

Midocean University
Faculty of Information Technology
Master of Modern Artificial Intelligence (MSMAI)

Assignment 2: Research Proposal Components

Course: Research Methodology in Computer Science (AI800)

Research Title:

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

Student Name: Moayad Khayat

Student ID: 261002503

Instructor: Dr.Mohammed Fawzy

Date: March 1, 2026

1 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, such as government circulars, legal manuscripts, and institutional archives, present unique visual and linguistic challenges that traditional OCR engines, like Tesseract, 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 (Abdelrahman et al., 2024). Furthermore, archival documents suffer from physical degradation, scanning artifacts, 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 (Hassan et al., 2023). Moreover, simple text extraction is no longer sufficient; modern data ecosystems require the extraction of structured data (e.g., JSON) including document types, entities, and dates for automated indexing and search (Al-Shuhail et al., 2025).

There is a critical need for a holistic research approach that leverages multimodal architectures to understand both the visual layout and the semantic context of Arabic documents simultaneously. Without a transition toward fine-tuned Vision-Language Models (VLMs) that can handle these complexities through parameter-efficient techniques like LoRA (Hu et al., 2021), a vast portion of Arabic intellectual and administrative heritage will remain digitally "dark" and inaccessible for automated analysis.

2 Research Objectives

The central goal of this research is to develop a robust, high-precision framework for the structured digitization of complex Arabic archival documents. To achieve this, the following specific objectives have been defined:

- **To leverage** advanced Vision-Language Models (VLMs) to simultaneously decode text and layout features, thereby overcoming the accuracy limitations of traditional OCR in noisy document environments.
- **To implement** a knowledge distillation pipeline using cloud-based models (such as Gemini 1.5 Flash) to generate high-fidelity synthetic datasets tailored for Arabic archival contexts.
- **To fine-tune** a localized, parameter-efficient model (such as Gemma 3 - 4B) using Low-Rank Adaptation (LoRA) to enable accurate structured data extraction from private archives without compromising data sovereignty.
- **To optimize** the inference throughput of the fine-tuned model using high-performance serving engines (like vLLM) to ensure the methodology is computationally viable for large-scale digitization tasks.
- **To evaluate** the effectiveness of splitting training tasks into content extraction and meta-data categorization to minimize hallucinatory outputs in the generated JSON results.

3 Literature Review

In this section, a comprehensive review of six core research papers relevant to the application of Vision-Language Models (VLMs) and advanced OCR techniques for Arabic script is provided.

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

Objectives: This study seeks to bridge the gap in structured text extraction for Arabic documents by adapting the Nougat architecture.

Methodology: The authors utilized a Vision Transformer (Donut encoder) combined with an mBART decoder. They fine-tuned the model on a specialized synthetic dataset, *arabic-img2md*, containing 13.7k pairs of document images and Markdown labels.

Key Findings: The model achieved state-of-the-art results in preserving document layout and converting complex Arabic book pages into clean Markdown format.

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

Objectives: To develop a domain-specific VLM optimized for large-scale Arabic document digitization and knowledge extraction.

Methodology: Baseer employs a decoder-only fine-tuning strategy (based on Qwen2.5-VL) using a mixed dataset of synthetic and real-world Arabic archival documents.

Key Findings: Baseer demonstrates superior performance in handling diverse Arabic fonts and diacritics while producing direct JSON outputs with 98% accuracy.

3.3 DeepSeek-VL2 (DeepSeek-AI, 2024)

Objectives: Investigating high-resolution optical context compression for long-context document understanding.

Methodology: The research introduces "optical context compression," allowing VLMs to process dense documents with fewer tokens via high-resolution vision encoders.

Key Findings: The model achieves high decoding precision (97%) in multilingual settings, though performance is sensitive to extreme compression in Arabic scripts.

3.4 Qwen2-VL (Qwen Team, 2024)

Objectives: Enhancing multimodal models to handle varying resolutions and document types natively.

Methodology: Qwen2-VL introduces Naive Dynamic Resolution to capture spatial features without fixed-size resizing.

Key Findings: The model rivals proprietary models in document understanding, showing high robustness against skewed layouts and handwritten script.

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

Objectives: Assessing the performance of CNN-LSTM architectures for historical Arabic manuscript recognition.

Methodology: The authors proposed a framework utilizing GAN-based data augmentation followed by a CNN-LSTM-CTC pipeline.

Key Findings: The study achieved 92.5% accuracy on manuscripts but revealed heavy dependency on the accuracy of the initial segmentation step.

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

Objectives: Proposing a parameter-efficient method for adapting large-scale pre-trained models.

Methodology: LoRA injects trainable low-rank decomposition matrices while freezing original model weights.

Key Findings: LoRA maintains full-model performance while requiring significantly fewer trainable parameters, enabling local adaptation of VLMs.

4 References

- Abdelrahman, A., et al. (2024). *Arabic-Nougat: Efficient Structured Text Extraction for Arabic Documents*. arXiv preprint arXiv:2404.09132.
- Al-Shuhail, B., et al. (2025). *Baseer: A Vision-Language Model for Arabic Document OCR*. Misraj Research Institute.
- DeepSeek-AI. (2024). *DeepSeek-VL2: Strategies for High-Resolution Document Understanding*. Technical Report.
- Hassan, M., et al. (2023). *Fine-tuning Multimodal LLMs for Complex Layout Analysis in Arabic Manuscripts*. MDPI Applied Sciences.
- Hu, E. J., et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR 2022.
- Qwen Team. (2024). *Qwen2-VL: Enhancing Vision-Language Models for Multilingual Document Analysis*. Alibaba Cloud Research.