ArVLM- a framework for Arabic text recognition using Multi-lingual VLMs (Draft, Final version will be available after running all experiments)

Mohamed Wael*, Youssef Ghallab*, Mohamed Ashraf Ragih*,
Omar Khaled Yehia*, Omar Ossama*, Mohamed Mahmoud Salem*
*Computer and Communication Engineering, Alexandria University, Alexandria, Egypt
Emails: es-mohamed.abdelmoneim2025@alexu.edu.eg, es-YoussifGhalab2025@alexu.edu.eg,
es-MohamedAshraf2025@alexu.edu.eg, es-omarkhaledyehia2025@alexu.edu.eg,
es-OmarOsama2025@alexu.edu.eg, es-MohamedMa.Salem2025@alexu.edu.eg

***Abstract*—Multimodal Large Language Models (LLMs) have significantly advanced various tasks, including Visual Question Answering (VQA) and Optical Character Recognition (OCR), due to their ability to understand language rules and patterns through extensive pretraining on vast datasets. However, these models often face challenges stemming from the limited availability of resources for underrepresented languages. Handwritten text recognition is particularly affected, with Arabic handwriting recognition (AHR) presenting unique difficulties due to the complexities of Arabic script, because the Arabic language is morphologically rich and written in a cursive script, characterized by the use of dots and diacritics positioned above and below the characters, and the use of diverse writing styles, including Riq'ah and Naskh fonts. This study compares the performance of the existing MLLMs in Arabic OCR tasks before and after Parameter Efficient Fine-Tuning (PEFT) in terms of recognition accuracy. We also introduce various augmentation techniques to enhance the generalization capability of our framework across both printed and handwritten documents.**

***Index Terms*—LLM, MLLM, OCR, Computer Vision,Arabic Optical Character Recognition (OCR); Arabic OCR software; Arabic OCR evaluation.**

I. Introduction

Optical Character Recognition (OCR) technology plays a pivotal role in digitizing and processing textual information across various languages. The Arabic language, with its complex script and diverse forms, presents unique challenges for OCR systems, particularly when dealing with handwritten and printed text. In this paper, we present the development of an offline OCR system designed specifically for Arabic text, leveraging two key datasets: the KHATT dataset for handwritten Arabic characters and an in-house dataset for printed Arabic scripts. The KHATT dataset, a comprehensive resource for handwritten Arabic script, includes diverse samples captured from a wide range of writers, reflecting the variability in handwriting styles. This diversity introduces significant challenges in terms of recognition accuracy, as handwritten characters can vary greatly in shape and structure. To complement this, we created an in-house Arabic Printed Characters dataset by utilizing the ground-truth text from the KHATT dataset and applying smart augmentation techniques. These augmentations simulate real-world variations in printed text, such as font styles, sizes, distortions, and skewness, which enhances the robustness of our OCR system.

In order to combat these challenges, we propose:

- Novel Augmentation techniques to simulate different document types and shooting conditions.
- A framework for PEFT Fine-tuning VLMs on OCR datasets.
- A python script to perform analysis on full-page handwritten scripts

By combining these two datasets, our OCR framework aims to achieve robust performance across both handwritten and printed Arabic text. We performed a comparative analysis on various MLLMs models and architectures and reported the results. The proposed system is designed to operate offline in the sense that it processes entire documents and scripts at once, rather than providing real-time predictions as text is being hand-written. This approach is suitable for applications where large-scale document processing is required, such as archival systems, document digitization, license plate detection, and automated content creation.

This work aims to fill a gap in an under-researched domain, while there have been plenty of studies on VLMs capabilities to perform OCR, there has been little focus on the field of Arabic OCR, which represents unique challenges due to the scarcity of data resources and the inherent complexity in Arabic scripts.

II. Literature Review

Deep learning approaches have revolutionized the field of handwritten text recognition, offering significant improvements over traditional methods. Notable among these are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have demonstrated high efficacy in capturing spatial and temporal

dependencies within data. Bissacco et al. [1] developed a CNN-based OCR system for recognizing Arabic handwritten characters, achieving an accuracy of 97.8%. Their approach involved a three-layer CNN architecture with max-pooling and dropout, trained using the Adam optimizer. Al-Hajj Mohamad et al. [2] used a Restricted Boltzmann Machine to pre-train a CNN for Arabic handwritten digit recognition, achieving an accuracy of 96.37%. Research by Alqasemi et al. [3] and Altwaijry et al. [4] utilized Generative Adversarial Networks (GANs) to augment training data for Arabic handwriting recognition. Their CNN models trained on both real and synthetic data achieved accuracies ranging from 92.8% to 97.7%. Similarly, Ahmed et al. [5] and Sahlol et al. [6] demonstrated that hybrid CNNRNN models could achieve recognition rates above 94Despite these advancements, challenges remain in achieving high accuracy for AHTR due to the script's inherent complexities. This paper takes a novel approach, leveraging VLMs to develop SOTA AHTR performance.

III. Methodology

Document images are often affected by distortions introduced during real-world processes. For example, folds, wrinkles, or tears can create color shifts and shadows in scanned pages. Variations in printer ink density may cause some parts of a document to appear overly dark or faint. Additionally, human annotations, such as pencil marks or highlighting, add another layer of noise. These imperfections can influence the performance of machine learning tasks that rely on document data. Despite this, OCR applications are required to operate reliably on noisy scanned images. considering the challenges mentioned above, we proposed:

- Three augmentation pipelines to simulate printed document setting, handwritten document setting, folded or wrinkled documents.
- Pre-processing techniques to standardize input to the model

A. Data Augmentation

Data augmentation plays a vital role in enhancing model robustness and preventing overfitting. In our methodology, we implemented a comprehensive pipeline for Arabic text image augmentation, with careful consideration given to dataset splitting to ensure unbiased evaluation. The dataset was initially split into training and test sets prior to any augmentation steps to prevent data leakage and maintain the independence of the evaluation sets. This strategic splitting ensures that the model remains completely blind to test data characteristics during training. Our augmentation pipeline consists of three main branches:

- Camera perspective distortion simulation
- Natural handwriting variation simulation through line deviation
- Printed text simulation using various Arabic fonts

Each branch undergoes affine transformations independently before being merged into the final training dataset. This multi-branch approach ensures diverse and realistic data representation, crucial for developing robust models for Arabic text recognition. The augmented training data was then used to fine-tune Vision Language Models, with subsequent evaluation performed on the pristine test sets to assess real-world performance. figure 1 shows the augmentation pipeline used

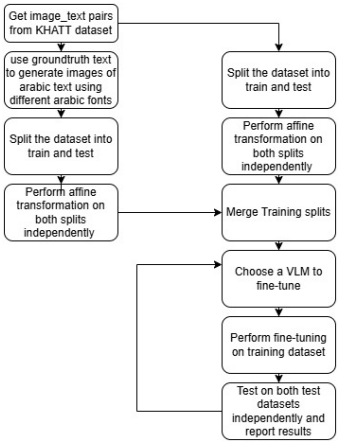


Fig. 1: Augmentation pipeline

1) Augmentation for handwriting: To simulate varying imaging settings, we apply four image augmentation techniques during training: geometric distortion, perspective transformation, affine transformation, and shear transformation. These augmentations increase training data diversity, helping the model learn invariant features for better performance across varying conditions.

A - Geometric Distortion: We apply random rotations within $[-3°, 3°]$ and random scaling along both axes, simulating variations in orientation and size as seen in figure 2.

B- Perspective Transformation: We perturb the four corners of the image to simulate different viewing angles, ensuring the model generalizes across various perspectives.

C- Affine Transformation: A random affine matrix is applied, combining scaling, rotation, and translation to simulate real-world variations due to camera movement or object displacement.

D- Shear Transformation: Random shear factors skew the image horizontally or vertically, simulating deformations caused by viewing angles or acquisition distortions. Since these augmentations involve matrix multiplications, and matrix multiplication is not commutative, different sequences of operations produce different augmentations. To introduce variability, augmentations are applied randomly in each training batch with a randomized order. Additionally, certain augmentations are dropped with a probability of 0.2. This approach ensures that the model learns robust features and generalizes

effectively to unseen data, particularly when the dataset lacks variability or requires handling various perspectives and spatial variations.

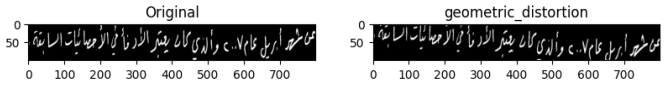


Fig. 2: Geometric distortion augmentation

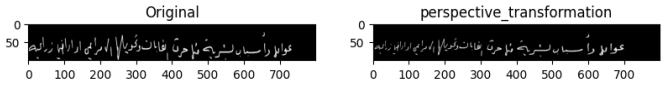


Fig. 3: Perspective transformation

2) Printed documents generation:

a) Font Selection: We chose the *Amiri* font, which is one of the most widely used fonts in printed Arabic documents. It is a classical Naskh style font that has been extensively adopted in printed materials, ensuring high legibility and uniformity. The *Amiri* font is ideal for OCR tasks involving printed Arabic text due to its widespread use in books, newspapers, and digital media.

Additionally, to capture the nuances of handwritten Arabic text, we selected the *Aref Ruqaa* font. The *Aref Ruqaa* font closely resembles Arabic handwriting, making it well-suited for applications that require OCR on handwritten or semi-handwritten documents. The Ruqaa script, as seen in figure 4 and figure 5, is characterized by its simplified and more angular forms compared to the traditional Naskh script, with some unique stylistic features, such as the use of a dash (_) instead of two dots (..) above and under the letters. To further simulate handwritten text, we applied a handwriting augmentation script to 50% of the generated *Aref Ruqaa* font samples. This augmentation process introduces variability in the text, mimicking natural handwriting variations such as irregular spacing, skewing, and distortion. This approach enhances the model's ability to generalize to real-world handwritten documents.

The combination of printed and handwritten font samples in our OCR model, along with the augmentation techniques, aims to improve the recognition accuracy of both printed and handwritten Arabic text, particularly in real-world scenarios where text may vary in style and appearance.

بيان معنوي كافي و لسيدي خلقت عالم عاطل و باطل او لورودي

Fig. 4: Generated text using *Aref Ruqaa*

B. Fine tuning

Fine-tuning large-scale vision-language models (VL models) such as QwenV2-VL-2B, QwenV2-VL-7B, and

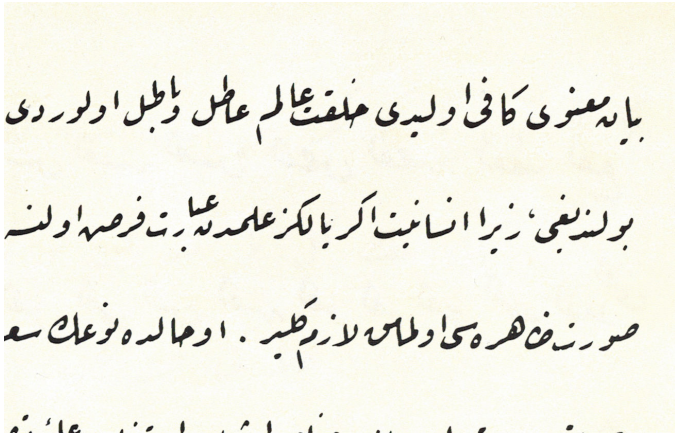


Fig. 5: a handwrittenscript written in Ru'qaa

LLaMA 3.2-VL-11B typically requires significant computational resources, especially on resource-constrained platforms like Google Colab and Kaggle. To address these challenges, we explore the use of Parameter-Efficient Fine-Tuning (PEFT) with Quantized Low-Rank Adapter (QLoRA), a method that allows for efficient adaptation of pre-trained models while minimizing memory consumption and computational complexity, we performed fine-tuning on vision layers, language layers, mlp module and attention module for all models.

1) PEFT with QLoRA: PEFT with QLoRA modifies only a small subset of model parameters, applying low-rank factorization and quantization to adapter layers, additionally, it quantizes model parameters to be represented using only 4 bits. This approach significantly reduces the number of trainable parameters and their size, enabling large models to be fine-tuned efficiently without retraining the entire network. By focusing on adapter parameters, QLoRA reduces memory usage and computational cost, making it feasible to fine-tune large models on platforms with limited resources.

2) Experiments and Results: We compared the fine-tuning performance of three vision-language models—QwenV2-VL-2B, QwenV2-VL-7B, and LLaMA 3.2-VL-11B on KHATT dataset, Qualitative results are shown in figures 6,7 and 8

TABLE I: Model Training Parameters

Parameter	Value
learning_rate	2×10^{-4}
per_device_train_batch_size	16
gradient_accumulation_steps	8
num_train_epochs	1
per_device_eval_batch_size	4

3) Qualitative analysis: In this section we preview some samples of Qwen2VL-2b output before and after fine tuning, it's evident that the fine-tuned model's performance significantly exceeds that of the base model

Ground Truth:

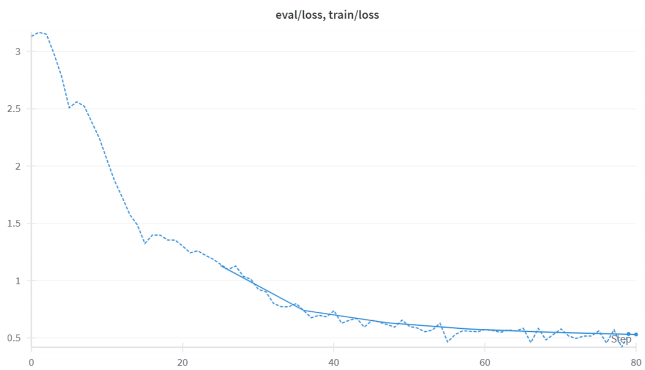


Fig. 6: Training loss(dotted) eval loss(solid) for Qwen 2b

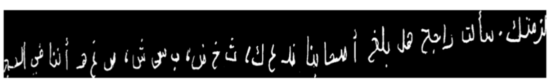
نطاق لفظةْ العربْ اليوم على سكّان بلاد واسعة، يكتبون ويؤلفون وينشرون ويخاطبون

Base Model Prediction:

نطاق لنفيةْ العربْ اليوم على سكان بلد واسع، بيكتون وبيولون وبيشرون وبياسبون

Fine-tuned Model Prediction:

نطاق لنظةْ العربْ اليوم على سكان بلاد واسعة، يكترون ويولون ويثيرون ويهاونون



Ground Truth:

لزمتك. سألت راجح هل بلغ أصحابنا ظ ع ك ، ث خ ض، ب س ش، ص غ هـ أننا في الحج.

Base Model Prediction:

لزمنك، سأنت راجع عل بلخ أبسا بنا ندعك، تجح ش، بس ش، س نم و أبنا في المديح

Fine-tuned Model Prediction:

لزمتك. سألت راجح هل بلغ أصحابنا ظ ع ك ، ث خ ض، ب س ش، ص غ هـ أننا في الحج.

Ground Truth:

إنتقالية ترتفع فيها درجة الحرارة حتى تصل الى درجة الصفر المئوي في مستوياتها العليا . تحتوي

Base Model Prediction:

في تقليدية ترتفع فيها درجة الحرارة حتى تصل إلى درجة الصفر المئوية في مستواها العليا. تحتوي على ترطيب طبيعي

Fine-tuned Model Prediction:

التقليدية ترتفع فيها درجة الحرارة حتى تصل إلى درجة الصفر المئوي في مستواهاتها العليا. تحتوي

Ground Truth:

أطماع الناس قد صارت بمعزل عني وآيسة مني وأني قد أحكمت هذا الباب وأتقنته

Base Model Prediction:

أنا أحبك، أنا أحب

Fine-tuned Model Prediction:

الكلمات التاليه لهذا النص: مشمش، بغلول غي رايق سر وناء الله أخى من هذا المكان أفهمه

Ground Truth:

يلبي. عند وصولنا طفنا و سعينا مع شيخ. كان جاري في الخيمة

Base Model Prediction:

ليلي. حندي وموالنا طفنا وسجنا مع سبع. كان بياري في الرفية

Fine-tuned Model Prediction:

يلبي. عند وصولنا طفنا وسعينا مع شيخ. كان جاري في الخيمة

References

[1] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[2] M. Al-Hajj Mohamad, J. Likforman-Sulem, and C. Mokbel, "Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1165–1177, 2009.

[3] R. Alqasemi, T. Shanableh, and A. R. Al-Ali, "Printed Arabic text recognition using artificial neural networks and hidden Markov models," *Signal Processing: Image Communication*, vol. 21, no. 7, pp. 570–577, 2006.

[4] H. Altwaijry and J. A. Parvez, "Offline Arabic handwriting recognition using a hybrid of multilayer perceptron and hidden Markov models," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2009, pp. 644–648.

[5] S. Ahmed, T. M. S. El-Sheikh, and M. S. El-Soud, "Arabic handwritten character recognition using modified chain code and neural network," in *Proceedings of the IEEE International Conference on Computer Engineering and Systems*, 2007, pp. 221–226.

[6] A. T. Sahlol, A. Y. A. Hegazy, and R. M. Ismail, "A new technique for Arabic handwriting recognition using neural network," in *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, 2007, pp. 1134–1139.

[7] P. Wang *et al.*, "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.
[8] A. Dubey *et al.*, "The LLaMA 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.