Latent Narratives at AR-MS NakbaNLP 2026: Reducing Character Errors in Arabic Manuscript Transcription: A CER Oriented System

Fatima Shaza*, **Sara Al-Desouky***, **Sarah Ayad**
Computer Science Department,
Arab Open University, KSA
24465909KSA@aou.edu.sa, 23463952KSA@aou.edu.sa, s.ayad@arabou.edu.sa
*Equal contribution

Abstract

Historic Arabic handwritten texts present significant challenges due to varied handwriting styles, cursive structure, diverse diacritics, and inconsistent character and word sizes. In this work, we introduce Historic-Arabic-OCR, a vision-language OCR system built upon Qari-OCR, which itself is based on Qwen2-VL-2B-Instruct, and further fine-tuned using Low-Rank Adaptation (LoRA) for Arabic manuscript transcription. The proposed approach incorporates contrast enhancement using CLAHE and deterministic decoding strategies to reduce character-level errors. Our model achieves competitive performance, with a Word Error Rate (WER) of 0.28 and a Character Error Rate (CER) of 0.10 on historical Arabic texts, including low-resolution images. The final submitted system uses CLAHE preprocessing with deterministic greedy decoding to minimize character-level errors.

Keywords: Arabic OCR, Vision-Language Models, Qwen2-VL, LoRA, CER Optimization

1. Introduction

Arabic Optical Character Recognition (OCR) plays a very important role for the digitization and preservation of Arabic texts. A task that is extremely important for cultural heritage, modern communication and information accessibility for over 400 million Arabic speakers worldwide. However, it comes up with multiple challenges such as complex morphology, contextual variation, ligatures and shaping etc. (Kasem et al., 2025)

This paper introduces Historic-Arabic-OCR, a fine-tuned Visual Language Model (VLM) based on Qwen2-VL-2B-Instruct specifically focused on historic Arabic texts. It was developed iteratively, while testing different configurations and different generation settings.

Our key contribution includes developing a competitive model for historic Arabic texts with diverse writing styles, varying image resolution, and faded ink. We also publicly release our trained model for reproducibility. The code and trained model weights are publicly available at https://github.com/sara-aldesouky/nakba-nlp-ocr-qwen2vl-lora.git and https://huggingface.co/Shaza2004/Historic-Arabic-OCR.

2. Literature Review

Building an accurate OCR system for Arabic is challenging due to its cursive handwriting, diacritics, and variation in character and word shapes (Al-Sheikh et al., 2020). These challenges are amplified in historical texts due to diverse writing styles and inherent script complexity. Additionally, Arabic datasets are less abundant compared to English, making Handwritten Text Recognition (HTR) more difficult (Chan et al., 2024). Despite recent advancements, OCR systems still struggle to produce accurate transcriptions for historical Arabic texts (Althobaiti and Lu, 2017).

OCR technology has evolved from early methods to deep learning architectures (Wasfy et al., 2025). Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely used, with CNNs showing strong performance in visual pattern recognition, achieving 94.9% accuracy on isolated Arabic characters (El-Sawy et al., 2017). Compared to RNNs, CNNs can capture spatial features effectively even with limited data (Alrobah and Albahli, 2022).

However, CNN- and RNN-based approaches often require Large Language Models (LLMs) for post-processing. This limitation led to transformer-based OCR, which enables end-to-end text recognition. Models such as TrOCR combine visual encoding with transformer decoding, improving performance on handwritten text recognition tasks (Li et al., 2023).

Further work, such as the MDLSTM-based approach by Ahmad et al. (Ahmad et al., 2020), highlights the importance of multi-directional feature extraction for Arabic scripts. Additionally, decoding strategies such as beam search and length constraints significantly impact Character Error Rate (CER) and Word Error Rate (WER), particularly for Arabic (Li et al., 2023; Chan et al., 2024).

More recently, Multimodal Large Language Models (MLLMs) have been applied to OCR by combining visual and textual understanding. However, general MLLMs are not optimized for Arabic OCR tasks (Wasfy et al., 2025). Qari OCR, built on Qwen2-VL-2B-Instruct (Wang et al., 2024),

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 275–279

adapts these models for Arabic OCR through fine-tuning, achieving a CER of 0.061 and WER of 0.160 (Wasfy et al., 2025).

Despite these advancements, most OCR and HTR systems fail to address the complexities of historical Arabic manuscripts. The script's cursive nature, positional variation, and reliance on diacritics increase ambiguity. Historical datasets further introduce stylistic diversity and degradation, including faded ink and irregular spacing. As a result, models trained on uniform datasets struggle to generalize, highlighting the need for approaches that account for the variability of historical Arabic texts.

3. Approach

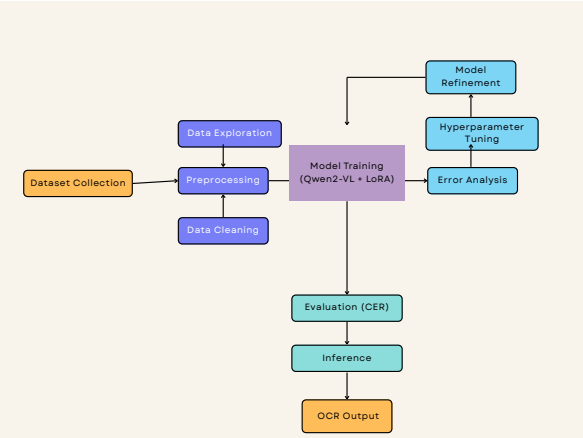

Figure 1: Overview of the proposed Arabic handwritten OCR pipeline, including data exploration, preprocessing, model training using Qwen2-VL with LoRA, evaluation using CER, and inference with iterative model optimization.

As shown in Figure 1, the proposed system begins with dataset collection and preprocessing... The development of our final model involved a two-stage process: first, the application of contrast-aware visual processing to improve image clarity, and second, parameter-efficient fine-tuning of the Qari-OCR model. The main goal was to minimize Character Error Rate (CER) on the official ARMS Subtask 2 dataset, derived from the Omar Al-Saleh manuscript collection.(Hamoud et al., 2026) This corpus consists of approximately 15,962 training, 1,774 development, and 2,671 held-out test lines. This dataset presented many challenges inherent to historical documents, including faded ink, varying illumination, and differing handwriting throughout the corpus. The inherent difficulties of Arabic script, such as cursive connectivity, positional glyph variations, and delicate features like dots and diacritics (*tashkeel*), added to the complexity of the task.

To improve the visibility of faded pages and aged paper, each image underwent Contrast Limited Adaptive Histogram Equalization (CLAHE) with a clipLimit of 2.0 and a tileGridSize of $(8, 8)$. This increased the clarity for minute details such as the dots, *hamza*, and *tashkeel* by reducing character substitution and deletion errors, thus reducing overall CER. We built our model on Qwen2-VL-2B-Instruct using an autoregressive formulation. This allows the model to predict the full conditional sequence likelihood $p(y|x)$, which naturally aligns with CER and WER evaluation metrics. To adapt the general-purpose model to the specifics of ancient manuscripts, we employed Low-Rank Adaptation (LoRA). Low-Rank Adaptation (LoRA) is applied for parameter-efficient fine-tuning using a rank of $r = 16$ and a scaling factor $\alpha = 16$, with no dropout. The adapters are inserted into the transformer's attention and feed-forward projection layers, including query, key, value, and output projections. Only LoRA adapter parameters are updated during training, while the base model weights remain frozen. All experiments were conducted on a single NVIDIA Tesla T4 GPU (16 GB VRAM). The training process was conducted in two different phases for a total of 4 epochs. First, the model was trained on the dataset for 2 epochs, followed by 2 epochs using CLAHE-enhanced images, allowing the model to refine and focus on the intricacies and minute details of the Arabic script. We used the AdamW optimizer with an initial learning rate of 1×10^{-5}, an effective batch size of 16, and a linear decay scheduler to manage the learning rate across the full training duration.

We used the prompt "Return the plain text in this image. Do not add anything" to enforce literal transcription behavior and to prevent any additions, as the model previously attempted to "correct" missing details in the images. We made our model deterministic by employing num_beams=1 (greedy decoding) and do_sample=False with a limit of 128 new tokens. Post-decoding normalization was minimal, including removing whitespace or English characters that were mixed with Arabic. No external dictionaries were used during this process.

4. Results

All experiments use a Qwen2-VL-2B-Instruct backbone fine-tuned via LoRA and are evaluated using Character Error Rate (CER) based on Levenshtein distance. As shown in Table 1, the baseline achieves a CER of 0.12. Incorporating beam search with dictionary correction from training dataset and normalization increases CER to

Table 1: CER results across different configurations.

Configuration	Preprocessing	Post-processing	CER
Fine-tuned model only	None	None	0.12
Beam search + dictionary correction + normalization	None	Dictionary + normalization	0.14
CLAHE + greedy decoding (final)	CLAHE	None	**0.10**

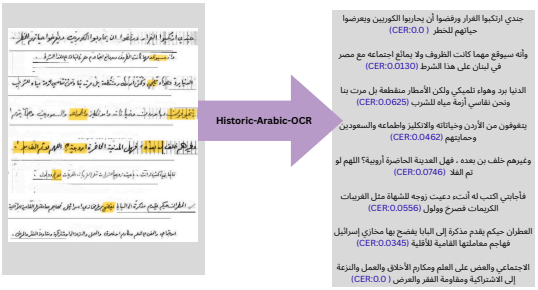

Figure 2: Qualitative comparison of OCR outputs across different configurations. The figure shows the input manuscript and the corresponding predicted text with their Character Error Rate (CER) values.

0.14. The final configuration, using CLAHE preprocessing with deterministic greedy decoding and no lexical correction, achieves the best performance with a CER of 0.10, highlighting the importance of evaluation-time design choices.

The configurations differ in preprocessing and decoding strategies. The baseline uses greedy decoding (num_beams=1, do_sample=False, repetition_penalty=1.05), while the second configuration introduces beam search (num_beams=5) with dictionary-based post-processing. The final configuration applies CLAHE (clipLimit=2.0, tileGridSize=(8,8)) to enhance contrast in degraded regions while omitting semantic post-processing.

Qualitative inspection shows that errors are primarily micro-orthographic, including hamza-related swaps, dot-sensitive confusions (e.g., Jeem/Ha), and terminal-form ambiguities such as Ya variants and Ta marbuta versus Ha. These patterns indicate that while the model captures base glyph structure, it remains sensitive to diacritic and dot-level distinctions in degraded regions.

5. Discussion

Qualitative analysis further elucidates the behavior of our model configurations. As shown in Figure 3, the base model produces several character-level errors. While heavy normalization with dictionary correction modifies the output, it often introduces additional distortions rather than improving accuracy. In contrast, the CLAHE-enhanced model produces more accurate transcriptions, aligning with our quantitative results where the CLAHE-enabled configuration achieved the lowest CER of 0.10. Contrast enhancement helps the model capture subtle character features, particularly in degraded manuscript regions.

Our findings indicate that deterministic greedy decoding outperforms beam search combined with dictionary correction. Because CER penalizes exact character mismatches, lexicon-driven substitutions or normalization often introduce unnecessary edit-distance penalties, suggesting that literal transcription tasks benefit more from deterministic decoding than from semantically motivated corrections. CLAHE preprocessing reduced CER from 0.12 to 0.10 by amplifying local stroke boundaries, reducing ambiguity in dot placement and fine glyph structure, demonstrating that domain-aligned preprocessing can complement model fine-tuning in historical OCR.

Transcription fidelity is critical for archival materials, as micro-orthographic errors—particularly in hamza placement and terminal characters—can fragment named entities and distort corpus analysis. Improving CER enhances the reliability of downstream tasks, including indexing and historiographic study. More broadly, language modeling alone cannot fully resolve fine-grained visual ambiguity in dot-based scripts, highlighting that evaluation-aligned decoding strategies and visually grounded preprocessing remain essential for robust historical manuscript transcription.

Table 2: Top 10 character confusion errors ($n = 500$)

Count	Original		Target
60	ا	→	أ
56	أ	→	ا
29	أ	→	إ
18	إ	→	أ
16	إ	→	ا
15	ا	→	إ
8	ح	→	خ
7	ي	→	ى
7	ه	→	ة
7	ى	→	ي

Errors remain concentrated in heavily degraded

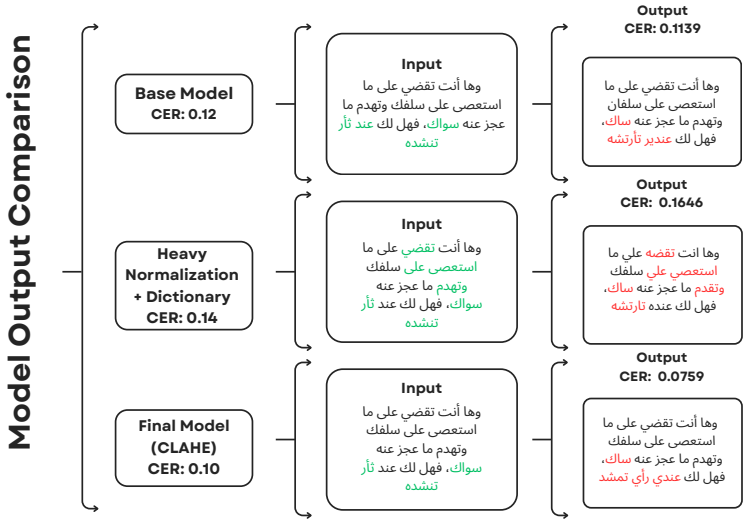

Figure 3: Qualitative comparison of OCR outputs across configurations. Incorrect characters highlight transcription errors.

regions where stroke visibility is reduced. Characters differing only by dot or hamza placement remain particularly challenging.

6. Conclusion

In conclusion, this paper presents the different stages of fine-tuning the Qwen2-VL model and improving CER performance. It also demonstrates that heavy normalization does not always lead to better results, and how this issue was addressed by enhancing image quality using CLAHE. Through these stages, the proposed system surpassed existing open-source solutions in accurately extracting and predicting complex layouts in handwritten Arabic text, while effectively handling diacritics and diverse fonts.

Limitations

Despite the performance of the Historic-Arabic-OCR system, several limitations remain. Errors are primarily concentrated in heavily degraded manuscript regions where ink fading and paper aging reduce stroke visibility. The model also struggles with fine-grained distinctions between characters that differ only in dot or hamza placement.

Additionally, while the fine-tuned Qwen2-VL model performs well on the NakbaNLP dataset, its generalizability to diverse regional calligraphic styles not represented in the training set has not been fully evaluated. The computational requirements of large Vision-Language Models may also limit deployment in resource-constrained environments compared to smaller OCR systems.

Future work may explore improved contrast modeling, dot-aware data augmentation, and hybrid visual-linguistic constraints tailored for historical Arabic scripts to further enhance transcription accuracy.

Acknowledgements

The authors would like to express their sincere gratitude to the Doha Institute for Graduate Studies for organizing such an illuminating experience, which provided us with valuable knowledge and experience.

Ethical Considerations

This work uses publicly available data provided by the shared task organizers. The dataset consists of historical Arabic manuscripts and does not contain sensitive or personally identifiable information. The system is intended to support the digitization and preservation of cultural heritage. OCR errors may occur, particularly in degraded manuscripts, and outputs should be verified before use in critical applications.

References

R. Ahmad, S. Naz, M. Afzal, M. Liwicki, and A. Dengel. 2020. A deep learning based ara-

bic script recognition system: Benchmark on KHATT. *The International Arab Journal of Information Technology*, 17(3).

Saleh Al-Sheikh, Masnizah Mohd, and L. Warlina. 2020. A review of arabic text recognition dataset. *Asia-Pacific Journal of Information Technology and Multimedia*, 9(1):69–81.

Naseem Alrobah and Saleh Albahli. 2022. Arabic handwritten recognition using deep learning: A survey. *Arabian Journal for Science and Engineering*, 47(8):9943–9963.

H. Althobaiti and C. Lu. 2017. A survey on arabic optical character recognition and an isolated handwritten arabic character recognition algorithm using encoded Freeman chain code. In *Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Adrian Chan, Anupam Mijar, Mehreen Saeed, Chau-Wai Wong, and Akram Khater. 2024. Hatformer: Historic handwritten arabic text recognition with transformers. *arXiv preprint arXiv:2410.02179*.

A. El-Sawy, M. Loey, and H. El-Bakry. 2017. Arabic handwritten characters recognition using convolutional neural network. *WSEAS Transactions on Computer Research*, 5(1):11–19.

Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. Nakba nlp 2026: Shared task on arabic handwritten manuscript understanding (palestine memory–omar al-saleh memoir). In *Proceedings 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)*, Palma, Mallorca, Spain.

Mahmoud Salaheldin Kasem, Mohamed Mahmoud, and Hyun-Soo Kang. 2025. Advancements and challenges in arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 58(4).

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

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 perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-OCR: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.