Misraj AI at AR-MS NAKBA-NLP 2026: A State-of-the-Art VLM in Arabic Handwritten Text Recognition

Khalil Hennara, Muhammad Hreden, Zeina Aldallal, Sara Chrouf, Safwan AlModhayan
Misraj AI
Khobar, Saudi Arabia
{hennara, hreden, aldallal, sara.chrouf, safwan}@misraj.ai

Abstract

Handwritten Text Recognition (HTR) for Arabic presents unique challenges due to the script's cursive nature, varying writer styles, and morphological complexity. While modern Vision-Language Models (VLMs) have significantly advanced document parsing, their direct application to highly specific cursive domains requires strategic adaptation. This paper details our submission to the Nakba OCR competition, which adapts a 3B-parameter VLM to recognize historical Arabic manuscripts. We employ a progressive training pipeline that utilizes domain-matched data augmentation to bridge the gap between standard printed Arabic OCR and historical handwritten manuscripts. Moving beyond standard decoder-only Supervised Fine-Tuning (SFT), we fine-tune the entire encoder-decoder architecture using differential learning rates. This approach, followed by a final checkpoint merge, allows the model to better resolve the fine visual details of cursive Arabic script. Our final unified model (submitted under the team name **Misraj AI**) establishes a new state-of-the-art (SOTA) on the Nakba dataset, achieving a Word Error Rate (WER) of 0.24 and a Character Error Rate (CER) of 0.08, and officially securing first place on the leaderboard.

Keywords: Arabic OCR, Handwritten Text Recognition, Vision-Language Models, Model Merging, Supervised Fine-Tuning

1. Introduction

Optical Character Recognition (OCR) for handwritten Arabic text presents unique challenges compared to printed text or Latin-based scripts. The cursive nature of Arabic, varying writer styles, and the presence of diacritics demand robust models capable of both precise visual feature extraction and deep linguistic understanding (Salaheldin Kasem et al., 2025; Lorigo and Govindaraju, 2006). Recently, Vision-Language Models (VLMs) have demonstrated state-of-the-art capabilities across a variety of multimodal tasks, moving beyond traditional machine learning architectures for OCR (Hennara et al., 2025; Team et al., 2025; Wu et al., 2025; Li et al., 2025; Comanici et al., 2025). In this work, we detail our submission to the Nakba OCR competition for Arabic handwritten text. We hypothesize that a pre-trained VLM, specifically tailored for Arabic text processing, can be effectively adapted for challenging handwritten domains through strategic data augmentation and progressive fine-tuning. Our contributions are threefold:

1. We demonstrate the efficacy of a two-stage Supervised Fine-Tuning (SFT) pipeline utilizing an Arabic handwritten dataset as an intermediary domain-adaptation step before fine-tuning on the Nakba dataset.
2. We show that unfreezing the vision encoder and training the full encoder-decoder architecture with differential learning rates yields a significant $> 5\%$ improvement over standard decoder-only tuning.
3. We apply checkpoint merging techniques to smooth the loss landscape, yielding our final state-of-the-art results (WER:0.24, CER:0.08) as shown in Table 1.
4. We open-source our adapted model weights [1] and inference pipeline[2] to support the broader research community and facilitate further advancements in handwritten Arabic OCR.

Team	CER	WER
Misraj AI	**0.079**	**0.244**
Oblevit	0.0925	0.3268
3reeq	0.0938	0.2996
Latent Narratives	0.105	0.3106
Al-Warraq	0.1142	0.378
Not Gemma	0.1217	0.3063
NAMAA-Qari	0.195	0.5194
Fahras	0.2269	0.5223
baseline	0.3683	0.6905

Table 1: Official Nakba Competition Leaderboard Results

2. Literature Review

The landscape of Optical Character Recognition (OCR) has undergone a paradigm shift in re-

[1] https://huggingface.co/Misraj/Baseer_Nakba

[2] https://github.com/misraj-ai/Nakba-pipeline

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 196–200

cent years, transitioning from traditional cascaded pipelines that rely on separate text detection, segmentation, and recognition modules to end-to-end multimodal architectures (Wei et al., 2024).

2.1. Vision-Language Models for Document AI

Vision-Language Models (VLMs) have recently set new benchmarks in document understanding by directly aligning visual features with autoregressive text generation (Wei et al., 2024). Frontier closed-source models, such as Gemini 2.5, have demonstrated remarkable zero-shot accuracy in reading dense text, interpreting complex spatial layouts, and extracting structured data without the need for traditional OCR engines (Comanici et al., 2025).

In the open-source domain, models like Qwen3-VL have significantly advanced the field by introducing dynamic resolution processing and robust spatial understanding, enabling the model to handle multi-lingual OCR tasks (Wu et al., 2025). Furthermore, specialized frameworks have been developed to handle the strict structural demands of document parsing. For instance, HunyuanOcr offers layout-preserving capabilities that faithfully convert complex PDFs into structured Markdown (Team et al., 2025). Similarly, models in the Monkey-OCR family employ advanced cross-modal alignment techniques specifically optimized for text-rich images, significantly reducing hallucination rates in document-based Question Answering (QA) (Li et al., 2025).

2.2. Advancements in Arabic OCR

Despite the massive leap in general VLM capabilities, handwritten Arabic text remains a notoriously difficult domain. (Salaheldin Kasem et al., 2025).

To address these challenges, a prominent line of work has converged on fine-tuning general-purpose vision-language models on targeted Arabic datasets, with each effort emphasizing a different facet of the problem. Some approaches prioritize diacritized text recognition and font diversity, as seen in Qari (Wasfy et al., 2025), while others focus on dialect-specific orthographic conventions, such as AtlasOCR (Imane Momayiz, 2025), which targets Moroccan Darija. Further along this spectrum, Baseer (Hennara et al., 2025) extends the scope beyond character-level transcription toward holistic Arabic Document Information Extraction, jointly addressing layout, context, and linguistic structure.

2.3. Positioning Our Work

Although models like Gemini 2.5 and Qwen3-VL provide immense general capabilities, there is a clear gap in applying modern VLM architectures to the highly specific domain of cursive Arabic Handwritten Text Recognition (HTR). Our work directly addresses this by adopting the Arabic-centric model as our foundation. Rather than relying on zero-shot capabilities, we demonstrate that a progressive, multi-stage fine-tuning curriculum incorporating domain-matched augmentation and advanced encoder-decoder tuning is essential to unlock SOTA performance on the Nakba dataset.

Metric	NTD	NDD	MD
Total Samples	15,962	2,095	21,196
Unique Characters	137	120	176
Avg. Text Length (Chars)	56.79	57.43	43.71
Mean Aspect Ratio	11.75	11.84	8.50

Table 2: Statistical comparison of datasets used for training.

3. Methodology

3.1. Data Acquisition

The development of our Arabic HTR system leveraged a combination of archival competition data and a large-scale enhancement dataset consisting of historical manuscripts.

3.1.1. Nakba Dataset (Competition Data)

The primary source of data is the Nakba dataset (**ND**), provided as part of the AR-MS NAKBA NLP 2026 shared task (Zaraket et al., 2026). See the shared task overview paper (Hamoud et al., 2026). This dataset consists of high-resolution line images extracted from the *Omar Al-Saleh* memoir collection. As an archival collection, it presents challenges typical of 20th-century historical documents, including varying ink density and age-related paper degradation.

The Nakba training dataset (**NTD**) contains 15,962 valid text samples, while the Nakba development set (**NDD**) includes 2,095 samples. The images are characterized by an extreme mean aspect ratio of approximately 11.75 (614×52 pixels).

3.1.2. Muharaf Dataset

To further enhance the model's ability to generalize across different cursive styles, we incorporated the Muharaf dataset (**MD**) (Saeed et al., 2025), a collection of 21,196 valid samples from historical manuscripts. Muharaf includes a wider variety of calligraphic styles (predominantly Ruq'ah). This external data provides critical exposure to older and more varied handwriting compared to the relatively consistent style of the **ND**. The statistical breakdown of *ND and MD* datasets is summarized in Table 2.

3.2. Data Pre-processing

A selective pre-processing strategy was applied to maintain computational efficiency while preserving the integrity of the competition data. While the original **ND** images were utilized in their provided format, the **MD was converted to grayscale**. This ensures that the supplementary data matches the tonal distribution and visual complexity of the competition set as illustrated in Figure 1.

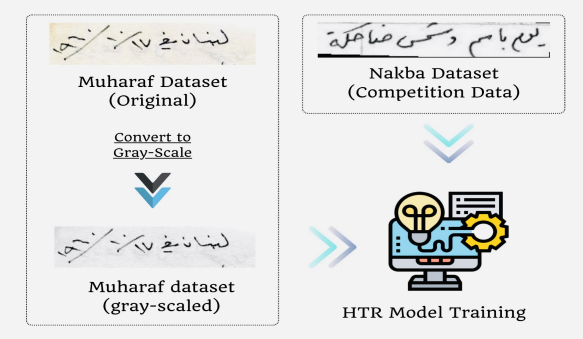

Figure 1: Visual comparison of the Nakba and Muharaf datasets, illustrating the grayscale preprocessing applied to the Muharaf enhancement data.

3.3. Proposed Method

Our methodology adapts a pre-trained Vision-Language Model (VLM) for cursive Arabic handwritten text recognition using a progressive pipeline.

3.3.1. Supervised Fine-Tuning (SFT) Approaches

Our core approach relies on adapting the VLM via standard autoregressive next-token prediction. Given an input image and a text prompt, we apply prompt masking so the cross-entropy loss is calculated exclusively on the generated transcription tokens. This guides the LLM decoder to map visual features to linguistic representations. We investigate two distinct SFT paradigms:

- **Decoder-Only SFT:** The vision tower is frozen, and only the text decoder weights are updated. This provides a computationally efficient mechanism for domain adaptation.
- **Full Encoder-Decoder Tuning:** To better align the vision encoder with the fluid nature of handwritten characters, we unfreeze the vision encoder. To prevent catastrophic forgetting of pre-trained visual representations, we employ *differential learning rates*, applying a significantly smaller learning rate to the vision encoder compared to the text decoder.

3.3.2. Checkpoint Merging

Drawing upon extensive prior research that demonstrates the efficacy of weight averaging in neural network optimization (Yang et al., 2026), we utilize checkpoint merging to maximize generalization. Studies have consistently shown that averaging the weights of multiple checkpoints smooths the model's functional loss landscape. By merging the weights of the top-performing epochs from our training runs, we construct a unified model that effectively mitigates local overfitting to the idiosyncrasies of the training data.

4. Experiments and Results

4.1. Implementation Details and Hyperparameters

All supervised experiments were conducted using the 3B-parameter Baseer (Hennara et al., 2025) across 2 NVIDIA H100 GPUs. To ensure fair comparison, training hyperparameters were standardized across all configurations. Models were trained for 5 epochs using the AdamW optimizer with a weight decay of 0.01 and a cosine learning rate schedule. We utilized a batch size of 128, a maximum sequence length of 1200 tokens, and an input image resolution of 644$\times$644 pixels. For decoder-only training, the text decoder learning rate was set to 1e-4. During full encoder-decoder tuning, the text decoder learning rate remained 1e-4, while the vision encoder utilized a reduced learning rate of 9e-6 to facilitate stable visual adaptation.

4.2. Zero-Shot Evaluation of General-Purpose Models

To establish a performance baseline and assess the out-of-the-box capabilities of frontier Vision-Language Models on this specialized task, we conducted a zero-shot trial using Gemini 3.1 Pro (Google, 2026). We evaluated the model's transcription accuracy directly on the *official development set*. The zero-shot approach yielded a Character Error Rate (CER) of 0.22 and a Word Error Rate (WER) of 0.47.

For comparison, our most basic supervised configuration training the 3B-parameter Baseer model for only three epochs exclusively on the Nakba Dataset (**ND**) achieved a CER of 0.19 and a WER of 0.42. Given that our simple, domain-specific SFT baseline comfortably outperformed a massive state-of-the-art general-purpose VLM, we concluded that the morphological complexity and cursive nature of Arabic handwritten text strictly necessitate dedicated fine-tuning. Consequently, we discontinued further zero-shot explorations with other proprietary models.

Architecture	Dataset / Curriculum	CER	WER
Decoder-Only	**ND**	0.19	0.42
	ND + **MD** (Joint)	0.18	0.42
	MD → **ND** (Sequential)	0.16	0.40
Encoder-Decoder	**MD** (Dec) → **ND** (Enc-Dec)	0.12	0.32
	MD (Enc-Dec) → **ND** (Enc-Dec)	0.10	0.29
Checkpoint Merging	**Encoder-decoder (Full) ep1 + ep5**	**0.08**	**0.24**

Table 3: Comprehensive Results of Training Strategies. *Note: **ND** refers to the Nakba Dataset, and **MD** refers to the Muhraf Dataset. All results reflect the performance on the official hidden test set.*

4.3. Decoder-Only SFT Evaluation

In our initial supervised experiments, we froze the vision encoder and tested three data curricula to evaluate the impact of domain-matched augmentation: (a) **ND** only, (b) **ND** and **MD** combined jointly, and (c) a sequential approach training first on **MD**, followed by **ND**.

As shown in Table 3, the data curriculum significantly impacts downstream performance. Training exclusively on **ND** established a baseline CER of 0.19 and WER of 0.42. Joint training yielded only a marginal improvement (CER: 0.18). However, adopting a sequential domain-adaptation approach (**MD** → **ND**) produced our strongest decoder-only results (CER: 0.16, WER: 0.40), confirming that pre-conditioning the language modeling head on structurally similar handwriting is highly beneficial.

4.4. Full Encoder-Decoder Tuning Evaluation

To better account for the structural variations of the script, we transitioned to full-parameter tuning. We initialized the model using the intermediate decoder-only checkpoint trained *exclusively* on **MD**. From this foundation, we unfroze the vision encoder and applied our differential learning rate strategy to train the full architecture on **ND**. We also tested applying this full encoder-decoder recipe sequentially from scratch (**MD** full-tuning → **ND** full-tuning).

This transition provided dramatic performance gains, proving that the vision tower requires domain-specific visual adaptation for cursive Arabic. Initializing with the **MD**-trained decoder and unfreezing the vision tower for **ND** training dropped the WER substantially to 0.32 and CER to 0.12. Applying the recipe sequentially from scratch further accelerated performance, pushing the metrics down to a CER of 0.10 and WER of 0.29 as show in Table 3.

4.5. Final Submission via Checkpoint Merging

During the evaluation of our best full encoder-decoder run on the hidden test set, we observed slight performance variance across different epochs. To stabilize and optimize our final predictions, we applied weight averaging to the Epoch 1 and Epoch 5 checkpoints, which individually exhibited the strongest blind test results.

As shown in Table 1, the results highlight the exceptional robustness of our progressive fine-tuning approach. Our model demonstrated a massive improvement over the competition baseline, reducing the CER by an absolute 28.9% (from 0.3683 down to 0.079) and the WER by 44.6% (from 0.6905 down to 0.244). Furthermore, our system significantly outperformed all other participating teams. We maintained a comfortable lead of over 5.5% in absolute WER compared to the next best submission (3reeq at 0.2996) and over 1.3% in absolute CER compared to the second-place CER score (Oblevit at 0.0925). This substantial performance gap underscores the necessity and effectiveness of combining domain-matched data augmentation with full encoder-decoder VLM tuning for cursive Arabic HTR.

5. Conclusion

In this paper, we presented a highly effective methodology for Arabic Handwritten Text Recognition (HTR) using Vision-Language Models. By leveraging the Baseer 3B model and implementing a multi-staged training curriculum incorporating domain-matched data from the Muharaf dataset, progressive full-parameter tuning with differential learning rates, and checkpoint merging, we achieved state-of-the-art results, securing top rankings in the Nakba OCR competition.

Our findings demonstrate that while frontier general-purpose models provide a strong baseline, the morphological complexities of cursive Arabic necessitate specialized fine-tuning for high-precision tasks. By meticulously engineering the supervised fine-tuning pipeline, we showed that VLMs can be robustly adapted to challenging historical document transcription, offering a stable solution for complex OCR applications.

6. Ethics Statement

All datasets utilized in this study were obtained through the official NAKBA NLP 2026 shared task and public research repositories, and were used strictly in accordance with their respective licensing agreements and ethical guidelines. Furthermore, we acknowledge that our model's capabilities are intrinsically tied to the specific calligraphic styles and historical periods present in the training distribution. As such, the system may exhibit representation biases if applied to documents from different demographic groups, regions, or eras. We strongly advocate for responsible deployment and emphasize the necessity of human-in-the-loop verification when applying this technology to broader historical archiving efforts.

Bibliographical References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google. 2026. Gemini (3.1 pro). Large language model.

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 memoryâ€“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.

Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Baseer: A vision-language model for arabic document-to-markdown ocr. *arXiv preprint arXiv:2509.18174*.

Abdeljalil Elmajjodi Haitame Bouanane Imane Momayiz, Soufiane Ait Elaouad. 2025. Atlasocr: Open-source ocr for moroccan darija with vision–language models. https://huggingface.co/atlasia/AtlasOCR.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*.

Liana Lorigo and Venu Govindaraju. 2006. Offline arabic handwriting recognition: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28:712 – 724.

Mehreen Saeed, Adrian Chan, Anupam Mijar, Joseph Moukarzel, Georges Habchi, Carlos Younes, Amin Elias, Chau-Wai Wong, and Akram Khater. 2025. Muharaf: Manuscripts of handwritten arabic dataset for cursive text recognition.

Mahmoud Salaheldin Kasem, Mohamed Mahmoud, and Hyun-Soo Kang. 2025. Advancements and challenges in arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys*, 58(4):1–37.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, et al. 2025. Hunyuanocr technical report. *arXiv preprint arXiv:2511.19575*.

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.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2026. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41.

Fadi Zaraket, Bilal Shalash, Hadi Hamoud, Ahmad Chamseddine, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, and Bernard Ghanem. 2026. Ar-ms: Arabic manuscript understanding. 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.