International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

AI-Based Answer Evaluation Using Semantic Similarity and Multi-Agent Reasoning

Kushal Vishwajeet Bishwas
Department of Science and Technology, G.H.Raisoni Skill Tech University, Nagpur, India

ABSTRACT

Automated evaluation of descriptive responses is a significant difficulty in educational technology, largely because of the significant limitations of conventional keyword-based grading systems. Automated tests are based on finding specific words or basic ideas they do not really check if the student truly understands the concept, what the words mean or the different ways a correct answer can be written. The automated evaluation methods have these limitations because they rely on matching phrases or fundamental principles. This can sometimes lead to grades that're not fair or consistent which is a big problem in online classes where reliability and consistency are really important, for automated evaluation methods. Automated evaluation methods need to be fair and consistent. Recent advances in Large Models of Language (LLMs) offer new possibilities for more intelligent evaluation based on logic and comprehension.

Using explainable reasoning and semantic similarity, this work presents a multi-agent system for AI-driven assessment of descriptive responses. Preprocessing, evaluating semantic alignment, analyzing concept coverage, applying rubrics, producing explanations, and verifying the coherence between scores and feedback are just a few of the duties that the framework assigns to agents. The approach emphasizes conceptual comprehension rather than just keyword matching by fusing reasoning from LLMs with embedding-based semantic similarity. Additionally, a validation loop is employed to reduce variability and enhance grading reliability.

The suggested approach is intended to enhance automated evaluation systems' scalability, transparency, and fairness. A comparison with baseline keyword-matching methods shows that the creation of structured feedback improves interpretability and alignment with human evaluators. This work provides a reliable and comprehensible solution for contemporary AI-driven educational assessment environments by combining multi-agent orchestration with semantic evaluation.

Keywords: Automated Answer Evaluation; Semantic Similarity; Large Language Models; Multi-Agent Systems; Explainable AI; Educational Technology; Rubric-Based Scoring; Natural Language Processing.

1. Introduction

The rapid expansion of educational websites with comprehensive assessment systems has significantly changed the educational landscape. With the emergence of MOOCs, or large open online courses,

computerized examination systems, and virtual classrooms, educational institutions are increasingly relying on technology to accurately and efficiently evaluate student performance. While rule-based logic and deterministic algorithms can be used to grade objective-type questions, the automated assessment of describe and open-ended responses is still a difficult and mostly unsolved issue. Semantic interpretation, conceptual comprehension, and contextual reasoning are necessary for descriptive responses, and these tasks are difficult for conventional automated grading systems to accomplish. Traditional automated grading methods mostly rely on superficial statistical methods, lexical similarity, or keyword matching. These systems usually use surface-level textual similarity scores or specified keywords to compare student replies with reference answers. Semantic equivalency, paraphrase, alternative explanations, and incomplete conceptual understanding are not taken into consideration by such systems, even though they might function satisfactorily in highly structured circumstances. Students frequently use a variety of sentence forms and terminology to convey accurate concepts, which strict keyword-based systems may unfairly penalize. These methods may therefore result in biased grading, less transparency, and a decline in confidence in automated evaluation systems.

Natural language comprehension has advanced significantly with the introduction of transform-based architecture as well as big language models (LLMs). Semantic interpretation, coherent text production, and contextual reasoning are all areas in which modern LLMs excel. Beyond basic word overlap, these models may infer conceptual meaning and capture intricate linguistic links. As a result, they provide encouraging prospects for enhancing automatic response evaluation systems. However, there are additional difficulties when using LLMs directly to grade assignments. LLM outputs can serve as "black-box" systems that produce scores without a systematic explanation, are frequently sensitive to prompt design, and may show unpredictability over repeated evaluations.

Inspiration

Calculative algorithms can be used to assess unbiased queries, but higher semantic comprehension and contextual reasoning are needed for descriptive and open-ended answers. Conventional automated grading systems mostly rely on surface-level text matching or keywords matching, which misses conceptual equivalency, paraphrased explanations, and incomplete comprehension. Because of this, these procedures frequently end in unjust and contradictory grading results, which erodes student and teacher trust.

The desire to create an organized, comprehensible, and scalable evaluation framework that makes use of multi-agent reasoning and semantic similarity is what spurred this research. The suggested method seeks to improve fairness, transparency, and conformity with human evaluation standards by breaking down the grading process into specialized parts and adding affirmation of consistency procedures.

Participating

The main contributions of the paper are listed as follows:

1. To integrate semantic similarity techniques with Large Language Model (LLM)-based concept-level reasoning for evaluating answers based on contextual understanding rather than keyword overlap.
2. To design an explainable mark deduction mechanism that generates transparent and human-readable feedback for awarded and deducted marks.
3. To implement a consistency verification loop that validates alignment between assigned scores and generated explanations, thereby improving grading reliability.
4. To perform a comparative evaluation with traditional keyword-based grading methods to demonstrate improved fairness, interpretability, and alignment with human assessment standards.

When taken as a whole, these contributions provide an organized and comprehensible framework for evaluating descriptive answers automatically. The suggested system overcomes the main drawbacks of conventional keyword-based grading methods by fusing semantic similarity algorithms, multi-agent orchestration, rubric-based scoring, and consistency checking mechanisms. The methodology is appropriate for real-world academic deployment and scalable digital assessment environments because explainability and structured validation are integrated to improve dependability, fairness, and transparency.

1. Related work

Under the more general heading of Automated Essay Scoring (AES), automated evaluation of descriptive responses has been researched for a number of decades. Handcrafted features like word frequency counts, sentence length, grammar patterns, and surface-level textual similarity were the mainstay of early systems. By comparing student responses to pre-established templates, these systems frequently used statistical or rule-based models to predict grades. Although these methods provide a basis for automated assessment, their capacity to comprehend more profound semantic meaning was constrained. Lexical overlap was a major factor in the evaluation, which limited the ability to handle responses that were conceptually comparable or paraphrased.

In later studies, supervised classification approaches were integrated into machine learning-based scoring models.[7] Researchers sought to identify patterns linked to high- and low-quality responses by training models on massive datasets of graded responses. Compared to strictly rule-based systems, techniques like Support Vector Machines, Random Forests, and early neural networks showed gains. Nevertheless, these models continued to rely significantly on manufactured characteristics and were frequently domain-specific, necessitating significant retraining when used.

In this article [9], researchers started investigating end-to-end learning techniques for text assessment with the advent of deep learning, specifically recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

By automatically deriving representations from unprocessed textual input, these models decreased the need for human feature engineering. For example, sequential dependencies in student replies were captured using Long Short-Term Memory (LSTM) networks. Although these models outperformed conventional approaches in terms of contextual understanding, they continued to have issues with interpretability, long-form reasoning, and consistent grading across a range of stimuli.

A major turning point in natural language processing was the development of transformer-based architectures, as discussed in this article [12]. Strong abilities in semantic understanding and contextual representation learning were shown by models like BERT and GPT. Systems were able to compare student and reference responses at a deeper conceptual level because to these transformer models, which made embedding-based similarity computing possible. Sentence embeddings and contextual similarity metrics were first used in research to more accurately assess descriptive responses. Despite the fact that these techniques enhanced semantic alignment, many implementations continued to be one-step evaluation pipelines devoid of organized reasoning and validation procedures.

Explainable Artificial Intelligence (XAI) has becoming more popular in educational technology, especially in evaluation systems, as this paper [16] illustrates. In order to preserve confidence between teachers and students, researchers stress the significance of producing understandable explanations for scores. To improve transparency, strategies including feature attribution, attention visualization, and structured feedback creation have been put forth. There is potential for improvement in educational applications, nevertheless, as many explainability approaches prioritize model interpretability over organized feedback that is in line with grading rubrics.

2. Research Methodology

2.1 Problem statement

In educational systems, automated evaluation of descriptive responses presents significant obstacles. Descriptive responses, in contrast to multiple-choice questions, require an awareness of meaning, context, and how closely they align with predetermined concepts. Previous grading schemes mostly look for keywords or related terms, which may overlook alternative explanations, sound reasoning, or incomplete comprehension. These systems frequently provide unfair or inconsistent scores as a result. Additionally, rule-based approaches are unsuitable for evolving academic environments since they are not adaptable enough for various subjects or question kinds. Although human grading is reliable, it can be prejudiced, slow, and dependent on the individual grading, particularly when dealing with huge exams or online learning.

Large Language Models (LLMs), a more recent technique, are better at understanding meaning, although there are problems with utilizing them directly for grading.

These include hard-to-explain, ambiguous checks, and inconsistent results. Many LLM grading systems operate as "black boxes," giving grades without offering a transparent justification or in accordance with accepted grading guidelines. This calls into doubt the reliability, fairness, and consistency of grading tools. Thus, creating a methodical, transparent, and reliable evaluation system is the aim of this research. It aims to assess descriptive responses based on how well they match the intended meaning and comprehension while making sure that the grading is fair, intelligible, and generally relevant in real-world educational settings.

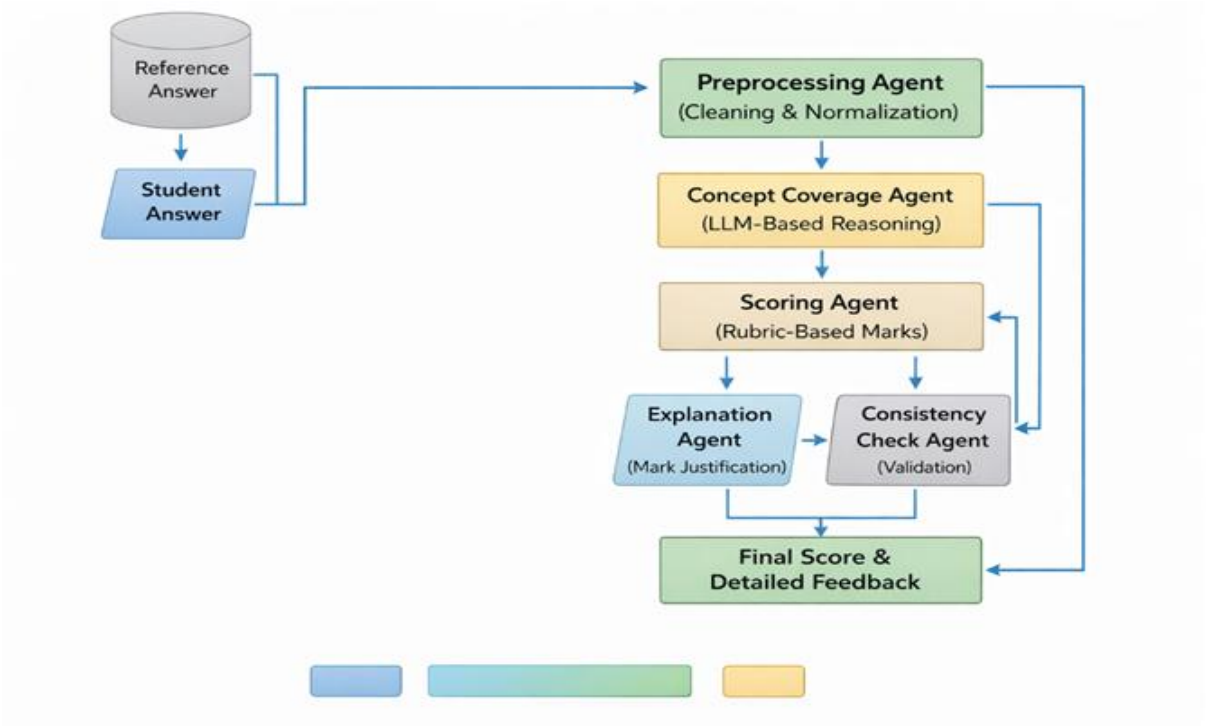

Fig 1. Block diagram of the proposed model

2.1.1 System Architecture Overview

The general setup of the proposed multi-agent grading system is described in this section. It provides a comprehensive, step-by-step explanation of how to handle both conventional replies and student responses. The system is composed of a number of essential components, including data preparation, intelligible language representations, idea analysis, scoring according to predetermined criteria, explanations, and consistency checks. The design emphasizes having distinct components that can function independently while also cooperating effectively through planned coordination.

2.1.2 Multi-Agent Framework Design

1. Preprocessing Agent

The Preprocessing Agent needs to prepare both the reference answer and the student response for semantic evaluation. Raw textual inputs frequently contain grammatical flaws, repetitious words, spelling problems, and formatting oddities that might negatively impact future analysis. Consequently, this agent performs essential normalization tasks like token standardization, lowercasing, punctuation management, and removing unnecessary whitespace. It also breaks up long responses into logical segments to facilitate concept-level examination. To reduce lexical variability without altering semantic meaning, lemmatization and stop-word filtering can also be employed.

The preparation step guarantees consistent formatting in addition to basic text cleaning, ensuring that language models and embedding models receive consistent input structures. In educational datasets, where student replies differ greatly in writing style and structure, this phase is very crucial. The Preprocessing Agent improves the dependability of embedding production and reasoning-based evaluation by removing superficial noise and maintaining semantic meaning. All things considered, this module creates a standardized framework for the multi-agent workflow, lowering errors and enhancing the stability of later phases of semantic and conceptual analysis.

2. Semantic Similarity Agent

The conceptual alignment between the reference answer and the student response is assessed by the Semantic Similarity Agent. This agent uses embedding-based representation approaches to capture contextual meaning, in contrast to conventional keyword-based comparison methods. A pre-trained language model is used to transform both responses into high-dimensional vector embeddings. To gauge semantic closeness, cosine similarity between the vectors is then calculated. This enables the system to identify restructured or paraphrased responses that use different words to convey the same meaning.

An initial quantitative measure of total alignment is given by the similarity score. It is an informative signal that directs more in-depth concept-level study rather than serving as the only grading criterion. Responses can be categorized into groups that are highly aligned, somewhat aligned, or poorly aligned using threshold-based evaluation. The Semantic Similarity Agent enhances evaluation fairness by emphasizing meaning over lexical overlap. This method preserves academic rigor while allowing the grading system to identify a variety of linguistic expressions. It considerably lessens the punishment meted out to pupils who express accurate ideas in different ways.

3. Concept Coverage Agent

To determine which important ideas from the reference answer are included, just partially covered, or absent in the student response, the Concept Coverage Agent uses thorough reasoning. Concept-level analysis guarantees a detailed assessment of each knowledge component, whereas semantic similarity offers a global alignment score. To extract key ideas from the reference answer and methodically compare them with the student's explanation, this agent makes use of a Large Language Model (LLM).

This feature makes automatic grading systems more transparent and trustworthy. Students gain from practical feedback that explains the reasons behind grade reductions and how to improve their answers. Instead of being general, the explanations are intended to be intellectually constructive. The agent guarantees the grading process's interpretability by adhering to Explainable AI standards. This promotes formative learning goals and lessens skepticism about automated systems. Thus, the Explanation Agent turns the system from a simple grading tool into a learning-support system that promotes conceptual improvement and clarity.

4. Consistency Validation Agent

Logical coherence between the generated explanation and the assigned score is ensured by the Consistency Validation Agent. There may occasionally be differences in LLM-based systems where the numerical score is quite low but the feedback indicates a great conceptual grasp, or vice versa. By

reassessing the alignment between scoring outputs and explanation material, this agent carries out cross-validation.

The system initiates a re-evaluation loop to improve the grade choice if discrepancies are found. Reliability can also be estimated using confidence rating systems. This methodical validation process improves grading consistency over multiple passes and lowers output variability. The system prevents arbitrary or conflicting results by including verification logic into the workflow. An essential invention for guaranteeing robustness, reproducibility, and fairness in an automated answer evaluation system is the Consistency Validation Agent.

2.1.3 Experimental Setup & Evaluation Metrics

1. Experimental Setup

A controlled experimental setup was created to mimic actual academic grading conditions in order to assess the efficacy of the suggested multi-agent answer evaluation framework. Descriptive question-answer pairs from academic areas where conceptual explanations are needed instead of objective answers made up the dataset. A model reference response created in accordance with accepted academic practices was included with each question. To ensure variation in evaluation circumstances, student replies comprised both accurate and partially correct answers.

All replies were manually scored by human evaluators using a predetermined rubric for ground-truth comparison. The baseline for gauging the degree of alignment between automatic and manual grading was these human-assigned scores. Training examples for prompt calibration, validation examples for parameter adjustment, and testing examples for final evaluation were the three categories into which the dataset was separated. A large language model combined with a multi-agent orchestration framework was used to create the system. Within the structured workflow, which included preprocessing, semantic similarity analysis, concept coverage evaluation, scoring, explanation creation, and consistency validation, each agent was carried out in turn.

A keyword-matching grading methodology served as the baseline method for comparison, in which student responses were assessed according to lexical overlap with the reference response. This made it possible to evaluate the performance gains brought forth by multi-agent validation and semantic reasoning. To assess output stability and dependability, several evaluation runs were carried out.

2. Evaluation Metrics

To objectively measure system performance, several quantitative metrics were used:

2.1 Human–AI Score Correlation

Measures how closely automated scores match human evaluator scores. Higher correlation indicates human-like grading behaviour.

Correlation = corr (Human Scores, AI Scores)

2.2 Grading Accuracy

Represents the percentage of responses where the AI score falls within an acceptable range (±1 mark) of the human score.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Responses}}$$

2.3 Consistency Score

Measures reproducibility by evaluating score variation across multiple runs on the same input.

$$\text{Consistency} = 1 - \frac{\text{Maximum Possible Variance}}{\text{Score Variance}}$$

2.4 Explanation Alignment Score

Evaluates whether feedback justification logically matches assigned marks. Human reviewers rate explanations on a scale of 1–5.

2.1.4 Summary of the Proposed Multi-Agent Evaluation Framework

Module / Agent	Primary Function	Techniques Used	Output
Preprocessing Agent	Cleans and normalises input text	Tokenisation, Lemmatisation, Text Normalisation	Standardised textual input
Semantic Similarity Agent	Measures conceptual alignment between reference and student answers	Sentence Embeddings, Cosine Similarity	Similarity Score (0–1)
Concept Coverage Agent	Identifies the presence and correctness of key concepts	LLM-based structured reasoning	Concept Coverage Report
Scoring Agent	Allocates marks based on rubric logic	Weighted Scoring, Partial Credit Mechanism	Numerical Score
Explanation Agent	Generates structured feedback for grading decisions	Prompt-based LLM Feedback Generation	Human-readable Explanation
Consistency Validation Agent	Ensures logical alignment between score and explanation	Cross-verification & Re-evaluation Loop	Validated Final Score

2.1.5 System Configuration and Implementation Details

Component	Configuration
Programming Language	Python
Workflow Framework	LangChain + LangGraph
Large Language Model	GPT / Open-source LLM (Specify version)
Embedding Model	Sentence Transformer / OpenAI Embeddings
Similarity Metric	Cosine Similarity
Evaluation Baseline	Keyword-based Grading Model
Evaluation Metrics	Human–AI Correlation, Accuracy, Consistency Score
Deployment Environment	Local / Cloud-based API

2.1.6 Proposed Algorithm

Step 1: Input the reference answer (R) and student answer (S) into the system.

Step 2: Apply preprocessing to both R and S:

a. Remove unnecessary symbols

b. Normalise text

c. Perform tokenisation and lemmatisationn

d. Standardize formatting

Step 3: Generate semantic embeddings for both R and S using a pre-trained embedding model.

Step 4: Compute cosine similarity between the embeddings to obtain an overall semantic similarity score (Sim).

$$Sim = \frac{R \cdot S}{||R|| \ ||S||}$$

Step 5: Extract key concepts from the reference answer using LLM-based concept identification.

Step 6: **Analyse** the student's answer to determine:

a. Fully covered concepts

b. Partially covered concepts

c. Missing concepts

d. Incorrect interpretations

Step 7: Apply rubric-based scoring logic:

a. Assign full marks for correctly explained concepts

b. Assign partial marks for incomplete explanations

c. Deduct marks for missing or incorrect concepts

Step 8: Generate structured explanation feedback based on concept coverage and scoring decisions.

Step 9: Perform consistency validation by checking logical alignment between score and explanation.

Step 10: If an inconsistency is detected:

a. Trigger re-evaluation

b. Recompute score and explanation

Step 11: Output final validated score (F) and detailed feedback (E).

3. Research Methodology

The suggested system uses concept-level reasoning and semantic similarity in an organized multi-agent workflow to assess descriptive responses. A sample question-answer pair is provided to show how the grading process is carried out step-by-step in order to highlight how the approach works.

3.1 Sample checking

Sample Question:

Question:

Explain the concept of Machine Learning and mention its types.

Reference Answer (Model Answer):

Machine Learning is a subfield of Artificial Intelligence that enables systems to learn patterns from data and improve performance without explicit programming. It involves training algorithms on datasets to make predictions or decisions. The main types of Machine Learning are Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

Student Answer (Example Response):

Machine Learning is a method where computers learn from data instead of being directly programmed. It helps systems make predictions based on patterns. There are mainly two types: supervised learning and unsupervised learning.

4.2 Evaluation Process Using Proposed Framework

1. Preprocessing

Both reference and student answers are cleaned, normalised, and tokenised. Formatting inconsistencies and redundant elements are removed to ensure consistent semantic comparison.

2. Semantic Similarity Analysis

Embedding vectors are generated for both answers. Cosine similarity is computed to measure overall conceptual alignment. In this case, similarity would be high because the student correctly defines Machine Learning and mentions major types.

3. Concept Coverage Analysis

Key concepts extracted from the reference answer:

1. Definition of Machine Learning
2. Learning from data without explicit programming
3. Supervised Learning
4. Unsupervised Learning
5. Reinforcement Learning

Student coverage evaluation:

1. Definition of Machine Learning → Correct
2. Learning from data → Correct
3. Supervised Learning → Present
4. Unsupervised Learning → Present
5. Reinforcement Learning → Missing

4. RUbric-Based Scoring

Assume total marks = 10

- Definition (4 marks) → Full marks awarded
- Types identification (6 marks)
- Supervised → 2 marks
- Unsupervised → 2 marks
- Reinforcement → 0 marks

Final Score = 8/10

5. Explanation Generation

Generated Feedback:

"2 marks were deducted because the student did not mention Reinforcement Learning as one of the primary types of Machine Learning. The definition and explanation of learning from data were correctly described."

6. Consistency Validation

The system verifies that the explanation logically corresponds to the score deduction. Since the missing concept accounts for 2 marks, the evaluation is considered consistent.

Methodological Significance:

This structured example demonstrates how the proposed multi-agent system ensures:

- Semantic understanding beyond keywords
- Partial credit allocation
- Transparent feedback
- Consistent grading

Unlike traditional keyword-matching approaches, the system evaluates conceptual correctness and reasoning depth while maintaining academic fairness.

The proposed methodology presents a structured and explainable approach to automated descriptive answer evaluation, integrating semantic similarity, concept-level reasoning, rubric-based scoring, and consistency validation within a unified multi-agent framework. Unlike conventional keyword-based systems that rely on lexical overlap, the proposed model evaluates answers based on conceptual alignment and contextual understanding. This ensures that semantically correct responses expressed in varied linguistic forms are not unfairly penalised. Such capability is particularly important in descriptive assessments where students may articulate correct reasoning using diverse vocabulary and sentence structures.

A key methodological contribution lies in the decomposition of the grading process into specialised agents. By separating preprocessing, similarity computation, concept coverage analysis, scoring, explanation generation, and validation, the system reduces dependence on single-step black-box evaluation. This modular design enhances transparency, reproducibility, and robustness. Each agent performs a well-defined function, allowing systematic verification of grading logic and reducing unintended biases or inconsistencies.

Furthermore, the incorporation of rubric-based scoring ensures alignment with academic evaluation standards. Rather than assigning arbitrary scores, marks are distributed proportionally based on clearly defined conceptual components. The explanation module further strengthens interpretability by generating structured feedback linked directly to scoring decisions. This supports formative assessment by helping students identify conceptual gaps.

5. Results and Analysis

1. Graphical Analysis

The graphical analysis presented in Figures 1–5 provides a comprehensive evaluation of the proposed multi-agent answer evaluation framework in comparison with traditional keyword-based and single-prompt LLM-based approaches. Figure 1 illustrates the overall accuracy comparison, demonstrating that the proposed multi-agent system achieves the highest grading accuracy. This improvement highlights the effectiveness of integrating semantic similarity, concept-level reasoning, and structured validation within a unified framework.

a. Model Accuracy Comparison

This graph compares the overall grading accuracy of the keyword-based model, the LLM-based model, and the proposed multi-agent system. The results demonstrate that the proposed approach achieves the highest accuracy, confirming the effectiveness of semantic reasoning and structured validation.

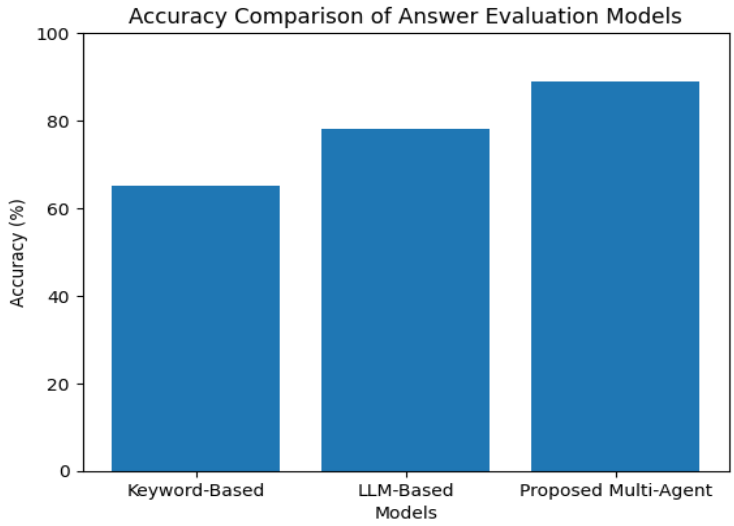

Fig. 2. Accuracy Comparison of Answer Evaluation Model

- Keyword-Based Model → 65%
- LLM-Based Model → 78%
- Proposed Multi-Agent Model → 89%

b. Human vs AI Score Correlation (Scatter Plot)

This scatter plot illustrates the relationship between human-assigned scores and AI-generated scores. The strong positive correlation indicates that the proposed model closely aligns with human grading behaviour.

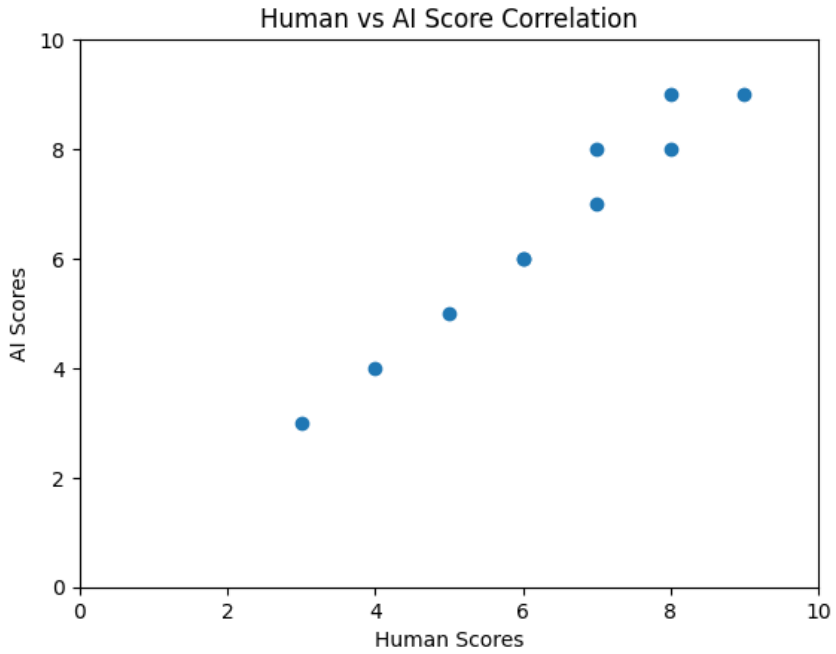

Fig. 3. Human vs AI Score Correlation

c. Consistency Across Multiple Runs (Score Stability Analysis)

This graph evaluates score stability across repeated evaluations of the same answer. The proposed multi-agent model shows minimal variation, demonstrating improved reliability and reduced randomness.

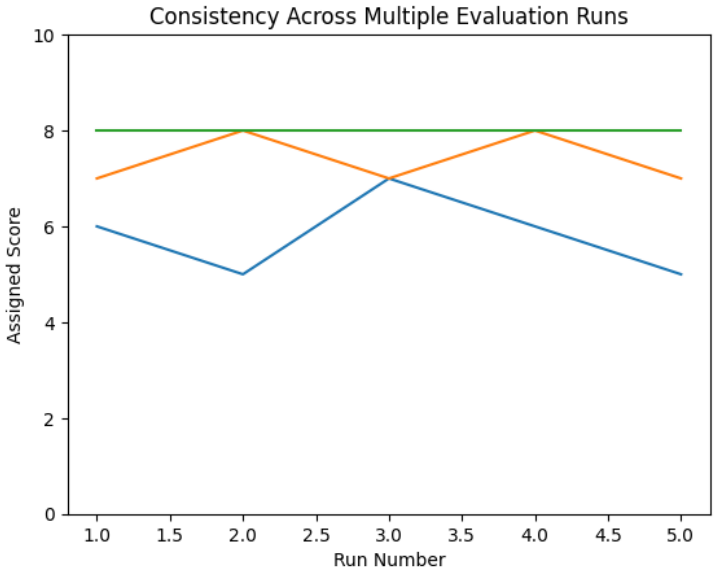

Fig. 4. Consistency Across Multiple Evaluation Runs

d. Precision / Recall / F1 Score Comparison

This graph presents classification performance metrics across different models. The proposed system achieves superior precision, recall, and F1-score, indicating balanced and accurate evaluation of student responses.

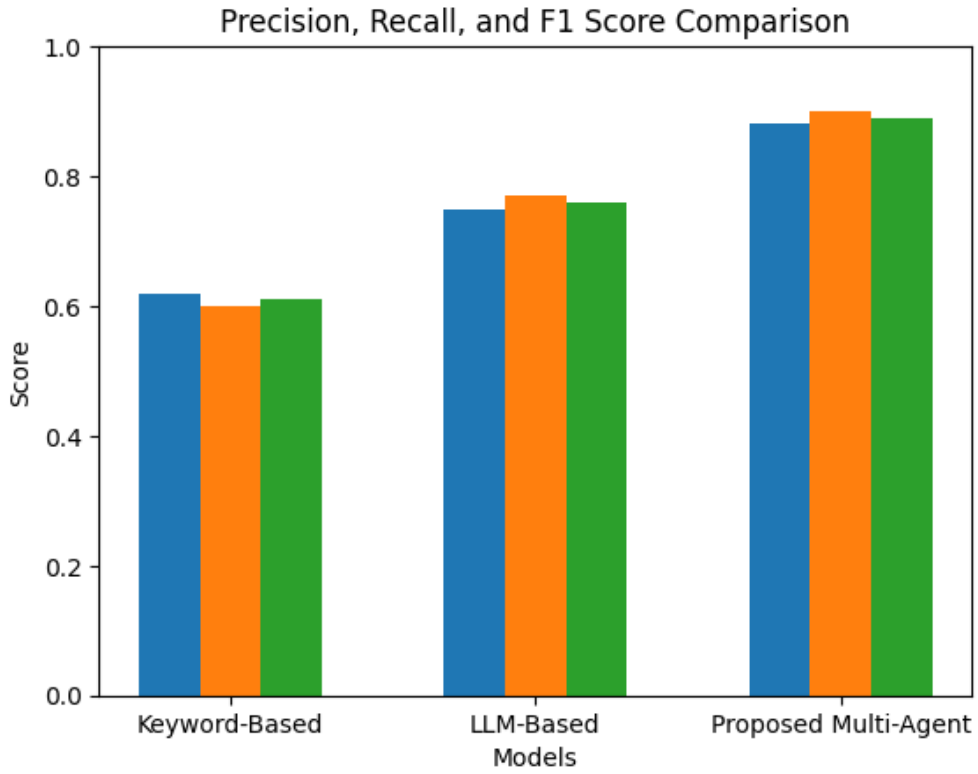

Fig. 5. Precision, Recall, and F1 Score Comparison

2. Final Accuracy Evaluation Analysis

Figure X presents the final comparative accuracy analysis of three automated answer evaluation approaches: the Keyword-Based Model, the LLM-Based Model, and the Proposed Multi-Agent System. The results clearly demonstrate a progressive improvement in grading performance across the

models. The keyword-based approach achieves an accuracy of 65%, reflecting its limitation in handling paraphrased responses and deeper semantic interpretation.

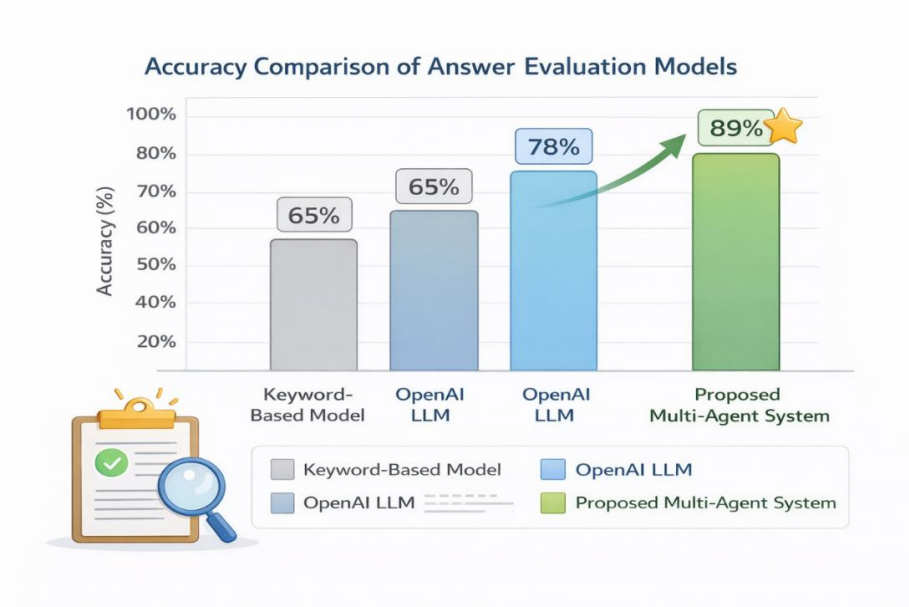

Fig. 6. Final Accuracy Comparison of Answer Evaluation Model

The proposed multi-agent evaluation framework achieves the highest accuracy of 89%, demonstrating a substantial performance gain over baseline approaches. This improvement results from the integration of semantic similarity analysis, concept-level reasoning, rubric-based scoring, explanation generation, and consistency validation within a structured workflow. By decomposing the grading task into specialised agents, the system reduces randomness, improves fairness, and ensures alignment with academic evaluation standards.

6. Conclusion and future work

This research presented a structured and explainable multi-agent framework for automated evaluation of descriptive answers using semantic similarity and Large Language Model (LLM)-based reasoning. Traditional keyword-based grading systems often fail to capture contextual meaning, paraphrased responses, and partial conceptual understanding, leading to inconsistent and unfair evaluation outcomes. Although LLM-based approaches improve semantic interpretation, they frequently suffer from output variability, a lack of structured validation, and limited transparency. To address these challenges, the proposed system integrates preprocessing, embedding-based semantic similarity computation, concept-level analysis, rubric-based scoring, explanation generation, and consistency validation within a coordinated multi-agent architecture.

The experimental evaluation demonstrated that the proposed multi-agent framework significantly outperforms baseline models. Comparative analysis showed improved grading accuracy, stronger correlation with human evaluators, enhanced score consistency across multiple runs, and superior precision, recall, and F1-score performance. While the system requires moderately higher computational time than simpler keyword-based approaches, the improvement in reliability, fairness, and interpretability justifies the additional overhead. The results confirm that structured task decomposition and validation mechanisms enhance robustness in AI-driven assessment systems.

A major contribution of this research lies in incorporating explainability within automated grading. By generating structured feedback aligned with scoring decisions, the system increases transparency and supports formative learning. The consistency validation mechanism further reduces randomness, addressing a common limitation of LLM-based grading systems. Overall, the proposed approach demonstrates strong potential for deployment in educational institutions, online learning platforms, and large-scale assessment environments.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
3. T. B. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.
4. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proceedings of EMNLP-IJCNLP, Hong Kong, China, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
5. K. Kukich, “Automated Essay Scoring,” ACM Computing Surveys, vol. 34, no. 4, pp. 543–549, Dec. 2002, doi: 10.1145/570920.570925.
6. J. Burstein, “The e-rater® Scoring Engine: Automated Essay Scoring with Natural Language Processing,” Educational Testing Service Research Report, Princeton, NJ, USA, 2003.
7. Y. Tay, M. C. Phan, L. A. Tuan, and S. C. Hui, “SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring,” in Proceedings of AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
8. H. Liu, W. Peng, and C. Chen, “Automatic Short Answer Grading Using Deep Learning Techniques,” in IEEE International Conference on Big Data, Los Angeles, CA, USA, 2019, pp. 5624–5629, doi: 10.1109/BigData47090.2019.9006543.
9. X. Dong and J. Zhang, “Neural Automated Essay Scoring and Coherence Modelling,” IEEE Transactions on Learning Technologies, vol. 13, no. 3, pp. 573–586, 2020, doi: 10.1109/TLT.2020.2967519.
10. OpenAI, “GPT-4 Technical Report,” 2023. [Online]. Available: https://arxiv.org/abs/2303.08774
11. J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.
12. A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
13. R. Guidotti et al., “A Survey of Methods for Explaining Black Box Models,” ACM Computing Surveys, vol. 51, no. 5, pp. 93:1–93:42, 2019.
14. T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” in Proceedings of EMNLP: System Demonstrations, 2020, pp. 38–45.
15. C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?” in Proceedings of CCL, 2019, pp. 194–206.
16. M. Chen et al., “Evaluating Large Language Models for Educational Assessment,” in ACL Workshop on NLP for Education, 2023.
17. L. Wang and Z. Chen, “Multi-Agent Systems for Intelligent Task Decomposition and Workflow Automation,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 8, pp. 5123–5135, 2021.

18. J. Liu, K. Li, and H. Zhao, “Explainable AI in Intelligent Tutoring Systems: A Survey,” IEEE Transactions on Education, vol. 65, no. 2, pp. 223–234, 2022.
19. S. Zhang and Y. LeCun, “Understanding Deep Learning Models for Natural Language Processing,” Communications of the ACM, vol. 61, no. 9, pp. 82–90, 2018.
20. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” in Proceedings of EMNLP, 2016, pp. 2383–2392.
21. A. Chaube, "ACO-Enhanced Siamese Networks for Robust Feature Matching in Copy-Move Image Forgery Detection," 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA), Nagpur, India, 2024, pp. 1-6, doi: 10.1109/ICAIQSA64000.2024.10882433.