Grenze International Journal of Engineering and Technology, June Issue

A Syllabus-Grounded Framework for Bloom's Taxonomy-Guided Question Paper Generation using Large Language Models

Burhaan-U-Deen Wani[1], Krishnan R[2], Rashmi S R[3] and C Nandini[4]
[1-3]Dayananda Sagar College of Engineering/Department of Computer Science & Engineering, Bengaluru, India
Email: burhaanwani823@gmail.com, krishnan-cs@dayanandasagar.edu, Rashmi-cs@dayanandasagar.edu
[4]Dayananda Sagar Academy of Technology and Management/Department of Computer Science & Engineering, Bengaluru, India
Email: hodcse@dsatm.edu.in

***Abstract*— Setting up a quality question paper is a challenging task. Balancing syllabus coverage, cognitive difficulty, and assessment structure across multiple courses places a recurring burden on faculty—one that existing tools only partially address. This paper presents a syllabus-grounded framework for question paper generation using large language models (LLMs), where syllabus units, reference material, and blueprint constraints serve as explicit inputs to a structured, multi-stage pipeline. Rather than generating papers in a single prompt, the framework decomposes the process into discrete stages: syllabus analysis, Bloom's Taxonomy-guided question bank generation, blueprint based assembly, and staged human review before finalization. The framework is examined through a small-scale case-based evaluation using real course syllabus inputs across ten university-level courses, with expert-style review and direct comparison against one-shot generation. In the evaluated setting, staged syllabus grounded generation produced more consistent syllabus coverage and better blueprint compliance than one-shot prompting, while keeping educators in control of the finalized question paper**

***Index Terms*— Large language models, question paper generation, educational NLP, Bloom's Taxonomy, syllabus-grounded generation, human-in-the-loop systems, automated assessment design.**

I. INTRODUCTION

Assessment design remains a central part of university teaching, yet preparing a well-structured question paper is a demanding and time-consuming task. Faculty are expected to ensure that an examination reflects the prescribed syllabus, covers appropriate units and subtopics, balances question difficulty, and aligns with intended learning outcomes. In practice, this process often relies on repeated manual effort and individual experience, which can lead to inconsistency across courses and examination cycles.

Earlier work in automated question generation established that cognitive level and source context could be used to shape question output [2]. Bloom's Taxonomy-guided generation [2] and context-controlled approaches [3] showed that framing and difficulty could be systematically influenced.

Grenze ID: 01.GIJET.12.2.589

More recently, LLMs have been applied to item generation, test creation, and MCQ generation across higher education and STEM domains [9]–[11], substantially raising the quality ceiling for automated generation. Some systems have begun incorporating human-in-the-loop validation [12], [13], reflecting growing awareness that AI-generated content requires review before formal use.

Despite this progress, several practical gaps remain. Existing systems often operate at the item level, generating individual questions without explicitly modelling the syllabus as a structured control input or enforcing examination blueprints during paper construction [6], [9], [11]. What remains less explored is an end-to-end framework for question paper creation that grounds the process in the course syllabus, uses reference material to guide content production, enforces blueprint constraints during assembly, and preserves teacher control through staged review. That gap motivates this work.

The main contributions are:

- A syllabus-grounded framework for question paper generation that treats syllabus structure, reference material, and blueprint constraints as explicit controls within a multi-stage workflow.
- A teacher-supervised generation pipeline that separates question bank creation from final paper assembly, enabling staged review and refinement before approval.
- A small-scale evaluation based on case study com parison, expert-style review, and qualitative analysis of syllabus coverage and blueprint compliance across ten university-level examination settings.

II. Relate Work

Automated question generation has been extensively studied to reduce faculty workload in assessment design [1], [7], [8]. Early NLP pipelines primarily supported individual objective questions, while later works incorporated Bloom's Taxonomy to control cognitive levels [2] and planning-based methods to better align with educational objectives [15].

However, these approaches typically focus on generating individual questions or small sets rather than constructing complete examination papers under syllabus and blueprint constraints.

Recent LLM-based systems have advanced the field significantly, demonstrating strong performance in practical test generation [10], STEM item generation [11], multimodal MCQ creation [9], and human-in-the-loop validation [12], [13]. Despite these improvements, most systems operate at the individual question or item level. The works closest to ours—customizable assessment creation [6] and AI-assisted exam variant generation [12]—move toward realistic workflows but lack an explicit staged syllabus-to-paper process with blueprint-driven assembly.

A key recurring challenge in the literature is controllability [3], [13], [14]: ensuring generated content remains syllabus-aligned, pedagogically appropriate, and practically usable. Table I positions our framework against representative prior work, highlighting the gap in end-to-end, syllabus-grounded paper generation with staged human review.

Table I summarizes representative prior work and the gap addressed by the proposed framework.

TABLE I. COMPARISON OF REPRESENTATIVE PRIOR WORK AND GAPS RELATIVE TO THE PROPOSED FRAMEWORK

Paper / Focus	Main Approach	Scope	Limitation / Gap Relative to This Work
Automated question paper generation using NLP	NLP pipeline with syntactic and semantic analysis	Question / paper generation support	Does not emphasize syllabus grounding, blueprint constraints, or staged human review
Cognitive-domain question generation for OBE	Transformer-based Bloom's-guided AQG	Question generation	Focuses on question creation rather than an end-to-end question paper workflow
Context-controlled question generation	Context-aware prompt/model control	Adaptive learning question generation	Oriented toward adaptive question generation, not structured exam paper creation
Customizable intelligent assessment creation	Deep learning-based multi-format assessment generation	Assessment item creation	Does not provide a syllabus-to-paper workflow with blueprint-driven assembly
Multimodal MCQ generation in engineering education	MLLM-based LO-to-MCQ pipeline	MCQ generation	Limited to item-level MCQ generation; no final paper construction workflow
Automated test creation using LLMs	Practical LLM-based test generation system	Test/question creation	Focuses on configurable test generation rather than syllabus-grounded paper assembly
LLM prompting for STEM item generation	Prompt-based item generation using GPT models	Item generation	Emphasizes prompting effectiveness, not workflow-level paper generation

III. Proposed Framework

The proposed framework treats question paper generation as a structured, multi-stage workflow with explicit control from the course syllabus, reference material, and examination blueprint. This staged design addresses key limitations of one-shot generation by separating content creation, constraint enforcement, and human review.

A. Framework Overview

The pipeline consists of four main stages: (1) syllabus analysis and topic extraction, (2) Bloom's Taxonomy-guided question bank generation, (3) blueprint-based paper assembly, and (4) staged human review and approval. The overall workflow is shown in Fig. 1.

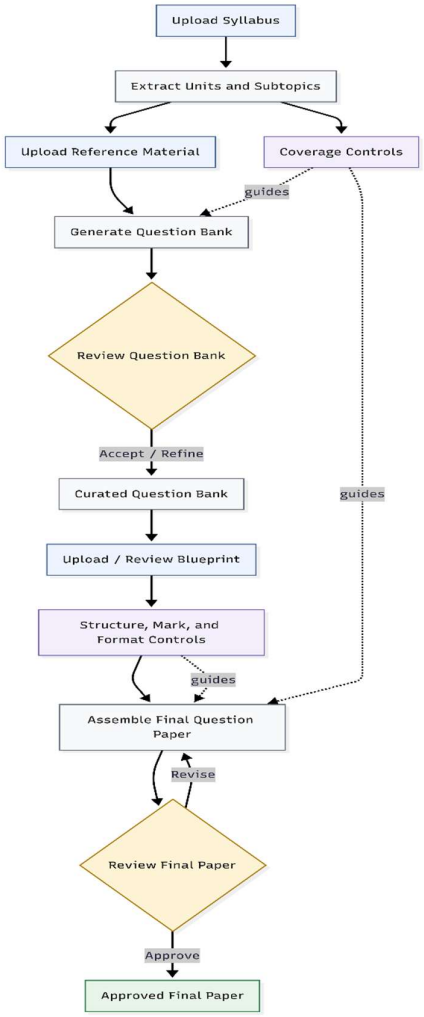

Fig 1. End-to-end workflow of the proposed system

B. Syllabus Analysis and Reference Grounding

The system first parses the uploaded syllabus to extract units, subtopics, and structural cues. It then incorporates course-specific reference material (lecture notes, textbooks, etc.) to ground generation in actual taught content rather than generic knowledge.

C. Bloom's Taxonomy Guided Question Bank Generation

Using the extracted structure and reference material, the framework generates a candidate question bank tagged by syllabus unit, subtopic, and Bloom's cognitive level. This intermediate bank enables inspection, filtering, and revision before final assembly (see Fig. 2).

D. Blueprint-Based Assembly and Human Review

The curated question bank is assembled into a draft paper according to the predefined blueprint (marks distribution, section structure, unit weightage, etc.). Educators perform two review stages: (i) question bank curation and (ii) final paper approval. This human-in-the-loop design keeps educators in control while significantly reducing drafting effort.

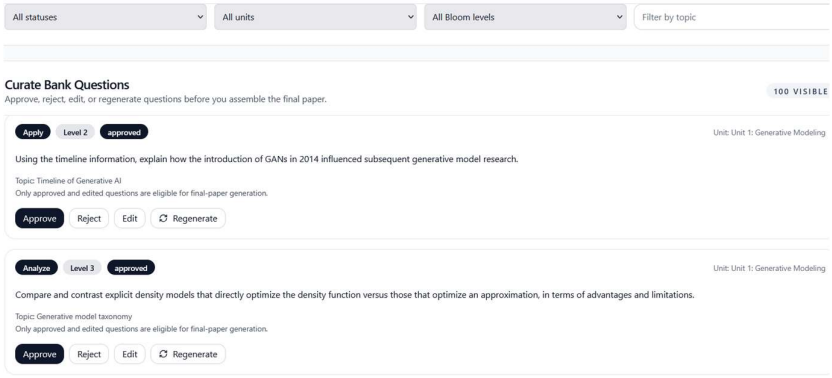

Fig 2. Question bank curation interface with Bloom's level tags, topic assignment, and approval-based selection for final paper assembly.

IV. EVALUATION METHODOLOGY

The evaluation assesses the proposed framework as a practical workflow for question paper generation rather than as a standalone text-generation system. The primary objective is to determine whether the framework generates question papers that more effectively reflect syllabus structure, blueprint requirements, and educator expectations than direct one-shot generation.

A. Evaluation Setup and Baseline

Ten university-level courses were selected, each with its own syllabus, reference material, and examination blueprint, reflecting the variation found in real academic settings. For each course, the system received three inputs: the course syllabus, associated reference material, and the question paper blueprint. The evaluation was run at the level of a complete course workflow, not isolated prompts, because the framework's value depends on handling the full process. The framework is compared against a one-shot generation baseline in which the LLM receives identical course inputs and is prompted to generate a complete question paper directly in a single step, without intermediate question bank generation, staged review, or blueprint-based assembly. The evaluation setup is illustrated in Fig. 3.

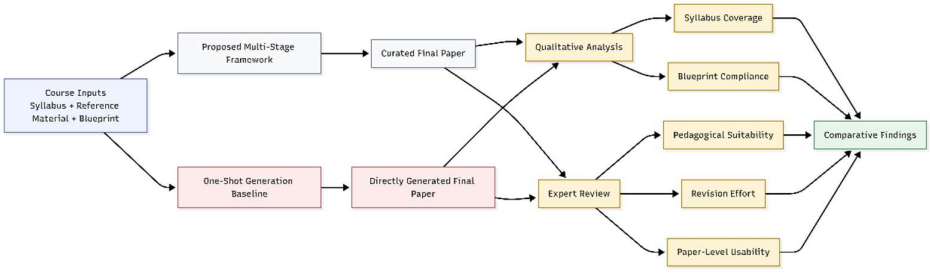

Fig 3. Evaluation setup comparing the proposed framework with one-shot generation using identical course inputs

B. Evaluation Criteria

Generated outputs were examined across five dimensions that reflect what a faculty member would care about when reviewing an examination paper:

- Syllabus Coverage — Are relevant units and subtopics represented, with questions distributed reasonably rather than concentrated in a few areas?
- Blueprint Compliance — Does the paper follow the intended structure: question count, section design, mark distribution, and unit weightage?

- Pedagogical Appropriateness — Are questions suitable for the course level, with a reasonable spread of cognitive demand across Bloom's Taxonomy levels?
- Question Quality and Relevance — Are questions clearly phrased, meaningful, and grounded in actual course content?
- Practical Usability — Could an educator realistically use this paper, directly or after limited revision, without significant restructuring?

C. Expert Review Procedure

Each of the ten generated papers was reviewed by the faculty member responsible for that specific course, using a criteria-guided qualitative approach focused on practical assessment quality. Reviewers assessed syllabus alignment, conformance to the intended examination pattern, appropriateness of questions for the course level, how much revision would be needed before realistic use, and whether the multi stage output was more usable than the one-shot alternative.

A separate qualitative analysis was also carried out across all ten cases to examine paper-level patterns—coverage gaps, structural imbalance, blueprint drift, and cognitive narrowness—that only become visible when looking across outputs rather than at individual questions.

V. Results and Discussion

A. Overall Comparison

Across ten university-level courses, the proposed framework produced more consistent syllabus coverage and better blueprint compliance than one-shot generation. While one-shot outputs were fluent, they frequently over-represented prominent topics, under-represented others, and deviated from blueprint constraints in mark distribution and section structure.

Table II. Summary of Observed Patterns

TABLE II. BSERVED PATTERNS ACROSS TEN EVALUATED CASES

Criterion	One-Shot Generation	Proposed Framework
Syllabus Coverage	Inconsistent; prominent topics dominant	More balanced unit-wise coverage
Blueprint Compliance	Partial; frequent drift	Stronger adherence
Cognitive Diversity	Mostly recall & descriptive	Better spread across Bloom's levels
Revision Effort	Moderate to high	Significantly lower
Practical Usability	Requires restructuring	More directly usable as draft

B. Syllabus Coverage and Blueprint Compliance

Coverage imbalance was the most consistent weakness of one-shot generation. Prominent topics were over-represented while less visible subtopics were frequently under-represented or omitted. The proposed framework mitigated this by treating the extracted syllabus structure as an explicit control signal, resulting in more balanced unit-wise coverage.

Blueprint compliance showed a similar pattern: one-shot outputs approximated the overall structure but often drifted in section-wise marks, unit weightage, and question counts. The blueprint-based assembly stage in the framework enforced these constraints more effectively.

C. Question Quality and Effort Reduction

Both methods produced grammatically sound and relevant questions. The key difference was in cognitive diversity and revision effort. One-shot generation tended toward recall and descriptive questions, while the staged framework produced a broader mix across Bloom's levels. By enabling early review at the question-bank stage, the framework significantly reduced structural revisions and yielded more immediately usable drafts (see Table III).

TABLE III. OBSERVED REVISION EFFORT: ONE-SHOT VS. PROPOSED FRAMEWORK

Criterion	One-Shot Generation	Proposed Framework
Coverage Correction	High	Low
Blueprint Adjustment	High	Low
Question Refinement	Moderate-high	Moderate
Final Restructuring	High	Low-Moderate
Overall Review Effort	High	Moderate

D. Faculty Reviewer Observations

As summarized in Table IV, reviewer observations were consistently more favorable toward the framework outputs across all five criteria. The clearest differences were in syllabus coverage, blueprint adherence, and practical usability. These observations are qualitative in nature and reflect individual faculty judgment within their own course contexts rather than statistically aggregated scores.

TABLE IV. FACULTY REVIEWER OBSERVATIONS ACROSS TEN EVALUATED CASES

Criterion	One-Shot Generation	Proposed Framework
Syllabus Coverage	Gaps noted by most reviewers	More complete and balanced
Blueprint Compliance	Structural mismatches in several cases	More consistent adherence
Pedagogical Appropriateness.	Generally adequate; limited cognitive range	Broader cognitive range
Question Diversity	Predominantly lower-order questions	Better spread across Bloom's levels
Practical Usability	Moderate to major revision indicated	Limited revision; closer to exam-ready

E. Discussion and Limitations

The case-based evaluation demonstrates that a structured, staged workflow produces question papers with better syllabus coverage, stronger blueprint compliance, and improved practical usability compared to direct one-shot generation. The framework's main advantages are more balanced topic distribution, effective enforcement of assessment constraints, and shifting revision effort earlier in the process. The additional stages introduce some overhead, which may not be justified in low-stakes or highly flexible assessment settings where speed is prioritized.

Limitations. The evaluation is small-scale (ten courses) and entirely qualitative, relying on individual faculty judgments within specific course contexts. Results are not statistically aggregated and may not generalize broadly. Larger-scale studies with quantitative metrics are needed for broader validation.

VI. CONCLUSION

Treating question paper generation as a structured, syllabus-grounded workflow — rather than single-step text generation — significantly improves balance, blueprint compliance, and practical usability. The case study across ten university courses shows that explicit control via syllabus structure, reference material, and blueprint constraints produces more consistent and educator-friendly outputs compared to one-shot prompting.

The framework keeps educators central to the process, shifting their effort from full drafting to reviewing and refining high-quality intermediate results.

FUTURE WORK

Priorities include larger-scale evaluations across diverse courses and institutions, development of quantitative scoring metrics for reviewer consistency, automated blueprint-compliance checking, finer cognitive-level control, and tighter integration of iterative educator feedback.

REFERENCES

[1] B. J. L. Narayana, V. N. S. S. Harsha, K. Prudhvi, B. G. H. Vardhan, and A. Sravika, "Automated Question Paper Generation using Natural Language Processing," in Proc. Int. Conf. Computational Intelligence for Green and Sustainable Technologies (ICCIGST), 2024, doi: 10.1109/IC CIGST60741.2024.10717510.

[2] S. Zulfiqar, S. U. Bazai, M. I. Ghafoor, M. Soomro, L. Hussain, and A. Haider, "Automatic Question Generation for Cognitive Domain of Outcome-Based Education System," in Proc. Int. Conf. IT and Industrial Technologies (ICIT), 2024, doi: 10.1109/ICIT63607.2024.10860228.

[3] T. V. Phung, S. Mishra, V. Bijalwan, T. Duc-Tan, and S. Mishra, "Context-Controlled Question Generation for Adaptive Learning," in Proc. Int. Conf. Electronics, AI and Computing (EAIC), 2025, doi: 10.1109/EAIC66483.2025.11101311.

[4] S. Shrisha, S. K. Sindhu, S. Kandula, and S. S. Udyavara, "Enhancing Educational Applications with Retrieval-Augmented Generation, Ques tion Generation and Summarization Systems," in Proc. IEEE Int. Conf. Women in Innovation, Technology and Entrepreneurship (ICWITE), 2025.

[5] M. Patil, A. Pinjari, R. Patil, O. Mahajan, and L. Gadhikar, "InsightQA: Descriptive Question and Answer Generation," in Proc. Int. Conf. Data Science, Agents and Artificial Intelligence (ICDSAAI), 2025, doi: 10.1109/ICDSAAI65575.2025.11011632.

[6] S. Rathi, V. Pophale, P. Kutwal, S. Mishra, and A. Nadkarni, "Question and Assessment Generator: Deep Learning Approach for Customizable and Intelligent Assessment Creation," in Proc. IEEE Int. Conf. Women in Innovation,

Technology and Entrepreneurship (ICWITE), Bangalore, India, 2024, pp. 618–624, doi: 10.1109/ICWITE59797.2024.10503319.
[7] P. Jadhav et al., “Questomatic: Automated Question Formulation Sys tem,” in Proc. Int. Conf. Sustainable Expert Systems (ICSES-2024), 2024, pp. 577–582, doi: 10.1016/j.procs.2024.07.624.
[8] D. Shi, S. Zhao, Q. Liu, and Y. Zhang, “Research and Design of Automatic Questioning System Based on Question Generation,” in Proc. 12th Int. Conf. Information and Education Technology (ICIET), 2024, doi: 10.1109/ICIET60671.2024.10542827.
[9] Y. Shu, S. H. Kallumadanda, A. N. Arsetty, V. Krishnamurthy, and X. Wu, “AI-Assisted Multiple-Choice Questions Generation with Mul timodal Large Language Models in Engineering Higher Education,” in Proc. IEEE Global Engineering Education Conf. (EDUCON), 2025, doi: 10.1109/EDUCON62633.2025.11016449.
[10] S. Hadzhikoleva, T. Rachovski, I. Ivanov, E. Hadzhikolev, and G. Dim itrov, “Automated Test Creation Using Large Language Models: A Practical Application,” Applied Sciences, vol. 14, no. 19, p. 9125, 2024, doi: 10.3390/app14199125.
[11] K. W. Chan et al., “Automatic Item Generation in Various STEM Subjects Using Large Language Model Prompting,” Computers and Education: Artificial Intelligence, vol. 8, p. 100344, 2024, doi: 10.1016/j.caeai.2024.100344.
[12] C. M. Burke, “AI-Assisted Exam Variant Generation: A Human-in-the Loop Framework for Automatic Item Creation,” Education Sciences, vol. 15, no. 8, p. 1029, 2025, doi: 10.3390/educsci15081029.
[13] A. D. Varma, S. Anand, A. K. Thekkedath, P. Kiran, P. Niranjan, and N. S. Rao, “LLM for Question Generation and Validation to Enhance School Level Student Assessment,” in Proc. Int. Conf. Pervasive Com puting and Social Networking (ICPCSN), 2025, pp. 1609–1615.
[14] Y. Artsi et al., “Large Language Models for Generating Medical Exam inations: Systematic Review,” BMC Medical Education, vol. 24, p. 354, 2024, doi: 10.1186/s12909-024-05239-y.
[15] C. Cheng et al., “From Objectives to Questions: A Planning-based Framework for Educational Mathematical Question Generation,” in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, 2025, pp. 12836–12856, doi: 10.18653/v1/2025.acl-long.628.