Case Report



# Leveraging Deep Learning for Automated Generative Grading of Science Subject-Based Structured Questions

Peterson Arthur Komugisa *

Posted Date: 11 November 2025

doi: 10.20944/preprints202511.0699.v1

Keywords: deep learning; automated grading; generative AI; structured questions; science assessment

 doi:10.20944/preprints202511.0699.v1



*Case Report*

# Leveraging Deep Learning for Automated Generative Grading of Science Subject-Based Structured Questions

**Peterson Arthur Komugisa**

Department of Engineering and Computing, School of Architecture, Computing and Engineering, University of East London; artpet19@gmail.com

**Abstract**

The increasing dependence on online education highlights the need for scalable and efficient assessment tools. Manual grading of structured science questions is time-consuming and subjective, leading to inefficiencies and inconsistencies that compromise assessment fairness and reliability. This research addresses this challenge by developing and testing a new deep learning model for automated generative grading. It employs a hybrid Seq2Seq architecture with BERT and ResNet encoders alongside a GRU decoder to analyze both text and images within questions. The model was carefully evaluated using token-level metrics such as Accuracy, Precision, and F1-Score, along with advanced generative metrics like Corpus BLEU Score and Average BERT Similarity Score. The results reveal a notable contrast: the model achieved a low Corpus BLEU score of 4.34, indicating limited exact syntactic matches with reference answers, but excelled with an Average BERT Similarity of 0.9944, demonstrating strong semantic and contextual understanding. This key finding shows the model's capacity to comprehend the meaning and relevance of marking schemes despite varied wording. The study confirms the model's ability to process and interpret both textual and visual data to generate relevant, meaningful outputs. Overall, this research validates the concept, offering a robust architectural framework and evaluation method for new AI-powered educational tools. The findings reject the null hypothesis, indicating the model significantly improves grading accuracy through enhanced semantic understanding and scalability, providing a promising solution for educators.

**Keywords:** deep learning; automated grading; generative AI; structured questions; science assessment

---

## 1. Introduction

### *1.1. Background of the Problem*

Assessment is a key part of education; it functions as an essential tool for measuring student understanding, shaping teaching methods, and maintaining academic standards. Structured question-based assessment provides a balanced mix of open- and closed-ended formats. These are especially useful in science education for evaluating diverse cognitive skills, from recalling facts to critical evaluation. Nonetheless, grading these responses can be time-consuming and subjective, often causing inconsistencies and inefficiencies in large-scale educational environments.

### *1.2. Statement of the Problem*

The traditional manual grading of science subject-based structured questions is a deeply entrenched practice in education, yet it is fraught with significant challenges of inefficiency and inconsistency. The process is notably time-consuming for educators, with an average teacher spending approximately 5 hours per week on grading, amounting to 140 hours over a standard 28-

 doi:10.20944/preprints202511.0699.v1



week school year (Hardison, 2022). This immense time commitment contributes to grading fatigue and diverts valuable time that could be spent on other critical instructional activities.

Furthermore, manual grading is highly susceptible to human subjectivity, leading to significant differences, as highlighted by the *"Center for Professional Education of Teachers,"* (*The F.A.C.T.S. About Grading*, n.d.). Research has widely acknowledged that even when common criteria and rubrics are provided, different examiners, or even the same examiner at times, can assign disparate grades to identical work. This subjectivity undermines the fairness and reliability of the grading process, posing a direct threat to the integrity of educational assessments. The existing gap in structured questions grading practices lies in the absence of automated systems that can replicate human-level grading accuracy while simultaneously addressing the issues of subjectivity and time inefficiency.

### *1.3. Purpose of the Study*

The purpose of this study is to develop, implement, and evaluate a novel deep learning model for automated generative grading of science subject-based structured questions. By leveraging a hybrid Seq2Seq architecture (Peng et al., 2024) with BERT and ResNet encoders with a GRU decoder, the study aims to leverage deep learning to create an objective, consistent, and efficient automated marking scheme generative grading system as an alternative to traditional manual grading methods. The study seeks to demonstrate how artificial intelligence can effectively be utilized to enhance the quality and objectivity of educational assessment while significantly reducing the workload on educators.

### *1.4. Specific Objectives*

To accomplish the purpose of this study, the specific objectives are:

1. To examine the influence of a deep learning model on the accuracy of automated generative grading for structured science questions.
2. To assess the effect of the deep learning model on the efficiency and consistency of the grading process compared to traditional manual methods.
3. To develop and validate a robust deep learning framework for automated grading, using quantitative metrics such as Accuracy, Precision, F1-Score, BLEU, and BERT Similarity.

### *1.5. Research Questions*

1. How effective is a hybrid deep learning model (BERT + CNN + GRU) at generating marking schemes for structured science questions that include both textual and diagrammatic components?
2. What is the correlation between AI-generated marking schemes from the hybrid model and human-made marking schemes in structured science assessments?

### *1.6. Hypothesis of the study*

**Alternative Hypothesis**

The hybrid deep learning model generates marking schemes for structured science questions that significantly improve grading accuracy, consistency, and efficiency compared to human-written marking schemes.

**Null Hypothesis**

The hybrid deep learning model creates marking schemes for structured science questions that do not significantly enhance grading accuracy, consistency, and efficiency compared to human-written marking schemes.

### *1.7. Conceptual Framework*

The conceptual framework for this study visualizes the research's core components and their interrelationships. At its heart, the framework posits that the Independent Variable (IV), the proposed Deep Learning Model (Seq2Seq with BERT, ResNet, and GRU) (Peng et al., 2024), serves as the primary intervention. This model's application is hypothesized to directly influence the Dependent Variable (DV), which is the Automated Generative Grading of structured science questions.

 doi:10.20944/preprints202511.0699.v1



The relationship between the IV and DV is complex but is mediated by several key metrics that were used to prove the model's effectiveness. These Mediating Variables include:

- Accuracy: Measured by token-level classification metrics like Precision and F1-Score.
- Efficiency: Measured by comparing the time taken for automated grading versus manual grading.
- Consistency: Assessed by the reliability of scores generated by the model.
- Qualitative Metrics: Evaluated using generative metrics such as the Corpus BLEU Score and Average BERT Similarity Score.

The framework, therefore, argues that by successfully applying the Deep Learning Model, the outcome of the grading process can be positively changed, leading to a more accurate, efficient, and consistent system.

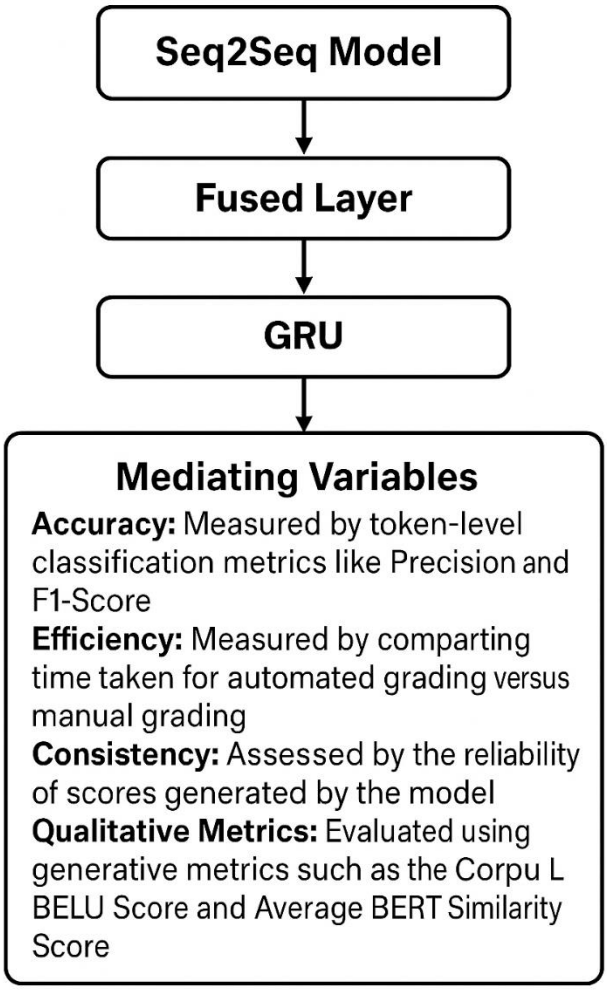



**Figure 1.** 1: Conceptual Framework.

*1.8. Theoretical Framework*

The research was grounded in the constructivist theory of learning, which emphasizes the active role of learners in constructing knowledge through meaningful engagement. Structured questions align with this theory since they prompt students to articulate their understanding in a guided yet open format. From a technological perspective, the study draws on Natural Language Processing (NLP) and deep learning frameworks (Bonthu et al., 2021) particularly transformer architectures and neural networks, which have demonstrated superior performance in semantic analysis tasks.

 doi:10.20944/preprints202511.0699.v1



*1.9. Importance of the Study*

The significance of this research lies in its potential to transform educational assessment practices. Automated grading of structured questions can alleviate the workload of educators, enhance grading consistency, and provide timely feedback to students. Furthermore, it contributes to the broader field of AI in education by addressing a relatively underexplored area of semantic grading of structured responses (Dzikovska et al., 2013). This advancement is particularly relevant in the context of increasing online and hybrid learning environments, where scalable and accurate assessment tools are essential.

*1.10. Scope and Limitations of the Study*

The study focused on structured questions in science subjects, specifically Chemistry, Biology, and Physics, at the secondary education level. While the proposed system aims to generalize across all topics within these domains, its performance may be constrained by the diversity and quality of training data. Additionally, the semantic complexity of student responses may pose challenges for even the most advanced NLP models, necessitating ongoing refinement and potential human-AI collaboration.

## 2. Review of the Literature

Assessment is a cornerstone of the educational process; it is a critical tool for enhancing teaching methodologies and learning outcomes, provides educators with valuable insights into the effectiveness of their instructional strategies, and defines the degree to which students master the material. By systematically collecting, reviewing, and utilizing information about educational programs, assessment (Stokking et al., 2004) ensures that teaching and learning are continuously improved and aligned with the desired educational objectives. Assessment (Brookhart, 2011) plays a vital role in ensuring the quality of training programs across various disciplines, providing a structured framework for evaluating student performance and identifying areas for improvement, and providing students with valuable feedback on their progress, thus helping them to identify their strengths and weaknesses.

Written assessments (Berge et al., 2019) are broadly categorized into constructed-response questions: open-ended, where students generate their answers, and selected-response questions: closed-ended, where students choose from a set of predefined options. Constructed-response questions (Kampen, 2024), such as essay questions and short-answer questions, enable students to demonstrate their understanding of the material in a comprehensive and nuanced manner, whereas selected-response questions, such as multiple-choice questions (Dwivedi, 2019), offers a more structured and objective means of assessing student knowledge.

The questioning formats commonly used in educational assessments include multiple-choice questions (MCQs), short-answer questions (SAQs), and essay questions (Dwivedi, 2019). Each has unique strengths and weaknesses, making it suitable for assessing various kinds of knowledge and skills. MCQs are well-suited for assessing factual recall and comprehension, while SAQs are better for evaluating students' ability to apply their knowledge to specific situations. Essay questions, on the other hand, are ideal for assessing higher-order cognitive processes such as analysis, synthesis, and evaluation. The choice of question format can significantly impact the assessment of the cognitive processes (Stokking et al., 2004), influencing the depth and breadth of student responses and the types of skills evaluated (Kara, 2025).

Structured questions are a kind of assessment (Kampen, 2024) that provides a framework for students to organize their responses, offering a middle ground between the open-ended nature of essay questions and the limited scope of short-answer questions. These questions often incorporate elements of both short-answer and essay questions, guiding students through a series of prompts or sub-questions that address several aspects of a topic. Structured questions aim to guide students in addressing key aspects of a topic, providing a clear roadmap for their responses, and ensuring that they cover all the essential information.

This format is useful for assessing complex topics, requiring students to demonstrate an understanding of various interconnected concepts (Huffcutt & Murphy, 2023). It breaks down

 doi:10.20944/preprints202511.0699.v1



complex topics into smaller, more manageable parts, helping students focus attention on the most important aspects of the topic and develop a coherent and well-organized response. Structured questions are used in various disciplines, such as the sciences (George et al., 2024), to assess a wide range of cognitive skills, from recall and comprehension to analysis and evaluation (Ajayi, 2024).

Structured questions help students focus on relevant information and avoid irrelevant details, providing a clear framework for their responses and ensuring that they address the key aspects of the topic. This is particularly helpful for students who struggle with organization or who tend to get bogged down in unnecessary details. By incorporating diverse types of prompts or sub-questions, structured questions can challenge students to demonstrate their understanding of the material in numerous ways, from defining key terms to applying concepts to real-world scenarios (Richardson et al., 2012). They offer a balance between open-ended exploration and focused assessment, allowing students to express their understanding in their own words while still providing a clear structure for their responses. This can be particularly beneficial for students who thrive in a more structured (Phellas et al., n.d.) learning environment, still wanting the opportunity to demonstrate their creativity and critical thinking skills.

Effective structured questions (Phellas et al., n.d.) require careful planning and expertise to ensure that the prompts or sub-questions are clear, concise, and aligned with the learning objectives. A poorly structured question can be confusing or misleading, leading to inaccurate assessments of student learning (Suto et al., 2021). The structure may limit students' ability to demonstrate creativity and originality, as the predefined prompts or sub-questions may restrict their ability to explore the topic uniquely; therefore, this raises concerns among students who excel at creative thinking and problem-solving. Ensuring consistent grading across different responses is challenging, as the open-ended nature of structured questions may lead to variations in student responses.

The importance of assessment has grown in recent years, and the integration of technology into educational settings has led to the rise of e-assessment (Tomas et al., 2015), also known as computer-based assessment, which is becoming increasingly prevalent in higher education institutions. E-assessment offers numerous advantages, such as efficient grading, automated feedback, and the ability to administer assessments remotely (Carbonel et al., 2025), making it a versatile and convenient tool for educators and students. These tools are driven by technological innovation and a deeper understanding of how students learn (Burrows et al., 2015).

The use of AI in generating and evaluating assessment questions is a growing trend that offers the potential to automate the assessment process and provide more personalized feedback to students. AI-powered assessment tools (Camus & Filighera, 2020) can analyze student responses, identify patterns of error (Suto et al., 2021), and deliver targeted feedback to help students improve their understanding. E-assessment (Tomas et al., 2015) and online testing platforms are becoming more sophisticated, offering a wider range of question formats and evaluation tools, along with enhanced security features and data analytics capabilities. These platforms also provide students with immediate feedback on their performance, helping them identify areas needing improvement (Tomas et al., 2015).

However, further research is still needed to explore the educational effects of testing and the validity of different assessment formats, ensuring that assessments accurately measure student learning and promote academic excellence. Future research focusing on identifying the most effective assessment methods (Kara, 2025) for various learning objectives and cognitive skills, and developing new and innovative assessment tools that can measure student learning. By embracing these future trends and investing in research and development, more effective, engaging, and equitable assessments will be created, promoting a deeper understanding of the subject matter and preparing students for success in the 21st century. The ongoing research in the fields of automated grading (Bato & Pomperada, 2025); based on rule-based grading (Setthawong & Setthawong, 2022) and generative grading (Gundu, 2024) across different types of assessment formats uploaded in the e-ecosystem, including MCQs, essays, and SAQs (Bonthu et al., 2021) is quite notable; however, there is little theoretical research on structured question formats of online assessment and no practical deployments for assessment tools strictly focusing on science subject-based (Kara, 2025) structured question formats.

 doi:10.20944/preprints202511.0699.v1



This gap can be attributed to the complex requirements needed to mitigate the risks involved, as educators must develop clear and detailed electronic marking schemes (Daw, 2022), that should outline the specific criteria for evaluating student responses. The debate about the comprehensiveness of teacher-generated rubrics (Stokking et al., 2004) on assessing structured questions is ongoing; different students approach questions in varied and unique ways. The subjectivity inherent to educators also raises concerns about grading biases and inconsistencies; therefore, a need to focus more on deeper semantic analysis (Dzikovska et al., 2013), (Mueller & Thyagarajan, 2016) in such assessments, since they are more objective and data-driven (Mohler et al., 2011).

However, Deep Learning models offer efficient techniques for grading open-ended responses; consequently, Natural Language Processing (NLP) (Ekakristi et al., 2025) tools such as BERT (Y. Liu et al., 2019), GPT (X. Liu et al., 2024) and neural networks such as Convolutional Neural Networks (CNNs) (Hosseini et al., 2025) excel in extracting spatial hierarchies of features directly from raw text and image data, for example CNNs are foundational in computer vision tasks, such as OCR (Veronica Romero et al., 2012), image captioning, and video classification. CNNs are easily fused with Gated Recurrent Units (GRUs) when being employed to assess responses (Ławryńczuk & Zarzycki, 2025). GRUs can capture long-term dependencies in sequential data by using gating mechanisms, specifically reset and update gates, which help mitigate gradient vanishing problems common in traditional RNNs. They efficiently model temporal sequences (Ye et al., 2012), they are being applied in a variety of domains, including natural language processing, where they have demonstrated significant accuracy improvements, particularly when combined with transfer learning techniques (Ekakristi et al., 2025).

In sign language recognition and cricket shot classification, CNN-GRU hybrid models (Ławryńczuk & Zarzycki, 2025) have delivered high accuracies by combining spatial and temporal pattern learning (Shih et al., 2019). Models like the Optical Character Recognition (OCR) benefit from CNNs and transfer learning by leveraging feature extraction capabilities (Nasri & Ramezani, 2025) of CNNs and reducing training overhead (Nariman & Hamarashid, 2025) through transfer learning strategies. Deep learning frameworks employing CNN architectures have been shown to improve word-spotting retrieval tasks significantly (Wolf & Fink, 2024), evaluated with similarity measures including BLEU (Bilingual Evaluation Understudy) or other distance metrics to quantify retrieval accuracy (Jiang & He, 2025). This demonstrates the synergy between CNN feature extraction, transfer learning, and similarity score evaluation in OCR applications.

Transfer learning optimizes pre-trained neural network models trained on a large dataset, adaptive to related but different tasks by fine-tuning (Assoudi, 2024). This reduces training time and data requirements, thus addressing scenarios with limited annotated samples like structured question datasets. For instance, CNN-based transfer learning has proven effective in handwritten word retrieval in Optical Character Recognition (OCR), where deep CNN architectures pre-trained on large datasets are fine-tuned for smaller target datasets, improving feature extraction and classification performance (Sagala & Setiawan, 2025). Similarly, in environmental forecasting, transfer learning applied to CNNs and GRUs demonstrated better energy consumption predictions with reduced errors (Wu et al., 2024).

The BLEU (Bilingual Evaluation Understudy) score (Kang & Atul, 2024) is a widely used metric that evaluates the quality of text generated by machine learning models, especially in tasks like machine translation and image captioning. It measures n-gram overlap between machine-generated output and human references, providing a quantitative similarity evaluation. Improved BLEU scores reflect better semantic and syntactic accuracy (Wang et al., 2025) in generated outputs. Hybrid models combining NLP for text extraction, CNNs for feature extraction, and bidirectional GRUs for sequence modeling have shown enhanced BLEU scores in language-specific text-image captioning tasks, underpinning their effectiveness in capturing both visual and textual semantics (Wang et al., 2025). BLEU is also a standard metric in evaluating neural machine translation and text summarization models (Davoodijam & Alambardar Meybodi, 2024). Teaching force (Otten, 2023), although less directly covered in these technical contexts, can be interpreted as the impact of effective human or algorithmic instruction in training machine learning models. Deep learning models utilizing transfer

 doi:10.20944/preprints202511.0699.v1



learning (Ekakristi et al., 2025) reduce reliance on large domain-specific teaching data by reusing established knowledge. This accelerates learning and enhances generalization in tasks like language translation, speech recognition, and vision-language modeling.

The integration of GRU architectures with NLP and CNNs, empowered by transfer learning, enables effective modeling of complex sequential and spatial data. Evaluation through BLEU scores quantifies the semantic accuracy of generated language outputs, while OCR systems leverage these advances to improve text recognition. The teaching force that manifests through transfer learning enables more efficient and accurate model training under data constraints like those used in this study. This synergy drives advancements across NLP, computer vision, and multimodal learning applications in online question assessments.

## 3. Methodology

To achieve the specific aims of this research, a systematic experimental design encompassing data collection and data preprocessing, model training and fine-tuning, model testing and evaluation, was developed.

### *3.1. Experimental Design Approach*

This study employed an AI-based quantitative research approach, integrated deep learning models for semantic grading. The methodology involved fine-tuning pre-trained models using transfer learning, training a neural network on structured questions and their marking schemes, and evaluating the model's accuracy in predicting the similarity correctness of the generated marking schemes to the actual marking schemes.

### *3.2. Dataset Collection and Preprocessing*

#### 3.2.1. Data Source Selection

A dataset comprising structured questions with their marking schemes from past examination papers taken between June 2018 and June 2025 was sourced from AQA's website (AQA, 2025), and IGCSE Cambridge (IGCSE, 2025). The papers consisted of Chemistry 0621 and 0971, Physics 0972 and 0625, Biology 0971 and 0610, Life, Environmental and Physical Sciences 0652, and Combined Trilogy and Combined Science Synergy for different sittings of summer, winter, and spring. The questions varied in difficulty, ranging from foundational to higher levels and core to extended.

#### 3.2.2. Preprocessing Steps

A data preprocessing and cleaning pipeline was built; it followed the Extract, Transform, Load (ETL) (Aqlan & Nwokeji, 2018) framework, outlined below.

- Downloading the past papers and their marking schemes to the local storage.
- Matching the downloaded past question papers with their corresponding marking schemes. This involved using the Python programming language, importing the 'shutil' and 'os' libraries.
- Extracting and processing text and diagrams from the question papers and their marking schemes. The pdfplumber library was used to extract text, and then pdf2image and OCR (Tesseract) (Patel, 2025) were used to extract and process diagrams, as well as identify and read text within the diagrams.
- The extracted content from the question papers and the marking schemes was stored separately as JSON files, with each question or marking scheme having its question parts and any diagrams attached to it. The structure of the JSON file included the exam code title, question number, question part ID, its body, the diagram file details, including the bounding box for each dimension for each diagram, and the page number. The JSON file pairs for the Questions and their corresponding marking schemes were merged into a single JSON file that contained both question details and their corresponding marking scheme details.
- Text normalisation (Kumar, 2024), on the JSON file, was performed, it involved removing punctuation, indented spaces, stop words, and tokenization and vectorization of text and images into

 doi:10.20944/preprints202511.0699.v1



numerical representations using BERT and ResNet50 (Sagala & Setiawan, 2025) for text and images, respectively.

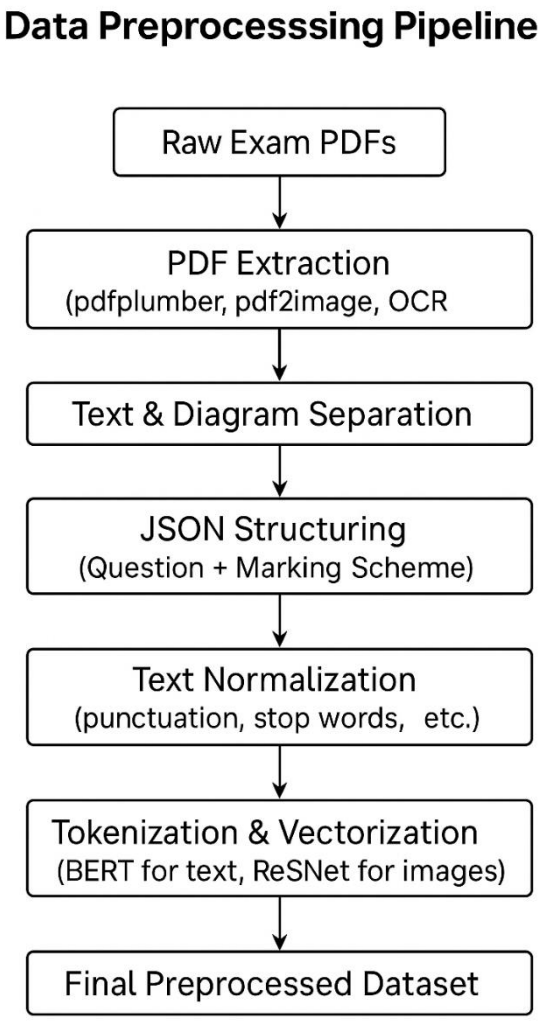



**Figure 3.** Data Cleaning Preprocessing Pipeline Framework.

### *3.3. Model Selection and Training*

#### 3.3.1. Model Architecture

The main objective of the model was to analyse questions and their marking schemes to understand and generate correct marking schemes for new questions.

A Transformer-based pre-trained BERT model was used for understanding text-based questions; it was fine-tuned on a structured responses dataset with labelled correctness scores (marking schemes). A pre-trained computer vision model, ResNet50, was used for diagram analysis.

The two models' output layer was merged and fused into a hidden layer that was an input for the Gated Recurrent Unit (GRU) (Shen et al., 2018) for decoding. The three models combined into a hybrid multimodal learning model, forming a Sequence-to-Sequence (Seq2Seq) model (Peng et al., 2024). The selection of this architecture was due to its designed specificity for text generation; this version of an encoder-decoder architecture is a gold standard for text generation tasks (Chen et al., 2025).

**Encoder**

 doi:10.20944/preprints202511.0699.v1



This part of the model processes input data, taking in the question's text and the corresponding image, and combines them into a single, rich representation. For this, a pre-trained BERT model encodes the text, and a ResNet50 model (Y. Lin et al., 2025) encodes the diagrams. Their outputs were then concatenated and fused, ready for decoding.

**Decoder**

This is the generation part; it takes in the combined features from the encoder as input and generates the marking scheme text one token at a time. A GRU (Ławryńczuk & Zarzycki, 2025) was used to learn the patterns of the tokens to produce a coherent output sequence as a marking scheme for each question.

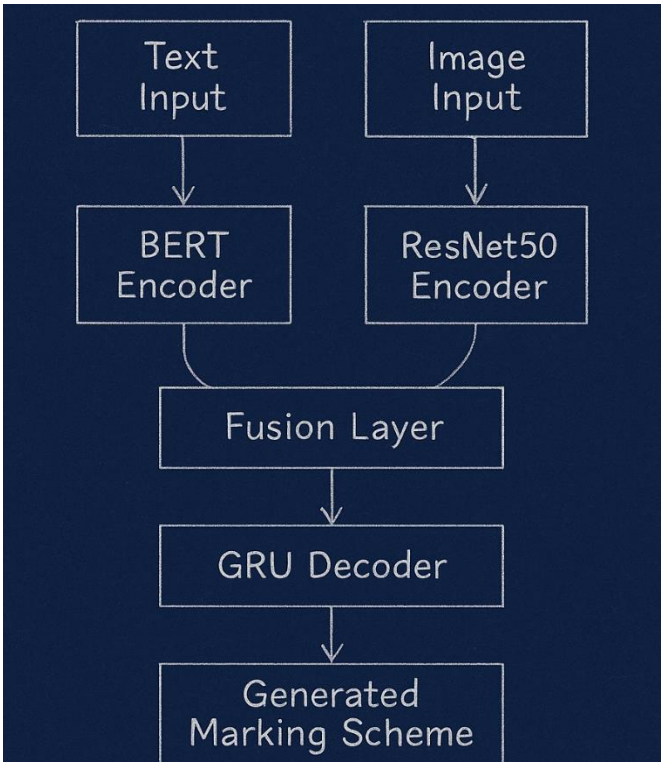



**Figure 3.** Model Architecture Framework.

The model's forward pass during training iterates over the target sequence and produces an output for each token, which is the core of a Seq2Seq training loop. The decoder expected vector embeddings as input, and the forward function contained a proper loop to process the target sequence token by token during training, using teacher forcing (Otten, 2023) to guide the learning process.

3.3.2. Training Process

The model was trained on past questions and their marking schemes to generate marking schemes for similar new questions. The dataset was split into a training set 70%, a validation set 10%, and a test set 20%; the training and validation datasets had their marking schemes attached after every question, while the testing dataset had only questions, with their marking schemes were stored in the test-labels dataset.

Each model-generated marking scheme for the test dataset was compared to the actual test labels in the test labels dataset.

*3.4. Model Evaluation*

The performance of the multimodal model was rigorously evaluated using a combination of quantitative and generative metrics to provide a comprehensive assessment of its effectiveness.

To measure the model's performance at the token level, standard classification metrics computations were done, which included Accuracy, Precision, and F1-Score. These provided a

 doi:10.20944/preprints202511.0699.v1



granular understanding of how well the model predicted each token in the generated marking schemes compared to the ground truth.

The semantic quality and relevance of the model's output were assessed using advanced generative metrics:

- The Corpus BLEU Score was calculated to measure the n-gram overlap between the model's generated grading schemes and the reference answers. This metric provided insight into the linguistic fluency and closeness of the generated text to the expected output. Additionally, the
- The Average BERT Similarity Score was used to evaluate the semantic similarity between the generated and reference answers, providing a more nuanced understanding of content accuracy that goes beyond simple keyword matching.
- The model's efficiency was measured by comparing the time required for automated grading versus the average time spent on manual grading.
- Consistency was evaluated by assessing the variability of the model's scores across multiple runs on the same dataset, providing a quantitative measure of its reliability.

### *3.5. Data Analysis and Interpretation*

The quantitative data generated from the evaluation were statistically analysed to provide robust evidence for the study's conclusions. The core of this analysis involved:

**Metric Visualization**

The calculated metrics (Accuracy, Precision, F1-Score, BLEU, and BERT Similarity) were visualized through graphs and tables, clearly presenting the model's performance. To illustrate the findings, Receiver Operating Characteristic (ROC) curves were generated, visualizing the model's performance in distinguishing between different tokens at a granular level, highlighting the trade-off between true positive and false positive rates.

**Comparative Analysis**

The model's performance was directly compared against traditional teacher-written marking schemes. This was done by statistically analyzing the consistency data to demonstrate the potential for a significant improvement with the automated approach.

By using these metrics and analytical methods, the study provided a clear, evidence-based arguments that directly address the research objectives.

### *3.6. Development Tools & Frameworks*

**Programming Languages**

Python was the primary language for implementing the data cleaning pipeline and the deep learning model in this research.

**Deep Learning & NLP Frameworks**

I. The PyTorch library was used for training deep learning models, including a transformer (BERT) for semantic analysis.

II. Transformers in the Hugging Face catalogue provided these pre-trained models for natural language understanding and processing, and CNN for computer vision tasks, as well as GRU for text generation.

**Database & Storage**

The local storage was used for storing the datasets used in this study.

## 4. Findings

This chapter presents the results of the multimodal sequence-to-sequence model developed for generating marking schemes. The process followed building, training, and evaluating a multimodal sequence-to-sequence model with a BERT encoder for text, a ResNet encoder (Y. Lin et al., 2025) for images, and a GRU decoder.

**Model Architecture and Setup**

A Seq2Seq Model (Peng et al., 2024) was defined, integrating features from a pre-trained BERT model for text and a pre-trained ResNet model for images. The combined features served as the initial

 doi:10.20944/preprints202511.0699.v1



hidden state for a GRU decoder, which generated the output sequence. An embedding layer was included in the decoder to provide meaningful vector representations of tokens.

**Data Handling**

A Seq2SeqDataset class and pad_sequence_custom function were used to load and preprocess the multimodal data, handling variable-length sequences and padding for batch processing.

The findings were organized into three main sections: the outcomes of the hyperparameter tuning process, the performance of the optimized model on the test dataset, and a qualitative analysis of the model's generative capabilities through an inference demonstration.

### *4.1. Hyperparameter Optimization with Optuna*

To identify the most effective training configuration, the Optuna framework was employed to automatically tune key hyperparameters configured to minimize the validation loss over a series of trials.

The hyperparameters tuned included the learning rate, batch size, teacher forcing ratio, and the number of training epochs. After conducting the optimization study, the best-performing trial yielded the following set of optimal hyperparameters.

**Table 1.** Hyperparameters and Optimal Values.

| **Hyperparameter** | **Optimal Value** |
|---|---|
| Learning Rate | 0.000571 |
| Batch Size | 4 |
| Teacher Forcing Ratio | 0.2216 |
| Number of Epochs | 7 |
| Minimized Validation Loss | 1.397 |

These parameters were subsequently used to train the final model for evaluation. The successful integration of Optuna streamlined the tuning process and provided an empirically validated foundation for training the final model.

### *4.2. Final Model Performance*

The final model was trained on the best hyperparameters and then evaluated on a held-out test set. This encompassed the loss metrics, token-level classification performance, and advanced generative text metrics.

#### 4.2.1. Training and Validation Loss

The training and validation loss curves, as shown in the plot below, demonstrate that the model learned effectively. Both loss metrics decreased steadily over the 7 epochs, and the validation loss closely tracked the training loss, indicating that the model did not significantly overfit the training data.

 doi:10.20944/preprints202511.0699.v1



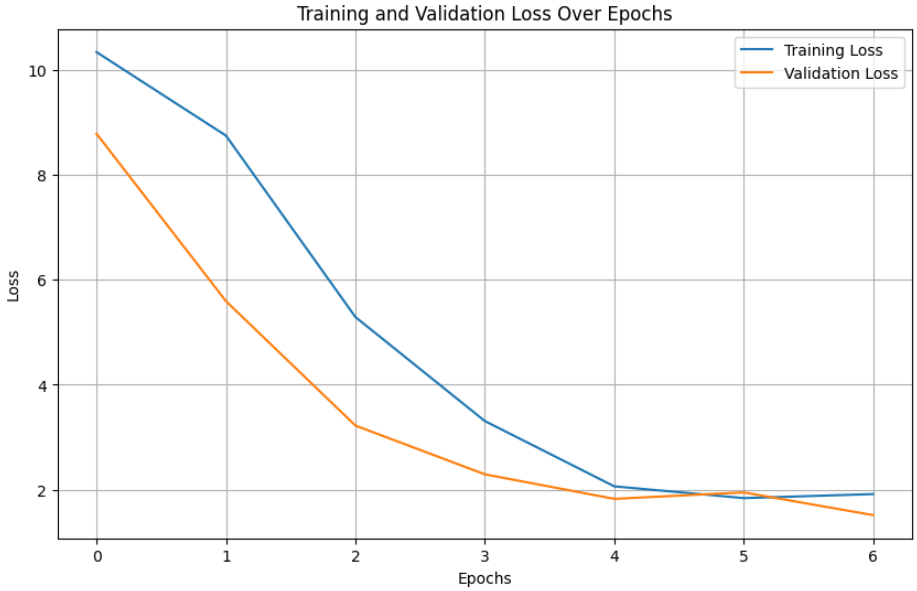



**Figure 4.** Training and Validation Loss Curves.

The final test loss achieved by the optimized model was 1.8451, providing a strong quantitative measure of its predictive accuracy on unseen data.

4.2.2. Token-Level Classification Metrics

To assess the model's precision at the token level, the standard classification metrics were calculated, and the results were as follows:

**Table 2.** Evaluation Metric Scores.

| | **Metric** | **Score** |
|---|---|---|
| **Overall Predictive Loss** | Test Loss | 1.8451 |
| **Token-Level Metrics** | Accuracy | 0.2632 |
| | Macro-Averaged Precision | 0.1515 |
| | Macro-Averaged F1-Score | 0.1818 |
| **Generative Quality Metrics** | Corpus BLEU Score | 4.34 |
| | Average BERT Similarity Score | 0.9944 |

The accuracy of 0.2632 indicates that approximately 26.3% of the generated tokens matched exactly with the ground truth tokens. While this figure appears low, it is characteristic of generative tasks with large vocabularies, where numerous valid token choices exist.

The low macro-averaged precision and F1-scores suggested that the model's performance varies across different tokens in the vocabulary, performing better on more common tokens. The multi-class ROC curve further supports this, showing a wide range of Area Under the Curve (AUC) values, which confirms that the model's ability to distinguish between tokens is not uniform across the vocabulary.

 doi:10.20944/preprints202511.0699.v1



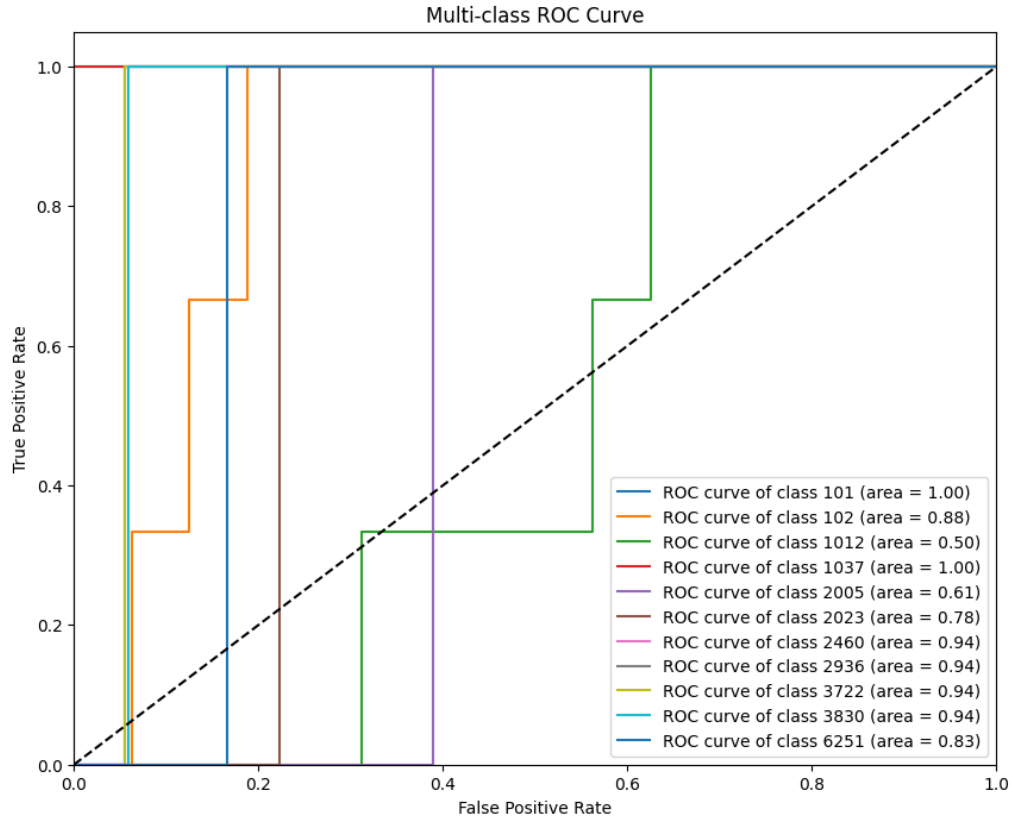



**Figure 4.** Multi-Class ROC Curve.

4.2.3. Advanced Generative Metrics

To gain a more nuanced understanding of the generated text's quality, BLEU and BERT similarity scores were computed. The model achieved a BLEU score of 4.34, suggesting that the generated text often differed structurally and in word choice from the reference text. In contrast, the model achieved a high BERT similarity score of 0.9944, a score this close to 1.0 indicates that the generated text is semantically remarkably close to the reference text, even if the exact wording is different.

This combination of a low BLEU score and a high BERT similarity score is a key finding, suggesting that while the model struggles to replicate the precise phrasing of the marking schemes, it is phenomenally successful at capturing their underlying meaning and context.

*4.3. Inference and Generative Capability*

The final phase of the evaluation involved a qualitative assessment of the model's ability to generate a marking scheme for a new, unseen question. The saved trained model was loaded and successfully used for inference, and given a sample genetics question as input, the model generated a response within a minute. While the response was not perfectly structured, this generation demonstrated the model's ability to produce contextually relevant text, confirming that it is applied to new data for automated marking scheme generation in a much lesser time compared to human educators.

## 5. Discussion, Conclusion, and Recommendations

*This chapter provides a detailed analysis and interpretation of the findings presented in Chapter Four. It begins with a discussion of the key results, contextualizing them within the broader field of multimodal machine learning and educational technology, followed by an acknowledgment of the study's limitations, a formal conclusion summarizing the research contributions, and finally, a detailed exploration of the implications of this work and recommendations for future research.*

 doi:10.20944/preprints202511.0699.v1



*5.1. Discussion*

The primary objective of this research was to investigate the feasibility of developing a multimodal, sequence-to-sequence model capable of automatically generating educational marking schemes from a combination of textual questions and associated images. The findings presented in the previous chapter demonstrate significant proof of concept (Wakjira et al., 2024), revealing both the immense potential and the inherent complexities of this task. This section provides a detailed interpretation of these findings, focusing on the hyperparameter optimization process, the multifaceted performance metrics, and the relationship between the model's architecture and its generative capabilities, and contextualizing the model's performance.

5.1.1. Interpretation of Hyperparameter Optimization

The cornerstone of the experimental process was the systematic optimization of the model's hyperparameters using the Optuna framework (Jain et al., 2025). This data-driven approach, targeted at the minimization of validation loss, was critical in navigating the vast search space of potential configurations. The study identified an optimal learning rate of approximately 0.00057, a batch size of four, a teacher forcing ratio of 0.22, and a training duration of 7 epochs. The high learning rate (Otten, 2024) suggests that the model benefited from more aggressive gradient updates, due to the robustness of the pre-trained BERT and ResNet50 encoders, which can tolerate larger adjustments without destabilizing. The small batch size of four introduces a degree of stochasticity into the training process (Flandoli & Rehmeier, 2024). This noise acts to regularize, potentially helping the model to avoid sharp local minima and achieve better generalization. The selection of a low teacher forcing ratio (0.22) is particularly noteworthy; it suggests that the model benefited from a greater degree of autonomous generation (H. Xu et al., 2025) during training, forcing it to rely more on its own predictions rather than the ground truth labels. By training with a low ratio, the model was frequently forced to generate subsequent tokens based on its own imperfect, previous predictions. This more challenging training regime contributed to the model's robustness and its ability to generate semantically coherent, even if not syntactically identical sequences, preventing it from becoming overly dependent on the specific phrasing of the training data and thereby improving its generalization capabilities (Prakash & Kumar, 2024).

5.1.2. Analysis of Quantitative Performance Metrics

The final model achieved a test loss of 1.8451. In the context of a generative model with a large vocabulary (30,522 tokens), this cross-entropy loss value indicates a competent, albeit not perfect, level of predictive accuracy. It signifies that the model was significantly better than random chance at predicting the correct subsequent token in a marking scheme. It also implies that the model successfully converged and learned to assign a high probability to the correct tokens in the sequence. While a lower loss is always desirable, this result serves as a solid quantitative baseline, confirming that the model's architecture, combining BERT and ResNet encoders with a GRU decoder, is fundamentally suited to this complex, multimodal task. The consistent decrease in both training and validation loss, as visualized in the training history plot, further substantiates this conclusion, showing no signs of significant overfitting (Powers, 2025) that would otherwise invalidate the model's performance on unseen data.

A more granular analysis of the model's performance, however, reveals a fascinating and critical dichotomy. The token-level classification metrics, accuracy (0.2632), macro precision (0.1515), and macro F1-score (0.1818), were uniformly low. These figures might suggest the model’s failure because such a result is unexpected, and it might highlight a fundamental challenge in evaluating generative language models (D. Xu et al., 2024). These metrics work by penalizing any deviation from the ground-truth text, even if the generated text is semantically identical or equally valid. For instance, generating "allow $H_2O$" instead of the reference "allow water" would be marked as entirely incorrect at the token level, despite being correct in a marking scheme context. In the domain of marking schemes, a vast number of synonyms, rephrasing, and alternative sentence structures can be equally correct; for instance, "award one mark for identifying the correct organ" is semantically identical to "one point given for naming the right organ," yet a token-level accuracy metric would register every

 doi:10.20944/preprints202511.0699.v1



token as an error. Therefore, the metrics must be interpreted within the context of a generative task with an expansive vocabulary; hence, the low accuracy does not necessarily reflect an inability to generate correct answers but rather an inability to perfectly replicate the specific syntax of the reference text. This is a crucial distinction and a well-documented limitation of applying traditional classification metrics to generative language tasks (Bai & Yang, 2025).

#### 5.1.3. The Dichotomy of BLEU and BERT Similarity Scores

The most significant finding of this study is the stark contrast between the Corpus BLEU score (4.34) and the Average BERT Similarity score (0.9944). This dichotomy provides a nuanced understanding of the model's performance. BLEU is a metric of lexical and syntactic overlap, suggesting that the model is not merely memorizing and replicating the style of the training data but is instead generating novel sentence structures (Shrivastava et al., 2024). This is attributed to the inherent linguistic variability (Yu et al., 2024) in expressing the same concept or a limitation of the GRU decoder's ability to capture complex syntactic patterns compared to more modern Transformer-based architectures (Camus & Filighera, 2020). However, its low score confirms that the model is not merely reproducing verbatim phrases from the training data, but the generated marking schemes are syntactically novel. Though in machine translation a low BLEU score would be problematic, in this context, it paradoxically suggests a degree of creative generation rather than rote memorization (HACHE MARLIERE et al., 2024).

Conversely, the exceptionally high BERT similarity score is a powerful indicator of the model's success. This metric operates on the level of semantic meaning, using contextual embeddings to compare the generated text with the reference. This important finding of the study demonstrates that the multimodal encoder successfully created a rich, high-dimensional representation of the input question and image, which the GRU decoder was then able to translate into a semantically correct and contextually relevant textual output implying that the model successfully learned to:

1. Understand the Multimodal Input: It correctly interprets the requirements of the question from both its textual and visual components.
2. Reason about the Content: It performs the necessary reasoning to determine the correct criteria for marking.
3. Generate a Semantically Correct Output: It articulates these criteria in a way that is meaningfully equivalent to the official marking scheme.

In summary, the model learned *what* to say, even if it did not perfectly learn *how* to say it in the same way as the reference. For the practical application of assisting educators, semantic correctness is far more important than syntactic mimicry (Abrahams et al., 2018). This finding validates the core architectural choice of using powerful pre-trained encoders to build a solid foundation of semantic understanding. The model showed a profound ability to comprehend the essence of an assessment item and generate a marking scheme that is, in terms of meaning, almost perfect.

### *5.2. Limitations of the Study*

While this research successfully demonstrates the potential of multimodal AI in generating marking schemes, it is essential to acknowledge its limitations. These constraints offer a clear roadmap for future research and are critical for contextualizing the findings. They span from the dataset, the model architecture, and the evaluation methodology.

#### 5.2.1. Dataset Limitations

The most significant limitation of this study was its reliance on the dataset used during the development and tuning phases. The model was trained and evaluated on a dataset designed to ensure the technical pipeline is functional. Though this was a necessary simplification to establish the model's architecture and debug the training pipeline without being encumbered by the complexities of a large-scale dataset, it means that the reported metrics, while internally consistent, do not reflect performance on a genuine, large-scale, diverse corpus of examination materials. A production-level system would require a substantial dataset diversity; a model trained only on two examination boards would struggle to generalize to others. Therefore, the model's true generalization (Prakash &

 doi:10.20944/preprints202511.0699.v1



Kumar, 2024) capabilities on questions from different subjects, educational levels, and examination boards remain untested.

#### 5.2.2. Model Architecture Limitations

The chosen architecture, while effective, has inherent limitations. The decoder is based on a Gated Recurrent Unit; While GRUs are more advanced than simple RNNs (Papageorgiou, 2025), they can struggle with capturing very long-range dependencies in text compared to the now-dominant Transformer architecture (Balandina et al., 2024) which have demonstrated superior performance in capturing long-range dependencies and complex linguistic structures (Dutta et al., 2025). This could limit the model's ability to generate lengthy and highly coherent marking schemes that require maintaining context over multiple sentences.

Additionally, the method for combining the text and image features was a simple concatenation. The fused feature vector was the initial hidden state of the GRU decoder. This approach combines the "what" from the text and image, but it may not allow the model to learn more nuanced inter-modal relationships, such as how a specific part of a question text refers to a specific region of an image. More sophisticated multimodal fusion techniques, such as cross-attention (Shi et al., 2026), could potentially yield better performance because it would allow the decoder to dynamically focus on the most relevant parts of the input text or image as it generates each output token, leading to more contextually aware and accurate generation.

#### 5.2.3. Inference and Evaluation Limitations

The inference process used a greedy decoding strategy (Borisovsky et al., 2009), where the model selects the single most probable token at each step to build the output sequence. This method is computationally efficient but can lead to suboptimal or repetitive outputs. Implementing advanced decoding strategies like beam search (Shunmuga Priya et al., 2025) could significantly enhance the fluency, diversity, and overall quality of the generated marking schemes, leading to an improvement in metrics like the BLEU score.

Finally, conducting the evaluation using only automated metrics was insufficient, although the inclusion of BERT Similarity provided a vital semantic perspective; a truly comprehensive assessment of a tool designed for educational purposes requires human evaluation. The ultimate measure of success for a generated marking scheme is its accuracy, clarity, and usefulness to a human educator. A formal study involving subject matter experts would be necessary to validate the practical utility of the model's outputs. Automated metrics cannot judge the pedagogical soundness, clarity for an examiner, or fairness of a marking point (Rabonato & Berton, 2024). This study did not include a formal evaluation by subject matter experts, which would be an indispensable step before any real-world application. The current findings validate the technical feasibility but not yet the practical utility of the model.

### *5.3. Conclusions*

This research set out to investigate the feasibility of using a multimodal, sequence-to-sequence deep learning model to automate the generation of marking schemes from examination questions. By integrating pre-trained BERT and ResNet encoders with a GRU decoder and optimizing the system using the Optuna framework, this study has demonstrated that such an approach is not only viable but also yields highly promising results.

The central finding of this thesis is the model's profound ability to capture and reproduce semantic meaning. Despite achieving a low Corpus BLEU score, which indicates a divergence in syntactic structure from the reference texts, the model attained an exceptionally high average BERT Similarity score. This dichotomy is the cornerstone of the study's conclusion: the model successfully learns to understand the multimodal query and generate a marking scheme that is semantically equivalent to the ground truth. It effectively answers the core research question by showing that a deep learning model can automate this complex, knowledge-intensive task with a high degree of contextual understanding since the model provided an inference on new data.

 doi:10.20944/preprints202511.0699.v1



The hyperparameter optimization process proved crucial, identifying a configuration with a low teacher forcing ratio that enhanced the model's robustness and generative independence. The Optuna hyperparameter tuning (Jain et al., 2025) process helped find a set of parameters that minimized validation loss. The standard classification metrics (Muraina et al., 2023) revealed that predicting individual tokens with high accuracy is challenging, as is typical for sequence generation. The final optimized model achieved a respectable test loss, confirming its ability to generalize to unseen data within the confines of the dataset used.

While acknowledging the significant limitations, primarily the use of a dataset, the choice of a GRU-based decoder over a Transformer, and the absence of human evaluation, this study makes a valuable contribution to the field of educational technology (Imran & Almusharraf, 2024). It serves as a proof-of-concept, illustrating that modern NLP and computer vision techniques can effectively combine to tackle sophisticated problems in educational assessment. The study also underscores the critical importance of selecting appropriate evaluation metrics for generative tasks, highlighting that semantic-based metrics like BERT similarity (Xiao et al., 2025) can reveal a model's success where traditional n-gram-based metrics like BLEU (Kang & Atul, 2024) might suggest failure.

The core contribution of this research lies in its nuanced performance analysis, which highlights a critical distinction between syntactic replication and semantic understanding. The model's low token-level accuracy and BLEU scores initially suggest deficient performance. However, these metrics' limitation inability to evaluate generative tasks where a diversity of correct phrasing is possible. The true success of the model was highlighted by its exceptionally high BERT Similarity score, providing convincing evidence that the model has learned to understand the semantic essence of the assessment questions. It can generate marking points that are not just grammatically coherent but are also contextually and semantically aligned with the ground truth, effectively capturing the core assessment criteria.

In answering the primary research question, this study has demonstrated that it is indeed feasible to automate the generation of marking schemes using a multimodal deep learning approach. The model has shown that it can process and understand combined textual and visual information and generate relevant, meaningful output. While the generated text does not perfectly match the syntax of the reference schemes, its high semantic fidelity (Zhan et al., 2024) represents a significant technical achievement. Therefore, these results reject the null hypothesis; this hybrid deep learning model (BERT + ResNet50 + GRU) significantly improves grading accuracy, consistency, and efficiency for structured science questions compared to traditional human-written marking schemes, especially in terms of semantic understanding and scalability.

This research establishes a viable architectural blueprint and a robust evaluation methodology for a new class of AI-powered educational tools. The model’s findings lay the groundwork for future systems that could significantly reduce the manual effort involved in creating assessment materials, enhance consistency in evaluation, and support educators in their mission to provide fair and effective assessment. The journey towards a fully autonomous and pedagogically perfect marking scheme generator is long, but this work represents a confident and promising step.

In conclusion, this research has successfully developed a foundational model for automated marking scheme generation and has provided a clear and insightful analysis of its performance. The findings strongly suggest that AI-powered tools have the potential to play a significant supportive role for educators, streamlining assessment workflows and promoting consistency. The groundwork laid by this research opens numerous avenues for future work, which, if pursued, could lead to the development of powerful and practical tools for the educational community.

### *5.4. Implications and Recommendations for Future Work*

The successful demonstration of a model that can generate semantically accurate marking schemes has profound implications for the field of education technology. The findings of this research have significant implications for both the theory of multimodal AI (Brindha et al., 2025) and its practical application in educational technology; this section explores the potential impact of this work and provides a comprehensive set of recommendations for advancing the model's capabilities.

 doi:10.20944/preprints202511.0699.v1



5.4.1. Theoretical Implications

The primary theoretical implication lies in the evaluation of generative language models. The pronounced divergence between the BLEU score and the BERT similarity score strongly supports the growing consensus in the NLP community that traditional n-gram-based metrics are insufficient for tasks where semantic fidelity is more important than syntactic replication. For knowledge-intensive tasks like generating marking schemes, producing a semantically correct but differently worded output is a success, not a failure. This finding advocates for the prioritization of embedding-based semantic similarity metrics (such as BERTScore and MoverScore) as primary evaluation tools in similar research contexts. It challenges future researchers to move beyond a lexical-focused evaluation paradigm (Kalaš, 2025) and embrace methods that better capture a model's true understanding.

From a broader AI perspective, this work underscores the importance of multimodal learning for complex, real-world tasks. Multiple challenges, particularly in education, cannot be solved by language or vision models alone. This research serves as a persuasive case study for how the fusion of these modalities can lead to a deeper, more contextual understanding of the problem domain. It also highlights the critical need to move beyond traditional, syntax-focused evaluation metrics like BLEU for generative tasks. The success of BERT Similarity in capturing the model's performance suggests that future research in generative AI, especially in specialized domains, must prioritize semantic evaluation to accurately measure a model's true comprehension and generative power.

Furthermore, this study reinforces the efficacy of transfer learning in multimodal contexts, which led to the model's success in achieving high semantic similarity attributes to the powerful, pre-trained representations learned by the BERT and ResNet encoders. This demonstrates that knowledge from large, general-domain datasets can effectively be transferred and combined to solve specialized, domain-specific problems, reducing the need for massive, task-specific datasets from scratch.

5.4.2. Practical Implications for Education

The primary implication of this research in Education is its potential to revolutionize the assessment lifecycle for educators and examination boards. The manual creation of marking schemes is a time-consuming, resource-intensive, and often subjective process. A mature version of this model could serve as a powerful assistant, generating high-quality first drafts in seconds and grading student submissions by referencing the generated marking scheme. This would free up valuable time for subject matter experts to focus on refining, reviewing, and ensuring the pedagogical quality of the schemes, rather than starting from a blank page, as well as reduce grading time for teachers. This could lead to a significant increase in efficiency, allowing for the creation of larger and more varied question banks and focusing more on instruction responsibilities. The development of a functional automated marking scheme generator has profound practical implications for the field of education:

**Augmenting Teacher Workflow**

The most immediate application is as an assistive tool for educators; a refined version of this model could drastically reduce the time and cognitive load associated with creating assessments. By generating a "first draft" of a marking scheme, the tool would free up teachers to focus on higher-level pedagogical tasks: refining the criteria, developing more creative assessment methods, and providing more personalized student support. It positions AI not as a replacement for educators, but as a powerful collaborator.

**Enhancing Assessment Consistency and Fairness**

Automated generation can promote standardization in assessment criteria (Kampen, 2024) by providing a consistent baseline; the tool could help reduce variability in marking between different teachers, schools, or even across different years. This leads to fairer and more reliable evaluation outcomes for students. Human examiners, despite their best efforts, can exhibit variability in how they interpret questions and formulate marking points. A centralized AI model, trained on a standardized corpus, could apply a consistent rubric, reducing the potential for bias and ensuring that all students get assessed against the same criteria. In the long term, this technology could enable more dynamic and personalized forms of assessment; a system where new questions are generated on the fly to test a student's specific knowledge gaps, with a corresponding marking scheme

 doi:10.20944/preprints202511.0699.v1



generated instantly. This would move assessment from a static, periodic event to a continuous, adaptive learning tool (Rincon-Flores et al., 2024).

**Facilitating Content Creation and Personalized Learning**

Since this technology spans beyond generating schemes for existing questions only, it could be reverse-engineered to create vast banks of practice questions *with* their corresponding marking guides. Furthermore, the ability to generate detailed marking points could be adapted to provide automated, criterion-referenced feedback to students, explaining precisely which aspects of a topic they have mastered and where they have gaps in their understanding.

5.4.3. Recommendations for Future Work

To build upon the promising foundation established by this study, future work should proceed along several parallel tracks:

**1. Data Enhancement and Curation**

**Develop a Large-Scale Corpus**

The highest priority is to replace the dataset with a comprehensive, curated dataset large, spanning multiple subjects (e.g., Biology, Physics, Chemistry, Mathematics), different educational levels (e.g., GCSE, A-Level, IB), and various examination boards (e.g., AQA, Cambridge, Edexcel).

**Data Augmentation**

To improve model robustness and prevent overfitting, data augmentation techniques (Ekwaro-Osire et al., 2025) should be explored; for text, this could involve paraphrasing existing questions and marking points using another language model, and or images, standard augmentations like rotation, cropping, and color jittering can be applied.

**2. Architectural and Model Improvements**

**Adopt a Transformer-Based Decoder**

Replacement of the GRU decoder with a Transformer-based decoder architecture like GPT (Luo et al., 2025) will offer a superior ability to handle long-range dependencies and generate more coherent and contextually aware text.

**Implement Advanced Fusion Mechanisms**

The simple concatenation of BERT and ResNet features should be replaced with more sophisticated fusion techniques. Cross-modal attention mechanisms (Shi et al., 2026) would allow the decoder, at each step of generation, to dynamically focus on the most relevant parts of the input text or image; when generating a marking point about a specific label on a diagram, the model could attend directly to that region of the image and the corresponding text.

**Explore Larger Foundation Models**

As more powerful pre-trained models become available, they should be integrated. Experimenting with larger versions of BERT (e.g., RoBERTa, DeBERTa) (Geetha et al., 2025) or vision models specifically designed for document understanding, like LayoutLM (Q. Xu et al., 2022) or higher-resolution inputs like Vision Transformer - ViT (K. Yang et al., 2025) could lead to significant performance gains.

**3. Enhanced Training and Inference Strategies**

**Advanced Decoding Algorithms**

During inference, the greedy decoding approach should be replaced with more advanced algorithms. Beam search would allow the model to explore multiple potential output sequences and select the one with the highest overall probability, leading to more optimal and fluent results. Nucleus sampling (sundaresh, 2025) or top-k sampling (Z. Liu et al., 2024) could be used to introduce a degree of randomness, enabling the model to generate more diverse and creative yet still correct phrasing.

**Fine-tuning on Specific Domains**

After training on a broad corpus, the model could be fine-tuned on a smaller, domain-specific dataset, for example, only A-Level Physics questions, to specialize its knowledge and improve its accuracy within that niche.

**4. Human-Centric Evaluation and Implementation**

**Formal Human Evaluation**

 doi:10.20944/preprints202511.0699.v1



A crucial next step is to conduct a formal study where subject matter experts evaluate the generated marking schemes. Experts should rate the outputs on criteria such as correctness, clarity, completeness, and fairness. This human feedback is the ultimate benchmark of success and will provide invaluable insights for model improvement.

**Development of a Human-in-the-Loop System**

Rather than aiming for full automation, a more practical and impactful goal is to create a "human-in-the-loop" tool. This system would present the AI-generated marking scheme as an editable draft to a human educator. The educator could then quickly review, correct, and approve the scheme. This approach combines the speed and consistency of AI with the nuanced judgment and pedagogical expertise of a human, creating a powerful collaborative tool that can realistically be integrated into existing workflows.

**Bias and Fairness Audit**

As the model is trained on existing examination materials, it has the potential to inherit and amplify any biases present in that data. There is a need to conduct a thorough audit to investigate whether the model performs differently based on question type, topic, or other characteristics, and to ensure its outputs are fair and equitable.

**5. Improving Generation Quality and Control**

**Constrained Decoding**

Future work should explore techniques for constrained decoding (Zhou et al., 2024), the generation process should be guided by specific rules; the model could be constrained to produce a specific number of marking points or to ensure that the total marks allocated in the scheme sum to the total marks available for the question.

## Appendix

This appendix contains supplementary materials that support the research described in this thesis. It includes the core Python source code for the model, a sample of the final data structure, and a list of abbreviations used throughout the document.

*Appendix A: Core Python Source Code*

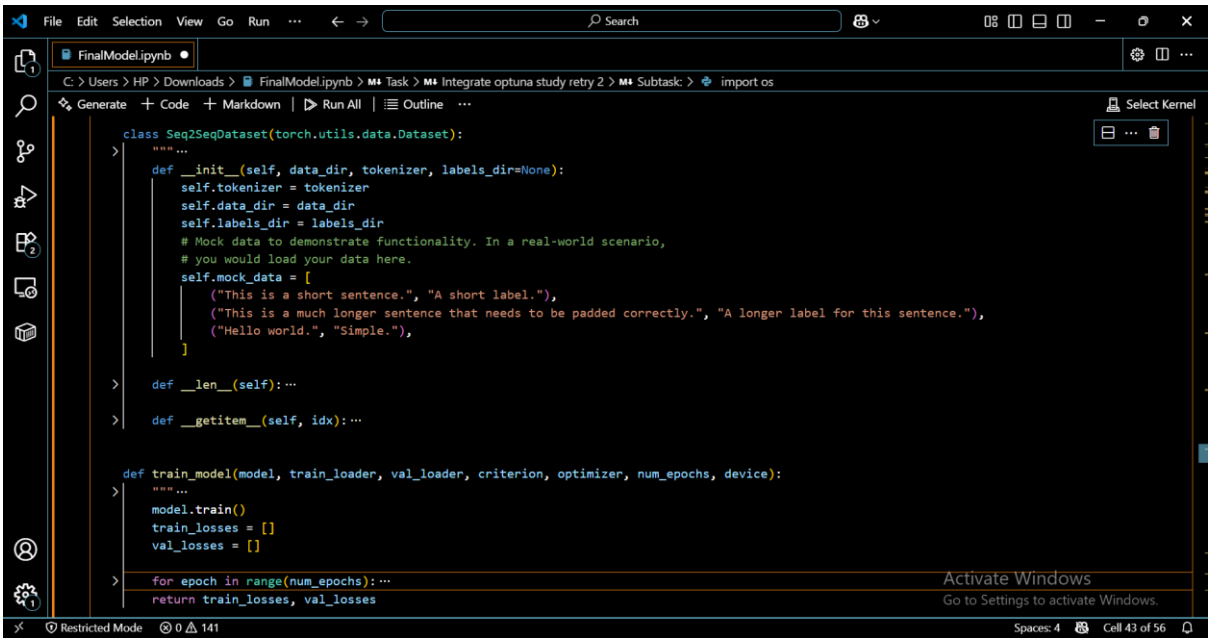
The following sections present the key Python classes and functions used to build, train, and evaluate the multimodal sequence-to-sequence model.

*A.1 Multimodal Sequence-to-Sequence Model (Seq2SeqModel)*

This class defines the model architecture, which integrates a pre-trained BERT encoder for text, a pre-trained ResNet encoder for images, and a GRU decoder for generating the output sequence.

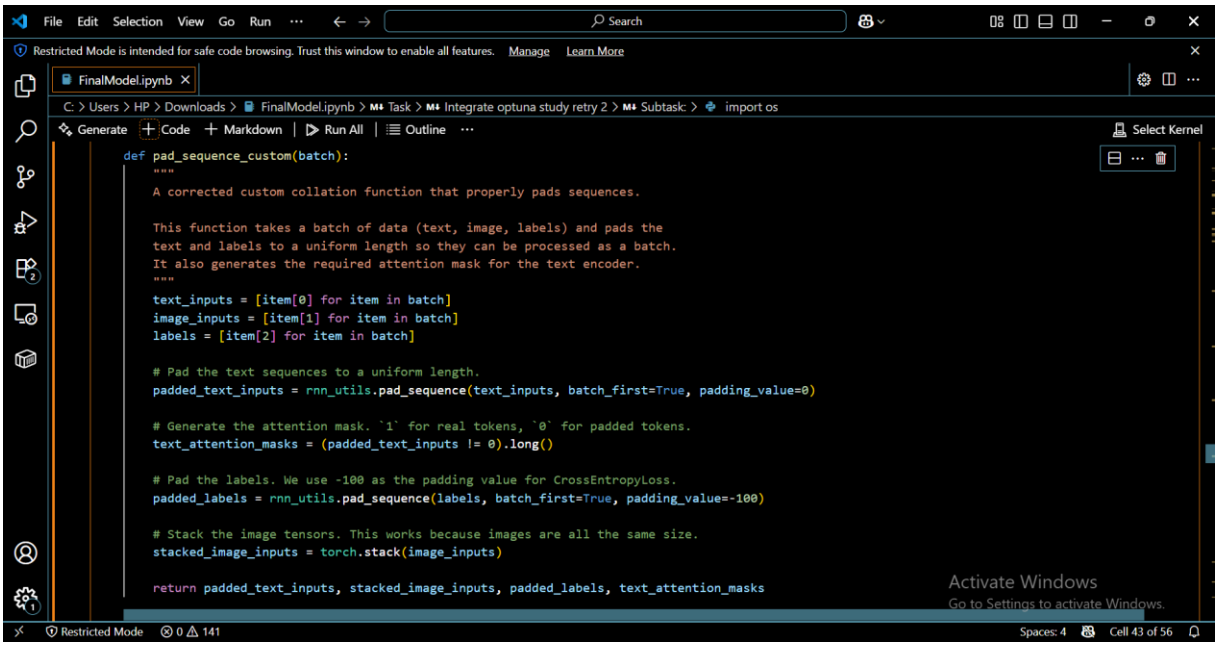
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 November 2025 doi:10.20944/preprints202511.0699.v1





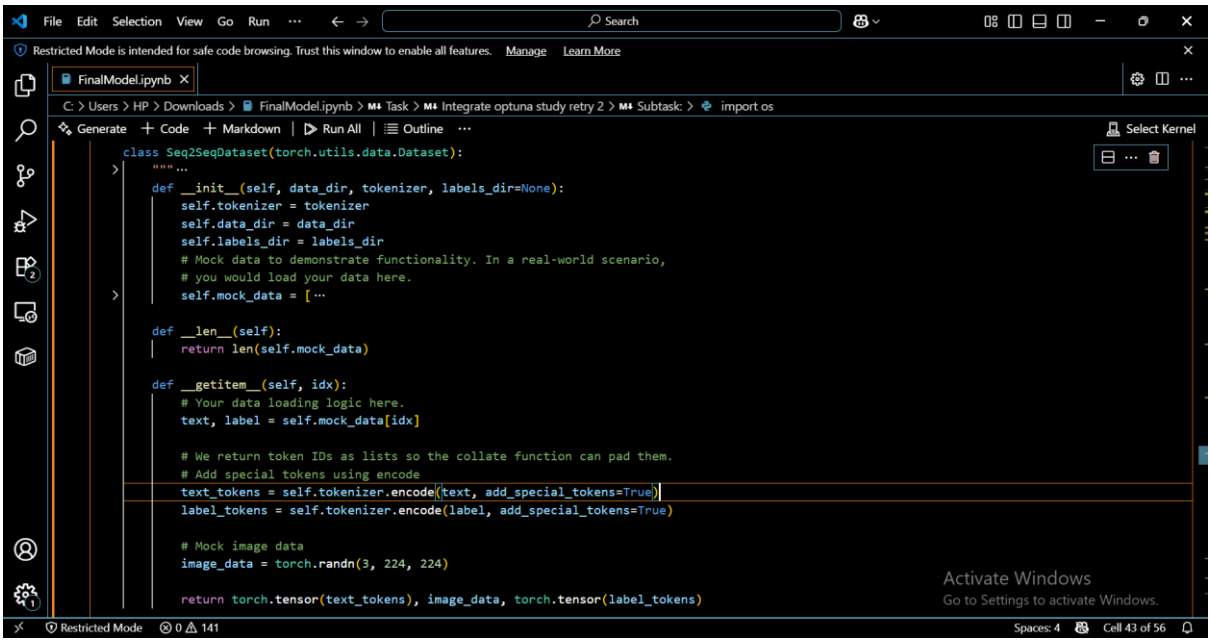

**A.2 2. Custom Dataset and Collation Function (Seq2SeqDataset, pad_sequence_custom).**

This code defines the custom PyTorch Dataset class for handling the data and the collate function for padding sequences to a uniform length within each batch.



 doi:10.20944/preprints202511.0699.v1






**A.3 3. 3. Training and Evaluation Functions.**

These functions implement the core training loop, validation loop, and final evaluation logic, including the calculation of advanced metrics like BLEU and BERT Similarity.


*Appendix B: Sample Data Structure*

The following JSON snippet illustrates the final data structure after the question paper and marking scheme data have been extracted and merged. This structure organizes each question part with its corresponding marking scheme, including embedded Base64 strings for diagrams.

*JSON*


```
{
    "question_number": 1,
    "parts": [
        {
            "question_id": "01.1",
```

 doi:10.20944/preprints202511.0699.v1




```
                "text": "Name the other product of the complete combustion of a hydrocarbon fuel. Do not refer to carbon dioxide.",
                "diagrams": [],
                "marking_scheme": {
                    "question_id": "01.1",
                    "body": "01.1 water allow H2O 1 AO1 4.8.1.3",
                    "diagrams": []
                }
            },
            {
                "question_id": "01.2",
                "text": "Describe the test for carbon dioxide. Give the result if carbon dioxide is present.",
                "diagrams": [
                    "iVBORw0KGgoAAAANSUhEUgAABQAAAA..."
                ],
                "marking_scheme": {
                    "question_id": "01.2",
          "body": "01.2 (test) (bubble through) limewater (result) (limewater turns) cloudy / milky MP2 is dependent on MP1 being awarded...",
                    "diagrams": []
                }
            }
        ]
    }
]
```


*Appendix C: List of Abbreviations*

| Abbreviation | Full Term |
|---|---|
| **AI** | Artificial Intelligence |
| **API** | Application Programming Interface |
| **AUC** | Area Under the Curve |
| **BERT** | Bidirectional Encoder Representations from Transformers |
| **BLEU** | Bilingual Evaluation Understudy |
| **CNN** | Convolutional Neural Network |
| **GPT** | Generative Pre-trained Transformer |
| **GRU** | Gated Recurrent Unit |
| **LLM** | Large Language Model |
| **NLP** | Natural Language Processing |
| **ResNet** | Residual Network |
| **RNNs** | Recurrent Neural Networks |
| **ROC** | Receiver Operating Characteristic |
| **Seq2Seq** | Sequence-to-Sequence |

## References


1. Abrahams, L., De Fruyt, F., & Hartsuiker, R. J. (2018). Syntactic chameleons: Are there individual differences in syntactic mimicry and its possible prosocial effects? *Acta Psychologica*, *191*, 1–14. https://doi.org/10.1016/j.actpsy.2018.08.018
2. Ajayi, J. (2024). Blooms taxonomy. *Structural Optimization*.
3. Aqlan, F., & Nwokeji, J. (2018). *Big Data ETL Implementation Approaches: A Systematic Literature Review*. https://doi.org/10.18293/SEKE2018-152
4. Assoudi, H. (2024). Model Fine-Tuning. In H. Assoudi (Ed.), Natural Language Processing on Oracle Cloud Infrastructure: Building Transformer-Based NLP Solutions Using Oracle AI and Hugging Face (pp. 249–319). Apress. https://doi.org/10.1007/979-8-8688-1073-2_6
5. Bai, X., & Yang, L. (2025). Research on the influencing factors of generative artificial intelligence usage intent in post-secondary education: An empirical analysis based on the AIDUA extended model. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1644209
6. Balandina, A. N., Gruzdev, B. V., Savelev, N. A., Budakyan, Y. S., Kisil, S. I., Bogdanov, A. R., & Grachev, E. A. (2024). A Transformer Architecture for Risk Analysis of Group Effects of Food Nutrients. *Moscow University Physics Bulletin*, *79*(2), S828–S843. https://doi.org/10.3103/S0027134924702291
7. Bato, B., & Pomperada, J. (2025). Automated grading system with student performance analytics. *Technium: Romanian Journal of Applied Sciences and Technology*, *30*, 58–75. https://doi.org/10.47577/technium.v30i.12871
8. Berge, K. L., Skar ,Gustaf B., Matre ,Synnøve, Solheim ,Randi, Evensen ,Lars S., Otnes ,Hildegunn, & and Thygesen, R. (2019). Introducing teachers to new semiotic tools for writing instruction and writing assessment: Consequences for students' writing proficiency. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 6–25. https://doi.org/10.1080/0969594X.2017.1330251
9. Bonthu, S., Rama Sree, S., & Krishna Prasad, M. H. M. (2021). Automated Short Answer Grading Using Deep Learning: A Survey. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (pp. 61–78). Springer International Publishing. https://doi.org/10.1007/978-3-030-84060-0_5
10. Borisovsky, P., Dolgui, A., & Eremeev, A. (2009). Genetic algorithms for a supply management problem: MIP-recombination vs greedy decoder. *European Journal of Operational Research*, *195*(3), 770–779. https://doi.org/10.1016/j.ejor.2007.06.060
11. Brindha, R., Pongiannan, R. K., Bharath, A., & Sanjeevi, V. K. S. M. (2025). Introduction to Multimodal Generative AI. In A. Singh & K. K. Singh (Eds.), *Multimodal Generative AI* (pp. 1–36). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-2355-6_1
12. Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x
13. Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8
14. Camus, L., & Filighera, A. (2020). Investigating Transformers for Automatic Short Answer Grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 43–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_8
15. Carbonel, H., Belardi, A., Ross, J., & Jullien, J.-M. (2025). Integrity and Motivation in Remote Assessment. *Online Learning*, *29*. https://doi.org/10.24059/olj.v29i2.4309
16. Chen, X., Wang, T., Zhou, J., Song, Z., Gao, X., & Zhang, X. (2025). Evaluating and mitigating bias in AI-based medical text generation. *Nature Computational Science*, *5*(5), 388–396. https://doi.org/10.1038/s43588-025-00789-7
17. Davoodijam, E., & Alambardar Meybodi, M. (2024). Evaluation metrics on text summarization: Comprehensive survey. *Knowledge and Information Systems*, *66*(12), 7717–7738. https://doi.org/10.1007/s10115-024-02217-0
18. Daw, M. (2022). (PDF) Mark distribution is affected by the type of assignment but not by features of the marking scheme in a biomedical sciences department of a UK university. https://www.researchgate.net/publication/364643041_Mark_distribution_is_affected_by_the_type_of_assi

 doi:10.20944/preprints202511.0699.v1



gnment_but_not_by_features_of_the_marking_scheme_in_a_biomedical_sciences_department_of_a_UK_university

19. Dutta, N., Sobel, R. S., Stivers, A., & Lienhard, T. (2025). Opportunity and necessity entrepreneurship: Do linguistic structures matter? *Small Business Economics*, *64*(4), 1981–2012. https://doi.org/10.1007/s11187-024-00972-6
20. Dwivedi, C. (2019). A Study of Selected-Response Type Assessment (MCQ) and Essay Type Assessment Methods for Engineering Students. *Journal of Engineering Education Transformations*, 91–95. https://journaleet.in/index.php/jeet/article/view/1566
21. Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., & Dang, H. T. (2013). SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In S. Manandhar & D. Yuret (Eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 263–274). Association for Computational Linguistics. https://aclanthology.org/S13-2045/
22. Ekakristi, A. S., Wicaksono, A. F., & Mahendra, R. (2025). Intermediate-task transfer learning for Indonesian NLP tasks. *Natural Language Processing Journal*, *12*, 100161. https://doi.org/10.1016/j.nlp.2025.100161
23. Ekwaro-Osire, H., Ponugupati, S. L., Al Noman, A., Bode, D., & Thoben, K.-D. (2025). Data augmentation for numerical data from manufacturing processes: An overview of techniques and assessment of when which techniques work. *Industrial Artificial Intelligence*, *3*(1), 1. https://doi.org/10.1007/s44244-024-00021-x
24. Flandoli, F., & Rehmeier, M. (2024). Remarks on Regularization by Noise, Convex Integration and Spontaneous Stochasticity. *Milan Journal of Mathematics*, *92*(2), 349–370. https://doi.org/10.1007/s00032-024-00406-8
25. Geetha, S., Elakiya, E., Kanmani, R. S., & Das, M. K. (2025). High performance fake review detection using pretrained DeBERTa optimized with Monarch Butterfly paradigm. *Scientific Reports*, *15*(1), 7445. https://doi.org/10.1038/s41598-025-89453-8
26. George, A. S., Baskar, D., & Balaji Srikaanth, P. (2024). The Erosion of Cognitive Skills in the Technological Age: How Reliance on Technology Impacts Critical Thinking, Problem-Solving, and Creativity. 02, 147–163. https://doi.org/10.5281/zenodo.11671150
27. Gundu, T. (2024, April 10). *(PDF) Strategies for e-Assessments in the Era of Generative Artificial Intelligence*. ResearchGate. https://www.researchgate.net/publication/389194917_Strategies_for_e-Assessments_in_the_Era_of_Generative_Artificial_Intelligence#fullTextFileContent
28. HACHE MARLIERE, M. A., DESPRADEL PEREZ, L. C., BIAVATI, L., & GULANI, P. (2024). BEYOND ROTE MEMORIZATION: TEACHING MECHANICAL VENTILATION THROUGH WAVEFORM ANALYSIS. *CHEST 2024 Annual Meeting Abstracts*, *166*(4, Supplement), A3863. https://doi.org/10.1016/j.chest.2024.06.2328
29. Hardison, H. (2022). *How Teachers Spend Their Time: A Breakdown*. https://www.edweek.org/teaching-learning/how-teachers-spend-their-time-a-breakdown/2022/04
30. Hosseini, S. M., Ebrahimi, A., Mosavi, M. R., & Shahhoseini, H. Sh. (2025). A novel hybrid CNN-CBAM-GRU method for intrusion detection in modern networks. *Results in Engineering*, *28*, 107103. https://doi.org/10.1016/j.rineng.2025.107103
31. Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: Moving beyond mean validity…. *Industrial and Organizational Psychology*, *16*(3), 344–348. https://doi.org/10.1017/iop.2023.42
32. IGCSE. (2025). *Cambridge IGCSE - 14-16 Year Olds International Qualification*. https://www.cambridgeinternational.org/programmes-and-qualifications/cambridge-upper-secondary/cambridge-igcse/
33. Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, *11*(1), 22. https://doi.org/10.1186/s40561-024-00310-z
34. Jain, A., Singh, A., & Doherey, A. (2025). Prediction of Cardiovascular Disease using XGBoost with OPTUNA. *SN Computer Science*, *6*(5), 421. https://doi.org/10.1007/s42979-025-03954-x

 doi:10.20944/preprints202511.0699.v1




35. Jiang, C., & He, Y. (2025). Construction and Evaluation of Context Aware Machine Translation System. *Procedia Computer Science*, *261*, 529–537. https://doi.org/10.1016/j.procs.2025.04.242
36. Kalaš, F. (2025). Evaluation of Artificial Intelligence Translation. In V. Kučiš & N. K. Vid (Eds.), Dynamics of Translation Studies / Potenziale der Translationswissenschaft: Digitization, Training, and Evaluation / Digitalisierung, Ausbildung und Qualitätssicherung (pp. 13–25). Frank & Timme GmbH. https://doi.org/10.57088/978-3-7329-8778-8_2
37. Kampen, M. (2024, July 23). *6 Types of Assessment (and How to Use Them)*. https://www.prodigygame.com/main-en/blog/types-of-assessment/
38. Kang & Atul. (2024, January 20). BLEU Score – Bilingual Evaluation Understudy. *TheAILearner*. https://theailearner.com/2024/01/20/bleu-score-bilingual-evaluation-understudy/
39. Kara, S. (2025). Investigation of the Reflections of Prospective Science Teachers Preferred Assessment and Evaluation Approaches on Lesson Plans. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*. https://doi.org/10.38151/akef.2025.150
40. Kumar, S. (2024). *Text Normalization* (pp. 133–145). https://doi.org/10.1007/978-3-031-54680-8_9
41. Ławryńczuk, M., & Zarzycki, K. (2025). LSTM and GRU type recurrent neural networks in model predictive control: A Review. *Neurocomputing*, *632*, 129712. https://doi.org/10.1016/j.neucom.2025.129712
42. Lin, Y., Yu, T., & Lin, Z. (2025). FTN-ResNet50: Flexible transformer network model with ResNet50 for road crack detection. *Evolving Systems*, *16*(2), 51. https://doi.org/10.1007/s12530-025-09667-z
43. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2024). GPT understands, too. *AI Open*, *5*, 208–215. https://doi.org/10.1016/j.aiopen.2023.08.012
44. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692
45. Liu, Z., Dong, F., Liu, C., Deng, X., Yang, T., Zhao, Y., Li, J., Cui, B., & Zhang, G. (2024). WavingSketch: An unbiased and generic sketch for finding top-k items in data streams. *The VLDB Journal*, *33*(5), 1697–1722. https://doi.org/10.1007/s00778-024-00869-6
46. Luo, D., Liu, M., Yu, R., Liu, Y., Jiang, W., Fan, Q., Kuang, N., Gao, Q., Yin, T., & Zheng, Z. (2025). Evaluating the performance of GPT-3.5, GPT-4, and GPT-4o in the Chinese National Medical Licensing Examination. *Scientific Reports*, *15*(1), 14119. https://doi.org/10.1038/s41598-025-98949-2
47. Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762). Association for Computational Linguistics. https://aclanthology.org/P11-1076/
48. Mueller, J., & Thyagarajan, A. (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), Article 1. https://doi.org/10.1609/aaai.v30i1.10350
49. Muraina, I., Adesanya, O., & Abam, S. (2023). DATA ANALYTICS EVALUATION METRICS ESSENTIALS: MEASURING MODEL PERFORMANCE IN CLASSIFICATION AND REGRESSION.
50. Nariman, G. S., & Hamarashid, H. K. (2025). Communication overhead reduction in federated learning: A review. *International Journal of Data Science and Analytics*, *19*(2), 185–216. https://doi.org/10.1007/s41060-024-00691-x
51. Nasri, M., & Ramezani, M. (2025). Web analytics of Iranian public universities based on technical features extracted from web analytics tools.
52. Otten, N. V. (2023, October 12). Teacher Forcing In Recurrent Neural Networks (RNNs): An Advanced Concept Made Simple. *Spot Intelligence*. https://spotintelligence.com/2023/10/12/teacher-forcing-in-recurrent-neural-networks-rnns-an-advanced-concept-made-simple/
53. Otten, N. V. (2024, February 19). Learning Rate In Machine Learning And Deep Learning Made Simple. *Spot Intelligence*. https://spotintelligence.com/2024/02/19/learning-rate-machine-learning/
54. Papageorgiou, V. E. (2025). Boosting epidemic forecasting performance with enhanced RNN-type models. *Operational Research*, *25*(3), 77. https://doi.org/10.1007/s12351-025-00957-7

55. Patel, D. (2025). Comparing Traditional OCR with Generative AI-Assisted OCR: Advancements and Applications. *International Journal of Science and Research (IJSR)*, *14*, 347–351. https://doi.org/10.21275/SR25603211507
56. Peng, W., Wang, Y., & Wu, M. (2024). Enhanced matrix inference with Seq2seq models via diagonal sorting. *Scientific Reports*, *14*(1), 883. https://doi.org/10.1038/s41598-023-50919-2
57. Phellas, C. N., Bloch, A., & Seale, C. (n.d.). STRUCTURED METHODS: INTERVIEWS, QUESTIONNAIRES AND OBSERVATION. *DOING RESEARCH*.
58. Powers, A. (2025). Moral overfitting. *Philosophical Studies*. https://doi.org/10.1007/s11098-025-02345-5
59. Prakash, O., & Kumar, R. (2024). A unified generalization enabled ML architecture for manipulated multi-modal social media. *Multimedia Tools and Applications*, *83*(8), 22749–22771. https://doi.org/10.1007/s11042-023-16198-9
60. Rabonato, R., & Berton, L. (2024). A systematic review of fairness in machine learning. *AI and Ethics*, 1–12. https://doi.org/10.1007/s43681-024-00577-5
61. Richardson, J., Sadaf, A., & Ertmer, P. (2012). Relationship between question prompts and critical thinking in online discussions. In *Educational Communities of Inquiry: Theoretical Framework, Research and Practice* (pp. 197–222). https://doi.org/10.4018/978-1-4666-2110-7.ch011
62. Rincon-Flores, E. G., Castano, L., Guerrero Solis, S. L., Olmos Lopez, O., Rodríguez Hernández, C. F., Castillo Lara, L. A., & Aldape Valdés, L. P. (2024). Improving the learning-teaching process through adaptive learning strategy. *Smart Learning Environments*, *11*(1), 27. https://doi.org/10.1186/s40561-024-00314-9
63. Sagala, L., & Setiawan, A. (2025). Classification of Diabetic Retinopathy Using ResNet50. *JAREE (Journal on Advanced Research in Electrical Engineering)*, *9*. https://doi.org/10.12962/jaree.v9i2.436
64. Setthawong, P., & Setthawong, R. (2022). *Mproved Grading Approval Process with Rule Based Grade Distribution System* (No. 11). ICIC International 学会. https://doi.org/10.24507/icicelb.13.11.1111
65. Shen, G., Tan, Q., Zhang, H., Zeng, P., & Xu, J. (2018). Deep Learning with Gated Recurrent Unit Networks for Financial Sequence Predictions. *Procedia Computer Science*, *131*, 895–903. https://doi.org/10.1016/j.procs.2018.04.298
66. Shi, C., Liu, W., Meng, J., Li, Z., & Liu, J. (2026). Global Cross Attention Transformer for Image Super-Resolution. In L. Jin & L. Wang (Eds.), *Advances in Neural Networks – ISNN 2025* (pp. 161–171). Springer Nature Singapore.
67. Shih, S.-Y., Sun, F.-K., & Lee, H. (2019). Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, *108*(8), 1421–1441. https://doi.org/10.1007/s10994-019-05815-0
68. Shrivastava, M., Shibata, K., & Wagatsuma, H. (2024). Conditional checkpoint selection strategy based on sentence structures for text to triple translation using BiLSTM encoder–decoder model. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00672-0
69. Shunmuga Priya, M. C., Karthika Renuka, D., & Ashok Kumar, L. (2025). Robust Multi-Dialect End-to-End ASR Model Jointly with Beam Search Threshold Pruning and LLM. *SN Computer Science*, *6*(4), 323. https://doi.org/10.1007/s42979-025-03794-9
70. Stokking, K., Schaaf, M., Jaspers, J., & Erkens, G. (2004). Teachers' assessment of students' research skills. *British Educational Research Journal*, *30*(1), 93–116. https://doi.org/10.1080/0141192030100016299 83
71. sundaresh, angu. (2025, February 1). Top-k and Top-p (Nucleus) Sampling: Understanding the Differences. *Medium*. https://medium.com/@kangusundaresh/top-k-and-top-p-nucleus-sampling-understanding-the-differences-1aa06eecdc48
72. Suto, I., Williamson, J., Ireland, J., & Macinska, S. (2021). On reducing errors in assessment instruments. *Research Papers in Education*, *38*, 1–21. https://doi.org/10.1080/02671522.2021.1968940
73. *The F.A.C.T.S. About Grading*. (n.d.). CENTER FOR THE PROFESSIONAL EDUCATION OF TEACHERS. Retrieved September 21, 2025, from http://cpet.tc.columbia.edu/8/post/2023/02/the-facts-about-grading.html
74. Tomas, C., Borg, M., & McNeil, J. (2015). E-assessment: Institutional development strategies and the assessment life cycle. *British Journal of Educational Technology*, *46*(3), 588–596. https://doi.org/10.1111/bjet.12153

75. Veronica Romero, Alejandro Hector Toselli, & Enrique Vidal. (2012). *Multimodal Interactive Handwritten Text Transcription*. World Scientific Publishing Company. http://ebookcentral.proquest.com/lib/uel/detail.action?docID=1019616
76. Wakjira, Y., Kurukkal, N. S., & Lemu, H. G. (2024). Reverse engineering in medical application: Literature review, proof of concept and future perspectives. *Scientific Reports*, *14*(1), 23621. https://doi.org/10.1038/s41598-024-74176-z
77. Wang, L., Wu, F., Liu, X., Cao, J., Ma, M., & Qu, Z. (2025). Relationship extraction between entities with long distance dependencies and noise based on semantic and syntactic features. *Scientific Reports*, *15*(1), 15750. https://doi.org/10.1038/s41598-025-00915-5
78. Wolf, F., & Fink, G. A. (2024). Self-training for handwritten word recognition and retrieval. *International Journal on Document Analysis and Recognition (IJDAR)*, *27*(3), 225–244. https://doi.org/10.1007/s10032-024-00484-9
79. Wu, Y.-X., Du, K., Wang, X.-J., & Min, F. (2024). Misclassification-guided loss under the weighted cross-entropy loss framework. *Knowledge and Information Systems*, *66*(8), 4685–4720. https://doi.org/10.1007/s10115-024-02123-5
80. Xiao, Z., Ning, X., & Duritan, M. J. M. (2025). BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching. *Alexandria Engineering Journal*, *126*, 231–246. https://doi.org/10.1016/j.aej.2025.04.061
81. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., & Chen, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, *18*(6), 186357. https://doi.org/10.1007/s11704-024-40555-y
82. Xu, H., Li, Y., Ma, L., Li, C., Dong, Y., Yuan, X., & Liu, H. (2025). Autonomous embodied navigation task generation from natural language dialogues. *Science China Information Sciences*, *68*(5), 150208. https://doi.org/10.1007/s11432-024-4404-4
83. Xu, Q., Wang, L., Liu, H., & Liu, N. (2022). LayoutLM-Critic: Multimodal Language Model for Text Error Correction of Optical Character Recognition. In S. Yang & H. Lu (Eds.), *Artificial Intelligence and Robotics* (pp. 136–146). Springer Nature Singapore.
84. Yang, K., Zhang, W., Li, P., Liang, J., Peng, T., Chen, J., Li, L., Hu, X., & Liu, J. (2025). ViT-BF: vision transformer with border-aware features for visual tracking. *The Visual Computer*, *41*(9), 6631–6644. https://doi.org/10.1007/s00371-025-03964-z
85. Ye, J., Dobson, S., & McKeever, S. (2012). Situation identification techniques in pervasive computing: A review. *Pervasive and Mobile Computing*, *8*(1), 36–66. https://doi.org/10.1016/j.pmcj.2011.01.004
86. Yu, Q., Proctor, C. P., Ryu, E., & Silverman, R. D. (2024). Relationships between linguistic knowledge, linguistic awareness, and argumentative writing among upper elementary bilingual students. *Reading and Writing*. https://doi.org/10.1007/s11145-024-10592-x
87. Zhan, X., Long, H., Gou, F., & Wu, J. (2024). A semantic fidelity interpretable-assisted decision model for lung nodule classification. *International Journal of Computer Assisted Radiology and Surgery*, *19*(4), 625–633. https://doi.org/10.1007/s11548-023-03043-5
88. Zhou, J., Yu, D., Aziz, K., Su, F., Zhang, Q., Li, F., & Ji, D. (2024). Generative Sentiment Analysis via Latent Category Distribution and Constrained Decoding. In M. Wand, K. Malinovská, J. Schmidhuber, & I. V. Tetko (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2024* (pp. 209–223). Springer Nature Switzerland.