# Integrating LLM for Peer Evaluation with OCR

## Introduction

The integration of Large Language Models (LLMs) into peer evaluation systems offers a transformative approach to automating the assessment of student submissions. By combining text extraction techniques with advanced natural language processing, such systems can provide consistent, scalable, and insightful evaluations. A significant challenge arises when dealing with documents containing handwritten text, as traditional text extraction methods like PyPDF are less effective. Recent advancements in Optical Character Recognition (OCR), particularly Mistral OCR, provide a potential solution for extracting text from handwritten documents. This report details a proposed system for integrating LLMs into a peer evaluation framework, focusing on text extraction strategies for both typed and handwritten documents, a comparison of OCR model accuracies, and the workflow for generating evaluation outputs in JSON format.

## Background

Peer evaluation systems benefit from automation to reduce manual grading efforts and ensure objectivity. For documents in PDF format, the nature of the content—typed or handwritten—determines the appropriate text extraction method. Typed PDFs, typically generated from document editors like Microsoft Word, Google Docs, or Apple iWork Pages, can be processed using libraries like PyPDF, which reliably extract machine-readable text. Handwritten documents, however, require OCR to convert images of text into digital formats. Mistral OCR, introduced by Mistral AI, is marketed as a state-of-the-art document understanding API capable of handling complex documents, including handwritten notes ([Mistral AI](https://mistral.ai/news/mistral-ocr)). However, its effectiveness for handwritten text, particularly cursive or disorganized writing, requires scrutiny. Once text is extracted, LLMs can evaluate it against an ideal answer, providing scores, plagiarism detection, and feedback.

## OCR Model Accuracy for Handwritten Text

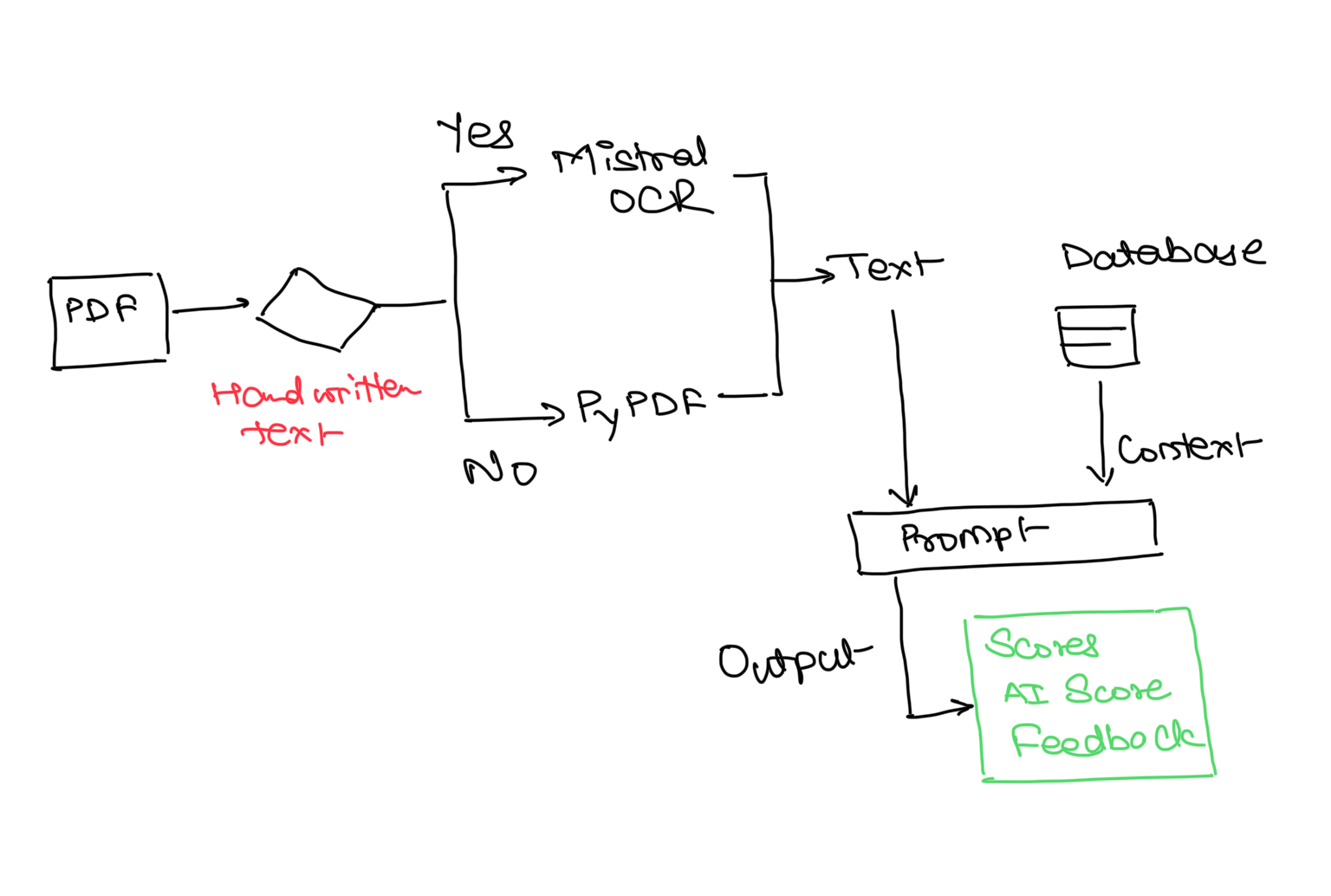
Accurate text extraction is critical for reliable LLM-based evaluation, especially for handwritten documents. A 2024 benchmark by HandwritingOCR.com provides a comparative analysis of OCR models for handwritten text, using Word Error Rate (WER) as the metric. WER measures the percentage of words incorrectly transcribed, with accuracy calculated as 100% - WER. The following table summarizes the results:

| **Model** | **Word Error Rate (WER) (%)** | **Accuracy (%)** |
| --- | --- | --- |
| Mistral OCR | 0.99 | 99.1 |
| Microsoft Azure Document AI | 8.67 | 91.33 |
| Amazon Web Services (AWS) Textract | 10.5 | 89.5 |
| Google Document AI | 23.3 | 76.7 |
| Transkribus | 47.7 | 52.3 |
| Tesseract | 95.4 | 4.6 |

### Notes on Mistral OCR

Mistral OCR’s performance on handwritten text is not included in the above benchmark, and specific accuracy figures are scarce. A 2025 benchmark by AIMultiple indicates that Mistral OCR struggles with cursive handwriting and disorganized text layouts, with accuracy ranging widely depending on the document ([AIMultiple Benchmark](https://research.aimultiple.com/ocr-accuracy/)). Another comparison by JigsawStack suggests that Mistral OCR failed to detect most handwritten text in a driver’s license example, while their vOCR model achieved high accuracy ([JigsawStack Comparison](https://jigsawstack.com/blog/mistral-ocr-vs-jigsawstack-vocr)). Mistral AI’s official documentation highlights a 98.96% accuracy for scanned documents, but it is unclear if this includes handwritten text ([Mistral AI](https://mistral.ai/news/mistral-ocr)). Given these findings, Mistral OCR’s suitability for handwritten text in peer evaluation systems should be validated through testing with representative documents.

## Proposed System Workflow



The peer evaluation system can be structured to handle both typed and handwritten documents efficiently, ensuring accurate text extraction and robust evaluation. The workflow is as follows:

### 1. Document Type Identification

* **Typed PDFs**: Documents created in digital editors (e.g., Word, Google Docs) contain machine-readable text. These can be identified by checking if PyPDF can extract coherent text without errors.
* **Handwritten PDFs**: Documents with handwritten content, typically scanned or photographed, require OCR. These can be identified if PyPDF fails to extract meaningful text or if metadata indicates an image-based PDF.

### 2. Text Extraction

* **For Typed PDFs**: Use the PyPDF library to extract text. PyPDF is highly reliable for machine-readable PDFs, as it directly accesses the text layer without needing image processing. This ensures near-perfect accuracy for documents generated from digital editors.
* **For Handwritten PDFs**: Employ Mistral OCR to extract text. Mistral OCR processes image-based PDFs, preserving document structure (e.g., tables, headings) and extracting text from handwritten content. Preprocessing steps, such as converting images to black-and-white or increasing contrast, can enhance accuracy, as noted in AIMultiple’s methodology ([AIMultiple Benchmark](https://research.aimultiple.com/ocr-accuracy/)).

### 3. LLM-Based Evaluation

Once text is extracted, it is fed into an LLM along with the context of the ideal answer provided by the professor. The LLM performs the following tasks:

* **Scoring**: Assigns a quantitative score based on the alignment of the extracted text with the ideal answer, considering content accuracy, completeness, and relevance.
* **Plagiarism Detection**: Compares the extracted text against known sources or the ideal answer to identify potential plagiarism, generating a plagiarism score.
* **Feedback Generation**: Provides qualitative feedback, highlighting strengths, weaknesses, and areas for improvement in the response.

### 4. Output Generation

The LLM returns a JSON object containing the evaluation results. An example structure is:

{  
 "score": 85,  
 "plagiarism\_score": 0.95,  
 "feedback": "The response addresses the main points of the ideal answer but lacks depth in discussing secondary concepts. Consider elaborating on the implications of the topic."  
}

This structured output can be integrated into the peer evaluation system for display or further processing.

## Implementation Considerations

To ensure the system’s effectiveness, several factors should be addressed:

### OCR Accuracy for Handwritten Text

While Mistral OCR is a robust tool for document understanding, its performance on handwritten text, particularly cursive or disorganized writing, may not match that of specialized models like Mistral OCR (99.1% accuracy) or Microsoft Azure Document AI (91.33% accuracy). If high accuracy is critical, consider evaluating alternative OCR models or combining Mistral OCR with preprocessing techniques to improve results.

### Preprocessing Handwritten Documents

Preprocessing can significantly enhance OCR accuracy. Techniques include:

* Converting images to black-and-white to reduce noise.
* Increasing contrast to make text more legible.
* Removing backgrounds using tools like CamScanner, as practiced in AIMultiple’s benchmark ([AIMultiple Benchmark](https://research.aimultiple.com/ocr-accuracy/)).

### Testing and Validation

Before deployment, test Mistral OCR with a sample set of handwritten documents representative of the peer evaluation context (e.g., student notes, exam answers). Compare its performance against other high-accuracy models to determine the best tool. This is particularly important given Mistral OCR’s reported challenges with cursive handwriting.

### LLM Configuration

The LLM used for evaluation should be configured to handle potential OCR errors, such as misrecognized characters or missing words. Robust models can infer meaning from imperfect text, but the system should flag low-confidence extractions for manual review. Additionally, the LLM should be trained or fine-tuned to provide educationally relevant feedback, aligning with the professor’s grading criteria.

### Scalability and Integration

Ensure the system can handle large volumes of documents efficiently. Mistral OCR’s API supports batch inference, potentially doubling the pages processed per dollar, making it cost-effective for large-scale applications ([Mistral AI](https://mistral.ai/news/mistral-ocr)). Integration with existing educational platforms should be seamless, with the JSON output easily parsed for display or storage.

## Recommendations

* **OCR Selection**: While Mistral OCR is a strong candidate due to its recent advancements and ability to handle complex documents, its performance on handwritten text should be validated. If cursive handwriting is common, consider using Mistral OCR or Microsoft Azure Document AI, which have demonstrated superior accuracy in benchmarks.
* **Hybrid Approach**: For maximum reliability, implement a hybrid text extraction strategy: use PyPDF for typed PDFs and an OCR model (Mistral OCR or an alternative) for handwritten PDFs. A decision tree can automate the selection process based on document characteristics.
* **Preprocessing Pipeline**: Incorporate a preprocessing pipeline for handwritten documents to enhance OCR accuracy. This can be automated using image processing libraries like OpenCV.
* **Pilot Testing**: Conduct a pilot test with a diverse set of documents to evaluate the system’s performance. Collect feedback from educators to refine the LLM’s scoring and feedback mechanisms.
* **Continuous Monitoring**: Monitor the system’s performance over time, updating the OCR model or LLM as new advancements emerge. Mistral OCR’s ongoing improvements may enhance its handwritten text capabilities in the future.

## Conclusion

Integrating LLMs into a peer evaluation system offers a powerful solution for automating grading while maintaining quality and fairness. By using PyPDF for typed PDFs and Mistral OCR for handwritten documents, the system can handle diverse inputs. However, the choice of OCR model is critical, as Mistral OCR’s performance on handwritten text may not match that of specialized models like Mistral OCR or Microsoft Azure Document AI. Thorough testing and preprocessing are essential to ensure accuracy. Once text is extracted, LLMs can provide comprehensive evaluations, delivering scores, plagiarism checks, and feedback in a structured JSON format. This approach promises to streamline peer evaluation, making it an invaluable tool for educational institutions.