Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 4 | April-2026
ISSN (O) 3107-6696

AssignMatch: Intelligent Matching of Student Submissions

[1] Aman Kumar Verma, [2]Kumail Mujtaba, [3]Asim Kaif, [4]Syed Tabish Sajjad

[1,2,3,4,] *Dept. Of CSE/Cloud Computing & AI, Integral University, Lucknow ,India*

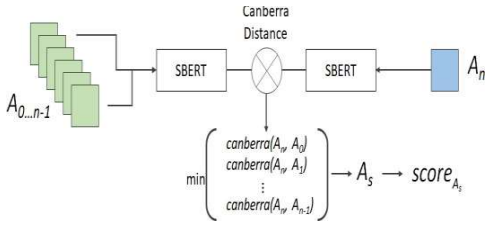
Abstract –Recent advances in digital learning environments and AI-assisted content generation have significantly changed how academic assignments are created and evaluated. Conventional manual grading approaches are increasingly difficult to scale, often requiring substantial instructor effort while remaining vulnerable to inconsistency and undetected semantic copying.

This paper introduces **AssignMatch**, an intelligent assignment evaluation framework designed to automate assessment through the integration of Optical Character Recognition (OCR), Natural Language Processing (NLP), large language models (LLMs), and semantic similarity analysis. The system accepts submissions in multiple formats, including scanned documents and images, extracts textual content using OCR, and performs structured preprocessing before evaluating responses against reference solutions using meaning-aware comparison techniques.

Beyond automated scoring, AssignMatch incorporates embedding- based plagiarism detection and cross- document similarity analysis to identify paraphrased or highly similar submissions. Experimental observations indicate that the proposed system substantially reduces grading time while maintaining strong agreement with instructor evaluation. The framework provides a scalable and adaptable solution for institutions seeking efficient, transparent, and AI-supported academic assessment.

I. Introduction

Assignment-based evaluation remains a central component of modern education, enabling instructors to assess conceptual understanding, analytical reasoning, and written communication skills. However, increasing class sizes and the widespread adoption of digital submission platforms have made manual grading progressively more demanding. Traditional evaluation workflows often require extensive instructor time, introduce subjective variations in scoring, and provide limited mechanisms for detecting intelligently paraphrased or AI-generated responses.



Recent accessibility of generative AI tools has further complicated academic assessment by enabling students to produce well-structured answers with minimal effort. As a result, institutions require evaluation systems capable of analyzing meaning rather than relying solely on keyword matching or superficial textual comparison.

Conventional grading approaches face several practical limitations:

- Significant evaluation time per submission
- Variability in scoring consistency among graders
- Limited capability to identify semantic plagiarism
- Difficulty processing handwritten or scanned assignments automatically

To address these challenges, this research proposes **AssignMatch**, an AI- driven framework that automates assignment evaluation using a multi- stage processing pipeline combining OCR, NLP-based preprocessing, semantic analysis, and similarity detection.

The primary objectives of this work are to develop a scalable system capable of:

- Extracting text from heterogeneous submission formats
- Performing meaning-based answer evaluation
- Detecting near-duplicate or plagiarized responses
- Generating structured feedback and grading summaries
- Supporting instructor oversight through human review mechanisms

By integrating automated intelligence with human supervision, the proposed system aims to improve grading efficiency while maintaining academic integrity and fairness.

1. Related Work

1.1 Automated Essay and Short-Answer Scoring

Research in automated assessment has evolved considerably over the past two decades, moving from objective testing toward evaluation of descriptive and analytical responses. Early automated grading systems primarily focused on structured question formats due to their predictable scoring criteria. However, growing demand for scalable evaluation of essays and short answers encouraged the development of computational approaches capable of assessing open- ended responses.

Initial Automated Essay Scoring (AES) methods relied on manually engineered linguistic indicators such as vocabulary usage, sentence length, grammatical patterns, and keyword frequency. Machine learning models, including regression algorithms and support vector machines, mapped these handcrafted features to predicted scores. Although these approaches demonstrated moderate agreement with human graders, they were limited in capturing deeper contextual meaning and reasoning quality.

The introduction of distributed word representations, including Word2Vec and GloVe, improved semantic comparison by representing words within continuous vector spaces. Subsequent transformer-based architectures marked a significant advancement by enabling contextual understanding through attention mechanisms. Models such as BERT and Longformer allow systems to evaluate responses based on semantic intent rather than surface similarity.

Recent studies applying transformer embeddings to academic datasets have reported strong correlations between automated and human grading outcomes. Benchmark datasets such as SemEval-2013 Beetle and SciEntsBank have played an important role in standardizing evaluation methodologies for short-answer assessment research. Despite these advances, challenges remain in ensuring fairness, rubric alignment, and robustness across diverse writing styles.

1.2 LLM-Based Grading Approaches

The emergence of large language models has introduced new possibilities for automated academic evaluation. Unlike earlier machine learning approaches that depended on predefined features, LLMs can interpret context, reasoning structure, and explanatory depth through large-scale pretraining.

Recent experimental studies have explored the use of models such as GPT-4 and open-source alternatives for grading short-answer assessments. When guided using structured prompts and clearly defined rubrics, these systems demonstrate strong agreement with human evaluators in both scoring accuracy and feedback generation quality. Importantly, research indicates that locally deployable open-source models can achieve competitive performance, enabling privacy- preserving academic applications.

Evaluation of LLM-based graders commonly relies on statistical agreement measures, including Pearson correlation, Spearman rank correlation, exact agreement rate, and mean absolute error. High-performing systems have reported strong alignment with instructor grading under controlled experimental settings.

Despite their advantages, LLM-based grading systems introduce challenges such as prompt sensitivity, variability across repeated executions, hallucination risks, and computational cost. To mitigate these limitations, modern approaches frequently incorporate retrieval grounding or rubric-based conditioning to stabilize evaluation behavior and improve reliability.

1.3 Handwritten and Image-Based Response Evaluation

In practical academic environments, student submissions often include handwritten answers, diagrams, or mathematical derivations, creating additional complexity for automated evaluation systems. To address this challenge, recent research combines Optical Character Recognition with computer vision techniques to transform visual documents into analyzable text.

OCR pipelines typically include preprocessing operations such as noise filtering, contrast enhancement, binarization, and segmentation to improve recognition accuracy. Once extracted, textual content can be processed using standard NLP workflows for semantic analysis and scoring.

Recent investigations have explored vision-language models and meta- learning techniques for evaluating handwritten mathematical responses and graphical solutions. Findings suggest that specialized architectures may outperform general-purpose models for structured visual tasks, while multimodal systems show promise for complex reasoning scenarios.

Although OCR technology has improved significantly, limitations remain when handling cursive handwriting, mathematical notation, or diagram interpretation. Consequently, integrating OCR outputs with semantic grading pipelines continues to be an active area of research.

1.4 Plagiarism and Academic Integrity Detection

Ensuring originality in student submissions is a critical requirement for automated grading systems. Early plagiarism detection approaches relied primarily on lexical comparison techniques such as n-gram overlap, TF- IDF similarity, and string matching algorithms. While effective for detecting direct copying, these methods often fail to identify semantically paraphrased content.

Recent advances emphasize embedding- based semantic similarity methods that represent sentences within high-dimensional vector spaces. Transformer- derived embeddings enable detection of conceptual similarity even when wording differs significantly. These approaches improve robustness against paraphrasing and AI-generated text variations.

Modern plagiarism detection frameworks increasingly combine lexical analysis, semantic embeddings, and cross-document comparison strategies. Integrated detection within grading pipelines allows simultaneous evaluation of answer quality and originality, improving academic integrity monitoring while reducing instructor workload.

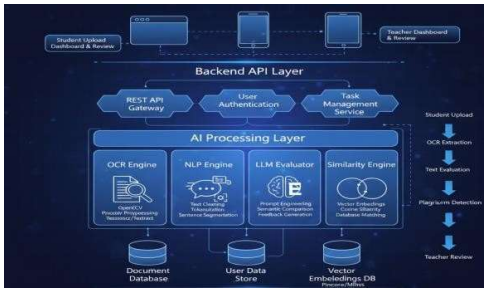
2. Problem Statement

The conventional assignment evaluation process is manual, time-intensive, and susceptible to inconsistency and academic dishonesty. There is a need for an automated system capable of extracting text from diverse formats, performing semantic evaluation, detecting plagiarism, and generating structured grading outputs while maintaining human oversight.

3. Proposed Approach

Building upon prior research in automated grading, semantic similarity modeling, and OCR-based document processing, AssignMatch is designed as a modular, end-to-end architecture for automated evaluation of scanned and digital assignments. The system integrates document ingestion, text extraction, semantic analysis, rubric- guided grading, plagiarism detection, and quality control into a unified pipeline.

The architecture follows a layered processing model to ensure scalability, maintainability, and extensibility across different academic domains.



3.1 Ingestion and OCR Layer

The first stage of the pipeline handles heterogeneous input formats, including:

- Scanned PDF documents
- Mobile-captured assignment images
- Printed handwritten submissions
- Editable digital documents

All uploaded files are stored with associated metadata (student ID, subject, timestamp, assignment ID) before processing begins.

For scanned or image-based submissions, Optical Character Recognition (OCR) is applied to convert visual content into machine-readable text. Prior to OCR execution, preprocessing steps are performed to improve recognition accuracy:

- Image deskewing to correct orientation
- Noise reduction using smoothing filters
- Contrast enhancement
- Adaptive thresholding for binarization
- Segmentation of text regions.

These preprocessing steps are essential because OCR accuracy significantly impacts downstream evaluation quality. The OCR engine (e.g., Tesseract or equivalent models) extracts textual content and, where supported, mathematical notation. The output is a structured digital representation of student answers. In cases involving equations, additional parsing mechanisms convert detected symbols into markup(e.g., LaTeX-like representations) or forward them to specialized math-processing modules.
This stage ensures that even handwritten submissions can enter the semantic grading pipeline.

3.2 Assignment Parsing and NLP Segmentation

Once textual content is extracted, AssignMatch performs intelligent segmentation of the assignment into structured units. This stage is critical because grading is typically question- specific rather than document-wide.

The parsing module performs:

- Question boundary detection
- Sub-question segmentation
- Sentence-level tokenization
- Paragraph grouping
- Identification of enumerated steps

Natural Language Processing techniques are applied to split responses into coherent semantic units. For assignments containing multiple parts (e.g., Q1(a), Q1(b)), rule-based and heuristic parsing methods isolate answer blocks.

For structured disciplines such as mathematics and programming:

- Code blocks are separated from explanatory text
  - Mathematical expressions are tagged and optionally processed by domain-specific evaluators

This modular segmentation enables targeted evaluation and rubric alignment per question.

- LLM-Based Evaluation Module

The core *intelligence of AssignMatch resides in the LLM Evaluation Module. Each segmented answer is evaluated* independently using a rubric-guided prompt. The evaluation prompt typically includes:

- The assignment question
- The official reference answer or marking scheme
- The student's response
- Explicit grading instructions
- Scoring scale (e.g., 0–5 or 0–10)

- A_s = Student Answer
- A_r = Reference Answer

The evaluation function is defined as:

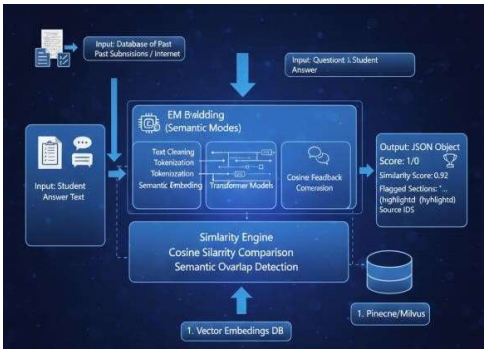
$$Score = \mathcal{G}(A_s, A_r, Q)$$

where $\mathcal{G}$ is an LLM-based grading function conditioned on question Q.

The prompt structure ensures that the model performs structured evaluation rather than free-form commentary.

To improve consistency and reduce stochastic variation, best practices are followed:

- Deterministic decoding (temperature = 0)
- Explicit output formatting instructions
- Constraint-based response templates
- Structured JSON-like output enforcement



The LLM performs:

- Semantic understanding
- Concept matching
- Logical reasoning validation
- Completeness assessment
- Error detection

Unlike traditional keyword matching, this module evaluates meaning and reasoning structure.

The output includes:

- Numerical score
- Detailed feedback
- Justification of deductions

In certain cases, retrieval grounding may be incorporated to provide the LLM with contextually relevant material from the reference solution corpus, reducing hallucination risk.

- Rubric Integration and Feedback Structuring

AssignMatch incorporates a predefined grading rubric aligned with instructor expectations. The rubric specifies:

- Key concepts required
- Step-wise marks allocation
- Partial credit conditions
- Common mistake penalties

After receiving the LLM output, a post- processing module parses the response to extract:

- Numeric grade
- Concept coverage indicators
- Feedback comments

If the LLM response deviates from the expected format, a lightweight validation engine triggers either:

- Prompt re-execution
- Output correction
- Rule-based fallback scoring

This validation layer ensures structural consistency and grading reliability.

The final output per question includes:

- Awarded marks
- Missed concepts
- Improvement suggestions

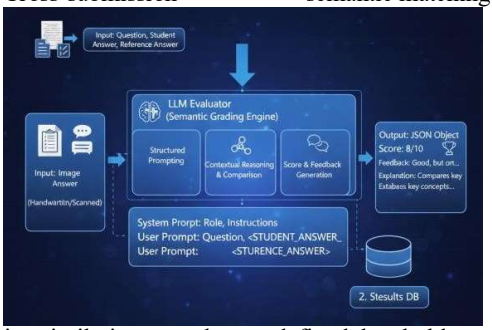
All question-level outputs are aggregated into a complete student evaluation report.

3.3 Plagiarism and Similarity Detection Module

Parallel to semantic grading, AssignMatch performs plagiarism detection.

Each answer undergoes similarity analysis using vector-space representations. The system computes:

- TF-IDF vector similarity
- Cosine similarity between embeddings
- Cross-submission semantic matching



If cosine similarity exceeds a predefined threshold (e.g., 0.8), the response is flagged for review. Embedding-based similarity helps detect paraphrased plagiarism beyond simple lexical overlap.

The system supports comparison against:

- Peer submissions
- Historical assignment archives
- Predefined reference corpora

Flagged responses are marked in the teacher dashboard with similarity scores and highlighted overlapping segments.

This dual evaluation (semantic grading + similarity detection) enhances academic integrity enforcement.

3.4 Quality Control and Human-in-the- Loop Layer

While LLM-based grading achieves high correlation with human scoring, full automation may introduce risks in edge cases. Therefore, AssignMatch integrates a quality control layer to maintain reliability.

Triggers for human review include:

- High plagiarism score
- Low LLM confidence
- Unusual grading patterns
- Ambiguous responses
- Extreme score deviations

In such cases, the system routes the response to an instructor dashboard for manual verification.

This hybrid AI + human workflow balances efficiency with accountability and aligns with modern educational AI deployment strategies.

3.5 System Modularity and Extensibility

The overall architecture follows this logical pipeline:

OCR → NLP Segmentation → LLM Evaluation → Plagiarism Detection → Report Generation → Human Review This modular design offers several advantages:

- Scalability across institutions
- Replaceable LLM backends
- Domain-specific plugin modules

$$Sim(A,B) = \frac{E_A \cdot E_B}{||E_A||\,||E_B||}$$

- Integration with Learning Management Systems (LMS)
- Adaptability to text, math, and code-based

assignments

Recent intelligent grading case studies emphasize similar architecture patterns involving OCR ingestion, subject classification, and routing to specialized grading agents. AssignMatch aligns with these emerging trends and supports future expansion into multimodal and domain-adaptive evaluation systems.

4. Data, Benchmarks, and Case Studies

To situate AssignMatch within established academic research, it is important to reference commonly used datasets and benchmarking standards in automated grading and plagiarism detection.

4.1 Benchmark Datasets in Automated Grading

Publicly available corpora such as the SemEval-2013 BEETLE dataset and university short-answer datasets have historically supported evaluation of automated essay scoring systems. Similarly, datasets such as the MIS scanned-answer corpus have been used to evaluate transformer-based grading models, where Longformer embeddings demonstrated strong correlation with human grading.

These datasets provide standardized evaluation frameworks for measuring agreement between automated systems and human graders using statistical metrics such as Pearson correlation and mean absolute error.

While AssignMatch is implemented as a system-level solution rather than a fine- tuned dataset-specific model, its evaluation methodology aligns with benchmark practices established in these studies.

4.2 Plagiarism Detection Benchmarks

The PAN plagiarism corpora are widely used benchmarks for evaluating plagiarism detection systems. These datasets include paraphrased and partially copied student responses, enabling measurement of recall and precision in similarity detection.

- AssignMatch employs cosine similarity and embedding-based comparison techniques consistent with methods validated in such corpora. Future evaluation phases may include formal benchmarking against these datasets to quantify detection performance.

4.3 Case Study-Based Validation

Beyond public datasets, real-world course-level deployment provides practical validation. Prior studies applying LLM-based grading in academic courses demonstrate that automated systems can achieve strong alignment with human grading while significantly reducing evaluation time.

AssignMatch is designed to support similar institutional pilot deployments, where system-generated scores can be compared with instructor grading to evaluate correlation, grading stability, and student feedback acceptance.

5. Experimental Methodology

To evaluate the effectiveness of AssignMatch, a structured experimental methodology was designed to measure grading accuracy, plagiarism detection capability, and overall system robustness.

5.1 Dataset Preparation

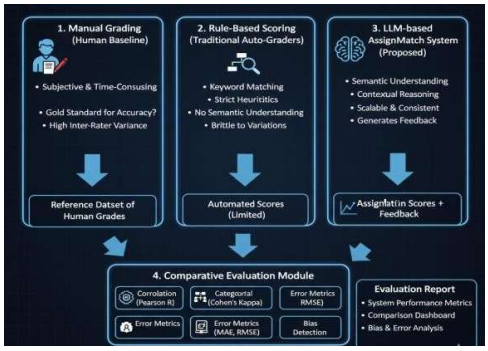
A test dataset consisting of scanned student assignments with instructor- assigned grades was compiled. The dataset included diverse question types such as short answers, descriptive responses, and mathematical expressions. Assignments were digitized through OCR preprocessing before being evaluated by the AssignMatch pipeline. Human grades served as the ground truth for performance comparison.

The AssignMatch prototype was implemented using a modular backend architecture integrating OCR processing, embedding generation, and LLM-based evaluation components. Preprocessing and segmentation modules were executed sequentially to ensure consistent text extraction before grading. Experimental runs were conducted on locally stored submissions to simulate institutional deployment conditions.

5.2 Baseline Comparison

AssignMatch's performance was compared against:

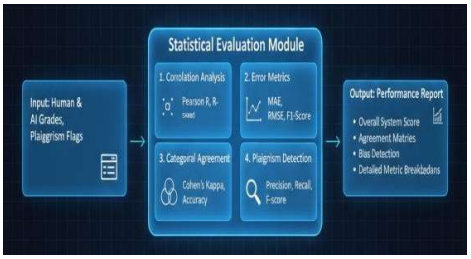
- Manual human grading (reference standard)
- Rule-based keyword matching approaches
- Similarity-based plagiarism detection tools



This comparison helped assess the value added by LLM-based rubric-guided evaluation.

5.3 Evaluation Metrics

To measure grading consistency and reliability, the following statistical metrics were used:



1. Pearson Correlation Coefficient (r)

   Measures linear agreement between AI-generated scores and human scores.

2. Mean Absolute Error (MAE)

   Computes the average absolute difference between AI and human grades.

3. Exact Match Rate

   Percentage of responses where AI-assigned scores exactly match instructor scores.

4. Cohen's Kappa

   Evaluates categorical agreement (e.g., grade bands such as A/B/C) while correcting for chance agreement.

5. Plagiarism Detection F1 Score For flagged cases, precision and recall were calculated to assess detection performance.

These metrics are commonly used in automated essay scoring and LLM- based grading research.

5.4 Ablation and Component Testing

To understand the contribution of different system components, controlled tests were conducted:

- Grading with and without rubric- enhanced prompts
- Comparison of deterministic decoding (temperature = 0) versus default settings
- Evaluation of OCR text quality impact on grading accuracy

During experimentation, multiple prompt configurations and segmentation strategies were evaluated to analyze grading stability. Minor variations in OCR output quality were intentionally introduced to observe downstream performance sensitivity. These implementation-level observations guided the selection of deterministic decoding and rubric-grounded evaluation settings used in the final system configuration.

5.5 Human Review and Qualitative Analysis

A subset of AI-generated feedback was reviewed by instructors to assess:

- Feedback clarity
- Fairness and bias
- Alignment with grading rubric
- Constructiveness of comments

This hybrid evaluation ensured that AssignMatch maintained academic integrity and instructional usefulness.

5.6 Robustness Testing

The system was tested against:

- Noisy OCR outputs
- Adversarial or irrelevant inputs
- Paraphrased plagiarism cases

These tests ensured resilience under real-world deployment conditions.

6. Discussion and Future Work

6.1 Discussion

The development of AssignMatch demonstrates the practical feasibility of integrating OCR, Retrieval-Augmented Generation (RAG), embeddings, vector databases, and large language models into a unified academic grading framework. The system successfully automates the evaluation of scanned assignments while maintaining structured, rubric-aligned feedback generation.

One of the key advantages observed during implementation is efficiency. Manual grading, particularly for large classes, is time-consuming and cognitively demanding. AssignMatch reduces turnaround time by automating text extraction, answer segmentation, scoring, and plagiarism analysis. This enables instructors to focus more on conceptual teaching rather than repetitive evaluation tasks. Additionally, students benefit from faster feedback cycles, which can support iterative learning and improvement.

The use of RAG significantly strengthens grading reliability. By retrieving relevant rubric criteria or reference answers from a vector database, the system grounds the LLM's evaluation in contextual information. This reduces hallucination risks and ensures that grading decisions are aligned with predefined academic standards.

However, several limitations remain. The system's performance depends heavily on OCR accuracy; poorly scanned documents or complex handwritten equations may affect downstream grading quality. Moreover, although deterministic decoding improves consistency, minor variations in qualitative feedback may still occur. Another important consideration is fairness—automated systems must ensure that diverse writing styles and expression patterns are evaluated equitably. To address these concerns, AssignMatch incorporates a human-in- the-loop review mechanism for low-confidence or flagged cases.

6.2 Future Directions

While AssignMatch demonstrates strong feasibility, several enhancements can further improve its robustness and scalability.

First, integrating multimodal or vision- language models could improve handling of handwritten mathematical expressions, diagrams, and structured problem-solving tasks. This would reduce dependence on traditional OCR pipelines.

Second, rubric optimization techniques could be explored. Iterative refinement of prompts or adaptive rubric tuning based on instructor feedback may enhance grading stability and reduce scoring variance.

Third, more advanced plagiarism detection mechanisms—such as stylometric analysis or cross-assignment similarity clustering—could strengthen detection of sophisticated paraphrasing and collusion.

Finally, large-scale benchmarking and institutional deployment studies would help validate AssignMatch in real classroom environments. Integration with learning management systems (LMS) could enable seamless assignment submission, grading automation, and analytics dashboards for instructors.

6.3 Concluding Remarks

AssignMatch highlights how AI-assisted grading systems can function as supportive educational tools rather than replacements for instructors. By combining automated evaluation with human oversight, the system balances efficiency with academic integrity. With continued refinement and responsible deployment, intelligent grading architectures have the potential to significantly enhance modern educational workflows.

7. Conclusion

AssignMatch presents a comprehensive approach to intelligent academic assessment by integrating Optical Character Recognition (OCR), Retrieval-Augmented Generation (RAG), embeddings, vector databases, and large language models into a unified grading pipeline. The system demonstrates how modern AI techniques can be practically applied to automate the evaluation of scanned student assignments while maintaining rubric alignment and structured feedback generation.

By combining text extraction, semantic retrieval, LLM-based evaluation, and similarity-driven plagiarism detection, AssignMatch addresses multiple stages of the grading workflow within a single modular architecture. This integration reduces manual effort, accelerates feedback delivery, and provides instructors with a decision-support tool that enhances efficiency without removing human oversight. The inclusion of a quality control layer ensures that uncertain or high-risk cases can be reviewed manually, preserving grading fairness and academic integrity.

The implementation of Retrieval- Augmented Generation plays a crucial role in grounding grading decisions within predefined rubrics and reference materials. This approach reduces hallucination risks and improves scoring consistency compared to standalone prompt-based evaluation. Additionally, embedding-based similarity analysis

strengthens the system's ability to detect high-overlap submissions and potential plagiarism.

While AssignMatch demonstrates strong feasibility as a scalable grading solution, continued refinement and large-scale validation will further strengthen its reliability. Future benchmarking against instructor grading across diverse subjects and question formats will provide deeper insights into system accuracy and alignment. Expanding the architecture to support multimodal inputs and enhanced plagiarism detection techniques will further improve robustness.

Overall, this project contributes an end- to-end system design and implementation framework for AI- assisted grading. AssignMatch illustrates how emerging AI technologies can be responsibly integrated into educational workflows to improve efficiency, consistency, and feedback quality. With ongoing development and institutional collaboration, such intelligent grading systems have the potential to become valuable tools in modern academic environments.

8. Reference

References: Cited sources include recent literature on AI grading systems, technical methods for OCR and NLP, datasets for automated assessment, and evaluation metrics.

[1] Automated assignment grading with large language models: insights from a bioinformatics course | Bioinformatics | Oxford Academic

https://academic.oup.com/bioinformatics/article/41/Supplement_1/i21/8199383

[2] A Structured Dataset for Automated Grading: From Raw Data to Processed Dataset | MDPI https://www.mdpi.com/2306- 5729/10/6/87

[3] An automated essay scoring systems: a systematic literature review - PMC

https://pmc.ncbi.nlm.nih.gov/articles/P MC8460059/

[4] LLM-Based Grader: Automated Assessment Overview

https://www.emergentmind.com/topics/l lm-based-grader

[5] Auto-scoring Student Responses with Images in Mathematics

https://educationaldatamining.org/edm2 023/proceedings/2023.EDM-short- papers.36/index.html

[6] Automated Grading of Students' Handwritten Graphs: A Comparison of Meta-Learning and Vision-Large Language Models https://arxiv.org/html/2507.03056v1

[7] Comparative analysis of text-based plagiarism detection techniques - PMC

https://pmc.ncbi.nlm.nih.gov/articles/P MC11977957/

[8] Plagiarism detection and prevention: a primer for researchers - PMC

https://pmc.ncbi.nlm.nih.gov/articles/P MC8436797/

[9] Intelligent Assignment Grading System https://app.readytensor.ai/publications/a aidc-capstone-intelligent-assignment- grading-system-rjTtP5fZfVdA