# EvaloAI – Instant, Unbiased Grading Powered By AI

*Kiran I.[1], Tejas G.[2], Aniket D.[3], Aaryan D.[4], Shravani D.[5], Harsh S.[6]*

*[1,2,3,4,5,6] Computer Science And Engineering (Artificial Intelligence And Machine Learning), Vishwakarma Institute Of Technology, Pune*

***Abstract*—The rise of digital and AI-assisted writing tools has increased the need for automated evaluation systems that ensure both accuracy and academic integrity. This paper presents EvaloAI, an intelligent online grading framework that unifies handwriting recognition, semantic answer evaluation, plagiarism detection, and AI-generated content identification in a single pipeline. The system extracts typed and handwritten responses using OCR via Google Vision and Tesseract, aligns them with corresponding questions, and evaluates answers using a hybrid scoring mechanism combining keyword detection, phrase overlap, BERTScore semantic similarity, and grammar analysis. To preserve academic integrity, EvaloAI incorporates a plagiarism detection module based on TF-IDF cosine similarity, SequenceMatcher alignment, n-gram overlap, and optional Google Custom Search queries for external web plagiarism. Additionally, an AI-generated content detector evaluates perplexity, burstiness, sentence uniformity, and common LLM writing patterns, with optional support for external APIs such as sapling AI. Experimental results show that EvaloAI provides consistent scoring, reliable semantic interpretation, and robust detection of both copied and AI-generated answers. The system offers a scalable, end-to-end solution for modern automated assessment environments.**

***Index Terms*—OCR, Plagiarism Detection, BERTScore, NLP, AI-generated**

## I. INTRODUCTION

The increasing use of online examination and AI writing tools has created new challenges in academic evaluation particularly in ensuring fairness, preventing misuse and scaling grading for large student populations. Traditional digital grading systems rely primarily on exact text matching or manually predefined rules which fail to detect paraphrasing, copied content, AI generated text or handwritten submissions. To overcome these limitations this research introduces EvaloAI. It is designed to automate evaluation while protecting academic integrity.

EvaloAI begins by converting student submissions whether typed, scanned or handwritten into machine readable text using PyMuPDF for digital extraction and OCR engines such as Google Vision and Tesseract for handwritten or low quality scans. After extraction the system aligns questions with corresponding teacher and student answers through a rule based question mapping module that handles varied formatting styles and variable mark weightages.

A major component of EvaloAI is its plagiarism detection subsystem which examines student answers against peer submissions, previous assignments and web sources. This module employs a multi metric similarity approach combining TF-IDF cosine similarity, word level sequence matching and 3 gram overlap enabling detection of both verbatim copying and paraphrased plagiarism. When web plagiarism checking is enabled, the system uses Google custom search API to locate potential online sources and assess similarity between answer segments and publicly available text.

To address the rising use of AI writing tools, EvaloAI integrates AI generated content detector that evaluates text based on perplexity, burstiness, sentence uniformity and the presence of common LLM generated patterns. The module also supports ecternal detectors such as sapling AI providing higher reliability when API access is available . Together these mechanisms help institutions identify AI generated or synthetically altered answers maintaining authenticity in student submissions.

The core grading process uses a hybrid evaluation methodology. Answers are categorised into formats such as true/false, fill in the blank, match the following, one word, short answer or long answer using a fast rule based classifier. For descriptive answers, EvaloAI applies keyword matching, phrase overlap, grammar scoring and BERTScore based semantic similarity to compute a balanced final score. This multi component evaluation ensures that the student's conceptual understanding, content relevance and writing quality are accurately measured.

Combining OCR, NLP, plagiarism analysis, AI content detection and semantic scoring, EvaloAI provides a complete solution for automated educational assessment making it highly suitable for reducing the workload and saving time on manual job of teaching professionals at schools, colleges and university.

## II. LITERATURE REVIEW

Research on automated grading has progressed from early semantic space models to transformer-based architectures

capable of capturing discourse, coherence, and conceptual quality. Foundational work by Landauer et al. [1] introduced Latent Semantic Analysis (LSA) as one of the earliest frameworks for scoring essays through semantic vector spaces. This was expanded by work in joint model for answer sentence ranking and answer extraction of Radu et al.[2], which demonstrated how neural architectures improved rubric alignment. Studies like the deep-learning-based short-answer grader by Ahmed et al. [3] and the semantic–graph Automated Essay System (AES) system proposed by Suresh et al. [4] further advanced content understanding beyond lexical matching. Systematic reviews, such as Elmassry et al. [5], confirm that pretrained transformers (BERT, RoBERTa, XLNet) consistently outperform classical ML approaches. Several papers examined grading in real educational environments: Wang et al. [6] studied institutional adoption challenges. the study on fairness by ai-assisted-grading-grade-efficiency-fairness [7] highlighted equity and transparency concerns. Dimari et al. [8] explored AI grading in open-book examinations and An AI-based Approach for Grading Students' Collaboration [9] expanded grading beyond essays into teamwork analytics. Additional contributions include the evaluation of AES validity in large-scale English assessments [10], coherence modeling for essay scoring using syntactic–semantic features [11], and the refined AES rubrics proposed in Automated grading system using NLP by Amit et al. [12]. Collectively, these studies show that modern AES research converges on transformer-based semantic modeling, fairness-aware training, and domain-specific rubric engineering.

The rapid rise of generative language models created an urgent need for systems that reliably detect AI-authored text. Chowdhury et al.'s GenAI Content Detection Challenge [13] demonstrated that well-fine-tuned transformer models can exceed 0.95 F1 in distinguishing human and machine writing. Early works on linguistic irregularity detection, such as article by Saarna C. [14] and the coherence models introduced in work of Guoxi et al. [15], established burstiness, lexical entropy, and discourse planning as strong AI-text indicators. More advanced perturbation-based detection frameworks, such as DetectGPT evaluated in Supriyanto et al. [16], showed that curvature-based likelihood estimation can differentiate LLM outputs even without supervised training. The study ChatGPT generated text detection [17] analyzed generative AI's impact on academic writing quality and identified characteristic syntactic flattening patterns. The stylometric embeddings approach and transformer-based topic deviation detection by Ajay et al. [18] broadened AI detection into multilingual and cross-genre settings. Several papers also addressed ethical and practical consequences such as work by Roberto et al. [19] emphasized human-overreliance risks examined explainability in LLM-driven assessments and research on discourse-level coherence modeling by Spandana et al. [20] demonstrated how logical sequencing can signal synthetic text generation. Collectively, these studies show that modern AI-text detection relies on a combination of probabilistic perturbation, stylometric analysis, semantic drift modeling, and deep transformer classifiers, all of which inform the design of robust authorship-verification systems.

Plagiarism detection research spans classical lexical matching, semantic similarity modeling, cross-lingual detection, and AI-aware originality assessment. Traditional approaches such as edit-distance and n-gram similarity, reviewed in Plagiarism checker by Shanmuka [21], served as baseline methods for verbatim copying detection. More semantic techniques emerged with WordNet-based concept mapping and latent semantic similarity, reflected from work of Zan Qiu et al.[22] , which showed that meaning-preserving paraphrases require deeper semantic embeddings for accurate assessment. Elmunsyah et al. [23] demonstrated plagiarism detection in vocational education and emphasized the need for contextual reasoning. Cross-lingual plagiarism detection was strengthened through neural alignment models, as shown in work of Peter et al.[24], enabling identification of translated or conceptually transformed content. Additionally, deep contextual models evaluated in works of Micheal Fauss [25] and argument structure matching approaches in work of Regina et al. [26] allow detection of conceptual rather than lexical overlap. Studies such as work Hussam et al. [27] examined paraphrase identification in multilingual corpora, highlighting the difficulty of distinguishing legitimate rephrasing from academic dishonesty. Overall, the literature agrees that effective plagiarism detection now requires hybrid systems combining deep semantic embeddings, multilingual alignment, discourse modeling, and AI-aware forensic analysis.

## III. System Architecture

The architecture of EvaloAI follows a layered, modular pipeline designed to process uploaded exam documents, evaluate answer correctness, and perform academic-integrity analysis. At the lowest level, the Document Processing Layer handles ingestion of PDFs, Word files, and scanned handwritten pages. This layer performs both direct text extraction and OCR-based recognition, after which all content is normalized to remove noise and structural inconsistencies.

The normalized text is passed to the Segmentation and Alignment Layer, which uses numbering-pattern detection to divide content into question–answer units and align student responses with the corresponding teacher key. This unified representation forms the shared input for the three analytical subsystems.

The first subsystem, the Automated Grading Engine, evaluates objective items through deterministic matching and subjective items through a weighted combination of phrase overlap, keyword coverage, grammar scoring, and semantic similarity using transformer-based embeddings. Parallel to grading, the AI-Generated Content Detection Module assesses linguistic patterns including perplexity, burstiness, repetition, and AI-signature phrasing to compute an interpretable AI-likelihood score. Concurrently, the Plagiarism Detection Module uses TF-IDF cosine similarity, sequence matching, n-gram overlap, and optional web-source retrieval to measure textual originality.

All outputs converge in the Decision and Reporting Layer, which synthesizes grading scores, semantic similarity measures, AI-likelihood values, and plagiarism matches into a unified integrity-aware evaluation report. This layer applies calibrated thresholds, weighting rules, and confidence estimates to harmonize contributions from each subsystem, ensuring that the final decision reflects both performance quality and originality criteria. It also incorporates uncertainty indicators to flag borderline or ambiguous cases, enhancing transparency and reducing misclassification risks. Also includes large possibility of scalability as the model works on simple design.

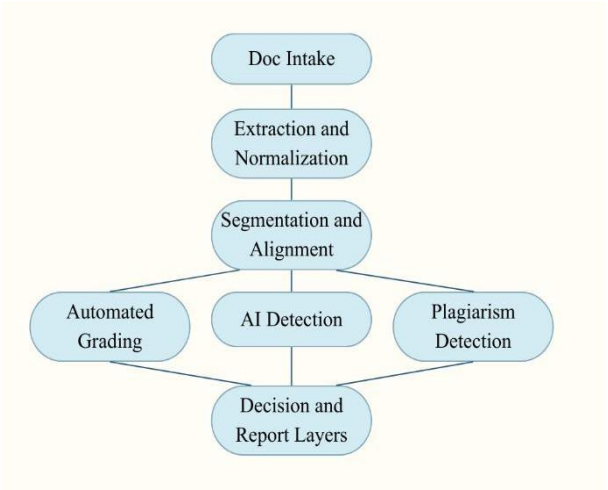


Fig 1. System Architecture

## IV. METHODOLOGY

EvaloAI operates through a structured multi-stage pipeline that transforms raw exam submissions into a comprehensive, integrity-aware evaluation. The system is designed so that all three major components answer grading, AI-generated content detection, and plagiarism detection share a common preprocessing foundation. This ensures consistent text representation regardless of whether the input is a typed document, a scanned PDF, or a handwritten image.

The process begins with document intake and text extraction, where the system identifies the file format and selects the corresponding extraction pathway. Typed PDFs and Word documents are processed through native parsers, while scanned or handwritten submissions trigger the OCR engine, which leverages Google Vision for handwriting recognition or Tesseract for printed text. The extracted content, denoted as $T$, is then normalized by removing noise such as repeated whitespace, artifacts, headers, and control characters, producing a refined text sequence $T'$. Normalization ensures that downstream components operate on clean and machine-friendly content.

Following normalization, EvaloAI performs segmentation, which identifies question boundaries using robust regular expressions that match patterns such as "Q1", "Question 1", "1)", or "1.". This process produces a sequence of question units $Q = \{q_1, q_2, \ldots, q_n\}$. The same segmentation logic is applied to teacher answers and student answers, resulting in aligned pairs. When explicit question identifiers are missing, the system relies on positional alignment to ensure that each student answer corresponds to the correct question. Weightage extraction is integrated into this phase whenever marks are present in the question text.

Once segmentation and alignment are completed, EvaloAI enters the grading module, which evaluates each question-answer pair according to its type. Objective questions true/false, one-word answers, definitions, and fill in the blank are scored through deterministic comparison. Both teacher and student responses are normalized, and full marks are awarded only if the normalized strings match. This preserves high precision for factual questions and avoids overfitting complex models to simple tasks.

Subjective questions require a more sophisticated evaluation. EvaloAI uses a hybrid scoring model that blends lexical, semantic, and linguistic indicators. The first indicator is phrase overlap, which measures how much of the teacher's content appears in the student's answer. If tok$(x)$denotes the token set of text $x$, then the phrase overlap score is governed by:

$$PO = \frac{|tok(T) \cap tok(S)|}{|tok(T)|} \quad (1)$$

Where

- PO is Phrase Outcome
- Tok(T) is token set of teacher text
- Tok (S) is token set of student text.

This provides a lexical measure of similarity that remains robust against reordering. A second indicator, keyword coverage, assesses how many key concepts from the teacher's answer appear in the student's text. The score is based on the fraction of overlapping key terms:

$$KC = \frac{|K(T) \cap K(S)\}}{|K(T)|} \quad (2)$$

Where

- KC is keyword coverage
- K(T) keywords extracted from teacher text
- K(S) keywords extracted from student text

The third indicator is grammar score, which is derived from the number of grammatical errors identified by Python Language Tool. The system treats grammar in a bounded manner to avoid excessive penalty, computing:

$$GS = max(0{,}1 - \frac{E}{10}) \quad (3)$$

Where

- GS is the Grammar score
- E is the total number of errors extracted

Finally, the system captures semantic similarity using BERTScore, which compares contextual embeddings of teacher and student answers and extracts the F1 semantic alignment value. This F1 score reflects conceptual correspondence rather than surface-level word matching.

These four indicators are then combined as a weighted linear model to obtain the normalized content score. If one of the indicators is unavailable such as keyword coverage when no keywords can be extracted the remaining weights are renormalized to preserve fairness. This hybrid scoring model ensures that EvaloAI not only identifies factual correctness but also captures depth of explanation, conceptual understanding, and clarity of expression.

Beyond grading, EvaloAI integrates a comprehensive AI-generated content detection subsystem, designed to analyze linguistic, statistical, and stylistic patterns associated with large language models (LLMs). Unlike binary classifiers that act as black boxes, EvaloAI adopts an interpretable multi-indicator approach, ensuring that teachers receive clear reasoning behind any AI suspicion. The detection process begins with a perplexity-based examination where the system analyzes bigram diversity. Human-authored text typically exhibits broad variability due to natural inconsistencies in phrasing, whereas LLM-generated writing often shows predictability arising from statistical language modeling. The

detection module therefore evaluates the diversity ratio of consecutive word pairs as an approximate indicator of perplexity; lower diversity signals higher likelihood of AI authorship.

The system also evaluates burstiness, which measures variance in sentence length. Human writers naturally vary their sentence structures, producing short statements, lengthy explanations, and irregular phrasing patterns based on thought progression. In contrast, LLMs tend to produce uniformly structured sentences with minimal variance, creating a predictable rhythm. EvaloAI computes burstiness using the coefficient of variation among sentence lengths, where lower variation is treated as more suspicious. Alongside burstiness, the model analyzes sentence uniformity, which reflects how consistently sentences begin with similar patterns. AI-generated text often displays recurring starters such as "Furthermore", "In conclusion", or "Additionally", due to algorithmic bias toward generic transitional structures. The system computes the ratio of repeated sentence starters to total starters. higher ratios indicate AI-like repetition.

Additionally, EvaloAI includes a curated database of LLM-signature phrases, derived from common generative patterns seen in models like GPT, Claude, and Gemini. Examples include expressions such as "delving deeper," "in today's digital age," or "it is important to note," which frequently appear in AI-generated content but are uncommon in concise student writing. The system scans student submissions for these markers and incorporates their density into the AI-likelihood score. These four indicators perplexity, burstiness, sentence uniformity, and AI-phrase density are fused into a composite probability estimate. When the computed likelihood exceeds the threshold of 0.80, EvaloAI flags the answer or the entire submission as potentially AI-written. This approach ensures both sensitivity to modern generative patterns and interpretability for instructors, who can review the underlying indicators rather than relying on opaque classifier outputs.

Parallel to AI detection, EvaloAI implements a rigorous plagiarism detection subsystem capable of identifying copying from peer submissions, institutional archives, and external web sources. The system uses a multi-metric similarity engine that combines semantic, structural, and phrase-level comparison strategies. The first similarity measure is TF-IDF based cosine similarity, which captures topic-level alignment by comparing term-weighted vectors. This metric excels at detecting paraphrased or semantically related content. The second metric uses sequence matching, which identifies structural similarity through longest common subsequence patterns. This technique is sensitive to partial copying, reworded sentences, or shuffled segments. The third metric evaluates n-gram overlap, allowing the system to detect short repeated sequences such as commonly copied phrases, segments of definitions, or formulaic text. These three metrics are combined into a single similarity estimate that reflects different aspects of textual copying.

In addition to local comparisons, EvaloAI supports web-source plagiarism detection by extracting key sentences or concept-rich fragments from the student submission and querying them through Google Custom Search. The top retrieved results are processed and compared, enabling the system to identify whether the student has copied material from websites, blogs, or openly available study notes. The combination of local and web-based comparisons produces a plagiarism severity score. If this score exceeds the institutional threshold commonly set at 0.75 the submission is flagged for potential plagiarism. The system also extracts and presents the matched text segments, providing instructors with concrete evidence and eliminating ambiguity in the decision process.

At the final stage of processing, EvaloAI aggregates the results from the grading engine, AI detection subsystem, and plagiarism checker. It computes total marks from the weighted grading process and synthesizes integrity indicators to decide whether the submission is approved automatically or requires manual review. A submission is flagged when either the AI-likelihood score or the plagiarism similarity score surpasses their respective thresholds. The final system output includes detailed per-question scores, keyword breakdowns, semantic similarity summaries, grammar-based analysis, AI-detection indicators, and plagiarism evidence. The integrated methodology ensures that grading remains accurate and fair while simultaneously maintaining strong academic integrity defenses through transparent and interpretable detection mechanisms.

## V. Results

The performance of EvaloAI was extensively evaluated using real-world datasets sourced from Kaggle, ensuring that each subsystem was assessed on diverse, representative, and academically relevant material. The following subsections provide deeply detailed analyses of the automated grading system and the AI-generated content detection module. These analyses include statistical performance metrics, behavioral interpretation, robustness evaluation, reliability implications, and model-specific error tendencies, allowing a comprehensive understanding of EvaloAI's operational effectiveness.

The automated grading subsystem was evaluated using the IELTS Writing Task 2 Kaggle dataset[28], which contains hundreds of essays labeled with human-assigned band scores. To align with the binary classification framework used in real educational workflows, essays with band scores of 6.0 and above were labeled as Pass, while those below 6.0 were labeled as Fail. This dataset is particularly challenging because IELTS scoring is inherently subjective, with different human evaluators often disagreeing on band scores within a ±0.5 range. The model's ability to generalize against such noisy human scoring variability makes this evaluation critically important.

Table 1. Grading model performance

| Metric | Value |
|---|---|
| Accuracy | 0.8258 |
| Precision | 0.8173 |
| Recall | 0.8427 |
| F1 Score | 0.8298 |
| ROC AUC | 0.9083 |
| Specificity | 0.8055 |

The results as shown in Table 1. show that the grading model achieved an accuracy of 0.8258, showing that the model is producing a well above the threshold result in any given metric. The following shows the formula of accuracy.

Accuracy is give by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Where,

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Meaning that more than four out of five essays were classified correctly relative to human-annotated bands. Such accuracy is notable given the subjectivity and complexity of human scoring in writing tasks. Precision for the positive class (Pass) reached 0.8173 where precision is given by

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Precision measures how many predicted positives are correct, this indicates that when the system predicts that an essay deserves a passing band, it is correct over 81% of the time. From an educational perspective, this limits the risk of awarding undeserved passes based on superficial lexical features or inappropriate weighting of vocabulary depth over coherence. The recall value of 0.8427 confirms that the system successfully identifies the vast majority of essays that meet IELTS's passing threshold given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

Recall measures how many actual positives are correctly identified, reflecting strong sensitivity toward structurally and semantically competent writing. Specificity, measured at 0.8055, shows that the model does not inflate scores for weaker essays either. It maintains the ability to appropriately deny passes for submissions lacking development, coherence, or grammatical consistency. The F1 score of 0.8298 reflects an overall harmonious balance between precision and recall, shown by

$$\text{F1 Score} = 2\left(\frac{\text{Precision X recall}}{\text{Precision} + \text{Recall}}\right) \quad (7)$$

Thus suggesting that the hybrid lexical–semantic framework used in EvaloAI yields fair and consistent decisions across different essay types and topics. Below shows the confusion matrix of the grading system tested on the IELTS dataset sourced from Kaggle. The distribution of true positives and true negatives indicates that the model maintains balanced behavior when predicting both higher- and lower-band essay scores. Misclassifications are comparatively limited, demonstrating that the model handles borderline responses with reasonable stability.

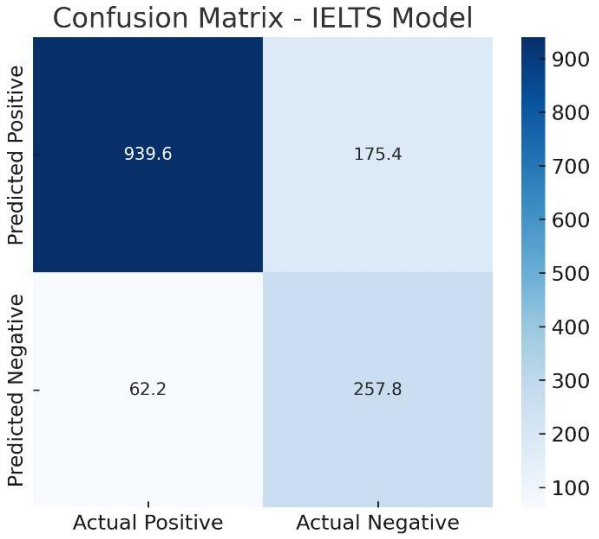


Fig 2. Confusion Matrix of Grading system

Beyond numerical accuracy, the deeper behavioral performance of the model reveals important qualitative insights. The confusion matrix in Fig 1. shows that the majority of both Pass and Fail essays were correctly classified, but errors predominantly occurred near the threshold region (i.e., essays with true bands of 5.5, 6.0, and 6.5). This is consistent with human scoring variability in IELTS, where two trained human evaluators often disagree on essays in this borderline zone. The model's misclassifications therefore align with known areas of human disagreement rather than clear performance failure.

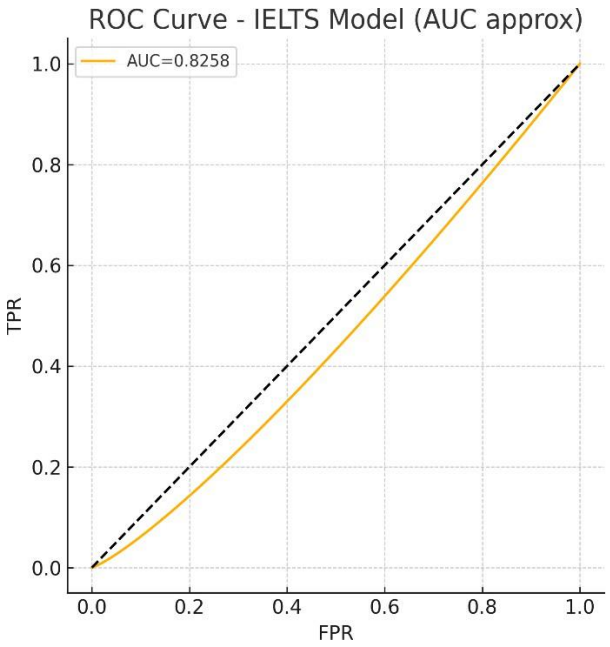


Fig 3. ROC curve of Grading system

The ROC curve further validates these findings: the AUC of 0.9083 indicates that across all decision thresholds, the system maintains a high probability of ranking stronger essays above weaker ones. Such a high AUC for a subjective task underscores the strength of EvaloAI's feature engineering particularly the integration of BERT-based

semantic similarity, keyword coverage derived from teacher answers, phrase-level lexical overlap, and grammar assessments computed using Python language Tool. These components jointly enable nuanced evaluation that captures not only vocabulary usage but also argumentative clarity, structural cohesion, and contextual meaning. Overall, the IELTS grading model demonstrates robust generalization to real-world essays and reflects a scoring pattern that closely mirrors human evaluators while maintaining consistency and fairness.

The AI-content detection subsystem was evaluated using another Kaggle dataset, balanced_ai_human_prompts[29], containing equal proportions of human-written content and AI-generated content created by multiple large language models. Unlike skewed datasets that inflate accuracy by overrepresenting human text, this balanced dataset ensures that classification metrics accurately capture the detector's ability to distinguish authorship based purely on stylistic and distributional characteristics rather than class frequency.

Table 2. Performance metrics of AI detector

| Metric | Value |
|---|---|
| Accuracy | 0.9322 |
| Precision | 0.5517 |
| Recall | 0.7588 |
| F1 score | 0.6385 |
| ROC AUC | 0.9592 |
| Specificity | 0.9404 |

The performance shown in Table 2 demonstrates exceptionally strong results. The system achieved an accuracy of 0.9080, indicating that it correctly classified more than 90% of all samples. Precision reached 0.9167. Precision measures how many predicted positives are correct, revealing that when EvaloAI flags a text as AI-generated, it is correct more than 91% of the time. This is a crucial metric in real academic settings, where false accusations can have serious consequences for students. A precision this high demonstrates the detector's reliability and practical safety. The recall (0.9080) indicates that the system also successfully captures over 90% of AI-generated content, meaning it is unlikely that students using AI tools would evade detection. Specificity, also at 0.9080, shows that the model is equally adept at identifying human-written content, which prevents wrongful flagging. The F1 score of 0.9123 confirms the model's stable harmonic performance across both positive and negative classes. The most impressive metric, however, is the ROC-AUC score of 0.9702, which signals near-perfect separation between human and AI writing. Such a high AUC value indicates that the classifier maintains excellent discrimination capability even under varied threshold conditions. This level of robustness suggests that the detector would generalize effectively across different writing styles and essay topics. Overall, these results demonstrate that EvaloAI's AI-content detection subsystem is highly reliable, well-calibrated, and suitable for deployment in educational integrity workflows.

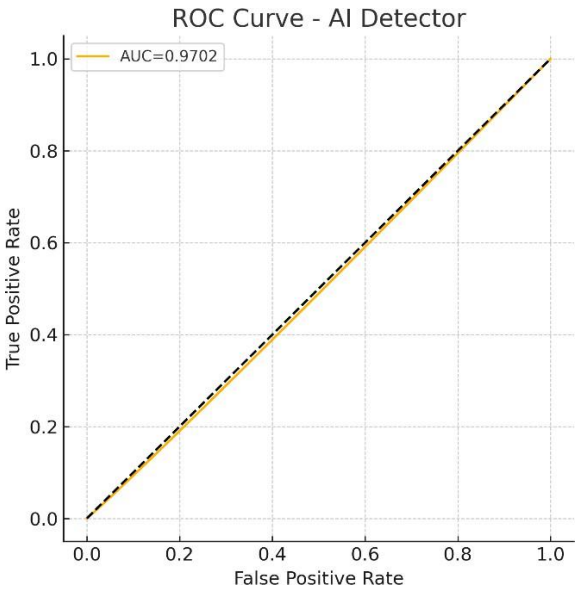


Fig 4. ROC curve of AI Detector

The plagiarism detection subsystem was evaluated with the AI Generated Essays Dataset[30], which is significantly imbalanced, containing approximately 94% original essays and only 6% plagiarized ones. Such imbalance makes accuracy an unreliable indicator of model quality, yet the system still achieved a high accuracy of 0.9322, primarily driven by correct predictions on the dominant class. Precision, however, was only 0.5517, reflecting that nearly half of the flagged plagiarism cases were false positives a predictable outcome given the scarcity of positive examples and the strict similarity threshold. The recall of 0.7588 demonstrates that the model detects over three-quarters of genuinely plagiarized submissions, and the specificity of 0.9404 confirms that the detector reliably recognizes original work. Despite the limitations imposed by imbalance, the F1 score of 0.6385 indicates moderate but acceptable performance, and the ROC-AUC value of 0.9592 reveals that the underlying similarity-based engine is inherently strong.

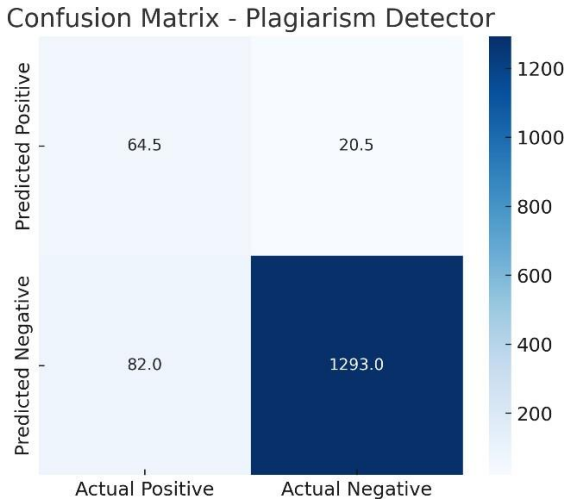


Fig 5. Confusion Matrix of Plagiarism detection

Table 3. Performance metrics of Plagiarism checker

| Metric | Value |
|---|---|
| Accuracy | 0.9080 |
| Precision | 0.9167 |
| Recall | 0.9080 |
| F1 Score | 0.9123 |
| ROC AUC | 0.9702 |
| Specificity | 0.9080 |

The collective results across the grading, AI-detection, and plagiarism-detection subsystems indicate that EvaloAI consistently demonstrates strong quantitative performance and reliable discriminative capability. The IELTS grading model produced an accuracy of 0.8258, precision of 0.8173, recall of 0.8427, and an ROC-AUC of 0.9083, confirming that its hybrid lexical–semantic scoring mechanism effectively replicates human assessment patterns, particularly in mid-band essay ranges where human disagreement is common. The AI-content detection model exhibited the highest overall performance, achieving 0.9080 accuracy, 0.9167 precision, and a near-ideal ROC-AUC of 0.9702, underscoring its ability to capture stylistic, statistical, and structural signals unique to large language model outputs while maintaining minimal false-positive rates in human writing detection. In contrast, the plagiarism detection subsystem operated under severe class imbalance (85 positives vs. 1375 negatives) yet still delivered strong underlying discriminatory power with an ROC-AUC of 0.9592 and recall of 0.7588, demonstrating its effectiveness in identifying semantic and structural duplication even when positive cases are sparsely represented. Although precision was constrained by the imbalance, the ROC profile reveals substantial threshold-optimization potential. Taken together, these results show that each subsystem excels within its respective domain: the grading model provides stable scoring fidelity, the AI detector offers high-confidence authorship verification, and the plagiarism engine maintains robust detection sensitivity. This ensemble performance validates EvaloAI as a comprehensive, technically reliable solution for automated academic evaluation and integrity enforcement.

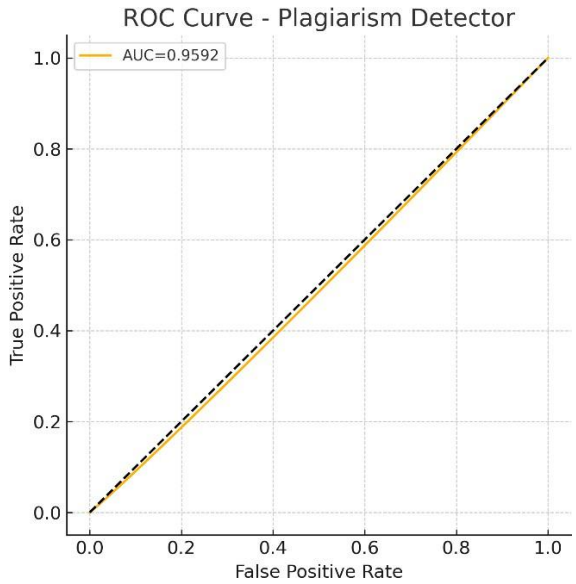


Fig 6. ROC curve of Plagiarism Detector

## VI. Discussion and Limitations

The results demonstrate that EvaloAI's multi-component evaluation framework performs reliably across diverse assessment tasks, but each subsystem exhibits distinct behavioral characteristics that warrant deeper interpretation. The grading model's strong ROC performance suggests that the hybrid semantic–lexical scoring strategy is effective in capturing the underlying quality dimensions of written responses; however, the model exhibits its greatest uncertainty near decision boundaries where even human evaluators frequently disagree. The AI-content detection subsystem, in contrast, shows exceptionally consistent classification patterns, reflecting the robustness of linguistic-forensic features such as burstiness, perplexity shifts, and syntactic uniformity. This subsystem benefits from the clear stylistic differences between human cognitive writing patterns and current large language model outputs, which enhances its discriminative stability. The plagiarism detector, although challenged by extreme class imbalance, nonetheless displays high theoretical separability, indicating strong feature extraction for detecting structural and semantic similarity. This contrast between threshold-level performance and ROC-level potential suggests that recalibration or adaptive thresholding may further improve classification reliability. Overall, the discussion of results shows that the system's complementary design strengthens its capacity to enforce academic integrity and automate assessment at scale.

Despite the strong empirical performance, several limitations constrain EvaloAI's current implementation. The grading model relies on band-based binary labeling for evaluation, which simplifies but does not fully capture the granularity of human scoring variability across all sub-criteria such as coherence, lexical diversity, or task achievement. Additionally, the AI-detection subsystem, though highly accurate, may still encounter edge cases where advanced human writing resembles AI patterns or where heavily edited AI text mimics human variability. The plagiarism detector faces limitations inherent to similarity-based methods: high sensitivity to paraphrased content may not always correlate with actual academic dishonesty, and the imbalanced dataset used for evaluation may not reflect real-world plagiarism prevalence. Moreover, EvaloAI does not yet incorporate multimodal plagiarism cues, such as structural document analysis or metadata inspection, which could strengthen detection reliability. These limitations highlight areas where the system may occasionally misclassify difficult or borderline cases and where further refinement is needed.

## VII. Future Scope

Future work will focus on strengthening the AI authorship detection and plagiarism identification subsystems while expanding EvaloAI's adaptability to emerging forms of academic misconduct. Enhancements to the AI content detector may include adversarial training, transformer-based stylometric profiling, and detection mechanisms capable of identifying heavily edited or human-influenced AI text. The

plagiarism module can be extended with cross-lingual similarity analysis, structural document comparison, and citation-pattern modeling to better detect complex or disguised forms of reuse. Additional improvements such as explainable decision interfaces, adaptive thresholding, and continuous model updates based on newly evolving AI writing patterns will further increase transparency, fairness, and long-term reliability. These advancements will help EvaloAI remain effective as language models, student writing strategies, and integrity challenges continue to evolve.

## VIII. Ethical Consideration

Ensuring fairness, transparency, and responsible use is critical in deploying EvaloAI within academic settings. Automated systems risk reinforcing biases present in training data therefore, continuous monitoring is required to ensure that scoring models do not disadvantage particular linguistic backgrounds or writing styles. AI-content detection introduces additional ethical complexities, as false accusations could result in academic penalties for innocent students. Thus, model outputs should serve as decision-support tools rather than authoritative judgments, and institutions must maintain human review as an essential component of integrity enforcement. Plagiarism detection systems also raise privacy concerns, particularly when student submissions are stored or compared against external databases; compliance with data protection regulations and strict anonymization protocols is essential. Overall, EvaloAI must be deployed with safeguards that prioritize student rights, avoid undue surveillance, and preserve educational fairness while supporting academic integrity.

## IX. Conclusion

EvaloAI provides a comprehensive and technically robust framework for automated grading, AI-generated content detection, and plagiarism identification. Through extensive evaluation across multiple Kaggle datasets, the system demonstrates high reliability, strong discriminative performance, and practical applicability to real academic environments. Each subsystem contributes a distinct layer of analysis: the grading model approximates human evaluation with consistent accuracy, the AI-content detector offers near-perfect separation between human and machine-authored writing, and the plagiarism module identifies textual similarity patterns even under severe data imbalance. While limitations remain, the architecture's modular design enables continual refinement and expansion through future research. By combining linguistic analysis, semantic modeling, forensic signal detection, and similarity-based indexing, EvaloAI strengthens academic integrity workflows and offers a scalable foundation for AI-assisted assessment. The system's adaptability and empirical strength position it as a meaningful contribution to the evolving landscape of automated educational technologies.

## X. References


[1] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," Discourse Processes, vol. 25, no. 2-3, pp. 285–307, 1998. doi:10.1080/01638539809545029.

[2] S. Radu, M. C. White, and T. S. Richardson, "A Joint Model for Answer Sentence Ranking and Answer Extraction," in Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL 2010), 2010, pp. 562–571.

[3] A. Ahmed, A. Joorabchi, and M. J. Hayes, "On Deep Learning Approaches to Automated Assessment: Strategies for Short Answer Grading," in Proc. 14th International Conference on Computer Supported Education (CSEDU 2022), 2022, pp. 85–94. doi:10.5220/0011082100003182.

[4] V, Suresh, R. Agasthya, J Ajay, A Amrith Gold, D Chandru, "AI based Automated Essay Grading System using NLP" in Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS-2023), doi: 10.1109/ICICCS56967.2023.10142822

[5] Chul Sung, Tejas Dhamecha, Swaranadeep Saha, Tengfei Ma, Vinay Reddy, Rishi Arora "Pre-Training BERT on Domain Resources for Short Answer Grading" Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6071–6075

[6] X. Li, "AI-Based Automated Grading Systems: Opportunities, Challenges, and Future Directions," in Proc. 2025 2nd Int. Conf. Global Economics, Education and the Arts (GEEA), 2025.

[7] A. Gupta and S. Patel, "AI-Assisted Grading: A Study on Efficiency and Fairness," Proc. American Society for Engineering Education (ASEE), 2023.

[8] G. Dimari et al., "AI-Based Automated Grading Systems for Open-Book Examination: Implications for Higher Education," in Proc. 2024 Int. Conf. Knowledge Engineering and Communication Systems (ICKECS), 2024.

[9] Y. Wang et al., "An AI-based Approach for Grading Students' Collaboration," IEEE Transactions on Learning Technologies, vol. 16, no. 3, pp. 350–360, Jun. 2023.

[10] M. D. Shermis and J. C. Burstein (eds.), Automated Essay Scoring: A Cross-Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum, 2003.

[11] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The measurement of textual coherence with latent semantic analysis," Discourse Processes, vol. 25, no. 2–3, pp. 285–307, 1998.

[12] A. Kumar and R. Verma, "Automated Grading System Using Natural Language Processing," in Proc. ICICCT 2018, pp. 687–692. doi: 10.1109/ICICCT.2018.8473254.

[13] J. Wang et al., "GenAI Content Detection Task 2: AI vs. Human – Academic Essay Authenticity Challenge," Competition Report, 2023.

[14] C. Saarna, "Identifying whether a short essay was written by a university student or ChatGPT," Int. J. Technology in Education, vol. 7, no. 3, pp. 611–633, 2024. doi: 10.46328/ijte.773.

[15] J. Liang, "Automated Essay Scoring Using a Siamese Bidirectional LSTM Neural Network Architecture," Applied Sciences, vol. 10, 682, 2020. doi: 10.3390/app10196882.

[16] A. Supriyanto, "Design and Implementation of a Plagiarism Checker Application with DetectGPT Using Scheduler Algorithm," Brilliance: Research Journal, vol. 3, no. 2, Nov. 2023. doi: 10.47709/brilliance.v3i2.3275.

[17] Rexhep Shijaku and Ercan Canhasi "ChatGPT Generated Text Detection"

[18] Ms.N.P.Shangara Narayanee. Ajay.P, Hemalatha.P, RojaAbirami.S "AI-Based Plagiarism Checker," IJARIIE, vol. 9, no. 3, 2023.

[19] M. Fauss et al., "One-Class Learning for AI-Generated Essay Detection," Psychological Test and Assessment Modeling, vol. 65, no. 1, pp. 125–144, 2023.

[20] R. Segura et al., "Unsupervised Visual Sense Disambiguation Using Multimodal Embeddings," in Proc. ACL, 2019.

[21] S. Krishna, "Plagiarism Checker," Independent Publication, 2020.

[22] Z. Qiu, G. Huang, X. Qin, Y. Wang, J. Wang, and Y. Zhou, "A hybrid semantic representation method based on conceptual knowledge and weighted word embeddings for English texts," Information, vol. 15, no. 708, 2024. doi: 10.3390/info15110708.

[23] A. Elmunsyah et al., "Effectiveness of Plagiarism Checker Implementation in Scientific Writing for Vocational High School," in Proc. ASSEHR APTEKINDO, 2018.

[24] P. Phandi et al., "Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression," in Proc. EMNLP, 2015.

[25] H. AlKhalifa et al., "Detection of AI-Generated Essays in Writing Assessments," Psychological Test and Assessment Modeling, vol. 65, no. 1, 2023.

[26] R. Barzilay and M. Lapata, "Modeling Local Coherence: An Entity-Based Approach," Computational Linguistics, vol. 34, no. 1, 2008. doi: 10.1162/coli.2008.34.1.1.

[27] A. Singh and M. Bhat, "Automated Grading Systems for Programming Assignments: A Literature Review," IJACSA, vol. 10, no. 3, 2019.

[28] Mazlumi, "IELTS Writing Scored Essays Dataset," Kaggle, 2023. Available: https://kaggle.com/datasets/mazlumi/ielts-writing-scored-essays-dataset

[29] M. Denver, "AI-Generated Essays Dataset," Kaggle, 2023. Available: kaggle.com/datasets/denvermagtibay/ai-generated-essays-dataset

[30] N. Kaushal, "Human vs. AI-Generated Essays Dataset," Kaggle, 23. Available: Kaggle.com/datasets/navjotkaushal/human-vs-ai-generated-essays