# Ketaba-OCR at AR-MS NakbaNLP 2026: Efficient Adaptation of Vision-Language Models for Handwritten Recognition

**Hassan Barmandah**[1,2], **Fatimah Emad Eldin**[1,3], **Khloud Al Jallad**[1,4], **Omer Nacar**[1,5]
[1] NAMAA Community, [2] Umm Al-Qura University, [3] Trouve Labs
[4] Syrian Society for Startups and Research, [5] Tuwaiq Academy
s445001043@uqu.edu.sa, Fatimah@trouve.works,
khloud.aljallad@syssr.org, o.najar@tuwaiq.edu.sa

**Abstract**

This paper presents Ketaba-OCR-LoRA, a system developed for the NakbaNLP 2026 Shared Task on Arabic Manuscript Understanding (Subtask 2), which targets the transcription of the historically significant Omar Al-Saleh Memoir Collection written in Ruq'ah and Naskh scripts. We propose a parameter-efficient adaptation of a publicly available pretrained Arabic-English Handwritten Text Recognition (HRT) model, originally trained on handwritten corpora including the Muharaf dataset. Instead of adapting general Vision-Language Models from scratch, we fine-tune the HRT backbone using Low-Rank Adaptation (LoRA) and 4-bit quantization (QLoRA), reducing memory requirements from 40GB to approximately 8GB. Our final submission combines multiple model variants through a novel Linear+Boost weighted ensemble strategy. Our approach achieves a CER of 0.0819 and WER of 0.2588 on the blind test set (per-line evaluation), ranking **1st** on per-line evaluation; on the official corpus-wide leaderboard, we rank **3rd** (CER 0.0938, WER 0.2996). This work demonstrates that specialized pretrained HRT models substantially outperform general-purpose Vision-Language Models for Arabic manuscript transcription, and that parameter-efficient fine-tuning provides a practical and reproducible approach for low-resource cultural heritage digitization.

**Keywords:** Arabic HRT, OCR, LoRA, Fine-tuning, Handwritten Text Recognition, Ensemble

## 1. Introduction

The NakbaNLP 2026 Shared Task (Subtask 2) addresses a critical challenge in Digital Humanities: automating the transcription of historical Arabic manuscripts. The task evaluates systems on the Omar Al-Saleh Memoir Collection (1951–1965), which comprises over 1.5 million words of culturally significant Palestinian testimonies (Hamoud et al., 2026). The manuscripts are written in historical Ruq'ah and Naskh scripts, for which robust OCR is essential for scholars and cultural preservation.

Rather than training a general-purpose Vision-Language Model (VLM) from scratch, our core strategy relies on parameter-efficient transfer learning. We utilize a specialized Handwritten Text Recognition (HRT) backbone (Sherif, 2025) pretrained on diverse datasets like Muharaf (Aladmani et al., 2024). To adapt this foundation to the specific calligraphy of the memoirs without prohibitive computational costs, we fine-tune the model using Low-Rank Adaptation with 4-bit quantization (QLoRA) (Hu et al., 2022; Dettmers et al., 2023). Our experiments yielded the following key outcomes:

- **Ranking & Performance:** Our system ranks **1st** on per-line evaluation (CER 0.082, WER 0.259) and **3rd** on the official corpus-wide leaderboard (CER 0.0938, WER 0.2996).
- **HRT vs. Generalist VLMs:** Specialized, Ketaba-OCR-LoRA models drastically outperform zero-shot generalist VLMs, which struggle with historical Arabic handwriting.
- **Parameter Efficiency:** QLoRA efficiently bridged the domain gap, reducing CER from 0.58 to 0.08 with minimal computational overhead.
- **Ensemble Innovation:** Our Linear+Boost weighting strategy improved CER by 7.4% over standard inverse-CER weighting.

Our model weights and source code are publicly available.[1]

## 2. Background and Related Work

Arabic Handwritten Text Recognition (HRT) is inherently complex due to its cursive script, context-dependent allographs, and diacritics (Wasfy et al., 2025). While early encoder-decoder models like TrOCR (Li et al., 2023) struggle with degraded historical manuscripts, recent state-of-the-art methods rely on synthetic data to overcome resource scarcity. For instance, Wasfy et al. (2025) introduced *Qari-OCR* using diverse synthetic fonts, and Hennara et al. (2025) developed *Baseer* (Hennara et al., 2025) to advance low-resource Arabic document understanding. Concurrently, Large Multimodal Models (LMMs) like Qwen-VL (Bai et al., 2023) offer powerful end-to-end vision-language generation. Multilingual models like GLM-OCR (Zai-org, 2026)

[1] Model weights and code: https://huggingface.co/HassanB4/Ketaba-OCR-LoRA



*Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026*, pages 160–170
11 May 2026. 

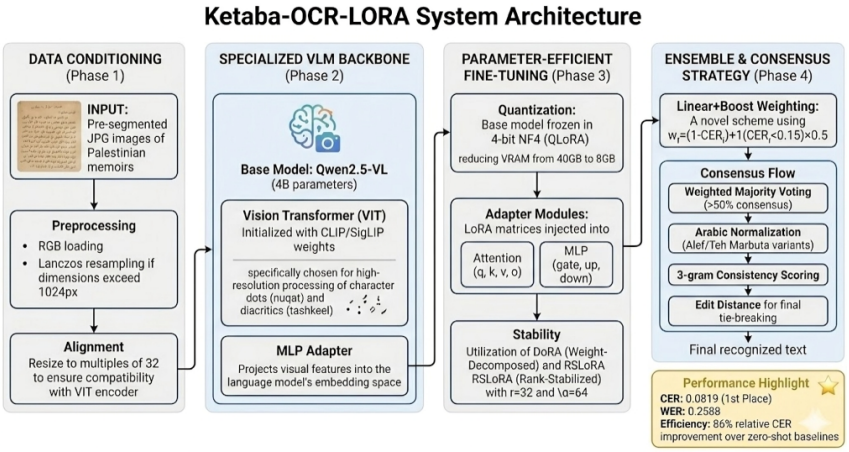



Figure 1: Overview of the Ketaba-OCR-LoRA pipeline.

also exist. However, without domain adaptation, these generalist models often hallucinate modern orthography when processing ambiguous historical texts.

### 2.1. Task Setup and Data

We participate in the NakbaNLP 2026 Shared Task (Subtask 2: Automatic Manuscript OCR) (Hamoud et al., 2026), which focuses on transcribing the Omar Al-Saleh Memoir Collection (1951–1965) in historical Ruq'ah and Naskh script. The input consists of pre-segmented JPG line images, and the target output is a character-level Unicode transcription that preserves orthography and diacritics (tashkeel). The shared-task dataset includes approximately 15,962 training samples, 1,774 development samples, and 2,095 test samples. In this paper, the 2,095 figure refers to the test split used during development, while the official shared-task ranking was computed on a final blind CodaBench evaluation with 2,671 hidden test images. Evaluation is based on Character Error Rate (CER) and Word Error Rate (WER).

Our submission is fine-tuned only on the official shared-task data. The Muharaf Dataset (Aladmani et al., 2024), the IAM Handwriting Database (Marti and Bunke, 2002), and other handwritten corpora mentioned with the Arabic-English HRT backbone (Sherif, 2025) were used only for backbone pretraining.

## 3. System Overview

Our proposed system, Ketaba-OCR-LoRA, is built upon Arabic-English HRT model based on Qwen2.5-VL (Sherif, 2025). The overall system design is illustrated in Figure 1. This model was selected for its high-resolution visual encoder, essential for capturing Arabic diacritics (tashkeel) and letter dots (nuqat).

### 3.1. Data Preprocessing

We preprocess manuscript line images before feeding them to the model. Each image is loaded in RGB; if either dimension exceeds 1024 pixels, we downscale it so that the longer side is capped at 1024 using Lanczos resampling. We then resize the image so that both dimensions are multiples of 32, as required by the vision encoder.

No additional augmentation or normalization is applied; transcriptions are used as provided.

### 3.2. Base Model: Arabic-English HRT based on Qwen2.5-VL

This 4B-parameter model (Bai et al., 2023) was fine-tuned on handwritten datasets including the Muharaf Dataset (Aladmani et al., 2024) and IAM Handwriting Database (Marti and Bunke, 2002). Its Vision Transformer encoder and MLP adapter provide robust priors for Arabic and Latin handwritten text recognition.

### 3.3. Low-Rank Adaptation (LoRA)

To efficiently adapt the pretrained HRT model to the Al-Saleh manuscript style without catastrophic forgetting, we employ LoRA (Hu et al., 2022). Instead of updating all model parameters $W$, we decompose the update $\Delta W$ into two low-rank matrices:

$$W' = W + \Delta W = W + BA \quad (1)$$

where $W$ is frozen in 4-bit precision (NF4 format, QLoRA) (Dettmers et al., 2023). We use rank $r = 32$ with scaling factor $\alpha = 64$, along with DoRA (Weight-Decomposed Low-Rank Adaptation) (Liu et al., 2024) and RSLoRA (Rank-Stabilized LoRA) (Kalajdzievski, 2023) for improved training stability.

| System / Configuration | Test Set CER ↓ | Test Set WER ↓ | Blind Test Set CER ↓ | Blind Test Set WER ↓ |
|---|---|---|---|---|
| ***1. Baselines*** | | | | |
| Organizer Baseline | 0.5840 | 0.8810 | 0.5910 | 0.8850 |
| ***2. Fine-Tuned Models (Non-Merged LoRA)***† | | | | |
| Ketaba-OCR-LoRA:Qwen2.5-VL† | 0.0810 | 0.1150 | 0.0884 | 0.2699 |
| **Ketaba-OCR-LoRA + Ensemble (Ours)**⋆ | – | – | **0.0819** | **0.2588** |
| ***3. Fine-Tuned Models (with Optimizations)***§ | | | | |
| Ketaba-OCR-LoRA:Qwen2.5-VL Improvement1§ | – | – | 0.1133 | 0.3104 |
| Fine-Tuned QARI-3 | 0.2782 | 0.5814 | 0.2635 | 0.5521 |
| ***4. Other / External Models*** | | | | |
| Arabic OCR 4bit Qwen2.5-VL-3B-v2 | – | – | 0.3234 | 0.6203 |

Table 1: Comparative results for fine-tuned models and external systems. Our row reports **per-line** CER/WER (1st on per-line); official corpus-wide ranking is 3rd (Table 2). Additional baseline experiments are in Table 6 (Appendix G).

LoRA adapters are injected into attention projections (q, k, v, o) and MLP layers (gate, up, down projections). We optimize with AdamW (Loshchilov and Hutter, 2019) using the Hugging Face Transformers library (Wolf et al., 2020).

## 4. Experimental Setup

### 4.1. Baseline Models

We evaluated two primary approaches:

1. **Zero-Shot Inference:** We tested inference without fine-tuning using Sherif's pretrained Arabic-English HRT model (Sherif, 2025) (`sherif1313/Arabic-English-handwritten-OCR-v3`[2]) based on Qwen2.5-VL (Bai et al., 2023). For comparison, we also evaluated general-purpose Vision-Language Models including Qwen2.5-VL directly (Bai et al., 2023), QARI-OCR (Wasfy et al., 2025), and GLM-OCR (Zai-org, 2026). A comprehensive list of all baseline models tested is provided in Table 7 (Appendix G.1).
2. **LoRA Fine-Tuning (Ours):** Fine-tuning Sherif's HRT model using the hyperparameters detailed in Appendix D. We used `peft` (Mangrulkar et al., 2022) for adapter injection and `bitsandbytes` (Dettmers et al., 2022) for quantization.

### 4.2. Evaluation Metrics

The main evaluation metrics are Character Error Rate (CER) and Word Error Rate (WER), detailed in Appendix A.

### 4.3. Ensemble Strategy

Our final submission employs a weighted ensemble combining predictions from six model configurations. After testing 30 weighting strategies, we found that a **Linear+Boost** scheme outperformed traditional inverse-CER weighting.

**Linear+Boost Weighting:** Instead of $w_i = \frac{1}{\text{CER}_i}$, we use:

$$w_i = (1 - \text{CER}_i) + \mathbf{1}[\text{CER}_i < 0.15] \times 0.5 \quad (2)$$

This applies a linear decay based on CER, plus a bonus of 0.5 for models with CER below 0.15. This gives weights [0.24, 0.24, 0.14, 0.14, 0.13, 0.12] for our six models, providing a more balanced distribution than inverse-CER while still favoring top performers.

The full ensemble algorithm—including weighted majority voting, Arabic normalization, n-gram consistency, and edit distance consensus—is detailed in Appendix E.

## 5. Results

### 5.1. Official Leaderboard and Per-Line Results

The shared task reports two evaluation schemes: **corpus-wide** (entire test set as a single sequence), used for the official ranking, and **per-line** (CER/WER computed per example then averaged). Our team ranks **3rd** on the official (corpus-wide) leaderboard (CER 0.0938, WER 0.2996) and **1st** on per-line evaluation (CER 0.0819, WER 0.2588; Table 1). The full leaderboards (corpus-wide and per-line) are in Appendix B (Table 2 and Table 3).

[2] https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-v3

As shown in Table 1, our fine-tuned system significantly outperforms all baselines, confirming the superiority of specialized HRT adaptation. Additional evaluations are in Appendix G.

### 5.2. Error Analysis

Detailed error analysis, including visual comparisons and architectural behavior across varying sequence lengths, is in Appendix F.

### 5.3. Ablation Study

To evaluate the impact of our parameter-efficient fine-tuning, we compared general-purpose VLMs against specialized Arabic OCR systems (Table 8).

Our zero-shot inference study (Appendix C) highlighted a significant domain gap (0.2032 CER), confirming the necessity of domain-specific adaptation. Base VLMs yielded over 40,000 substitutions, while MSA-centric models (Qari series) suffered catastrophic over-segmentation ($>$200,000 insertions). Our LoRA adaptation reduced these errors by an order of magnitude (Table 8).

#### 5.3.1. The Role of Ensembling

Applying our Linear+Boost ensemble strategy provided a critical final performance boost. It improved the blind test CER from 0.0884 (achieved via standard inverse-CER weighting) to 0.0819, representing a 7.4% relative improvement by effectively bypassing the generative hallucinations of individual models. More details on the discrepancy analysis of the blind test set, and how ensembling actively corrects catastrophic looping and contextual drift, are provided in Appendix H (specifically Table 9 and Table 10).

## 6. Discussion

Specialized pretrained HRT models substantially outperform general-purpose VLMs for Arabic manuscripts, as they explicitly encode morphological priors and possess invariances to noise, fading, and ink bleeding—properties absent in models trained on clean internet images. QLoRA provides a practical adaptation paradigm, achieving 86% relative CER improvement (0.58 to 0.08) with minimal overhead (8GB vs. 40GB). Our Linear+Boost ensemble strategy, outperforming 29 alternative configurations including inverse-CER and exponential decay, demonstrates the importance of balanced weighting over extreme disparities.

Limitations include system specialization to the Al-Saleh memoirs with unexplored generalization to other historical periods, and disproportionately high WER (0.26) versus CER (0.08) due to Arabic agglutination complexity. Future work should address morphologically-aware postprocessing and constrained decoding against historical dictionaries.

## 7. Conclusion

We presented Ketaba-OCR, which ranks 1st on per-line evaluation (CER 0.0819, WER 0.2588) and 3rd on the official corpus-wide leaderboard (CER 0.0938, WER 0.2996) at NakbaNLP 2026, by fine-tuning a specialized pretrained HRT model via QLoRA and combining predictions through our novel Linear+Boost ensemble strategy. Our key contributions are: (1) demonstrating that parameter-efficient adaptation of domain-specific HRT models substantially outperforms generalist VLMs for historical Arabic manuscripts; (2) introducing a Linear+Boost weighting scheme that outperforms traditional inverse-CER ensemble methods by 7.4%; and (3) providing a systematic comparison of 30 ensemble configurations.

While our open-source system advances digital humanities—specifically enabling the preservation of historical Palestinian narratives—it remains optimized for the Al-Saleh memoirs. Consequently, it requires further adaptation for other calligraphic styles, and historians should validate transcriptions prior to scholarly use. Ultimately, this work provides a reproducible, computationally efficient blueprint for digitizing low-resource cultural heritage texts.

## Ethics Statement

This work supports the digitization of historical Arabic manuscripts and the accessibility of culturally significant Palestinian testimonies (Omar Al-Saleh Memoir Collection), with potential benefit for scholars and cultural preservation. The shared task data consist of historical testimonies; consent and curation are the responsibility of the task organizers, and we use the data only as provided. Automated transcriptions may contain errors; we recommend validation by domain experts before scholarly or public use. Our model weights and code are released to support reproducibility (see footnotes).

## Acknowledgements

We thank the NakbaNLP 2026 organizers for access to the Omar Al-Saleh Memoir Collection. We acknowledge the Arabic-English HRT model based on Qwen2.5-VL (Sherif, 2025), and the Hugging Face community for PEFT and bitsandbytes libraries.

## References


Mehreen Aladmani, Erin Woertz, David Juen, Aashish Agarwal, Stephen Houdek, Gerik Jager, Razan Abuaita, and Nizar Habash. 2024. Muharaf: Manuscripts of handwritten Arabic dataset for cursive text recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6651–6661, Torino, Italy.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.

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.

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*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

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.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32100–32121. PMLR.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Urs-Viktor Marti and Horst Bunke. 2002. The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Ahmed Sherif. 2025. Arabic-english handwritten ocr v3. https://huggingface.co/sherif1313/Arabic-English-handwritten-OCR-v3. Vision-Language Transformer model fine-tuned for Arabic and English handwritten text recognition. Apache-2.0 License. Accessed: 2026-01-10.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zai-org. 2026. Glm-ocr: A multimodal large language model for ocr and document understanding. https://github.com/zai-org/GLM-OCR. Accessed: 2026-02-19.

# A. Appendix: Evaluation Metrics

$$\text{CER} = \frac{\text{Levenshtein}(\text{Pred}, \text{Ref})}{\text{Length}(\text{Ref})} \tag{3}$$

Word Error Rate (WER) assesses transcription usability:

$$\text{WER} = \frac{S + D + I}{N} \tag{4}$$

where $S$, $D$, $I$ are substitutions, deletions, insertions, and $N$ is reference word count.

# B. Appendix: Official Leaderboards (Corpus-Wide and Per-Line)

The shared task reports both corpus-wide and per-line CER/WER. Table 2 gives the official (corpus-wide) leaderboard; Table 3 gives the per-line leaderboard (CER/WER computed per example and averaged). Ketaba-OCR (ours) ranks 3rd on corpus-wide and 1st on per-line (Table 1).

# C. Appendix: Zero-Shot Ablation Study

# D. Appendix: Hyperparameters and Model Details

The hyperparameters and model configuration used for the LoRA fine-tuning experiment are listed in Table 5. These were selected based on extensive hyperparameter search on the validation set, prioritizing Character Error Rate minimization.

## D.1. Key Configuration Justifications

**Choice of Base Model:** Sherif1313's Arabic-English handwritten OCR model was selected because:

1. It is pretrained on Arabic and English handwritten text from multiple sources (Muharaf, IAM, proprietary collections)
2. It exhibits stronger robustness to manuscript-specific degradations compared to general VLMs
3. With 4-bit quantization, the model fits in ∼8GB VRAM despite having 4B+ parameters, enabling efficient fine-tuning on consumer hardware

**LoRA Rank Selection:** Ablation studies indicated that rank 32 with DoRA and RSLoRA provides the optimal trade-off between model capacity and computational efficiency. Ranks below 16 underfit; higher ranks show diminishing returns in CER improvement.

**Quantization Strategy:** 4-bit NF4 quantization via QLoRA reduces memory usage from ∼40GB (full precision) to ∼8GB, enabling training on consumer-grade GPUs while preserving model quality.

# E. Appendix: Ensemble Method Details

Our final winning submission employed our Linear+Boost weighted ensemble strategy. The ensemble achieved CER 0.0819 and WER 0.2588 on the blind test set.

**Models Combined with Linear+Boost Weights:**

- Ketaba-OCR-LoRA: CER 0.09, weight 0.24
- Ketaba-OCR-LoRA variant): CER 0.11, weight 0.24
- Zero-shot HRT with LoRA: CER 0.18, weight 0.14
- Zero-shot HRT baseline: CER 0.20, weight 0.14
- Fine-tuned QARI: CER 0.26, weight 0.13
- Arabic OCR 4-bit: CER 0.32, weight 0.12

**Weight Computation (Linear+Boost):**

$$w_i = \text{normalize}\left((1 - \text{CER}_i) + \mathbf{1}[\text{CER}_i < 0.15] \times 0.5\right) \tag{5}$$

**Why Linear+Boost Outperforms Inverse-CER:**

- Inverse-CER ($w = 1/\text{CER}$) creates extreme weight disparities (best model gets 3.5× weight of worst)
- Linear decay provides smoother distribution, allowing weaker models to contribute diversity
- The boost term (+0.5 for CER $< 0.15$) ensures top 2 performers have sufficient influence
- Combined effect: better balance between trusting top models and leveraging ensemble diversity

**Ensemble Algorithm:**

1. Apply Linear+Boost weights to all model predictions

| Rank | Team | CER ↓ | WER ↓ |
|---|---|---|---|
| 1 | Misraj Ai | 0.079 | 0.244 |
| 2 | Oblevit | 0.0925 | 0.3268 |
| 3 | **Ketaba-OCR (Ours)** | **0.0938** | **0.2996** |
| 4 | Latent Narratives | 0.105 | 0.3106 |
| 5 | Al-Warraq | 0.1142 | 0.378 |
| 6 | Not Gemma | 0.1217 | 0.3063 |
| 7 | NAMAA-Qari | 0.195 | 0.5194 |
| 8 | Fahras | 0.2269 | 0.5223 |
| 9 | baseline | 0.3683 | 0.6905 |

Table 2: Official leaderboard (corpus-wide evaluation; used for ranking). All participating teams and organizer baseline. Ketaba-OCR (ours) ranks 3rd.

| Rank | Team | CER ↓ | WER ↓ |
|---|---|---|---|
| 1 | **Ketaba-OCR (Ours)** | **0.0819** | **0.2588** |
| 2 | Misraj Ai | 0.0895 | 0.2516 |
| 3 | Latent Narratives | 0.1002 | 0.2845 |
| 4 | Oblevit | 0.1049 | 0.3316 |
| 5 | Al-Warraq | 0.1059 | 0.3468 |
| 6 | Not Gemma | 0.11 | 0.3126 |
| 7 | Fahras | 0.1819 | 0.4303 |
| 8 | NAMAA-Qari | 0.2032 | 0.5031 |
| 9 | baseline | 0.2811 | 0.5884 |

Table 3: Per-line leaderboard (CER/WER computed per example, then averaged). Ketaba-OCR (ours) ranks 1st (Table 1).

2. If weighted consensus $>$50% for any prediction, select it
3. Normalize Arabic text (alef variants, teh marbuta) and re-vote
4. Score remaining candidates by 3-gram consistency with other models
5. Final tie-breaking: minimum average edit distance to all candidates

**Ablation of Weighting Strategies (30 configs tested):** Config 18 (Linear+Boost$<$0.15) achieved the best blind test CER of 0.0819, outperforming inverse-CER (0.0884), exponential decay (0.0856), and rank-based methods (0.0871).

# F. Appendix: Error Analysis Details

## F.1. Qualitative Error Analysis

Figure 2 presents representative failure cases highlighting the differences between the zero-shot baseline, QARI-3, and our proposed Ketaba-OCR-LoRA model. In the first example, the ground truth word is العنوان. The zero shot model produces an incomplete prediction لعنو, omitting both the initial definite article ال and the final character indicating insufficient adaptation to Arabic morphological structure. In contrast, both QARI-3 and our proposed method successfully recover the complete word. The second example illustrates a more challenging scenario, where a horizontal artifact partially occludes the upper region of the text. The zero-shot model again fails to reconstruct the word correctly, while QARI-3 produces a near-correct output. Our proposed model accurately recovers the full word despite visual corruption, indicating improved robustness to noise and structural distortions.

Moreover, we expanded the error analysis for long sentences with punctuation. For instance, Figure 3 presents a challenging long-text example containing multiple words, punctuation, and complex Arabic morphology. Such samples require accurate character recognition, proper word boundary modeling, and contextual coherence. The zero-shot baseline shows character-level distortions and incomplete words, while QARI-3 partially improves structural consistency but remains sensitive to noise. In contrast, our proposed model, Ketaba-OCR-LoRA, accurately reconstructs the full sentence, demonstrating enhanced robustness and superior context-aware modeling for long text sequences.

## F.2. Quantitative Error Analysis by Sequence Length

The comparative analysis of Character Error Rate (CER) against sequence length shows distinct ar-

| Model | Test Set | | Blind Test Set | |
|---|---|---|---|---|
| | CER ↓ | WER ↓ | CER ↓ | WER ↓ |
| HRT: Zero-Shot Qwen2.5-VL | 0.1690 | 0.4987 | 0.2032 | 0.5031 |
| HRT: Zero-Shot Qwen2.5-VL (LoRA)[‡] | – | – | 0.1834 | 0.4644 |

Table 4: Zero-shot inference ablation study. [‡]Merged adapters. These results demonstrate the baseline performance of the pretrained HRT model without fine-tuning, highlighting the domain gap for unseen historical manuscripts.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| *Base Model Configuration* | | | |
| Base Model Name | Arabic-English-handwritten-OCR-v3 | Model Architecture | Vision-Language Transformer |
| Base Model Size | ∼4.07B parameters | Pretraining Data | Muharaf, IAM, Custom |
| *Quantization Configuration* | | | |
| Quantization Scheme | 4-bit NF4 (QLoRA) | Compute Dtype | bfloat16 |
| Double Quant | True | – | – |
| *LoRA Adapter Configuration* | | | |
| LoRA Rank ($r$) | 32 | LoRA Alpha ($\alpha$) | 64 |
| Target Modules | q, k, v, o, gate, up, down | LoRA Dropout | 0.05 |
| DoRA | True | RSLoRA | True |
| *Training Hyperparameters* | | | |
| Learning Rate | $2 \times 10^{-5}$ | Optimizer | AdamW (fused) |
| LR Scheduler | Cosine | Warmup Steps | 200 |
| Batch Size | 1 (per GPU) | Gradient Accumulation | 4 |
| Effective Batch Size | 4 | Number of Epochs | 1 |
| Max Gradient Norm | 1.0 | Weight Decay | 0.01 |
| Max Sequence Length | 2048 | Max Image Size | 1024 |
| Evaluation Strategy | steps | Eval Steps | 500 |
| Save Steps | 500 | Files per Chunk | 1000 |

Table 5: Comprehensive hyperparameters for QLoRA fine-tuning of Ketaba-OCR-LoRA. The base model (Arabic-English HRT) was pretrained on diverse handwritten datasets.

chitectural behaviors among the evaluated models (see Figure 4). Ketaba-OCR-LoRA demonstrates stability, maintaining a near-zero CER regardless of text length, indicating robust handling of long-range spatial dependencies. In contrast, the Zero-Shot model demonstrates a notable short-text penalty, where error rates are highest for sequences under 20 characters and improve as length increases, indicating heavy reliance on surrounding linguistic context to disambiguate characters, a luxury not afforded by shorter snippets. Conversely, QARI-3 shows a classic performance decay, with CER scaling positively with sequence length before plateauing between 0.15 and 0.20. This degradation in QARI-3 is likely caused by attention saturation or error accumulation in its recurrent components. Overall, while Ketaba-OCR-LoRA is the most versatile, the results highlight a critical trade-off between contextual reliance in Zero-Shot models and the scaling limitations of the QARI-3 architecture.

## F.3. Model Comparison Analysis

A comparative performance analysis of various models based on Character Error Rate (CER) and Word Error Rate (WER) is presented in Table 6 and Table 8. Qwen2.5-3B and Qwen2.5-7B models significantly outperform the remaining models, maintaining both CER and WER below the 1.0 threshold. Interestingly, while the Qwen series generally shows lower error rates, the Qwen2.5-VL-4B model shows a notable spike in error metrics compared to its smaller counterparts, indicating that model scaling or architectural changes in that specific configuration may have caused degradation.

Among the Qari models, performance is more volatile; Qari-0.3 serves as the strongest baseline in its group, while Qari-OCR-2B and Qari-0.1 show the highest error rates. A key observation is the relationship between character and word accuracy: in most models, WER tracks closely with CER, but Qari-0.1 and Qwen2.5-VL-4B show wider divergence, potentially indicating issues with linguistic consistency or spacing despite relatively better character recognition. Overall, the Qwen2.5 archi-


Figure 2: Qualitative Samples

tecture remains the most robust for this specific task, providing the highest levels of transcription accuracy.

## G. Appendix: Additional Experiments (Test Set Only)

### G.1. Baseline Models

| Model / Configuration | CER ↓ | WER ↓ |
|---|---|---|
| Qwen2.5-VL-7B | 0.6808 | 0.9198 |
| Qwen2.5-VL-3B | 0.6213 | 0.8628 |
| QARI v0.3 | 0.5293 | 0.8772 |
| QARI v0.1 | 0.7127 | 0.9296 |
| QARI-OCR 2B | 0.6840 | 0.9126 |
| GLM-OCR | 0.9999 | 0.9999 |

Table 6: Additional baseline model experiments evaluated on the test set only (blind test set results unavailable). These include general-purpose Vision-Language Models (Qwen2.5-VL), domain-specific Arabic OCR systems (QARI variants), and multilingual models (GLM-OCR).

The high insertion counts for Qari models (>100,000) indicate aggressive over-segmentation, where stylistic ligatures are misinterpreted as word boundaries. Our LoRA-adapted system eliminates these structural hallucinations.

| Model / Configuration | Type |
|---|---|
| Arabic-English HRT (v3) | Specialized HRT |
| Qwen2.5-VL-7B | General VLM |
| Qwen2.5-VL-3B | General VLM |
| QARI v0.3 | Arabic OCR |
| QARI v0.1 | Arabic OCR |
| QARI-OCR 2B | Arabic OCR |
| GLM-OCR | Multilingual LMM |
| Arabic OCR 4-bit Qwen2.5-VL-3B-v2 | Quantized VLM |

Table 7: Comprehensive list of baseline models evaluated. Specialized (HRT) and general-purpose (VLM) models were compared to assess domain adaptation benefits. Detailed performance metrics are provided in Table 1 and Table 4.

## H. Appendix: Blind Set Discrepancy and Hallucination Analysis

Since official ground truth annotations were not released for the blind test set, we constructed a *Pseudo-Ground Truth* (Pseudo-GT) from the ensemble's final predictions via weighted majority voting (Linear+Boost, Section 4.3). We consider this a reliable proxy: our ensemble achieved a corpus-wide CER of 0.0938 on the blind test set, meaning the predictions deviate from the true references by less than 10% at the character level. The resulting Pseudo-GT thus closely approximates the actual ground truth, making it suitable for qualitative failure analysis across individual models.

This analysis reveals systematic failure modes in individual models—specifically generative hallucination loops and contextual drift—and demonstrates the corrective power of ensemble consensus.

### H.1. Catastrophic Hallucinations (Total Disagreement)

Table 9 illustrates the most severe failure modes observed in individual models, specifically generative repetition loops. When the autoregressive decoder loses attention alignment, it collapses into infinite lexical looping (e.g., repeating the word وجنرالهم) or numerical hallucinations (e.g., looping ١١٨). Our ensemble strategy successfully mitigates these errors, relying on the robust consensus of the Linear+Boost weighting to extract coherent sequences.

### H.2. Ensemble Correction in Lower Disagreement

Table 10 demonstrates cases of "Lower Disagreement," where two models agree and one dissents. Here, the standalone HRT_LoRA_Merged model


Figure 3: Qualitative Sample 3

Table 8: Detailed Error Metrics for Baseline Models and Variants

| Model | CER | WER | Subs. | Dels. | Ins. |
|---|---|---|---|---|---|
| Qwen2.5-3B | 0.7868 | 0.9625 | 42,215 | 9,147 | 35,724 |
| Qwen2.5-7B | 0.8052 | 0.9922 | 46,953 | 15,977 | 21,065 |
| Qwen2.5-VL-4B* | 3.7699 | 3.0734 | 56,407 | 30,202 | 48,983 |
| Qari-0.3 | 2.6551 | 1.8641 | 49,762 | 3,053 | 145,225 |
| Qari-0.1 | 4.5836 | 2.7653 | 49,612 | 15,009 | 111,262 |
| Qari-OCR-2B | 4.2499 | 4.5788 | 45,142 | 7,381 | 223,041 |

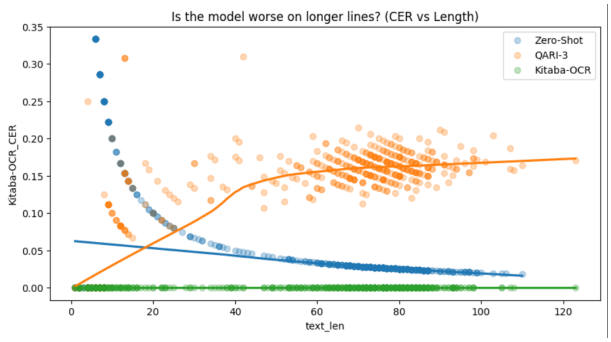



Figure 4: CER vs Sequence Length

occasionally suffers from premature truncation and contextual drift, hallucinating shorter, semantically unrelated phrases. The ensemble leverages consensus to output the structurally accurate transcription.

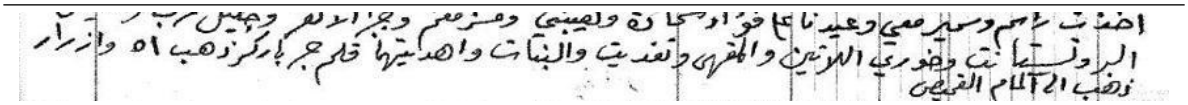
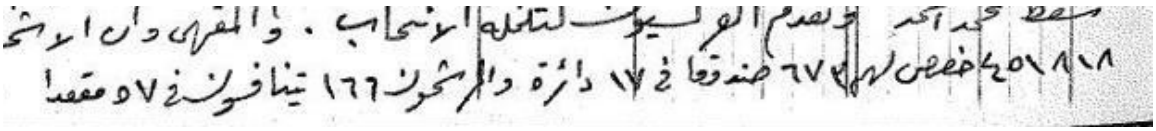
| Model | Prediction (Truncated) |
|---|---|
|  | |
| **Image ID:** `1962_p064_10035` | |
| HRT LoRA Merged | أجداد سَرم وسَمير معي وبعدنا عاقو… |
| Ketaba-OCR-LoRA_1 | ذهب إلى إكمال الغربة وسمر معي… وجنرالهم وجنرالهم… *[Infinite Loop]* |
| Ensemble | اخذت راسم وسمير معني وعيدنا على فؤاد… |
|  | |
| **Image ID:** `1962_p176_10069` | |
| HRT LoRA Merged | ٨٨١ ٣٥،٤ عدد واحد وعدم الموسيقى… |
| Ketaba-OCR-LoRA | ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١ ٨١… *[Numerical Loop]* |
| Ensemble | ٨١ ١٥٤ خصوص لهم ٣٧٦ صندوقا في… |

Table 9: Examples of severe generative hallucinations in single models compared to the stable ensemble output. The autoregressive loops in Submission_1 are completely bypassed by the ensemble consensus.

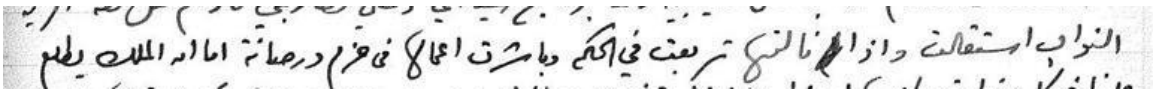
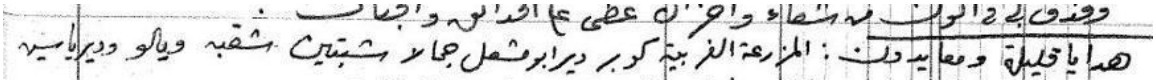
| Source | Text Sequence (Truncated) |
|---|---|
|  | |
| **Image ID:** `1956_p080_10012` | |
| Pseudo-GT | النواب استقالت وأوامرنا لنها تربعت… |
| HRT_LoRA_Merged | النواب استقالت واذا فانها تربعت في… *[Drift]* |
| Ketaba-OCR-LoRA | النواب استقالت وأوامرنا لنها تربعت… |
| Ensemble | النواب استقالت وأوامرنا لنها تربعت… |
|  | |
| **Image ID:** `1962_p077_10032` | |
| Pseudo-GT | وقد قبى وألوس من سكاء واحزن… |
| HRT_LoRA_Merged | وفوق في والونك من سطاء واخزان… *[Truncation]* |
| Submission_1 | وقد قبى وألوس من سكاء واحزن… |
| Ensemble (Ours) | وقد قبى وألوس من سكاء واحزن… |

Table 10: Correction of contextual drift and truncation. When individual models fail to capture the full manuscript line, the ensemble defaults to the more complete, text-anchored sequence.