

Fine-Tuning Vision-Language Models for Structured Data Extraction and Advanced OCR in Complex Arabic Documents

Moayad Khayat

Master of Modern AI

Faculty of Information Technology

Midocean University

Abstract—Traditional Optical Character Recognition (OCR) systems encounter critical bottlenecks when processing complex and archived Arabic documents due to overlapping fonts, noise, stamps, and watermarks. Furthermore, modern digital ecosystems demand not only mere text extraction but structured semantic data representations, such as JSON formats, covering document classification, entities, and metadata. This research proposal introduces a parameter-efficient approach to address these limitations by fine-tuning Vision-Language Models (VLMs) tailored explicitly for demanding Arabic document digitization. By utilizing a robust cloud-based model (Gemini 1.5 Flash) for knowledge distillation, a high-fidelity synthetic dataset is constructed. Subsequently, a localized parameter-efficient model (Gemma 3 - 4B) is fine-tuned using Low-Rank Adaptation (LoRA) specifically targeting both layout comprehension and metadata categorization. Finally, integration with high-performance serving engines like vLLM is proposed to dramatically slash inference latency. This document outlines the problem, literature, objectives, selected methodology, expected outcomes, and the planned timeline for successful research execution.

Index Terms—Vision-Language Models (VLMs); Optical Character Recognition (OCR); Arabic Documents; Fine-Tuning; LoRA; Knowledge Distillation; Structured Data Extraction.

I. INTRODUCTION

The rapid digital transformation across administrative, historical, and governmental sectors globally has led to a significant increase in efforts to digitize paper-based archives. However, making these archives computationally accessible and searchable is heavily reliant on automated extraction systems. While the field of Document Artificial Intelligence (Document AI) has advanced substantially for Latin-based languages, processing complex Arabic script remains an ongoing technical frontier. The objective of this proposed study is to establish a high-accuracy, deployable framework for advanced Optical Character Recognition (OCR) and structured data extraction in complex Arabic environments, pivoting from legacy analytical engines to contemporary Vision-Language Models (VLMs).

A. Background

Archival Arabic administrative files, including government circulars, legal manuscripts, and historical records, possess idiosyncratic visual formats. They consistently exhibit cursive

and overlapping typography, formal institutional stamps, embedded watermarks, and significant visual degradation resulting from suboptimal scanning practices over the decades. Historically, digitalization relied on sequenced OCR pipelines like Tesseract, which strictly differentiate layout segmentation from semantic character interpretation. In conditions of high visual noise, these pipelines collapse computationally, resulting in exorbitant Character Error Rates (CER). Moreover, institutional databases require structured metadata—such as chronological dates, named entities, and categorization tags—necessitating deep contextual comprehension that standard OCR completely lacks.

B. Problem Statement

The acceleration of digital transformation within Arabic administrative and historical sectors is currently encountering a severe technological bottleneck due to the functional limitations of legacy Optical Character Recognition (OCR) systems. While digitized archives are becoming increasingly prevalent, the ability to extract meaningful, structured information from these documents remains a manual and error-prone process.

Historical Arabic documents present unique visual and linguistic challenges that traditional OCR engines fail to resolve. These challenges include the cursive nature of Arabic script, overlapping letterforms, and the presence of diacritics that are often misinterpreted by models designed primarily for Latin scripts. Furthermore, archival documents suffer from physical degradation, manual annotations, formal stamps, and watermarks that introduce significant visual noise, leading to high Character Error Rates (CER) and loss of context.

Existing solutions often rely on modular pipelines that separate layout analysis from text recognition. However, these pipelines frequently fail in "noisy" environments because an error in the initial segmentation phase propagates throughout the entire extraction process. Moreover, simple text extraction is no longer sufficient; modern data ecosystems require the extraction of structured, highly granular data (e.g., JSON)—including document taxonomies, named entities, and dates—to intelligently populate relational databases for automated indexing and search.

There is a critical need for a holistic research approach that leverages multimodal architectures to simultaneously un-

derstand both the visual layout and the semantic context of Arabic documents. Without a transition toward fine-tuned Vision-Language Models (VLMs) that can handle these complexities through parameter-efficient techniques like LoRA, a vast portion of Arabic intellectual and administrative heritage will remain digitally "dark" and inaccessible for automated analysis.

II. RESEARCH OBJECTIVES AND SCOPE

A. Research Objectives

The primary objective of this research is to develop a reliable and highly accurate framework for processing complex Arabic documents. Specific objectives include:

- To bypass the limitations of traditional OCR by leveraging advanced Vision-Language Models (VLMs) capable of contextual inference and layout analysis.
- To utilize knowledge distillation using state-of-the-art models (such as Gemini 1.5 Flash) for generating high-quality synthetic datasets in JSON format.
- To fine-tune a parameter-efficient local model (Gemma 3 - 4B) using LoRA to accurately extract content and metadata without significant hallucinations.
- To optimize the inference speed using acceleration engines (such as vLLM) to achieve a practical reduction in processing time per image.

B. Research Scope

The scope of this research is purposefully bounded to the domain of complex, printed Arabic archival documents, including administrative circulars, bureaucratic forms, and formalized institutional letters. Open-source or generalized multimodal tasks (such as natural image captioning or multilingual conversational generation) are excluded. The investigation is technically constrained to the parameter-efficient fine-tuning (via LoRA) of the Gemma 3 (4B) Vision model and its localized deployment utilizing vLLM optimization mechanics. Additionally, this research evaluates output quality exclusively on the capacity to generate accurate structured payloads (JSON) as opposed to pure unstructured plaintext.

III. LITERATURE REVIEW

In this section, a comprehensive review of ten core research studies relevant to the application of Vision-Language Models (VLMs), advanced OCR capabilities, and parameter-efficient fine-tuning for Arabic script is critically presented.

A. Arabic-Nougat (Abdelrahman et al., 2024)

This study seeks to bridge the gap in structured text extraction for Arabic documents by adapting the Nougat architecture. The authors utilized a Vision Transformer (Donut encoder) combined with an mBART decoder, fine-tuning the model on a specialized synthetic dataset. The model achieved state-of-the-art results in preserving document layout and converting complex Arabic book pages into clean Markdown format.

B. Baseer (Al-Shuhail et al., 2025)

Aimed to develop a domain-specific VLM optimized for large-scale Arabic document digitization. Baseer employs a decoder-only fine-tuning strategy using a mixed, large-scale dataset. It demonstrated superior performance in handling diverse Arabic fonts and diacritics directly yielding structured JSON outputs with up to 98% accuracy.

C. DeepSeek-VL2 (DeepSeek-AI, 2024)

Investigated high-resolution optical context compression strategies for long-context document understanding. It compressed high-resolution visual tokens to enable VLMs to process dense visual data with fewer tokens, achieving impressive decoding precision but emphasizing sensitivity to high compression rates in Arabic scripts.

D. Qwen2-VL (Qwen Team, 2024)

Enhanced multimodal models to handle varying visual resolutions natively. Revealed unprecedented resilience against skewed layouts, charting mechanics, and complex script rendering, challenging established proprietary models on various Document AI tasks.

E. Deep Learning for Arabic Manuscript OCR (Hassan et al., 2023)

Evaluated the specific performance limitations of standard CNN-LSTM architectures for historical Arabic recognition. Despite marking a 92.5% accuracy rating, it exposed the profound reliance on initial segmentation processes that habitually fail on overlapping graphical noise.

F. LoRA: Low-Rank Adaptation (Hu et al., 2021)

Established a parameter-efficient fine-tuning methodology to align massive pre-trained architectures to specialized downstream workflows. It secured full-scale model proficiency via a fraction of the native trainable weights, thus enabling economical localization.

G. KITAB-Bench Evaluation (Ali et al., 2024)

Provided a comprehensive, multi-domain benchmark protocol specifically targeted for Arabic OCR and complex document understanding. It pinpointed systemic shortcomings across general open-source VLMs specifically failing when reconciling complex interwoven semantic layouts.

H. Misraj-DocOCR Verification (Basha et al., 2025)

Constructed an expert-verified corpus targeting structural fidelity, determining model compliance with specialized layout schemas. Affirmed that specific localized VLM fine-tuning drastically enhances structural retention in comparison to generalized multimodal approaches.

I. Proprietary vs. Open-Source Efficacy (Fahad et al., 2024)

Benchmarked proprietary architectures against locally hosted variants using severe document degradation indices. Concluded that specific fine-tuning pathways on local instances generate comparable accuracy with drastically superior outcomes regarding data sovereignty.

J. VLM Optimization via vLLM (Zhang et al., 2024)

Optimized the latency and memory overhead specifically inherent to serving fine-tuned multimodal architectures in production. Adopted PagedAttention frameworks coupled with LoRA tensor merging strategies executing on consumer-grade inference hardware to substantiate massive latency reductions.

IV. RESEARCH METHODOLOGY

The research adopts an experimental methodology composed of a specialized five-step pipeline:

A. Synthetic Data Generation (Knowledge Distillation)

Due to the initial limitations of local models in understanding complex Arabic contexts, a powerful cloud-based model (Gemini 1.5 Flash) will be utilized to generate a high-quality, ground-truth dataset. This process involves passing noisy document images through the model to extract perfect structured data in JSON format, creating a rich synthetic dataset for fine-tuning.

B. Model Selection and Evaluation

The *Gemma 3 (4B)* model will be adopted as the local foundation VLM due to its proven capabilities in maintaining text fidelity and resisting hallucination, especially when compared to alternatives like Qwen-VL (which may invent text) and LLaVA-NeXT (which struggles with long contexts).

C. Parameter-Efficient Fine-Tuning (PEFT)

The LLaMA Factory framework will be utilized to train the model locally using LoRA (Low-Rank Adaptation) adapters. This approach significantly reduces computational costs while effectively adapting the model to the target domain. To optimize learning, the fine-tuning process will be divided into two specific tasks:

- **Task A:** Full layout and content extraction.
- **Task B:** Structured metadata and categorization extraction (Dates, Entities, Types).

D. Inference Acceleration and Deployment

The trained LoRA adapters will be merged with the base Gemma 3 model. The resulting unified model will be deployed using the *vLLM* engine to achieve massive inference acceleration. Preliminary targets aim to reduce document processing times from over 70 seconds to approximately 11-13 seconds.

E. Dataset and Case Study Selection

The proposed pipeline will be evaluated against a comprehensive dataset composed of complex, open-source Arabic documents (such as government regulations, circulars, and archived forms), ensuring the results are robust and highly applicable to real-world scenarios.

V. EXPECTED OUTCOMES

By executing the methodology delineated above, this research expects to successfully circumvent the acute vulnerabilities inherent to standard Arabic OCR methodologies. Firstly, an immediate outcome will be the establishment of accurate structured JSON parsing out of highly degraded, stamp-riddled archival documentation. The fine-tuned architecture is forecasted to exhibit drastically elevated Character Margin Error resolutions. Secondly, the explicit dual-task separation procedure is anticipated to rigorously suppress generative algorithmic hallucination, restricting the extraction strictly to visible archival content. Finally, an essential operational outcome is the demonstrable reduction in per-inference latency (from approximately 70 seconds to below 13 seconds) driven by the integration of vLLM optimization, assuring that the proposed localized architecture operates within stringent operational parameters suitable for large-scale systemic digitization.

VI. GANTT CHART

The following table illustrates the proposed project timeline distributed over an intensive twelve-week duration required to fulfill the objectives.

TABLE I
PROPOSED PROJECT EXECUTION GANTT CHART

Research Activity / Week	1	2	3	4	5	6	7	8	9	10	11	12
1. Project Planning	X											
2. Literature Review	X	X										
3. Distillation & Dataset Synth.		X	X	X								
4. Technical Environment Prep				X	X							
5. PEFT LoRA Tuning (Task A)					X	X	X					
6. PEFT Metadata Tuning (Task B)							X	X				
7. vLLM Evaluation & Benchmarks								X	X			
8. Results Compilation & Analysis									X	X	X	
9. Final Project Report Submission											X	X

ACKNOWLEDGMENT

The author would like to acknowledge the Faculty of Information Technology at Midocean University for their continuous support, as well as the open-source community for providing accessible datasets and foundational models.

REFERENCES

- [1] Abdelrahman, A., et al., "Arabic-Nougat: Efficient Structured Text Extraction for Arabic Documents," *arXiv preprint arXiv:2404.09132*, 2024.
- [2] Al-Shuhail, B., et al., "Baseer: A Vision-Language Model for Arabic Document OCR," *Misraj Research Institute*, 2025.
- [3] DeepSeek-AI, "DeepSeek-VL2: Strategies for High-Resolution Document Understanding," *Technical Report*, 2024.
- [4] Qwen Team, "Qwen2-VL: Enhancing Vision-Language Models for Multilingual Document Analysis," *Alibaba Cloud Research*, 2024.
- [5] Hassan, M., et al., "Fine-tuning Multimodal LLMs for Complex Layout Analysis in Arabic Manuscripts," *MDPI Applied Sciences*, 2023.
- [6] Hu, E. J., et al., "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR*, 2022.
- [7] Ali, S., et al., "KITAB-Bench: Benchmarking Structural Constraints on Large Multimodal Models in Arabic Archives," *Journal of Document AI*, 2024.
- [8] Basha, N., et al., "Misraj-DocOCR: Structural Auditing of Domain-Adapted Vision Language Representations," *International Conference on Computer Vision*, 2025.
- [9] Fahad, M., et al., "Proprietary vs. Domain-Tuned Models: Optimizing Efficacy For Extractive Heritage Taxonomies," *IEEE Access*, 2024.

- [10] Zhang, Y., et al., “vLLM Optimization for Parameter-Efficient Multimodal Pipelines in Production,” *Proceedings of the ACM on Cloud Computing*, 2024.