Righting Writing: Enhancing OCR Output Using NLP

MJ Corey and **Akshat Ghoshal** and **Yassin Ali** and **Jordan Johnson**
University of Minnesota
corey094@umn.edu, ghosh159@umn.edu, ali00740@umn.edu, joh20376@umn.edu

Abstract

This study explores the enhancement of Optical Character Recognition (OCR) outputs for Arabic text through post-processing with large language models (LLMs). We used ...

1 Introduction

Optical Character Recognition (OCR) is critical for digitizing printed and handwritten texts, but its performance on Arabic script remains challenging due to complex morphology, diacritics, and diverse writing styles. Errors in OCR outputs, such as misrecognized characters or incorrect word segmentation, hinder downstream applications like text analysis and archival. Recent advances in natural language processing (NLP), particularly large language models (LLMs), offer a promising approach to post-process and correct OCR errors by leveraging contextual understanding. This project, "Righting Writing: Enhancing OCR Output Using NLP," aims to improve the accuracy of OCR systems for Arabic text through LLM-based post-processing, addressing a gap in NLP applications for low-resource languages.

1.1 Background & Motivation

Document reading is primarily dominated by two types of models: OCR models and Vision Language Models. OCR models take an image as input and inference bounding boxes around where it believes each individual character lies. The OCR model will then infer the most likely character in each bounding box based on a prior deep-learning training process. VLM's on the contrary read documents by analyzing both the image layout and the text together, rather than treating them separately. They combine what they see on the page with what the words mean to understand structure, content, and context all at once.

Both models have limitations, however. OCR models struggle when a document does not lend itself well to casting bounding boxes around characters which is especially prevalent in the cursive script of Arabic text. In addition, OCR models inherently struggle with unique handwriting styles since unique handwriting could not match any of the training done prior to a document's reading. VLM's, while one of the most powerful tools available for document reading, is costly to train due to the necessity for concurrent visual and textual understanding. For the same reason, VLM's also struggle at each individual document read because they need to hold both the image context and text content which makes VLMs inefficient to use for document reading at scale.

For these reasons, this study believes that utilizing an LLM post-processing step on basic OCR output has the potential to outperform basic OCR framework and be a more cost-effective alternative to VLMs, both monetarily and computationally. If this pipeline does yield competitive results it will be most impactful for those who need to do large document digitization at scale, like teachers with essays or government employees with citizen's documents for example.

2 Methodology

2.1 Dataset

For this study, we utilized the KITAB-Bench dataset. KITAB-Bench is a comprehensive Arabic OCR benchmark dataset that features various sub-datasets for different OCR tasks. For the purposes of our study we will be specifically looking at the datasets within KITAB-Bench that focus on the Image to Text task. This refined our data to 12 sub-datasets that featured images across various domains such as handwritten text, font written text, paragraph-length passages, sentence-length passages, and even word-length images.

The KITAB-Bench dataset provided a corpus with a breadth of image types to understand areas where LLM post-processing was most effective and areas where it may struggle.

2.2 OCR Systems

Four OCR systems are evaluated:

- **Tesseract**: An open-source OCR engine known for its versatility.
- **PaddleOCR**: A deep learning-based OCR system optimized for multiple languages.
- **EasyOCR**: A lightweight OCR tool supporting Arabic script.
- **AIN**: A Large Multimodal Model that specializes in Arabic, and excels in Arabic OCR tasks

2.3 Starting with a Large-Resource Language (English) for Pipeline Development

To develop a robust OCR post-processing pipeline, we initially prototyped our approach using English, a high-resource language with abundant datasets and well-established OCR and NLP tools. This strategy allowed us to design and test the pipeline in a controlled environment before adapting it to Arabic, a low-resource language with unique challenges such as right-to-left script, diacritics, and limited training data. We used the IAM Handwriting Database, which contains handwritten English text samples, and applied Google Vision OCR to generate raw OCR outputs.

2.4 Post-Processing with LLMs

We employ three LLMs for post-processing:

- **GPT-4o**: A state-of-the-art model, tested with zero-shot and five-shot prompts in both languages.
- **Allam**: An Arabic-focused model, tested with zero-shot and five-shot prompts in English and Arabic.

Prompts are designed to instruct the LLMs to correct OCR errors, with few-shot prompts providing five examples of correct text. Outputs are cleaned to remove diacritics and irrelevant prefixes (e.g., "Corrected:") using regular expressions, ensuring fair comparison with diacritic-free ground truth.

2.5 Evaluation Metrics

We compute the following metrics:

- **Word Error Rate (WER)**: Measures word-level errors (insertions, deletions, substitutions).
- **Character Error Rate (CER)**: Measures character-level errors.
- **Edit Distance**: Quantifies the minimum edits needed to match the ground truth.

Metrics are calculated using the jiwer and edit-distance libraries, comparing raw OCR outputs and post-processed texts against ground truth.

3 Initial Results

4 Discussion Points

4.1 Replicability

This study is inherently quite replicable. Since all of the datasets and OCR models are open source they are easily accessible for all looking to repeat the experiment. In addition, a temperature of 0.0 was used for the LLMs so that there was a lower amount of spontaneity within the responses of the LLM post-processing step. Finally, due to there being thousands of samples it is likely from a statistical perspective to reach very similar conclusions if the temperature were to be adjusted.

There is a monetary cost to using ChatGPT-4o via the API which also means the experiment becomes more expensive with scale. However, it stands with reason to believe that any LLM that has the capability for instruction-tuned prompts and Arabic comprehension would be a valid substitute to GPT-4o. This allows other open-source LLMs that are competitive with GPT-4o to be used for this experiment as well.

4.2 Ethical Implications

This experiment brings up one primary ethical concern in that of the OCR or LLM misinterpreting the actual text within a personal document. If there is poor confidence in a word from the baseline OCR it is likely that the LLM will take some agency in finding the best replacement word for the unclear one. While this is an ethical concern, the current limitations of OCR already share the same ethical concern of misinterpreting an unclear document. In addition, as shown earlier via BERTscores, a vast majority of documents had a stronger relationship semantically to the original text after post-processing versus the original OCR result. This argues that the post-corrected result is a better representation of the digitized text versus the baseline OCR. Finally, as OCR models improve, the baseline result is more likely to capture the majority of correct characters initially which will make the LLM post-correction more likely to capture the true word given the context.

4.3 Study Implications & Future Research

While there has been no direct inspiration on others' research currently, the results from our study offer an alternative option for document reading. The results of the experiment on our pipeline indicate a viable alternative to state-of-the-art VLMs and OCR models.

There are still some limitations to this area of study. The most obvious is that the pipeline is heavily dependent on the initial capabilities of OCR models to prov

Despite our positive findings, this study still has areas that can be improved upon. The first thing that was not in the scope of this project that would almost guarantee an improvement in results is fine-tuning an LLM to this specific task of text repair. If an LLM became accustomed to the common errors produced by OCR results through extensive training it is likely that future uses of the pipeline will better identify those errors. In addition, a properly trained LLM is more likely to preserve actual errors found in the documents instead of improperly correcting them.

Another potential path for this research could be adjusting both ends of the pipeline to reduce the number of options for replacement could further improve the change in WER. Many OCR models have the capability of displaying the confidence values of a character being any given character. If an LLM was trained on the task of selecting the best-fitting character among the most confident OCR results based on the surrounding context of the unconfident character there is potential for improved accuracy.