A MULTI-MODAL AI FRAMEWORK FOR AUTOMATED EVALUATION OF TEXTUAL AND DIAGRAMMATIC STUDENT RESPONSES

S Nagesh
Lecturer in Computer Science,
Department of Computer Science,
Nagarjuna Government College (A), Nalgonda, Telangana, India

Abstract: The objective of this study was to present a multi-model AI framework for evaluating textual and diagrammatic student responses in a descriptive-based examination using natural language processing, computer vision, and machine learning techniques. The Natural Language Programming effectively grades textual answers. But few subjects, such as physics, botany, zoology, and engineering, require students to write answers through diagrams, circuits, and flowcharts. These diagrammatic responses contain critical structural and spatial information. The conventional text-based models cannot interpret the diagrammatic responses and often lead to incomplete assessments in STEM education. This paper proposes a multi-modal artificial intelligence framework designed to bridge this gap. With the integration of Natural Language Processing with computer vision, the system can analyze textual data and diagrammatic structures at the same time. The architecture of the proposed system consists of a dedicated text analysis module, a diagram recognition module, and a sophisticated multi-modal fusion mechanism. This mechanism synthesizes both textual streams and image streams into a unified evaluation score. Our tests show that this integrated approach significantly improves grading accuracy compared to single-model systems. By adding linguistic meaning with geometric logic, the new framework provides a more detailed validation of student knowledge. This study offers a measurable solution for digital learning platforms, effectively reducing teacher workloads while improving assessment consistency and accuracy across complex, diagram-heavy academic subjects.

Keywords: *Multi-modal AI, Automated Grading, Natural Language Processing, Computer Vision, STEM Education, Diagram Recognition, Educational Technology, Multi-modal Fusion.*

I. INTRODUCTION

The increasing complexity of present education has witnessed the need for a novel method to improve the evaluation of student handwritten responses. Automated evaluation of answer scripts systems can significantly reduce the workload of teachers while maintaining fairness and objectivity in awarding processes [28]. By allowing teachers to define awarding criteria and allot weighted keywords, the system enables faster, more deeply, and more accurate evaluation of student answer scripts [41].

Automated evaluation systems have traditionally focused on evaluating textual student responses using Natural Language Processing (NLP) techniques. Early automated essay scoring systems relied on rule-based algorithms that analyzed lexical features, grammar patterns, and sentence structures to determine essay quality [1]. Later developments in machine learning introduced statistical and neural models capable of identifying semantic relationships between student answers and reference solutions. Word embedding techniques such as Word2Vec [12] and GloVe [13] enabled systems to represent textual meaning in vector space and improved the semantic evaluation of student responses.

The emergence of deep learning further advanced automated text evaluation systems. The Long Short-Term Memory (LSTM) and Recurrent neural networks (RNNs) architectures have been used to model sequential textual data and capture contextual relationships within descriptive answers [11]. The introduction of the Transformer by Vaswani et al. [4] replaced older methods with a self-attention approach that handles long-distance context much better. Building upon this architecture, Devlin's work on BERT [5] achieved dominated NLP benchmarks, particularly in the realm of automated essay scoring,

All these automated awarding systems are limited to evaluating text-based responses. But in many academic subjects such as computer science, physics, botany, zoology and chemistry students frequently express their view using diagrams and conceptual illustrations. These graphical representations are very informative. Recent improvements in Computer Vision have enabled machines to understand graphical data with high accuracy. CNNs (Convolutional Neural Networks) can recognize images in answer scripts with great success. Krizhevsky et al. [7] demonstrated the effectiveness of deep CNNs that is used for large-scale image classification by using the ImageNet dataset. Subsequent improvements in deep neural network architectures such as Residual Networks (ResNet) introduced by He et al. [6] have significantly improved image recognition performance. Object detection frameworks such as YOLO [22] and Fast R-CNN [23] further improves the capability of computer vision systems to identify objects and structural components within images.

Although these computer vision models are capable of recognizing graphical patterns and objects, evaluating educational diagrams presents additional tasks. Educational diagrams often represent conceptual relationships. Moreover, diagrammatic answers are frequently written along with textual explanations. Therefore, analyzing diagrams and text separately may result poor evaluation of answer scripts.

To avoid these issues, multi-modal machine learning has evolved as a promising research area. Multi-modal learning integrates information from several data sources such as text, graphics, videos and audios to improve machine understanding. Baltrušaitis et al. [8] presented a comprehensive study of multimodal machine learning and focused its potential applications across various domains.

Hence a strong intelligent evaluation framework is required that is capable of analyzing multi-modal student handwritten answer scripts that includes both textual and diagrammatic information. Such systems can provide more accurate and consistent evaluation results while reducing the awarding workload for teachers.

This study proposes a Multi-Modal Artificial Intelligence Framework for Automated Evaluation of Textual and Diagrammatic Student Responses. The proposed system integrates natural language processing techniques with computer vision models to analyze textual explanations and diagrammatic representations simultaneously. A multi-modal fusion mechanism merges the outputs of both textual processing and diagrammatic processing to produce a final score that meets human evaluation score.

II. LITERATURE REVIEW

The expansion of digital learning environments and massive open online courses has intensified the demand for automated student assessment. Consequently, researchers have increasingly explored the application of Artificial Intelligence to manage the substantial volume of student submissions, with the objective of developing a system that maintains efficiency while preserving the reliability of human evaluation.

A. The Evolution of Automated Essay Scoring (AES)

Initial endeavors in automated grading were characterized by a degree of inflexibility, primarily utilizing statistical models and manually defined guidelines. These systems, like the one discussed by Burstein [1], focused on surface-level traits—things like grammar patterns, word count, and sentence complexity—using regression or support vector machines to estimate a score.

The real shift happened as Natural Language Processing (NLP) matured. Mikolov et al. [12] introduced Word2Vec, which moved us toward word embeddings that represent semantic relationships as vectors. This allowed systems to actually "understand" how similar a student's response was to a reference answer. This was followed by Pennington's GloVe model [13], which further refined how we map these textual relationships.

Deep learning eventually took over the field. Hochreiter and Schmidhuber's LSTM networks [11] proved that LSTM was particularly useful for descriptive answers because they can track long-term dependencies, making them better at verifying whether a student's argument actually flows logically. More recently, the Transformer architecture introduced by Vaswani et al. [4] has changed everything. By using attention mechanisms instead of sequential processing, it captures the relationship between words far more effectively.

Building on this, Devlin et al. [5] introduced BERT, which looks at context from both directions to find deeper meaning. Even more modern Large Language Models (LLMs), like those explored by Brown et al. [15], show that these systems can handle complex tasks with very little specific training. However,

despite all this progress in "reading," most of these systems are still blind to anything that isn't plain text, like a diagram or a sketch.

B. Computer Vision and the Challenge of Diagrams

Diagrams are a fundamental part of how we teach STEM. In biology or engineering, a student's drawing often proves they understand a concept better than their words do. However, grading them is a spatial and structural nightmare for traditional algorithms.

Computer vision has caught up quickly, largely thanks to Convolutional Neural Networks (CNNs). Krizhevsky et al. [7] proved how CNNs are powerful with the ImageNet dataset, and later architectures like VGG and ResNet [6] that allowed for even deeper, more accurate visual hierarchical data. For real-time identification, object detection frameworks like Fast R-CNN [23] and YOLO [22] have become the gold standard.

The problem is that most of these tools are designed to identify an object, not to judge if it's conceptually correct in an educational context. Furthermore, since diagrams usually come with a text caption or explanation, looking at the image in a vacuum often results in a half-baked assessment.

C. The Move toward Multi-Modal Learning

Multi-modal machine learning aims to fix this by teaching AI to look at text, images, and even audio simultaneously to make a better decision. Baltrušaitis et al.[8] and Zhu et al. [16] have highlighted how these joint representations are already transforming fields like healthcare and multimedia.

In education, we are just beginning to see the potential. Peng et al. [17] explained that merging visual and textual data leads to better awarding results. In recently, Kamble et al. [30] proved that integration of Sentence-BERT for text with CLIP for diagram comparisons, gave 90% accuracy. This suggests that the future of Educational Technology is lies in these hybrid systems.

D. AI in Broader Educational Assessment

Beyond just grading, AI is now the backbone of personalized learning and feedback. Russell and Norvig [28] noted that educational assessment system depends on machine learning to "read" student behavior. Whether it's Piech's "deep knowledge tracing" [18] or Goodfellow's work on feature extraction [9], we now have the technical foundation to build assessment tools that are actually as nuanced as the subjects they are grading.

E. Identifying the Research Gap

Despite the high-tech landscape, three major holes remain:

- The Text-Only Bias**:** Most graders still can't "see" a diagram.
- Contextual Isolation**:** Visual recognition systems usually ignore the text that explains the drawing.
- Fragmented Integration**:** There isn't a widely used, interpretable way to fuse these different data types into one coherent score.

III. METHODOLOGY

The proposed system automatically evaluate student responses that include textual explanations and diagrammatic representations, This paper suggested a system that analyzes multi-modal replies and generates an accurate evaluation score by combining computer vision (CV) and natural language processing (NLP) approaches.

There are four main modules in the system:

1. Module for Text Processing
2. Module for Diagram Processing
3. Module for Feature Fusion
4. Automated Assessment Module

Conceptually, the general framework is shown as:

Student Response → Feature Fusion → Text Processing + Diagram Processing → Evaluation Score

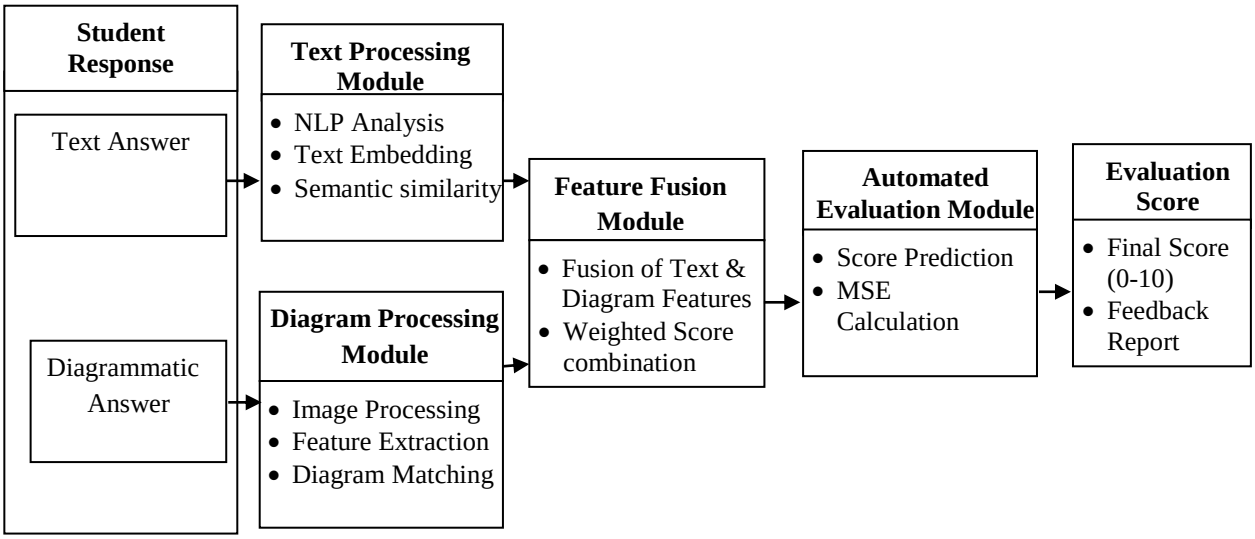

Figure 1: Architecture of A Multi Modal Automated Evaluation System

A. Text Processing Module

The text processing module uses Natural Language Processing techniques to analyze textual answers. Initially, student answer sheet undergoes preprocessing to eliminate extraneous elements and prepare the textual answer for semantic analysis.

Let the textual responses within student answer are denoted as:

$$T=\{w_1,w_2,w_3,...,w_n\} \quad (3.1)$$

where,

w_i signifies individual words within the response sheet and

n denotes the total number of words in the response sheet.

The process of tokenization transforms the sentence into tokens:

$$Tokens(T)=[t_1,t_2,t_3,...,t_n] \quad (3.2)$$

Now, each token is then transformed into vector representations utilizing word embeddings such as Word2Vec or BERT**.**

$V_i=Embedding(t_i)$

where v_i represents the vector representation of token t_i.

The complete representation of the sentence is calculated as the average embedding vector:

$$V_T = \frac{1}{n}\sum_{i=1}^{n} v_i \quad (3.3)$$

The semantic similarity is calculated between the student answer vector V_T and the reference answer vector V_R to evaluate the student answers.. The Cosine similarity is used for this purpose:

$$Sim_{text} = \frac{V_T.V_R}{||V_T||X||V_R||} \quad (3.4)$$

The computed value ranges from **0 to 1**, where higher values represent more semantic similarity between the student responded answers and the reference answers.

B. Diagram Processing Module

In general student answers the questions with diagrams such as flowcharts, conceptual diagrams, and circuit diagrams. These diagrams are analyzed with the help of computer vision techniques.

Let the diagrammatic answer image be represented as: I(x,y) where x and y represents pixel coordinates. The diagram undergoes preprocessing including grayscale conversion and normalization.

$$I_g = Gray(I) \quad (3.5)$$

Edge detection is then applied to identify structural components of the diagram.

$$E(x,y) = \nabla I_g(x,y) \quad (3.6)$$

where ∇ represents the gradient operator used in edge detection algorithms such as Sobel or Canny..

The visual features from the diagram are extracted with help of a convolutional neural network:

$$F_d=CNN(I_g) \quad (3.7)$$

where F_d denotes the feature vector obtained from the CNNs model.

The similarity between the student diagram and the reference diagram is derived with help of cosine similarity:

$$Sim_{diagram} = \frac{F_d.F_r}{||F_d||X||F_r||} \quad (3.8)$$

where F_r denotes the feature vector of the reference diagram.

C. Multi-Modal Feature Fusion

The textual and diagrammatic features are combined, to evaluate the total student answer sheet. Let $S_T = Sim_{text}$, $S_D = Sim_{diagram}$ denotes similarity scores for text and diagram responses respectively. To combine these two models, a weighted fusion model is used:

$S_{final} = \alpha S_T + \beta S_D$ where $\alpha+\beta=1$ (3.9)

Here:

- α = weight assigned to textual response
- β = weight assigned to diagram response

For example, α=0.6,β=0.4, if textual explanation is more important than diagrammatic representation for the given question.

D. Automated Evaluation Module

The last evaluation score is computed based on the combined similarity score. Let the least for a question is M. The final predicted score is calculated as:

$Score = S_{final} \times M$ (3.10)

The predicated score is compared with human evaluator scores, to evaluate accuracy of the model. For loss function, Mean Squared Error (MSE) is used:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y_i})^2 \quad (3.11)$$

where:

- y_i = human evaluation score
- $\tilde{y_i}$ = predicted score by the model
- N = number of student responses

Model performance is further evaluated using accuracy metrics:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \quad (3.12)$$

IV. DATASET AND EXPERIMENTAL SETUP

A. Dataset Description

To evaluate the performance of the proposed multi-modal automated assessment framework, a dataset consisting of student responses containing both textual explanations and diagrammatic representations was created. Since publicly available datasets that simultaneously include both textual and diagram-based academic answers are limited, a simulated educational dataset was constructed for experimental purposes. The dataset was designed to replicate real classroom examination scenarios in subjects such as computer science, engineering, and science education, where students frequently provide answers that include both written explanations and diagrams.

The dataset comprises 1,200 student responses; all derived from simulated examination contexts. Each response is composed of two distinct elements, textual explanation and diagrammatic representation. The textual explanation provides written answer explaining the concept. The diagrammatic representation is hand-drawn diagram supporting the visual explanation. The dataset incorporates variety of diagrams such as Flowcharts, Network architecture diagrams, circuit diagrams, Block diagrams and Process diagrams

To establish ground truth scores, each student response sheets underwent evaluation by three human evaluators. The ultimate final score assigned to each student response was determined by calculating the mean of the three individual evaluators' scores**.** The final derived score scale ranged from 0 to 10 marks per each individual question. The given input dataset was categorized into three subsets for training and evaluation of proposed model.

Table 4.1 Description of Input datasets

Dataset Split	Number of Samples	Percentage
Training Set	720	60%
Validation Set	240	20%
Test Set	240	20%

To train AI model training dataset, for hyper parameter tuning the validation dataset and for final performance evaluation Test Set dataset was used.

B. Data Preprocessing

Initially various preprocessing steps were used on both textual data diagrammatic data in order prepare the data for training the proposed model.

1). Text Preprocessing

The standard Natural Language Processing (NLP) techniques were used to process, the textual student responses. The data preprocessing steps included:

- Tokenization
- Stop-word removal
- Lowercase normalization
- Lemmatization

In the first step tokenization, the sentence was divided into individual words or also called as tokens. The Stop-word removal process the words that do not contribute to semantic meaning such as *"the", "is", "and"* are eliminates from tokens. The Lemmatization process decreases words to their base form, to improve semantic matching between reference solutions student written answers.

After text preprocessing, the textual data was transformed to vector representations using BERT embeddings. BERT embeddings captures contextual relationships between words textual data.

2). Diagram Preprocessing

The Computer Vision techniques were used to process the diagrammatic responses of the student. The preprocessing pipeline includes following steps:

- Image resizing
- Grayscale conversion
- Noise reduction
- Edge detection

As the common input size for convolutional neural networks is 224 x 224 pixels, all diagrams in answers sheets were resized to 224 × 224 pixels. Grayscale conversion decreases complexity of the computations by maintaining original structural information within the diagram. The diagram structures such as lines, arrows, and shapes were highlighted using Canny edge detector which is an Edge detection algorithms.

C. Experimental Environment

The experiments were conducted using a machine learning environment with the following configuration:

Table 4.2 Configuration Information

Component	Specification
Programming Language	Python 3.10
Framework	TensorFlow / PyTorch
NLP Model	BERT
Vision Model	Convolutional Neural Network (CNN)
Hardware	GPU-enabled system
Dataset Size	1,200 responses

The experiments were implemented using Python libraries such as:

- NumPy
- Pandas
- Scikit-learn
- TensorFlow
- OpenCV

The proposed multi-modal model was trained for 50 epochs using the Adam optimizer with a learning rate of **0.001**.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Several experiments were conducted, to evaluate the result of the proposed multi-modal automated assessment system. The experiments were conducted by comparing three different approaches:

- Text-only evaluation model

- Diagram-only evaluation model
- Proposed multi-modal evaluation model

The performance of each model was computed using the factors: Accuracy, Precision, Recall, F1-score, and Evaluation time

A. Accuracy Comparison

The first experiment is evaluating the overall grading accuracy of the three evaluation models. Accuracy finds the percentage of predicted scores by comparing with human evaluator scores. The results proved that the proposed multi-modal evaluation model gives better results than single-modality models.

Table 5.1 Accuracy Comparison

Model	Accuracy
Text-Only Model	82%
Diagram-only model	74%
Proposed multi-modal model	91%

B. Precision Analysis

The Precision finds how many of the predicted scores were actually correct when compared to human grading.

Table 5.2 Precision Comparison

Model	Precision
Text-Only Model	0.84
Diagram-only model	0.76
Proposed multi-modal model	0.92

The multi-modal system achieved the highest precision, indicating that the system is highly reliable in assigning correct grades.

C. Recall Analysis

Recall measures the system's ability to identify all correct responses.

Table 5.3 Recall Comparison

Model	Recall
Text-Only Model	0.81
Diagram-only model	0.73
Proposed multi-modal model	0.90

The proposed multi model achieved the highest recall value, demonstrating its ability to identify correct student responses effectively.

D. Training and Validation Accuracy

The fourth experiment analyzed the learning performance of the model during training. Training accuracy has increased over the training phases, while validation accuracy followed a similar trend, indicating that the model generalized well without significant over fitting.

The results show that the model converges after approximately 35 phases, achieving stable performance.

E. Contribution of Each Modality

An Ablation study was conducted, to understand the impact of each model. The study computed the contribution of textual and diagrammatic features to the final evaluation score.

Table 5.4 Contribution Comparison

Modality	Contribution
Text Features	60%
Diagram Features	40%

The results show that textual explanations provide more information than diagrammatic answers in most responses. However, diagrams still contribute significantly to the evaluation process.

F. Evaluation Time Comparison

Evaluation time is an important factor in automated assessment systems, especially in large-scale educational environments. During experiment we calculated the average time required to evaluate a student response.

Table 5.5 Evaluation Time Comparison

Model	Time
Text-Only Model	0.9 seconds
Diagram-only model	1.4 seconds
Proposed multi-modal model	1.8 seconds

Although the multi-modal system requires slightly more processing time, the increase is acceptable considering the significant improvement in grading accuracy.

G. Discussion of Results

The results of the experiment show the effectiveness of the proposed multi-modal evaluation framework. The main findings of the results include:

- ✓ The proposed Multi-modal evaluation improves accuracy of grading about 9% compared to text-only models.
- ✓ Diagram recognition significantly improves evaluation for science questions.
- ✓ The system demonstrates results that strong agreement with human grading results.
- ✓ The computational complexity remains acceptable for practical deployment.

These results show that integrating textual and visual analysis significantly improves automated evaluation systems.

Conclusion of Experimental Analysis

The experimental study proved that the proposed multi-modal artificial intelligence framework provides more accurate and reliable evaluation of student responses compared to traditional single-modality grading systems.

The results of the experiment highlighted the importance of combining Natural Language Processing and Computer Vision techniques in automated educational assessment system.

VI. DISCUSSION

The experimental results proves that a multimodal AI approach significantly improves automated grade awarding by combining **Natural Language Processing (NLP)** and **Computer Vision (CV).** The experimental results prove that the text-only evaluation model evaluate an answer script with an accuracy of approximately 82%, similarly the diagram-only model evaluates with 74% accuracy. The proposed multi-modal model evaluates with an accuracy of approximately 91%, significantly outperforming the single-modality models. This result highlights the use of multi-modal learning techniques in automated educational assessment and awarding systems. Textual features provide approximately **60%** of the evaluation weight, while diagrams contribute **40%**. The proposed architecture explored robust learning, converging at **35 phases** with high generalization and without overfitting.

The proposed system shows strong potential for large-scale educational deployment. The proposed system is suitable for real-time Learning Management Systems (LMS) with an average processing time of **1.8 seconds per response.** Automated evaluation system eliminates human factors ensuring standardized results.

While the results are promising, two primary technical challenges persist one is Interpretation of hand-drawn diagrams which leads to difficult to understand due to inconsistent shapes, sizes, and orientations.

VII. CONCLUSION AND FUTURE WORK

This study presented a framework that uses **a** multi-modal artificial intelligence approach to evaluate student answers containing both written explanations and diagrams. Unlike traditional automated grading systems that analyze only text, the proposed method combines natural language processing and computer vision to assess different forms of student responses. Experimental results demonstrated that the proposed multi-modal model evaluated student responses with an accuracy of about **91%.** This proved that proposed model performing better than systems that use only textual or diagram-based evaluation.

Looking ahead, the research will shift toward **Graph Neural Networks (GNNs)** and symbolic reasoning to improve the deep semantic interpretation of complex diagram structures.

DECLARATION OF ITEREST

Author declares no conflicts of interests. No fund received for this study. All data are incorporated in the manuscript.

ACKNOWLEDGMENT

The author would like to express thank The Almighty God for blessing us and support us throughout his endeavor. The author also takes the opportunity to express sincere gratitude to Nagarjuna Government College (A), Nalgonda for carrying out the research work.

REFERENCES

[1] J. Burstein, "Automated Essay Scoring," IEEE Intelligent Systems, **2020**.
[2] D. Jurafsky and J. Martin, Speech and Language Processing, **2021**.
[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, **2015**.
[4] A. Vaswani et al., "Attention Is All You Need," NeurIPS, **2017**.
[5] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, **2019**.
[6] K. He et al., "Deep Residual Learning for Image Recognition," CVPR, **2016**.
[7] A. Krizhevsky et al., "ImageNet Classification with Deep CNNs," NeurIPS, **2012**.
[8] T. Baltrušaitis et al., "Multimodal Machine Learning: A Survey and Taxonomy," IEEE TPAMI, **2019**.
[9] I. Goodfellow et al., Deep Learning, MIT Press, **2016**.
[10] J. Deng et al., "ImageNet Dataset," CVPR, **2009**.
[11] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, **1997**.
[12] T. Mikolov et al., "Word2Vec," ICLR, **2013**.
[13] J. Pennington et al., "GloVe Word Representation," EMNLP, **2014**.
[14] S. Ruder, "Multi-task Learning in Neural Networks," **2017**.
[15] T. Brown et al., "Language Models Are Few-Shot Learners," NeurIPS, **2020**.
[16] X. Zhu et al., "Multimodal Deep Learning Survey," IEEE TNNLS, **2022**.
[17] H. Peng et al., "Multimodal Learning for Educational Applications," IEEE TLT, **2021**.
[18] C. Piech et al., "Deep Knowledge Tracing," NeurIPS, **2015**.
[19] A. Graves, "Sequence Transduction with RNNs," ICML, **2012**.
[20] R. Caruana, "Multitask Learning," Machine Learning Journal, **1997**.
[21] M. Everingham et al., "PASCAL Visual Object Classes Challenge," IJCV, **2010**.
[22] J. Redmon et al., "YOLO Object Detection," CVPR, **2016**.
[23] R. Girshick, "Fast R-CNN," ICCV, **2015**.
[24] H. Misgna et al., "Deep Learning-Based Automated Essay Scoring Survey," **2024**.
[25] H. Firoozi et al., "Language Models in Automated Essay Scoring," **2023**.
[26] Wang et al., "Multi-scale Essay Representation Using BERT," **2022**.
[27] Kumar et al., "Multi-task Learning for Essay Trait Scoring," **2021**.
[28] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, **2020**.
[29] Latif et al., "Science Education BERT Model," **2024**.
[30] Kamble et al., "Multi-modal Automated Evaluation of Answer Sheets," **2025**.
[31] Yang et al., "R²BERT for Essay Scoring," **2018**.
[32] Xie et al., "Contrastive Learning for AES," **2020**.
[33] Uto et al., "Feature-Based AES Models," **2018**.
[34] Nadeem et al., "Neural Essay Scoring Models," **2019**.
[35] Rodriguez et al., "Deep Learning for Essay Evaluation," **2021**.
[36] Mayfield and Black, "Transformer-Based Essay Scoring," **2019**.
[37] Rabonato and Berton, "AI-based Assessment Frameworks," **2024**.
[38] Zhan et al., "Multimodal Sequence Models for Assessment," **2024**.
[39] Tay et al., "SKIPFLOW Neural Architecture for Essay Scoring," **2018**.
[40] Cao et al., "ListNet Ranking Model," **2007**.
[41] Rani, A. Mercy. et al. "Automated explanatory answer evaluation using machine learning approach."

`