International Research Journal of Education and Technology

Peer Reviewed Journal, **ISSN: 2581-7795**

Elio AI Question Paper Generator Using Natural Language Processing

JEEVAN PRASATH J[1], NITHIN K[2] and JELCY M[3]

[1]*UG Student(III B.Sc. COMPUTER SCIENCE), Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

[2]*UG Student(III B.Sc. COMPUTER SCIENCE), Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

[3]*Assistant Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, India*

Abstract - The rapid advancement of artificial intelligence and natural language processing (NLP) technologies has created new opportunities to automate and enhance academic assessment processes. This paper presents Elio, an AI-powered question paper generator that leverages state-of-the-art NLP models to automatically produce high-quality, contextually relevant, and pedagogically sound examination questions from any academic or research topic. The system integrates transformer-based language models, domain-specific fine-tuning, and a structured generation pipeline to create diverse question types including conceptual, analytical, application-based, and critical-thinking questions. Experimental results demonstrate that Elio achieves a semantic relevance score of 91.4%, a human evaluator acceptance rate of 87.6%, and significantly reduces question paper preparation time by 78% compared to manual processes. This work bridges the gap between AI-assisted education and scalable assessment generation, with direct implications for universities, research institutions, and online learning platforms.

Keywords: *Natural Language Processing, Question Generation, Transformer Models, AI in Education, Automated Assessment, BERT, GPT, Elio System*

1. INTRODUCTION

Academic assessment is a cornerstone of knowledge evaluation in educational institutions worldwide. The process of designing examination question papers, however, is both time-consuming and subjective, relying heavily on the expertise and availability of domain instructors. With the growing demand for scalable, adaptive, and personalized learning experiences, there is an urgent need to automate this process using intelligent systems.

Natural Language Processing (NLP) has emerged as one of the most transformative branches of artificial intelligence, enabling machines to understand, interpret, and generate human language with increasing sophistication. Recent advances in transformer architectures — most notably BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) — have pushed the boundaries of text comprehension and generation to near-human levels. These models offer immense potential for educational applications, particularly in the automated generation of assessment content.

Elio is a novel AI-powered question paper generation system designed to address the limitations of manual question paper creation. By accepting a topic, research paper, or abstract as input, Elio leverages fine-tuned transformer models to generate structured, domain-specific questions across multiple cognitive levels as defined by Bloom's Taxonomy (Bloom, 1956). The system supports diverse question types including short-answer, descriptive, analytical, and application-based formats, making it suitable for use in undergraduate programs, postgraduate courses, and research journal review workflows.

This paper makes the following key contributions: (1) A complete architectural design of the Elio AI question paper generation pipeline; (2) Fine-tuning methodology for domain-adaptive NLP models; (3) Evaluation framework assessing semantic relevance, diversity, and human acceptance; (4) A comparative analysis against existing automated question generation tools. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the system architecture, Section 4 outlines the methodology, Section 5 presents experimental results, Section 6 discusses implications, and Section 7 concludes the paper.

2. LITERATURE REVIEW

Automated question generation (AQG) has been an active research area since the early 2000s. Rule-based systems dominated early efforts, relying on syntactic transformations of input sentences to produce questions. Heilman and Smith (2010) proposed one of the earliest data-driven approaches using overgenerate-and-rank strategies, which paved the way for machine-learning-based AQG systems.

With the advent of deep learning, neural sequence-to-sequence models brought a paradigm shift in AQG. Du et al. (2017) introduced an attention-based neural network model for reading-comprehension question generation that significantly outperformed earlier rule-based methods. The model demonstrated that contextual understanding of passages was critical for generating meaningful questions.

The introduction of transformer models further accelerated progress in this field. Pan et al. (2019) applied BERT for answer-aware question generation, showing superior performance on the SQuAD benchmark. Similarly, Dong et al. (2019) proposed a unified pre-training approach for both question answering and question generation using shared model weights, highlighting the synergistic relationship between the two tasks.

More recently, large language models such as GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020) have demonstrated remarkable zero-shot and few-shot capabilities in question generation. Studies by Lopez et al. (2021) and Fabbri et al. (2022) showed that prompt-engineered LLMs could

generate contextually rich, pedagogically diverse questions without task-specific fine-tuning, albeit with trade-offs in consistency and domain specificity.

Despite these advances, existing systems face notable limitations: (1) They often generate questions with limited cognitive diversity; (2) Domain adaptation remains a challenge for specialized academic fields; (3) Most systems lack structured output for ready-to-use examination formats. Elio addresses these gaps through a domain-adaptive fine-tuning strategy and a multi-tier question taxonomy aligned with Bloom's Taxonomy.

3. SYSTEM ARCHITECTURE

The Elio system is designed as a modular, scalable AI pipeline comprising four primary components: the Input Processing Module, the NLP Core Engine, the Question Generation Module, and the Output Formatting Layer. Figure 1 illustrates the high-level architecture of the system.

3.1 INPUT PROCESSING MODULE

The Input Processing Module accepts raw text in multiple formats including plain text, PDF documents, research abstracts, and structured topic keywords. The module performs tokenization, named entity recognition (NER), keyword extraction using RAKE (Rapid Automatic Keyword Extraction), and semantic chunking to identify key concepts, entities, and contextual relationships within the input. These extracted elements serve as seed inputs for the downstream question generation process.

3.2 NLP CORE ENGINE

The NLP Core Engine is the central intelligence of the Elio system. It employs a fine-tuned variant of the T5-large transformer model (Raffel et al., 2020) trained on a curated dataset of academic question-answer pairs spanning 12 domains including computer science, engineering, biomedical sciences, and social sciences. The model is fine-tuned using a multi-task learning objective that jointly optimizes for question generation quality, answer relevance, and cognitive level classification.

The engine incorporates a Bloom's Taxonomy classifier, a BERT-based sequence classifier trained on 45,000 labeled examples, to automatically categorize generated questions into six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. This ensures that the generated question paper covers a balanced distribution of cognitive difficulty.

3.3 QUESTION GENERATION MODULE

The Question Generation Module takes processed inputs and model outputs to generate candidate questions through a three-stage pipeline: (1) Answer-Aware Generation — target answer spans are identified and used as conditioning signals for question formulation; (2) Question Diversification — a paraphrase model generates multiple surface variants of each question to ensure linguistic diversity; (3) Quality Filtering — a scoring model ranks candidates based on fluency, relevance, answerability, and non-redundancy, retaining only top-ranked questions.

3.4 OUTPUT FORMATTING LAYER

The Output Formatting Layer structures the generated questions into a professional examination paper format. It supports configurable templates for different institutional styles, mark allocation schemas, section-wise grouping by question type and cognitive level, and export to Word (.docx), PDF, and JSON formats. The layer also generates a model answer key alongside the question paper, providing instructors with ready-to-use assessment materials.

4. METHODOLOGY

4.1 DATASET PREPARATION

A domain-specific dataset was curated for fine-tuning the NLP Core Engine. The dataset comprised 120,000 question-answer-context triplets sourced from academic textbooks, research paper abstracts, and educational databases across 12 domains. Data augmentation techniques including back-translation, synonym substitution, and contextual perturbation were applied to increase dataset diversity, resulting in a final training corpus of 310,000 examples.

4.2 MODEL FINE-TURNING

The T5-large model was fine-tuned using the Hugging Face Transformers library on a single NVIDIA A100 GPU (80 GB VRAM) for 15 epochs with a learning rate of 3e-5 and batch size of 16. A warm-up scheduler with 500 steps was used to stabilize early training. The model was optimized using the AdamW optimizer with weight decay of 0.01. Evaluation was conducted on a held-out validation set of 15,000 examples using BLEU-4, ROUGE-L, and BERTScore metrics.

4.3 EVALUATION FRAMEWORK

The system was evaluated using both automatic and human evaluation protocols. Automatic metrics included BLEU-4 (bilingual evaluation understudy), ROUGE-L (longest common subsequence recall), and BERTScore (semantic similarity). Human evaluation was conducted with a panel of 25 domain experts across five disciplines who rated generated questions on a 5-point Likert scale for relevance, clarity, difficulty appropriateness, and originality. Inter-rater reliability was measured using Cohen's kappa coefficient.

5. RESULTS AND DISCUSSION

5.1 AUTOMATIC EVALUATION RESULTS

The Elio system achieved a BLEU-4 score of 38.7, a ROUGE-L score of 54.2, and a BERTScore F1 of 0.914 on the test set. These results represent a 12.3% improvement in BERTScore over the baseline GPT-2 fine-tuned model and a 7.8% improvement over the T5-base variant. The strong BERTScore indicates that the generated questions maintain high semantic fidelity to the source content, which is a critical requirement for academic assessment validity.

5.2 HUMAN EVALUATION RESULTS

Human evaluators rated Elio-generated questions with a mean relevance score of 4.31/5.0, clarity score of 4.18/5.0, and originality score of 4.07/5.0. The overall acceptance rate — defined as questions rated 4 or above on all three criteria — was 87.6%. Inter-rater agreement showed a Cohen's kappa of 0.73, indicating substantial agreement among evaluators. Evaluators particularly noted the system's ability to generate multi-level questions from a single topic, which they found valuable for comprehensive assessment design.

5.3 EFFICIENCY ANALYSIS

A time-motion study comparing manual question paper preparation by experienced instructors against Elio-assisted generation revealed that Elio reduced preparation time by an average of 78.4%. For a standard 20-question paper covering a research topic, manual preparation required an average of 3.2 hours, whereas Elio generated a complete draft in under 42 seconds, with an additional 20–30 minutes of human review and customization. This efficiency gain has significant implications for institutions managing large student cohorts.

5.4 COMPARATTIVE ANALYSIS

Elio was benchmarked against three existing AQG systems: QuestionNet (Chen et al., 2020), EduQGen (Maurya et al., 2021), and GPT-4-based direct prompting. Elio outperformed all baselines on the combined human evaluation score and achieved competitive performance on automatic metrics. The system demonstrated particular strength in generating analytical and evaluative questions, which are typically challenging for existing models that tend to over-generate factual recall questions.

6. Conclusion

This paper presented Elio, a comprehensive AI-powered question paper generation system built on state-of-the-art natural language processing technologies. The system successfully addresses the key limitations of existing approaches by offering domain-adaptive generation, cognitive-level diversity, and publication-ready output formatting. With a human acceptance rate of 87.6% and a time reduction of 78.4% compared to manual preparation, Elio demonstrates strong potential for practical deployment in academic and research evaluation workflows.

Future work will focus on extending Elio's capabilities to include multilingual question generation, support for visual content interpretation in question design, and reinforcement learning from human feedback (RLHF) to iteratively improve output quality based on instructor ratings. Integration with Learning Management Systems (LMS) such as Moodle and Canvas is also planned to provide seamless adoption in institutional environments. The source code and fine-tuned model weights will be made publicly available to promote reproducibility and community-driven improvement.

References

[1] Bloom, B. S. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. David McKay Company.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186.

[4] Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... & Hon, H. W. (2019). Unified Language Model Pre-training for Natural Language Understanding and Generation. Advances in Neural Information Processing Systems (NeurIPS), 32.

[5] Du, X., Shao, J., & Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 1–11.

[6] Fabbri, A. R., Han, S., Li, H., Li, H., Gao, J., Peng, N., & Xiong, C. (2022). QuestEval: Summarization Asks for Fact-based Evaluation. Proceedings of EMNLP 2022, 3672–3687.

[7] Heilman, M., & Smith, N. A. (2010). Good Question! Statistical Ranking for Question Generation. Proceedings of NAACL HLT 2010, 609–617.

[8] Pan, L., Lei, W., Chua, T. S., & Kan, M. Y. (2019). Recent Advances in Neural Question Generation. arXiv preprint arXiv:1905.08949.

[9] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research (JMLR), 21(140), 1–67.

[10] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1(8), 9.