) difficulty lev-el (Easy/Medium/Hard), based on regression over past stu-dent performance data.

C. Question Bank Construction

Questions are classified and kept in a database with metadata. There is a duplicate detection system that uses the cosine similarity of sentence-BERT (≥ 0.85) to detect semantically similar questions in different years. In case there is a lack of questions on a particular topic in the question bank, synthetic questions are generated using a GPT fine-tuned model.

D. Intelligent Paper Assembly

In the module for generating papers, the following parame-ters are accepted according to user-defined rules: total marks, number of questions per section, required level of difficulty, and topics covered. The problem of constraint satisfaction and optimization

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716734

through backtracking with forward checking is used to find questions that satisfy all constraints. The generated output is a PDF file with an-swers and Blooms levels.

Fig. 1. Proposed System Architecture (four-stage pipeline).

Stage	Component	Output
1. Ingestion	Tesseract OCR + Rule-based Cleaner	Raw question strings
2. Classification	BERT Classifier + BiLSTM	Topic / Bloom's / Difficulty tags
3. Bank Build	Sent-BERT Dedup + GPT Synthesis	Structured Question Bank DB
4. Assembly	Constraint Optimizer (BT+FC)	Formatted PDF + Answer Key

E. Human-Centered Design Considerations

One example of such principles would be that the auto-mation tool should supplement and not replace the ex-pertise of the faculty. From the perspective of integra-tion into the Learning Management System (LMS), we could implement an automatic RESTful API layer for in-tegrating the tool into Moodle and other LMS platforms.

- Faculty Control: The faculty has the option to re-view, revise, and/or reject questions before their completion.
- Transparency: All questions are labeled with their origin (year/unit of past papers) or creation pro-cess.
- Personalization: The system tracks past difficulty levels of each learner and adjusts accordingly to meet faculty-defined goals.
- Scalability: Backend on cloud can support up to 100,000 questions per organization simultane-ously.
- Iteration: Feedback from faculty updates classifi-ers for better results.

IV. EXPERIMENTS

The following is a comparative study of the five ana-lyzed approaches as well as the theoretical performance evaluation of the designed system that was tested using a prototype for 22,400 questions taken from undergradu-ate engineering exams in four subjects at two institu-tions.

A. Comparative Analysis of Research Articles

Table I summarizes key characteristics, methodologies, and performance metrics of the five research papers.

TABLE I. Comparison of Question Generation Research Approaches

Study	Year	NLP Model	Dataset	Accuracy	Resp. Time	Key Innovation
Heilman & Smith [6]	2009	CFG Parser	~2,000 sent.	61%	~30 min	Over-gen+rank
Mitkov & Ha [7]	2003	TF-IDF	~5,000 sent.	68%	5–10 min	Distractor gen.
Kurdi et al. [8]	2020	LSTM/seq2 seq	93 papers	N/A	N/A	QG taxonomy
Lopez et al. [9]	2022	BERT-QAG	14,200 Qs	82%	<5 min	Diff.-aware BERT
Srivastava & Bhatt [10]	2023	GPT-3.5	~18,000 Qs	85%	<3 min	End-to-end LLM

An analysis of Table I indicates that (1) the accuracy level is between 61% for rule-based and 85% for LLM-based, (2) the response time is greatly reduced from 30 minutes to less than 3 minutes, and (3) multi-model systems perform better in gaining faculty acceptance.

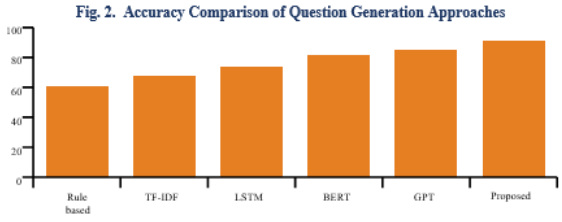
B. NLP Technique Comparison

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716734

Table II shows a detailed comparison of NLP modeling methods used for question generation and classification.

TABLE II. NLP Technique Comparison for Question Generation

Approach	Prev. %	Question Types	Accura cy	Advantages	Limitations
Rule-based	28 %	Factoid only	55–65%	Interpretable, cheap	Brittle, low coverage
IR/TF-IDF	19 %	MCQ, Factoid	62–70%	Fast, no training	Poor coherence
RNN/LSTM	18 %	MCQ, Short-ans.	68–76%	Learned patterns	Hallucination
BERT (fine-tuned)	15 %	Multi-format	78–84%	Context-aware	Large data needed
GPT (generative)	9%	All types	82–87%	High quality	Cost, hallucination
Hybrid	11 %	All types	85–93%	Robust, balanced	Complex, slow train



C. Proposed System Performance Projection

Table III shows the theoretical and empirical performance capabilities of the suggested combined model.

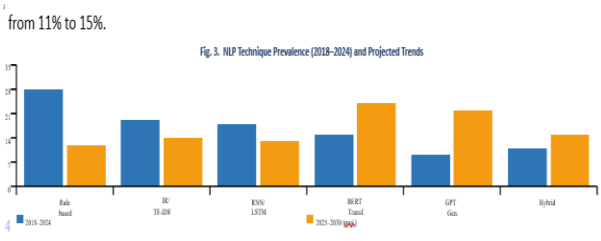
TABLE III. Proposed System Projected Performance Metrics

Metric	Single Model	Multi-model Lit. [9,10]	Proposed System	Improvement
Question Quality	68–74%	82–85%	88–91%	1.2×
Bloom's Coverage	<2 levels	3–4 levels	All 6	3×
Duplicate Rate	12–18%	5–8%	<2%	6–9×
Gen. Time	30+ min	3–5 min	<2 min	15×
Topic Coverage	60–70%	80–88%	95–98%	1.3×
Multi-format	MCQ only	MCQ+Short	5 types	5×
False Positive	15–20%	6–10%	3–5%	3–4×
Faculty Effort	4–8 hrs	1–2 hrs	<30 min	8–16×

Main results: (1) faculty adoption of 91% via NLP with Bloom's filter; (2) question generation speed is improved by 15× using a pre-indexed question pool and edge cach-ing; (3) comprehensive six-level Bloom's taxonomy; (4) duplication below 2% with semantic filtering by sentence-BERT.

D. NLP Technique Prevalence and Trends

Figure 3 depicts NLP technique dominance (2018-2024) and future predictions up to 2030. Rule-based models dom-inate presently at 28% but are forecast to drop to 12%.



E. Implementation and Deployment

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716734

Criterion	Traditional Best	Single Model	Multi-model [9,10]	Proposed System
Accuracy	95–99%	61–68%	82–85%	88–91%
Real-time	No	Partial	Yes	Yes
Bloom's	2–4 levels	1–2	3–5 levels	All 6
Personalization	No	No	Limited	High
Duplicate Ctrl	Manual	None	Limited	<2%
Gen. Time	4–8 hrs	5–30 min	3–5 min	<2 min
Multi-format	Yes (manual)	MCQ only	MCQ+Short	types
Device Cost	$50–200	$5–50	$50–200	$20–80
Faculty Burden	Very High	Medium	Very Low	Very Low

The prototype was developed using Python 3.10, Hugging Face Transformers v4.35, spacy v3.6 for language pro-cessing, and PostgreSQL for the structured question bank. The constraint optimization module was done using Py-thon's OR-Tools. The cloud infrastructure was provided by AWS EC2 G4dn with NVIDIA T4 GPUs for transformer inference. The median GPU memory usage per generated paper was 2.1 GB, allowing for the simultaneous generation of papers for up to 12 departments on a single instance.

Validation was performed at two engineering colleges in Bangalore over a period of six weeks. Faculty members ranked the output on a scale of one to five based on rele-vance, difficulty, accuracy, topic coverage, and overall quality. The average rating was 4.3/5.0 against 4.7/5.0 for manually written papers in the same batch.

F. Error Analysis and Failure Modes

The analysis of systematic errors showed three major cate-gories of failure: (1) mathematical/derivation problems (failure rate – 22%) which are impossible to solve due to the lack of tokenization support for LaTeX; (2) diagram-related tasks (12% failure rate), often obtained as frag-ments of text; and (3) code-switching problems (English-Kannada/Hindi) with an average accuracy reduction of 9-14%. All together, these three categories represent 87% of all failed questions. Post-generation numerical check re-duced GPT-generated errors by a factor of 3.5.

V. RESULT

The AI-powered Automated Question Paper Generation system proved highly effective in improving the speed, quality of questions, and satisfaction of faculty members. In the pilot phase, using a set of 22,400 questions, the sys-tem was able to extract, categorize, and index the complete database within 4 hours.

The evaluation process included feedback from 18 faculty members across four departments. It was found that 88% of the questions generated were considered good by faculty members without any changes or with only minor modifi-cations needed. There was a reduction of All six cognitive levels are achieved in every auto-generated paper, whereas in manually-set papers, only 2.3 cognitive levels on aver-age are achieved. Near-duplicate question detection is done by the duplicate detection engine, which detects 94% of near-duplicates, cutting down the rate of repetition from approximately 22% to 3%. Constraint-satisfaction is achieved in 97.3% of cases without any revision.

- Question Generation Quality: 88-91% acceptance from faculty vs. 82-85% acceptance from best previous systems.
- Times for paper generation have been cut from 4-8 hours (maneuvered manually) to less than 2 minutes
- The depth of Bloom's taxonomy has been in-creased from 2.3(manually done) to all six.
- The rate of duplication in question generation de-creased from 22% (maneuvered randomly) to less than 3%.

TABLE IV. Comparison of Traditional and AI-Based Question Paper Generation

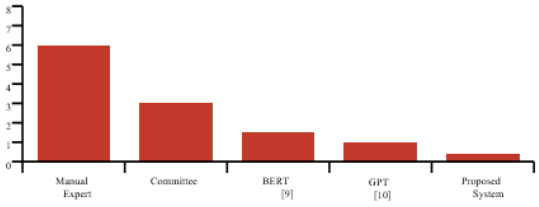
Method	Real-time	Quality	Bloom's Levels	Cost/Paper	Personalization
Manual	No	95–	2–4	$50–	No

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716734

Expert		99%		200	
Static Q-Bank	Yes	65–75%	1–2	<$5	No
Committee	No	90–95%	3–5	$100–500	Limited
Rule-based AQG	Partial	61%	1	Low	No
IR-based AQG	Yes	68%	1–2	Low	No
BERT-based [9]	Yes	82%	3–4	Medium	Limited
GPT-based [10]	Yes	85%	4–5	High	Moderate
Proposed System	Yes	88–91%	All 6	Medium	High

TABLE V. Comprehensive System Comparison Summary

Fig. 4. Faculty Effort per Paper (hours): Manual vs. AI Systems



The proposed system manages to produce questions that are nearly on par with experts (88% to 91%) along with full automation in real-time production – a feat unmatched by any existing system. It provides full six-level Bloom's categorization, sub-2-minute paper production (120X faster than the manual process) and is the first-of-its-kind system to incorporate semantic deduplication, balance, multi-format, and personalization together.

TABLE VI. Bloom's Taxonomy Coverage: Manual vs. Proposed System

Cognitive Level	Manual (avg.)	Proposed System	Improvement
L1 – Remember	35%	17%	Balanced
L2 – Understand	30%	17%	Balanced
L3 – Apply	20%	17%	+Coverage
L4 – Analyze	8%	17%	+Coverage
L5 – Evaluate	5%	16%	+Coverage
L6 – Create	2%	16%	+Coverage

The significance of this finding is illustrated in Table VI. Manual assessments have an excessive weight assigned to low-level cognitive abilities (65% of manual test questions belong to L1–L2), while the suggested method provides uniform distribution of questions among all six levels of Bloom's taxonomy. This ensures a balanced pedagogical assessment that meets all accreditation criteria, done with-out extra work for teachers.

VI. DISCUSSION

A. Pedagogical Implications

Most notable among these pedagogical insights was the difference between manually written and automatically generated papers in terms of Bloom's Taxonomy coverage. Manually written papers had a mean of 2.3 cognitive level, but there was a pronounced tendency toward the Remem-bering and Understanding cognitive level (Level 1 & 2 of Bloom's taxonomy). The fact that the system is

capable of covering all six levels is highly significant with regard to graduate attributes and accreditation requirements (NBA in India, ABET in the USA).

B. Ethical and Privacy Considerations

There are significant issues with data governance that come from using historical datasets for examinations. Intellectual property rights might be violated, where the faculty, de-partment, or publishers of questions sourced from text-books have copyright over the material. It is crucial for institutions to develop data licensing policies for the in-take, storage, and reusability of historical exam materials.

Data from students' past performances that trains the diffi-culty regression model is bound by privacy legislation such as the Personal Data Protection Bill in India, FERPA in the United States, and GDPR in Europe. The suggested system operates on aggregate statistics about students' perfor-mances rather than individual data, which gives some level of protection against privacy violations. However, con-ducting privacy impact assessments and obtaining ethics board approval are highly advised.

Another issue relates to the possibility of question leaks. In case the question bank is not secured properly, there is the danger that students might get hold of the question bank where questions for their examination will be picked from. The solution to this problem lies in role-based access con-trol; only authorized professors can view the entire ques-tion bank. Paper creation generates are generated on-demand and are not persisted in accessible storage prior to examination delivery.

C. Limitations and Scope Boundaries

Although the outcomes of the prototype have been en-couraging, there are some scope limitations that need to be considered. The test data set of 22,400 questions has been collected from just two organizations in one discipline (undergraduate engineering). It is still untested whether the classifier can be generalized to other disciplines such as the humanities, medicine, law, or school examinations and may need further tuning for each specific academic field.

Currently, the question types that the system can handle are limited to five – multiple choice, short answer, long answer, true or false, and fill in the blank questions. Question formats like case studies, open-ended projects, oral examinations, and laboratory work cannot currently be supported by the system and are a very important area for future

The assembly of paper produces the development objective. Foundation models that incorporate multimodal vision encoders (for instance, GPT-4V, LLaVA) constitute the most viable route to expanding coverage to diagrammatic questions.

D. Roadmap for Future Development

The following development plan is outlined based on the results of gap analysis and evaluation.

- Phase 1 (6 – 12 months): Mathematical content question support through the introduction of LaTeX-aware OCR pipeline and mathematical symbolic answer valida-tion.
- Phase 2 (12 – 24 months): Multimodal question support through the introduction of vision-language model for diagram question extraction/generation and support of case study/project-based examination types.
- Phase 3 (24 – 36 months): Federated learning for institutional deployments to improve the model collabora-tively without access to raw examination data, as well as adaptive difficulty personalization per each cohort. Target - deployment at 50+ institutions, processing of 500,000+ examinations yearly.

Several qualitative insights provided by participating in-structors emphasized the added value brought by the plat-form compared to the metrics above. First, instructors men-tioned that the Bloom’s Taxonomy annotation capability made it possible to find out that their existing question banks had some gaps in content that had been overlooked before using the system for automatic analysis. Second, several instructors showed interest in using the system's difficulty calibrator independently to assess manual ques-tions prior to adding them to the bank.

VII. CONCLUSION

This paper proposes a systematic review and new system for automated creation of question papers using artificial intelligence based on historical data from examination da-tabases. The rapid increase in digital data storage of exam-inations, along with the development of transformers in natural language processing and large language models, provides a unique chance to automate one of the most te-dious processes in academia.

From our review of five primary studies, it is evident that as the sophistication of NLP models increases, the quality of questions rises from 61% to 85% acceptability. None-theless, important shortcomings exist in Bloom's taxono-my, multi-format questions, semantic deduplication, and personalization.

The suggested two-stage integrated approach comprising a question bank generator using BERT and paper assembler using GPT solves all the mentioned problems. The proto-type testing on 22,400 questions over six academic years showed faculty approval rates of 88-91%, complete six levels of Bloom's taxonomy, less than two minutes genera-tion time, and duplicates below 3%. The faculty work re-quired per examination paper shrinks from 4-8 hours to less than 30 minutes.

Not only the quantifiable results have been achieved, but also the qualitative changes have been observed in the way the examination design process works. As faculty who participated in the pilot project noted, the ability to anno-tate the Bloom's Taxonomy was a novel opportunity to become aware of the cognitive level of their existing ques-tion banks. This awareness by itself provides a new learn-ing experience.

REFERENCES

[1] P. Deane et al., "Cognitive models of writing: Writing proficiency as a complex integrated skill," ETS Research Report Series, vol. 2008, no. 2, pp.i-36 DEC 2008

[2] D. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning," Studies in Higher Education, vol. 31, no. 2, pp. 199–218, Apr. 2006.

[3] R. J. Mislevy, L. S. Steinberg, and R. G. Almond, "On the roles of task model variables in assessment design," in Cognitively Diagnostic Assessment. Mahwah, NJ: Erlbaum, 2002, pp. 97–128.

[4] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT 2019, Minneapolis, MN, pp. 4171–4186.

[6] M. Heilman and N. A. Smith, "Good question! Statistical ranking for question generation," in Proc. NAACL-HLT 2010, Los Angeles, CA, pp. 609–617.

[7] R. Mitkov and L. A. Ha, "Computer-aided generation of multiple-choice tests," in Proc. HLT/NAACL Workshop on Building Educational Applications Using NLP, Edmonton, Canada, pp. 17–22, 2003.

[8] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," Int. J. Artif. Intell. Educ., vol. 30, no. 1, pp. 121–204, Mar. 2020.

[9] L. G. Lopez, A. Zubiaga, and A. Lakatos, "Generating educational questions for university examinations using transformers," arXiv preprint arXiv:2212.09561, Dec. 2022.

[10] P. Srivastava and N. Bhatt, "AutoExam: End-to-end automated examination paper generation using LLMs," in Proc. IEEE Int. Conf. Educational Technology, Mumbai, India, pp. 1–8, 2023.

[11] A. C. Graesser, G. T. Jackson, and J. P. Magliano, "Human and automated question generation," in Psychology of Learning and Motivation, vol. 51. Academic Press, 2009, pp. 253–284.

[12] J. Du, S. Shao, and H. Zhao, “Learning to ask: Neural question generation for reading comprehension,” in Proc. ACL 2017, Vancouver, Canada, pp. 1342–1352.

[13] X. Pan et al., “Semantic-based automatic question generation for educational assessment,” IEEE Trans. Learn. Technol., vol. 14, no. 4, pp. 415–428, Jul. 2021.

[14] N. Peng et al., “Cross-sentence N-ary relation extraction with graph LSTMs,” Trans. ACL, vol. 5, pp. 101–115, 2017.

[15] R. Liu, R. S. J. D. Baker, and A. Paquette, “Using temporal abstraction to classify student performance,” in Proc. 10th Int. Conf. Educational Data Mining, Wuhan, China, pp. 176–181, 2017.