



# NATIONAL CONFERENCE ON INTEGRATING TECHNOLOGIES, IDEAS AND DISCIPLINES FOR ENGINEERING INNOVATION (NCITIDE 2025)

ScoreMatrix: A Parameter-Guided Framework for Automated Evaluation of Answer Sheets

**Paper ID - 10**

**Pulkit Kapur | MSIT | 18-19 Nov. 2025**

Prakhar Chandra, Department of Computer Science and Engineering

Jaspreet Singh, Department of Information Technology

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5



# Abstract

## ScoreMatrix: Transparent & Configurable Automated Grading

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

### 1) **Objective:**

- a) Score Matrix: a modular, educator-configurable framework that balances interpretability and grading accuracy.

### 2) **Methodology:**

- a) Uses a weighted ensemble of core metrics:
  - i) Semantic alignment (meaning match)
  - ii) Keyword and concept coverage (main points/terms)
  - iii) Coherence and logical flow (organization)
  - iv) Grammar and readability (clarity, error detection)
- b) Evaluated on 847 student responses spanning descriptive, opinion, and objective questions across multiple subjects.
- c) Human expert ratings used as the gold standard.

### 3) **Results:**

- a) Achieved Pearson correlation of 0.54 and MAE of 3.37 relative to expert graders—outperforming traditional rule-based and single-metric AI baselines.
- b) Delivers criterion-level, interpretable feedback; educators can adjust metric weighting for local needs.

### 4) **Conclusion:**

- a) Score Matrix offers a practical, scalable, and transparent approach to automated answer grading, bridging a key gap in existing systems by empowering both teachers and students with actionable feedback.
- b)



# Introduction

## Introduction: The Challenge of Automated Grading

- 

### Background & Current Landscape:

1. Rapid expansion of digital education and large class sizes has created demand for efficient, scalable, and consistent grading systems.
2. Automated grading is quick and unbiased for multiple-choice questions (MCQs), but often fails to handle subjective, open-ended, or creative responses reliably.
3. Manual grading is still the norm for such questions, but it is slow, inconsistent (due to individual grader judgment and fatigue), and not feasible at large scale.

### Existing Approaches & Their Limitations:

4. Rule-based Automated Systems:
  - a. Grade by matching answers to keywords, patterns, or rubrics—transparent but inflexible to varied writing styles and creative responses.
5. Neural/AI Systems (e.g., LLMs):
  - a. Achieve better accuracy and adapt to diverse student answers, but act as 'black boxes,' lacking transparency and raising trust issues with educators.
  - b. Require quality training data and can sometimes introduce bias or grading inconsistencies across models.
6. Both approaches can underperform in assessing deep understanding, higher-order thinking, or originality.

### Motivation for ScoreMatrix:

7. A clear need exists for systems that:
  - a. Balance accuracy and interpretability, providing clear, explainable grading logic that teachers and students trust;
  - b. Allow teacher control and give fine-grained, actionable feedback for learning improvement.
  - c. Our Contribution:
8. Score Matrix addresses this gap by integrating a modular, weighted framework of interpretable metrics for grading.
9. Combines semantic similarity, content coverage, logical flow, and language quality for robust assessment.
10. Proven effective on 847 diverse student responses, enabling educator customization and transparent, criterion-level feedback.

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5



# Literature Survey and Problem Statement

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

## Evolution of Automated Grading Systems

- • Early Rule-Based Approaches:
  - Systems relied on keyword matching and fixed rubrics for grading () .
  - Offered transparency and easy configuration, but struggled with paraphrasing, synonyms, and assessing deeper understanding.
- Semantic Methods:
  - Adoption of vector-space models and sentence embeddings (e.g., Sentence-BERT) improved the ability to capture semantic similarity and handle diverse student expressions () .
  - Helped with paraphrased content but still lacked holistic evaluation.
- Neural & Transformer Models:
  - Introduction of RNNs, LSTMs, and Transformers like BERT enabled systems to learn patterns from large datasets and perform context-aware grading () .
  - Significantly improved accuracy but created "black box" systems with limited transparency for educators.
- Hybrid & Multi-Metric Approaches:
  - Recent efforts combine several interpretable metrics (semantic, keywords, structure, grammar) using weighted frameworks to balance accuracy and explainability.
  - This hybridization enables some customization, but often lacks flexible teacher control and may not fully close interpretability gaps () .
- Recent Developments – Large Language Models:
  - LLMs (e.g., ChatGPT, GPT-4) show impressive scoring ability.
  - However, they pose issues of calibration, explainability, and may introduce bias or fairness concerns, making adoption challenging for high-stakes assessment contexts () .

## Identified Research Gap:

- Most current systems force a tradeoff: high interpretability with lower accuracy, or black-box models with limited trust and teacher control.
- There remains a lack of configurable, criterion-level feedback systems that educators can tune for their objectives and students can learn from directly.

## Our Position — ScoreMatrix:

- ScoreMatrix is designed to combine the best of both worlds: delivering strong grading performance through a modular, weighted approach that is transparent, interpretable, and fully configurable for different teaching contexts.
- By offering detailed, adjustable criterion-level scoring across semantic, factual, structural, and language-related dimensions, ScoreMatrix addresses the shortcomings identified in the literature.



# Proposed Methodology

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

## Parameter-Guided Multi-Metric Evaluation Framework

### Mathematical Formulation:

Given: Question q, reference answer r, student response s

Goal: Generate overall score  $\hat{y} \in [0, S_{\max}]$  and criterion scores  $c = [c_1, \dots, c_K]$

### Composite Score:

$$\hat{y} = S_{\max} \cdot \sum_{k=1}^K w_k m_k(s, r), \quad \text{with} \quad \sum_{k=1}^K w_k = 1, \quad w_k \geq 0,$$

### Five Core Interpretable Metrics:

1. Semantic Similarity - Sentence-BERT embeddings with cosine similarity
2. Keyword & Point Coverage - NLP tools (spaCy, WordNet) for concept matching
3. Coherence & Logical Flow - Entity-grid modeling and sentence-similarity scoring
4. Grammar & Readability - LanguageTool + Flesch-Kincaid readability metric
5. Factual Consistency (Optional) - NLI models for contradiction detection

Extended Features: 15 additional features including TF-IDF similarity, structural alignment, technical term usage



# Experimental Setup

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

## Technical Stack:

- Language: Python 3.9
- NLP Libraries: spaCy, Sentence-Transformers (SBERT, MiniLM), NLTK, WordNet
- Grammar Check: LanguageTool
- Hardware: NVIDIA GPU, 32GB RAM

## Dataset:

- Size: 847 student responses across 15 questions
- Domains: Science, social studies, literature (undergraduate level)
- Question Types: 45% descriptive, 30% opinion-based, 25% objective
- Gold Standard: Averaged scores from two expert human raters
- Split: 60% training/calibration, 20% validation, 20% test

## Baselines:

- TF-IDF + Cosine Similarity
- Sentence-BERT (SBERT)
- Exact Match
- Unweighted Ensemble

## Evaluation Metrics:

- Pearson correlation ( $r$ ), MAE, RMSE,  $R^2$ , Kendall's tau ( $\tau$ )



# Results and Discussion

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

## Performance Comparison & Ablation Studies

Table: Performance on Test Set

Method	Pearson r	MAE	RMSE	R <sup>2</sup>	Kendall's T
TF-IDF + Cosine	0.43	3.87	4.62	-0.52	0.28
Sentence-BERT	0.47	3.55	4.28	-0.41	0.31
Exact Match	0.12	5.21	5.90	-1.85	0.08
Unweighted Ensemble	0.51	3.42	4.10	-0.28	0.35
ScoreMatrix (Weighted)	0.54	3.37	4.02	-0.15	0.38

Ablation Study Results:

Configuration	Pearson r	Kendall's T
Full system	0.54	0.38
- Grammar	0.49	0.34
- Coherence	0.50	0.35
- Keyword coverage	0.47	0.32
- Semantic similarity	0.45	0.30

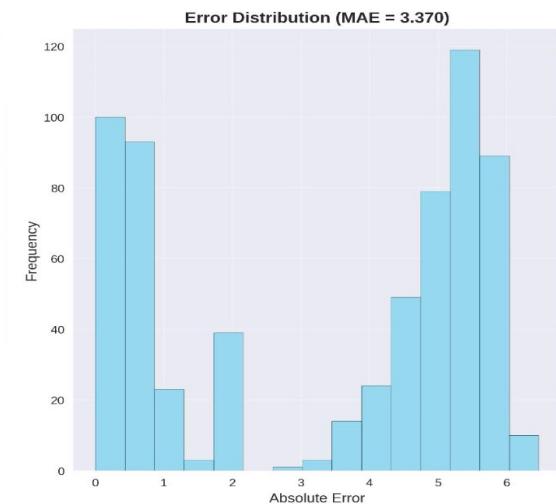
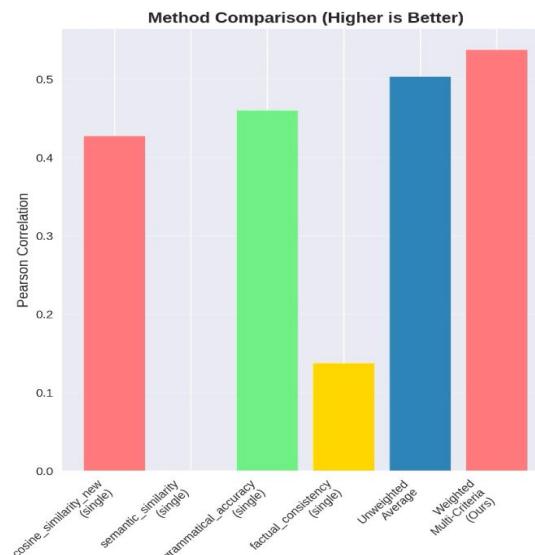
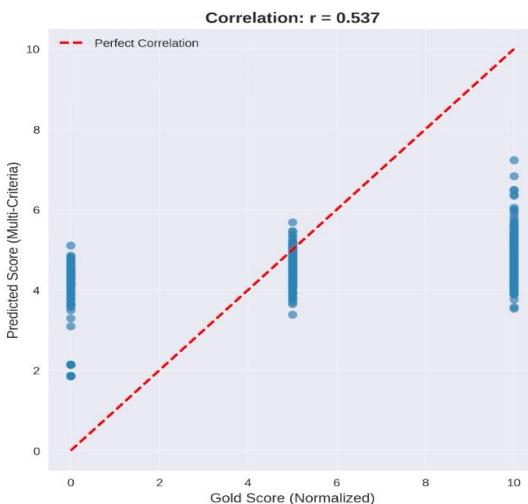
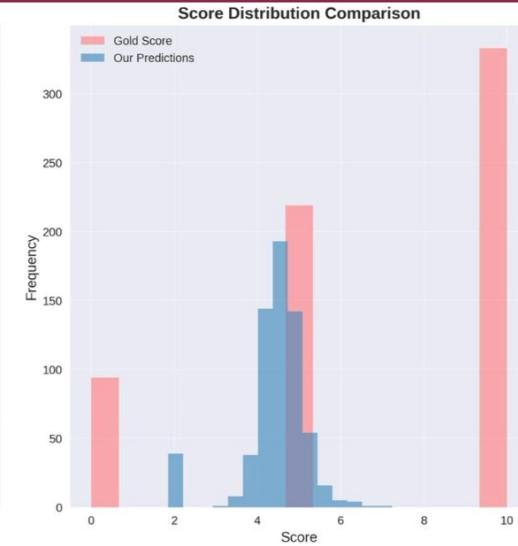
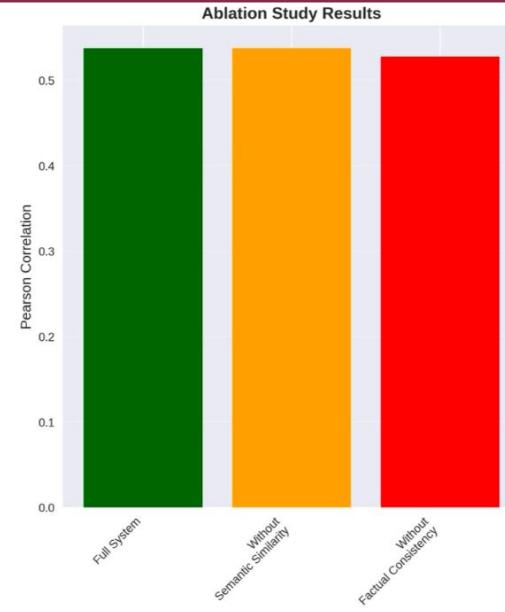
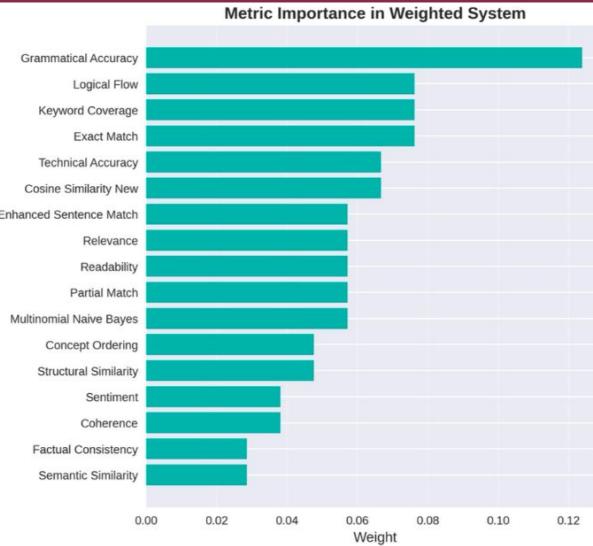
Key Findings:

- Each metric contributes meaningfully (5-9% drop when removed)
- Average runtime: 0.8s per response (95% faster than manual grading)
- Moderate rank-order agreement with human graders



N  
C  
I  
T  
I  
D  
E

# Results and Discussion





# Conclusion and Future Scope

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5

## Error Analysis & Quality Assessment

Sample Error Cases:

Q	Student Answer	ScoreMatrix / Human	Analysis
1	"The transistor amplifies current by controlling flow..."	4 / 6	Underscored: Concise but correct; low keyword match
2	"Newton's 3 laws are inertia, force, reaction..."	9 / 9	Perfect alignment: Both systems converged
3	"Photosynthesis happens in animals by light energy"	6 / 2	Overscored: Factually incorrect but keyword-rich

Accuracy Metrics:

- Correlation with humans: 0.54 (moderate alignment)
- Inter-rater agreement: 0.76 (Cohen's Kappa) between human evaluators
- Consistency: Perfect (100%) - same rules applied uniformly

Efficiency Gains:

- Grading time: 0.8s vs 90s manual (95% reduction)
- Feedback quality: Personalized, actionable, concept-level insights

Pedagogical Value:

- Highlights missing concepts, structural issues, grammar corrections
- Supports formative assessment beyond summative scoring



# References

## Key Citations

- A. Rudner and J. Liang, "Automated essay scoring using Bayes' theorem," *The Journal of Technology, Learning and Assessment*, vol. 1, no. 2, 2002.
  - J. Burstein, D. Harris, and K. Kukich, "Automated essay evaluation: the criterion online service," *AI Magazine*, vol. 25, no. 3, pp. 27–36, 2004.
  - C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 389–405, 2003.
  - M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," *Proc. 12th European chapter of the ACL*, 2011.
  - N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *Proc. Conf. Empirical Methods in NLP*, 2019.
  - M. A. Sultan, C. Salazar, and T. Sumner, "Fast and easy short answer grading," *Proc. NAACL*, 2016.
- [7-15] Additional references on neural models, LLMs, and hybrid approaches (see paper for complete list)

N  
C  
I  
T  
I  
D  
E

2  
0  
2  
5