Al-Warraq at AR-MS NAKBA-NLP 2026: Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR

Sarah Al Hamed, Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood,

Zainab Kamil, Jana Mukhles Alqasem, Sarah Ayad
Arab Open University, Riyadh, Saudi Arabia
24414598ksa@arabou.edu.sa, 23418335ksa@arabou.edu.sa, 22410231ksa@arabou.edu.sa,
23413390ksa@arabou.edu.sa, 24414598KSA@aou.edu.sa, 24414015KSA@aou.edu.sa,
s.ayad@arabou.edu.sa

Abstract

We present our submission to the NAKBA NLP 2026 Automatic Manuscript OCR/HTR shared task on Arabic manuscripts. The task aims to transcribe manuscript line images into machine-readable Arabic text. Our approach followed an iterative pipeline including model selection, training, error analysis, test-time augmentation, and postprocessing. After evaluating several OCR/HTR models, we selected and trained the most suitable model on the provided manuscript line images and transcriptions. Error analysis showed better character-level performance than word-level performance, which motivated the use of test-time augmentation and text cleaning to improve robustness. The final system achieved a CER of 0.1142 and a WER of 0.378, placing fifth in the shared task. These results show that simple but targeted improvements can support effective Arabic manuscript transcription.

Keywords: OCR, Transformer Model, Vision-Language Model

1. Introduction

Arabic manuscripts are an important part of cultural and scientific heritage. They preserve history, religion, medicine, literature, and many forms of knowledge written over long periods of time. However, much of this material still exists only as handwritten documents or scanned images, which makes access, search, and analysis difficult. Automatic transcription can help convert these manuscript images into machine-readable text and make them easier to preserve, index, and study. Recent work on historical Arabic manuscript datasets also highlights the importance of accurate OCR systems for improving access to these documents.

The goal of this subtask is to develop automatic systems that transcribe Arabic manuscript page images into text. This task is important, but it is also difficult. Arabic script is naturally cursive, and letter shapes change depending on their position in the word. Many letters differ only by dots, and handwritten manuscripts often contain irregular spacing, faded ink, noise, deformation, and inconsistent writing styles. These issues make Arabic manuscript transcription more difficult than modern printed OCR and also more difficult than many Latin-script settings. Recent Arabic HTR studies continue to describe segmentation, overlapping characters, and context dependency as core challenges.

Another challenge is that success at the character level does not always mean success at the word level. A system may correctly recognize many letters but still produce incorrect words because of missing spaces, merged words, or decoding mistakes. For this reason, evaluation in OCR and HTR commonly uses both Character Error Rate (CER) and Word Error Rate (WER), since they reflect different aspects of recognition quality. In manuscript settings, where handwriting is irregular and word boundaries are not always clear, this distinction becomes especially important.

In recent years, deep learning has changed the field of text recognition. Traditional pipelines often relied on hand-crafted features, segmentation rules, and multi-stage processing. In contrast, modern systems can learn directly from image-text pairs and perform end-to-end transcription. This progress is especially relevant for historical manuscripts, where visual variation is high and manual rule design is difficult. Transformer-based OCR and HTR models have shown strong results by learning visual and textual patterns jointly, and large vision-language models are now being explored for OCR and document understanding in multilingual settings.

In this paper, we focus on Arabic manuscript OCR/HTR under the shared task setting. Our work is motivated by the need for practical and robust transcription systems that can handle difficult handwritten line images. The approach combines model selection, training, recovery from technical issues, and simple inference-time improvement strategies. The broader aim is not only to improve recognition accuracy, but also to contribute to the digital preservation and accessibility of Arabic manuscript heritage.

2. Literature Review

Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are closely

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 191–195

related, but they are not the same. OCR is usually used for printed or well-structured text, where character shapes are more regular. In contrast, HTR focuses on handwritten text, where writing styles, letter shapes, and spacing can vary greatly. For manuscript collections, HTR is the more suitable setting because the input is handwritten and of- ten historical. Recent surveys show that HTR has moved from rule-based and segmentation- dependent methods to end-to-end neural models that can work at word, line, and document level (Garrido-Munoz et al., 2025).

Arabic OCR and Arabic HTR are more difficult than many other text recognition tasks. Arabic script is cursive, character forms change according to their position in the word, and many letters differ only by dots or small marks. These properties make segmentation harder and increase confusion between visually similar characters. The task becomes even more difficult in historical manuscripts because of image degradation, faded ink, irregular spacing, writer variation, and non- standard orthography. For these reasons, recent work in Arabic handwritten recognition of- ten prefers line-based and end-to-end approaches instead of fragile character-level segmentation (Garrido-Munoz et al., 2025).

Deep learning has become the main direction in OCR and HTR research. Earlier systems often relied on separate stages such as pre-processing, segmentation, hand-crafted feature extraction, and decoding. More recent approaches replace these stages with neural architectures that learn directly from image-text pairs. A representative example is TrOCR, which uses a Transformer-based visual encoder and a Transformer-based text decoder in an end-to- end framework. TrOCR showed that pre-trained Transformer models can achieve strong results on both printed and handwritten text recognition tasks, while also reducing the need for manually designed OCR pipelines (Li et al., 2021).

This move toward pre-training has also connected OCR and HTR with multimodal vision-language modeling. Recent vision-language models can take an image as input and generate text directly, which makes them attractive for OCR in zero-shot and few-shot settings. Qwen2-VL is one example of this trend. It introduced dynamic resolution handling to better process images of different sizes and reported strong performance on several multimodal understanding tasks, including OCR-related benchmarks (Wang et al., 2024). Such models suggest that large pre-trained multimodal architectures may support multilingual OCR without requiring a fully task-specific design.

Another example is DeepSeek-VL, which was introduced as an open-source vision-language model for real-world multimodal understanding. It was designed to handle diverse visual inputs, including OCR-style content, charts, PDFs, and screenshots (Lu et al., 2024). This is relevant to OCR research because it shows that general-purpose multimodal models are increasingly capable of reading and reasoning over text in images. However, general vision-language models are not automatically sufficient for difficult handwrit- ten Arabic manuscripts. In such settings, image quality, handwriting variation, and rare word forms still require careful adaptation, decoding, and post- processing.

Overall, the literature highlights four main points. First, manuscript transcription should be treated mainly as an HTR task rather than a simple printed OCR task (Garrido-Munoz et al., 2025). Second, Arabic manuscripts remain challenging because of the script structure and the degraded nature of historical documents.

Third, deep learning, especially Transformer-based models such as TrOCR, has significantly improved OCR and HTR performance (Li et al., 2021). Fourth, multimodal pre-trained vision-language models such as Qwen2-VL and DeepSeek-VL are promising for OCR-related tasks, but they still need careful evaluation when applied to handwritten Arabic manuscript data (Wang et al., 2024; Lu et al., 2024).

3. Approach

Our approach focused on building an automatic Arabic manuscript transcription pipeline for the Systems Track. The main goal was to convert cropped manuscript line images into machine-readable Arabic text. Because Arabic manuscripts are visually complex and many existing OCR mod- els are not designed for this type of data, we followed an iterative approach based on model selection, training, error analysis, and inference-time improvement.

We started by testing several OCR and handwritten text recognition models. At the beginning, some of the selected models were not suitable for Arabic script. In particular, one early model was mainly designed for English text, so its performance on Arabic manuscript images was weak. The generated text contained many recognition errors, and the image understanding quality was limited. For this reason, the team compared multiple models and configurations before selecting a better model for Arabic transcription. In total, several candidate models were explored by different team members until a more suitable solution was found. After model selection, we trained the system on the provided manuscript line images and their gold transcriptions. The training process required a long time because the dataset contained around 16,000 training images, and the model needed repeated tuning to improve its ability to recognize Arabic characters and words. During the early stage, the system achieved only about 30% accuracy. However, after repeated experiments, continued training, and changing unsuitable

models, the performance improved significantly and reached about 80% recognition accuracy during development.

The training stage also involved several technical difficulties. Some team members had limited hardware resources, and training on large numbers of images was computationally expensive. In some cases, the training process lasted for many hours or even days. To reduce the risk of losing progress, we adapted the code so that it could save intermediate checkpoints and continue from previous training stages. This was important because interruptions such as power failure could stop the process and lead to loss of many hours of work. By using checkpoint-based recovery, we were able to continue training without restarting from the beginning each time.

Another important challenge appeared after training, when we tried to reload the trained model. A KeyError occurred because the checkpoint file contained only the trained weights, while the processor files were missing. These processor files included essential resources such as the tokenizer and dictionary. To solve this problem, we reused the original processor files from the base model and combined them with the trained weights. This recovery step allowed us to restore the model and continue the experiments successfully.

After obtaining a working model, we evaluated its outputs and analyzed the error types. The results showed an important difference between character-level and word-level performance. The Character Error Rate (CER) was relatively strong, while the Word Error Rate (WER) remained higher. This indicated that the model was often able to rec- ognize many characters correctly, but it still had problems producing correct full words. In particular, the model sometimes merged adjacent words, inserted extra spaces, or produced incorrect spacing before punctuation marks. Therefore, improving word-level quality became an important part of our approach.

To reduce prediction variance and improve robustness at inference time, we applied test-time augmentation (TTA). Each input line image was processed three times using small rotation variations: 0◦, +2◦, and −2◦. These small transformations helped the model handle slight differences in alignment and orientation. After that, a voting strategy was used to select the most frequent predicted text among the three outputs. This simple ensemble-style step helped reduce random recognition variation without the need for additional re- training.

In addition to TTA, we designed a post-processing step to clean the generated Arabic text. This step aimed to correct common formatting and orthographic issues that appeared in the predictions. The cleaning function removed unnecessary spaces before punctuation marks, normalized some Arabic letter forms, and corrected frequent output inconsistencies. It also attempted to separate merged words when possible by using simple Arabic text patterns. This post-processing stage was useful because even when the model recognized characters correctly, the final text still re- quired small corrections to become more readable and closer to the gold transcription.

Time efficiency was another major concern during the final submission stage. Running TTA on all test images increased inference time significantly. At one point, the estimated runtime was longer than the remaining time before the submission deadline. To solve this problem, we reduced the beam search width from 4 to 2 during decoding. This decision decreased the processing time substantially while still preserving acceptable output quality. As a result, the full inference process finished in time and the final submission was uploaded only a few minutes before the competition deadline.

Overall, our approach combined four main elements: careful model selection, repeated training and recovery under limited hardware conditions, inference-time augmentation, and Arabic text post- processing. This pipeline allowed us to improve the system step by step and handle both technical and linguistic challenges. Using this approach, the team achieved competitive results in the shared task, including a CER of 0.1142 and a WER of 0.378, which placed the team among the top-ranked submissions.

4. Discussion

The results show that our system was able to transcribe Arabic manuscript lines with good character-level performance. The final system achieved a CER of 0.1142 and a WER of 0.378, which placed our team in fifth position in the competition. These results show that the selected model was able to learn important visual patterns from the manuscript images and produce useful transcriptions as shown in Table 1.

However, the results also show a clear gap between character recognition and word recognition. The CER was much better than the WER. This means that the model was often able to recognize many characters correctly, but it still had difficulty producing correct full words. This happened because Arabic manuscripts contain connected writing, unclear spacing, and visually similar letters. In some cases, the model merged words, inserted wrong spaces, or produced small orthographic errors. Therefore, word-level prediction remained more difficult than character-level prediction.

To improve robustness, we applied test-time augmentation (TTA). Each image was processed three times using small rotations: 0°, $+2^{\circ}$, and -2°. Then, a voting mechanism selected the most

frequent output. This step helped reduce random variation in the predictions without retraining the model. It provided a simple way to improve stability during inference.

We also applied algorithmic post-processing to clean the generated text. This step normalized some Arabic forms, removed unnecessary spaces before punctuation, and corrected common output inconsistencies. It also helped separate merged words in some cases. This was important because the main weakness of the system was not always at the character level, but often at the word and spacing level.

A practical challenge appeared during the final inference stage. Running TTA on the full test set increased the total processing time, and the estimated runtime became longer than the remaining time before the submission deadline. To solve this problem, we reduced the beam search width from 4 to 2. This decision reduced the runtime from about 15 hours to about 7 hours, which allowed us to finish the full inference process and submit successfully only a few minutes before the deadline. This step shows that system design in shared tasks is not only about accuracy, but also about efficiency and completion under time constraints.

Overall, our work shows that Arabic manuscript OCR/HTR can benefit from a combination of model selection, repeated training, test-time augmentation, and text post-processing. The results are promising, but they also confirm that word-level recognition remains a major challenge in Arabic handwritten manuscripts. In future work, we plan to improve decoding, use stronger language-aware correction methods, and test more specialized Arabic manuscript models in order to reduce WER further.

Table 1: Final system performance and ranking.

Item	Value
Team Name	Al-Warraq
Final Rank	5
Character Error Rate (CER)	0.1142
Word Error Rate (WER)	0.3780
Line Error Rate (LER)	0.3468
Training Effort	100+ hours

5. Conclusion

In this work, we developed an Arabic manuscript OCR/HTR system for the shared task and explored different ways to improve its performance. During this process, we faced several challenges, including unsuitable models, long training time, hardware limitations, and the difficulty of recognizing Arabic handwritten text from unclear manuscript images. Despite these difficulties, we were able to build a working system through repeated model testing, training, inference-time improvement, and text post-processing.

Our results show that automatic transcription of Arabic manuscripts is possible, but it remains a difficult task, especially at the word level. The system achieved better character-level recognition than word-level recognition, which shows that Arabic manuscript OCR still needs stronger handling of spacing, word formation, and noisy handwriting. Even so, the final system achieved competitive results in the task and demonstrated that practical improvements can be achieved through careful experimentation.

This competition also highlights the importance of digital tools for preserving handwritten heritage. With the help of artificial intelligence, historical Arabic manuscripts can become more accessible, searchable, and usable for future research. In future work, we plan to test more specialized Arabic models and improve the text correction stage in order to further reduce recognition errors and improve transcription quality.

6. Acknowledgments

The authors would like to express their sincere gratitude to the organizers of the NAKBA NLP 2026 for providing this valuable and enriching experience. We also would like to thank Dr. Sarah Ayad for her continuous supervision, support, and guidance throughout this work. In addition, we express our appreciation to the Arab Open University for providing an encouraging academic environment that supported learning, research, and student participation in this competition.

7. Bibliographical References

Carlos Garrido-Munoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. 2025. Handwritten text recognition: A survey. arXiv preprint arXiv:2502.08417.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yi- juan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. arXiv preprint arXiv:2109.10282.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2- vl: Enhancing vision-language model's percep-

tion of the world at any resolution. arXiv preprint arXiv:2409.12191

Sarah Al Hamed, Ahmad Edris Youssef, Aya Hafiz Faris, Alhasan Hamood, Zainab Kamil, Jana Mukhles Alqasem, Adapting Vision-Language and Transformer Models for Automatic Manuscript OCR/HTR of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC), 2026.