Not Gemma at AR-MS NakbaNLP 2026: Mubsir OCR: End-to-End Recognition of Arabic Handwritten Text

Ali Adel Sayed Ahmed[1,*], **Mona Khaled Ali**[4,*], **Mohamed Emad Sayed**[3,*], and **Ibrahim Nasser** [2,*]

[1,2,3]Helwan University, Ain Shams University, AAST University, Cairo, Egypt
[4]Mansoura University, Mansoura, Egypt
[*]{aliadelsayed2003, monakhaledd55, mohamedemad66553, ibrahim.nasser.2322}@gmail.com

Abstract

Historical Arabic handwritten OCR is difficult because of cursive script, fine diacritics, mixed numerals, and degraded media; classical segmentation pipelines compound errors, whereas end-to-end vision-language models can adapt when fine-tuned on in-domain data.

We present **Mubsir OCR**, a systematic evaluation on the NAKBA dataset: an annotated set (15,962 training line crops and 2,095 val lines with ground truth, used for all nine experiments) and a separate blind AR-MS (Subtask 2) set (2,671 images; scores only via official submission). We compare external vs. in-house VLMs (Qwen2.5-VL-3B, Qwen3-VL-4B-Instruct, Gemma3), inference backends (vLLM/bf16 vs. HuggingFace/bf16), training length (16 vs. 32 epochs), and test-time preprocessing (CLAHE+unsharp). Best on the annotated val set: **8.59% CER** / **25.87% WER** (HuggingFace bf16); the same configuration attains **11.00% CER** / **31.26% WER** on the blind set. Domain-specific fine-tuning beats general-purpose checkpoints; preprocessing helps only marginally and is not recommended without train-time augmentation.

Source code: NAKBA Mubsir OCR

Keywords: Arabic OCR, historical documents, vision-language models, fine-tuning, NAKBA dataset, character error rate, word error rate.

1. Introduction

Optical Character Recognition for historical Arabic handwritten documents remains one of the most challenging problems in document analysis. Arabic script is fully cursive, with each letter taking one of four context-dependent forms (initial, medial, final, and isolated). Diacritics (harakat) are semantically meaningful yet visually fine-grained, and historical manuscripts frequently mix Eastern Arabic and Western numeral systems. Physical degradation — fading ink, paper yellowing, variable calligraphy, and uneven illumination — further compounds these difficulties.

Classical OCR pipelines based on binarization, segmentation, and character classification degrade sharply on historical material because each stage accumulates errors and handwritten Arabic offers few reliable segment boundaries. End-to-end vision-language models (VLMs) bypass explicit segmentation; when fine-tuned on in-domain data they capture corpus-specific script styles that off-the-shelf models miss.

This paper reports nine controlled experiments (January–February 2026) on the NAKBA corpus: external vs. in-house checkpoints, training duration (16 vs. 32 epochs), inference backend and precision (vLLM/bf16 vs. HuggingFace/bf16), architecture (Qwen3-VL-4B vs. Gemma3), and test-time preprocessing (CLAHE+unsharp). All use the same 2,095-image annotated val set.

2. Related Work

Arabic handwriting recognition datasets. Muharaf (Saeed et al., 2024) is the largest public corpus of historic Arabic handwriting (1,600+ page images: personal letters, diaries, church and legal records), with line transcriptions and polygon annotations supporting line- and page-level HTR. AR-MS (Subtask 2) (Zaraket et al., 2026) is concurrent work on Arabic manuscript understanding in NAKBA-NLP 2026.

Vision-language models for OCR. Qwen3-VL (Bai et al., 2025) stacks DeepStack ViT features with interleaved MRoPE and reports state-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 269–274

of-the-art results on document-understanding benchmarks; the 4B dense model balances accuracy and compute. Gemma3 (Gemma Team, 2025) is Google's multimodal family (1B–27B) with high local-to-global attention for long-context KV caches. Both support instruction-following fine-tuning for OCR-style tasks.

Inference serving. vLLM (Kwon et al., 2023) uses PagedAttention to limit KV-cache fragmentation and deliver 2–4$\times$ throughput gains; bf16 cuts memory at modest precision cost. Sherif community checkpoints (Sherif, 2025a,b) show Qwen2.5-to-Qwen3-VL transfer for Arabic handwriting.

Prompt optimization. GEPA (Agrawal et al., 2025) is a genetic-Pareto prompt optimizer that beats GRPO by up to 20% with up to 35$\times$ fewer rollouts—useful for automated prompt search in OCR.

3. Dataset and Experimental Setup

3.1. The NAKBA Dataset

The NAKBA shared task (Hamoud et al., 2026) provides two distinct image sets. The **annotated set** (released at competition start for training and self-evaluation) contains 15,962 training line crops and **2,095 JPG val line crops** with human-annotated ground truth in `annotations_test.csv`. All nine experiments in Table 1 are evaluated on this annotated val set; none of the reported CER/WER values are computed on an internal training split.

The separate **blind set** of **2,671 images** (JPG and PNG) is the official AR-MS (Subtask 2) competition set; ground truth is not released and scores are obtained only via submission; widths range from short fragments to full-sentence lines. The language is primarily Arabic, with occasional Latin annotations, Eastern Arabic numerals, and date strings.

Images are used at native dynamic resolution; internal resizing is handled by the model's default `AutoProcessor` (i.e., no custom resizing or preprocessing is applied), ensuring consistent preprocessing across most experiments. In-house models are trained on the merged NAKBA training and validation splits.

3.2. Evaluation Metrics

Character Error Rate (CER) measures the Levenshtein edit distance at character level between the model hypothesis and the reference transcription, normalized by the total number of reference characters:

$$\text{CER} = \frac{S + D + I}{N_{\text{ref}}} \times 100\,\%$$

where S, D, I are character-level substitutions, deletions, and insertions. CER is micro-averaged across the test set (total edits divided by total reference characters), giving longer lines proportionally more weight and avoiding bias from short-line outliers. For Arabic, single-character errors such as a missing dot (distinguishing, e.g., *ba* from *ya*) or a misread diacritic change word meaning.

Word Error Rate (WER) applies the same Levenshtein framework at the word level, with whitespace tokenization used consistently for both hypothesis and reference. WER is more sensitive to word-boundary errors (spurious spaces, merged tokens) and to entirely dropped or hallucinated words. In Arabic OCR, WER is typically higher than CER because a single character-level error can invalidate an entire word token. No case normalization or punctuation stripping is applied; metrics reflect raw model output.

3.3. Models

Qwen2.5-VL-3B (external baseline). Sherif (2025b), fine-tuned on general Arabic and English handwriting via Qwen2.5-VL-3B, is the domain-mismatch baseline.

Qwen3-VL-4B-Instruct (Bai et al., 2025) is our primary in-house model, with dynamic-resolution vision encoding, DeepStack ViT feature integration, and strong multilingual instruction-following. We fine-tune it on NAKBA data for 16 epochs (Exp. 2) and 32 epochs (Exp. 3, 4, 6, 7, 8, 9) using the standard instruction prompt. All fine-tuning runs share the same training/validation split, data format (image path + transcription), and generation parameters (near-zero temperature).

Fine-tuning details. All Qwen3-VL-4B runs use **full parameter fine-tuning** (no LoRA/PEFT) with AdamW (learning rate 2×10^{-5}, cosine schedule, 0.1 warmup ratio, weight decay 0.01), batch size 2 with gradient accumulation 8 (effective batch 8), max sequence length 512, bf16 mixed precision with gradient checkpointing, on a single NVIDIA H100 GPU.

Gemma3 (Gemma Team, 2025). Fine-tuned on the same NAKBA training data under identical conditions (Exp. 5). Gemma3 uses a different vision encoder and language backbone for architectural comparison.

3.4. Inference Backends

vLLM (Kwon et al., 2023) processes images in batches using PagedAttention memory management and bf16 weight quantization. Total inference time for 2,095 (val set) images on a single H100 GPU is on average 830 s (≈0.4 s/image).

HuggingFace Transformers runs images sequentially, one at a time, in full bf16 precision. Total inference time is on average 6,846 s (≈3.3 s/image), ≈8× slower than vLLM. Sequential processing avoids variable-length padding artefacts that arise in batched inference.

Backend-controlled comparison (Exps. 6 vs. 9). Experiments 6 and 9 use the same checkpoint, instruction prompt, and decoding settings (temperature $= 0.0$, max new tokens $= 512$, greedy decoding); only the inference backend differs (vLLM/bf16 vs. HuggingFace/bf16).

4. Results

Table 1 reports CER and WER for all nine experiments on the 2,095-image annotated val set. The best self-evaluated result is **8.59% CER** / **25.87% WER** (Exp. 9, HF bf16); this configuration was used for our official AR-MS (Subtask 2) blind test submission, yielding **11.00% CER** / **31.26% WER** on the 2,671-image blind set. The best vLLM result is **10.49% CER** / **29.02% WER** (Exp. 6).

4.1. Domain Adaptation and Training Duration

Sherif (2025b) on the full val set (Exp. 1) gives 20.20% CER and 51.52% WER, the worst result overall.

Sixteen epochs of in-house NAKBA fine-tuning (Exp. 2) reduce CER to 12.34% and WER to 30.66%. Doubling to 32 epochs (Exp. 4) further reduces CER to 11.43%, and Exp. 6 reaches 10.49% CER / 29.02% WER. The model had not converged at 16 epochs; further gains may be possible beyond 32 epochs (e.g., with early stopping).

4.2. Prompt Sensitivity

Experiments 3 and 4 use the same 32-epoch Qwen3-VL-4B checkpoint but differ in the prompt used during training (modified vs. standard). The CER difference is 0.26 pp (11.69% vs. 11.43%) and the WER difference is effectively zero (29.19% vs. 29.23%). The standard instruction prompt is robust; prompt engineering is not a bottleneck for this task.

4.3. Backend and Numerical Precision

Experiment 9 evaluates the same 32-epoch Qwen3-VL-4B checkpoint in HuggingFace Transformers (bf16) instead of vLLM bf16. CER drops from 10.49% (Exp. 6, best vLLM) to 8.59%, a gain of 1.9 pp; WER drops from 29.02% to 25.87%, a gain of 3.15 pp. Two mechanisms explain this: (i) *Quantization errors* — bf16 reduces the dynamic range available for probability estimates, disproportionately affecting fine-grained Arabic distinctions such as diacritics and dotting patterns; bf16 preserves the model's full numerical range. (ii) *Batching artefacts* — vLLM pads sequences to uniform length within each batch, which can subtly alter attention patterns for shorter images; sequential single-image inference avoids this effect. Wall-clock time is $8\times$ longer (6,846 s vs. 830 s)—suitable for final evaluation but costly for rapid iteration.

4.4. Architecture Comparison

Gemma3 fine-tuned on NAKBA (Exp. 5) achieves 11.10% CER / 29.80% WER via vLLM, comparable to Qwen3-VL-4B at 32 epochs. Gemma3's vLLM inference takes ≈2,175 s (≈2.6× slower than Qwen3-VL-4B), and WER is 0.78 pp higher. Qwen3-VL-4B offers the better accuracy–speed tradeoff. Gemma3 was evaluated on the 2,671-image blind set.

Table 2 compares the two backends on the same 32-epoch Qwen3-VL-4B checkpoint.

4.5. Preprocessing Ablation

Experiments 7 and 8 form a controlled pair for image preprocessing: identical checkpoint and vLLM backend, with (Exp. 8) and without (Exp. 7) CLAHE contrast enhancement and unsharp masking. The *aggressive* variant (v1) applied CLAHE with 8×8 tiles (clipLimit 2.0) and strong unsharp masking (radius 1, 120%, threshold 2) with JPG re-encoding at quality 75. On ≈50 px-tall line crops, 8×8 tiles are only ≈6 px per side, distorting fine Arabic strokes such as dots and small connecting ligatures; 848 images worsened and 466 improved, so the net effect was negative and v1 was abandoned. The *conservative* variant (v2)

#	Experiment	Backend	Preproc.	CER %	WER %	Notes
1	Qwen2.5-VL-3B (Sherif, 2025b)	vLLM	None	20.20	51.52	—
2	Qwen3-VL-4B (16 ep in-house)	vLLM	None	12.34	30.66	—
3	Qwen3-VL-4B (32 ep, alt. prompt)	vLLM	None	11.69	29.19	—
4	Qwen3-VL-4B (32 ep, std. prompt)	vLLM	None	11.43	29.23	—
5	Gemma3 fine-tuned (32 ep)	vLLM	None	11.10	29.80	$\approx 2.6\times$ slower
6	Qwen3-VL-4B (32 ep)	vLLM	None	10.49	29.02	—
7	Qwen3-VL-4B (32 ep, ctrl.)	vLLM	None	11.15	28.60	—
8	Qwen3-VL-4B + CLAHE+Unsharp v2	vLLM	CLAHE+Unsharp	10.18	28.51	Marginal improvement
9	Qwen3-VL-4B (32 ep, HF bf16)	HF	None	**8.59**	**25.87**	Best; $\approx 8\times$ slower

Table 1: All nine experiments on the 2,095-image annotated val set. Exp. 9 (bold) is overall best on the annotated set and is the configuration used for the official AR-MS (Subtask 2) blind test submission (CER 11.00% / WER 31.26% on 2,671 images). Exp. 6 is best under vLLM.

Backend	Precision	CER %	WER %	Speed
vLLM (Exp. 6)	bf16	*10.49*	*29.02*	$\approx$0.4 s/img
HuggingFace (Exp. 9)	bf16	**8.59**	**25.87**	$\approx$3.3 s/img
HF gain over vLLM		-1.90 pp	-3.15 pp	$8\times$ slower

Table 2: Backend comparison on the same checkpoint (Qwen3-VL-4B, 32 ep). Bold = highest accuracy; italic = best throughput option.

uses 16×16 tiles (clipLimit 1.0), gentler unsharp (radius 1, 80%, threshold 3), JPG quality 95, and a small-image skip rule ($h < 45$ px or $w < 150$ px).

v2 yields a modest CER improvement (11.15% $\rightarrow$ 10.18%) and a negligible WER improvement (28.60% $\rightarrow$ 28.51%), but the gain is inconsistent across images. The fine-tuned model was never trained on CLAHE-enhanced or sharpened images, so test-time preprocessing introduces train–test distribution mismatch. **We do not recommend preprocessing by default.** Future work should consider training on augmented data (raw + enhanced) or employing learned test-time normalisation to improve robustness without mismatch.

Parameter	v1 (aggressive)	v2 (conservative)
CLAHE clipLimit	2.0	1.0
CLAHE tileGridSize	(8, 8)	(16, 16)
Unsharp radius / %	1 / 120%	1 / 80%
Unsharp threshold	2	3
JPG re-encode quality	75	95
Tiny image skip	No	Yes ($h<45$ or $w<150$ px)

Table 3: Preprocessing parameter comparison. v1 degraded net performance; v2 was used in Exp. 8.

5. Conclusions

We presented a systematic nine-experiment OCR study on the NAKBA Arabic historical handwriting dataset (2,095 line crops with human annotations, fine-tuned on 15,962 training and validation crops). Findings: **(1) Domain fine-tuning is indispensable** — the external 3B checkpoint reaches 20.20% CER; in-house Qwen3-VL-4B fine-tuned for 32 epochs on NAKBA reduces CER to 8.59%, an improvement of 11.61 pp absolute. **(2) Training length matters** — moving from 16 to 32 epochs yields a further 0.9–1.8 pp CER reduction. **(3) Precision beats throughput for accuracy** — bf16 sequential inference improves CER by 1.9 pp and WER by 3.15 pp over bf16 batched inference, at $8\times$ the wall-clock cost. **(4) Test-time preprocessing without training-time**

augmentation is not recommended — conservative CLAHE+Unsharp v2 gives a marginal 0.97 pp CER gain but introduces train–test mismatch. **(5) Qwen3-VL-4B outperforms Gemma3** on this corpus in both WER and inference speed. We release inference scripts and configurations for reproducibility and benchmarking on Arabic historical corpora such as Muharaf (Saeed et al., 2024).

Ethical Considerations

This paper studies handwritten OCR on public NAKBA data without human subjects or personally identifiable information, following standard responsible-use norms.

6. References

6.1. Bibliographical References

Agrawal, L. A., Tan, S., Soylu, D., Ziems, N., Khare, R. et al. (2025). GEPA: Reflective prompt evolution can outperform reinforcement learning. *arXiv:2507.19457*. https://arxiv.org/abs/2507.19457

Bai, S. et al. (2025). Qwen3-VL Technical Report. *arXiv:2511.21631*. https://arxiv.org/abs/2511.21631

Gemma Team (2025). Gemma 3 Technical Report. *arXiv:2503.19786*. https://arxiv.org/abs/2503.19786

Hamoud, H., Chamseddine, A. A., Shalash, B., Abid, F. B., Jarrar, M., Chakra, C. A., Ghanem, B., & Zaraket, F. A. (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 LREC 2026*. Palma, Mallorca, Spain, May 2026.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. In *Proc. SOSP*, pp. 611–626. https://arxiv.org/abs/2309.06180

6.2. Language Resource References

Saeed, M., Chan, A., Mijar, A., Moukarzel, J., Habchi, G., Younes, C., Elias, A., Wong, C.-W., & Khater, A. (2024). Muharaf: Manuscripts of Handwritten Arabic Dataset for Cursive Text Recognition. *NeurIPS 2024 Datasets and Benchmarks Track*. https://arxiv.org/abs/2406.09630

Sherif (2025a). Arabic-English-handwritten-OCR-Qwen3-VL-4B [Fine-tuned model, HuggingFace Hub]. https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-Qwen3-VL-4B

Sherif (2025b). Arabic-handwritten-OCR-4bit-Qwen2.5-VL-3B-v3 [Fine-tuned model, HuggingFace Hub]. https://huggingface.co/sherif1313/Arabic-handwritten-OCR-4bit-Qwen2.5-VL-3B-v3

Zaraket, F., Shalash, B., Hamoud, H., Chamseddine, A., Abid, F. B., Jarrar, M., Chakra, C. A., & Ghanem, B. (2026). AR-MS: Arabic Manuscript Understanding Dataset (Subtask 2, Nakba-NLP 2026 Shared Task). *Proc. 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026*, Palma, Mallorca, Spain.

7. Appendices

7.1. The Value of Domain-Specific Fine-Tuning

The gap between the best external checkpoint (20.20% CER, Exp. 1) and the best in-house model (8.59% CER) is an 11.61 pp improvement attributable to NAKBA-specific training data. Sherif (2025b) still reaches 20.20% CER on this corpus because historical manuscripts differ from modern handwriting in ink, vocabulary, ligature density, and physical condition.

7.2. Precision–Throughput Tradeoff

The 1.9 pp CER gain from bf16 sequential inference (Exp. 9 vs. Exp. 6) is substantial. For interactive annotation, vLLM's $\approx$0.4 s/image latency is preferable; for batch archival transcription, HuggingFace bf16 is better suited.

7.3. Residual Errors and Error Analysis

At 8.59% CER / 25.87% WER, errors remain substantial; the CER-to-WER ratio ($\approx$1:3) suggests many whole-word substitutions from ambiguous or partially visible glyphs. Three main error types: (1) *Diacritic confusion* — the model frequently omits or misplaces harakat (short vowel marks). Diacritics such as fatha, kasra, and sukun are fine-grained single-pixel marks; the 4B model's visual patch tokenizer loses some spatial detail at this scale. (2) *Digit script substitution* — Eastern Arabic (Indic) and Western Arabic (European, 0–9) numerals are confused in both directions, particularly for isolated digits, reflecting mixed numeral usage in the original documents and inconsistent annotation conventions. (3)

Ligature splitting/merging — uncommon Arabic ligatures, especially in older calligraphic styles, are occasionally split into two character hypotheses or two adjacent characters merged into one; these errors disproportionately affect WER since they typically corrupt an entire word token.

Future directions include: (i) training beyond 32 epochs with validation-based early stopping; (ii) scaling to larger model variants (Qwen3-VL-8B or -32B); (iii) joint training on raw and augmented images to mitigate preprocessing distribution shift; (iv) post-correction re-scoring using an Arabic language model; (v) automated prompt optimization via GEPA (Agrawal et al., 2025); and (vi) cross-corpus evaluation on Muharaf (Saeed et al., 2024) to assess generalisation.