IMPLEMENTASI MULTIMODAL *LARGE LANGUAGE MODEL* UNTUK EKSTRAKSI DAN TRANSLITERASI *PRINTED JAWI SCRIPT*

TUGAS AKHIR

Diajukan Oleh:

ZIDAN MUBARAK
220705071

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI AR-RANIRY
BANDA ACEH
2026 M/1448 H

LEMBAR PERSETUJUAN

IMPLEMENTASI MULTIMODAL *LARGE LANGUAGE MODEL* UNTUK EKTRAKSI DAN TRANSLITERASI *PRINTED JAWI SCRIPT*

TUGAS AKHIR

Diajukan Kepada Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh
Sebagai Salah Satu Beban Studi Memperoleh Gelar Sarjana (S1)
Dalam Ilmu Teknologi Informasi

Oleh:
ZIDAN MUBARAK
NIM. 220705071

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui Untuk Dimunaqasyahkan Oleh:

Pembimbing I,	Pembimbing II,
Bustami, M.Sc.	**Dr. Hendri Ahmadian, S.Si.,M.I.M.**
NIP. 198604082014031001	**NIP. 198301042014031002**

Mengetahui,
Ketua Program Studi Teknologi Informasi

Malahayati, M.T.
NIP. 198301272015032003

LEMBAR PENGESAHAN

IMPLEMENTASI MULTIMODAL *LARGE LANGUAGE MODEL* UNTUK EKSTRAKSI DAN TRANSLITERASI *PRINTED JAWI SCRIPT*

TUGAS AKHIR

Telah Diuji Oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Jumat, 14 Agustus 2026 M
1 Rabiul Awal 1448 H
Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,	Sekretaris,
Bustami, M.Sc.	**Dr. Hendri Ahmadian, S.Si.,M.I.M.**
NIP. 198604082014031001	**NIP. 198301042014031002**
Penguji I,	Penguji II,
Khairan AR, M.Kom.	**Mulkan Fadhli, S.T., M.T.**
NIP. 198607042014031001	**NIP. 198811282020121006**

Mengetahui:
Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh

Prof. Dr. Ir. Muhammad Dirhamsyah, M.T., IPU
NIP. 196210021988111001

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama	: Zidan Mubarak
NIM	: 220705071
Program Studi	: Teknologi Informasi
Fakultas	: Sains dan Teknologi
Judul	: Implementasi Multimodal *Large Language Model* Untuk Ekstraksi dan Transliterasi *Printed Jawi Script*

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggungjawabkan;
2. Tidak melakukan plagiasi terhadap naskah karya orang lain;
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli ataupun tanpa izin pemilik karya;
4. Tidak memanipulasi dan memalsukan data;
5. Mengerjakan sendiri karya ini dan mampu bertanggungjawab atas karya ini.

Bila dikemudian hari ada tuntutan dari pihak lain atas karya saya, dan telah melalui pembuktian yang dapat dipertanggungjawabkan dan ternyata memang ditemukan bukti bahwa saya telah melanggar pernyataan ini, maka saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh.

Demikian pernyataan ini saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh, 24 Juli 2026

Yang Menyatakan,

Zidan Mubarak

ABSTRAK

Nama	:	Zidan Mubarak
NIM	:	220705071
Program Studi	:	Teknologi Informasi
Judul	:	Implementasi Multimodal *Large Language Model* Untuk Ekstraksi dan Transliterasi *Printed Jawi Script*
Tanggal Sidang	:	14 Agustus 2026
Jumlah Halaman	:	68
Pembimbing I	:	Bustami, M.Sc.
Pembimbing II	:	Dr. Hendri Ahmadian, S.Si., M.I.M.
Kata Kunci	:	Aksara Jawi, *Multimodal Large Language Model*, Qwen2.5-VL, QLoRA, Transliterasi, *Fine-Tuning*

Aksara Jawi merupakan warisan tulisan Nusantara yang kemampuan membacanya di masyarakat terus menurun, sehingga menyulitkan upaya digitalisasi naskah-naskah lama beraksara Jawi. Penelitian ini bertujuan mengimplementasikan Multimodal *Large Language Model*, yaitu Qwen2.5-VL-3B-Instruct, yang di-*fine-tuning* menggunakan metode *Quantized Low-Rank Adaptation* (QLoRA) untuk melakukan ekstraksi dan transliterasi aksara Jawi cetak ke huruf Latin (Rumi) secara *end-to-end*. Dataset penelitian terdiri atas 34.380 citra gabungan data sintetik dan data autentik. Hasil pengujian menunjukkan model dasar (*baseline*) menghasilkan *Character Error Rate* (CER) ekstraksi Jawi sebesar 0,825 dan CER transliterasi Rumi sebesar 0,906, sedangkan setelah *fine-tuning* kedua nilai tersebut turun menjadi 0,174 dan 0,165, atau setara penurunan kesalahan sekitar 79%. Studi ablasi lebih lanjut menunjukkan bahwa kombinasi data sintetik dan autentik pada proses pelatihan memberikan performa terbaik dibandingkan penggunaan salah satu jenis data saja, sehingga membuktikan bahwa penggabungan kedua sumber data tersebut merupakan strategi komposisi data yang paling optimal untuk tugas ekstraksi dan transliterasi aksara Jawi.

ABSTRACT

Name	:	Zidan Mubarak
Student ID	:	220705071
Study Program	:	*Information Technology*
Title	:	*Multimodal Large Language Model Implementation for Extraction and Transliteration of Printed Jawi Script*
Date	:	*14 August 2026*
Number of Pages	:	68
Supervisor I	:	Bustami, M.Sc.
Supervisor II	:	Dr. Hendri Ahmadian, S.Si., M.I.M.
Keywords	:	*Jawi Script, Multimodal Large Language Model, Qwen2.5-VL, QLoRA, Transliteration, Fine-Tuning*

Jawi script is a written heritage of the Nusantara region whose readership among the public continues to decline, complicating efforts to digitize historical manuscripts written in this script. This study aims to implement a Multimodal Large Language Model, namely Qwen2.5-VL-3B-Instruct, fine-tuned using the Quantized Low-Rank Adaptation (QLoRA) method to perform end-to-end extraction and transliteration of printed Jawi script into Latin script (Rumi). The dataset consists of 34,380 images combining synthetic and authentic data. Testing results show that the baseline model produced a Character Error Rate (CER) of 0.825 for Jawi extraction and 0.906 for Rumi transliteration, whereas after fine-tuning these values decreased to 0.174 and 0.165 respectively, corresponding to an error reduction of approximately 79%. A further ablation study demonstrates that combining synthetic and authentic data during training yields the best performance compared to using either data type alone, confirming that combining both data sources is the most optimal data composition strategy for the Jawi script extraction and transliteration task.

KATA PENGANTAR

Puji syukur senantiasa penulis panjatkan ke hadirat Allah SWT, Tuhan Yang Maha Esa, atas segala rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "Implementasi Multimodal *Large Language Model* Untuk Ekstraksi dan Transliterasi *Printed Jawi Script*". Salawat serta salam tak lupa senantiasa tercurahkan kepada junjungan Nabi Besar Muhammad SAW, beserta keluarga, sahabat, dan para pengikutnya.

Penyusunan skripsi ini merupakan salah satu syarat akademik untuk menyelesaikan pendidikan program sarjana (S1) pada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, UIN Ar-Raniry Banda Aceh.

Penulis menyadari sepenuhnya bahwa penyelesaian skripsi ini tidak lepas dari dukungan, bimbingan, arahan, serta doa dari berbagai pihak sejak awal masa perkuliahan hingga tahap penyusunan akhir. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Malahayati, M.T., selaku Ketua Program Studi Teknologi Informasi.
2. Bapak Bustami, M.Sc. selaku Dosen Pembimbing I dan Bapak Dr. Hendri Ahmadian, S.Si.,M.I.M. selaku Dosen Pembimbing II, yang telah meluangkan banyak waktu, tenaga, dan pikiran untuk memberikan bimbingan, masukan berharga, serta motivasi yang tak ternilai harganya kepada penulis selama proses penyusunan skripsi ini.
3. Seluruh Bapak/Ibu Dosen serta Staf Akademik Program Studi Teknologi Informasi yang telah membagikan ilmu yang sangat bermanfaat selama masa perkuliahan dan banyak membantu urusan administrasi penulis.
4. Teristimewa untuk kedua orang tua tercinta, serta keluarga besar yang tidak pernah putus memberikan doa, kasih sayang, dukungan moral, maupun material sehingga penulis dapat menyelesaikan pendidikan dengan baik.
5. Dua rekan tim penelitian penulis yang telah bekerja sama membagi tugas dalam pengumpulan dan pengolahan data, serta seluruh rekan-

rekan mahasiswa Teknologi Informasi angkatan 2022 yang telah bersama-sama berjuang, belajar, dan saling memberikan semangat.

6. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang secara langsung maupun tidak langsung telah memberikan bantuan, dukungan, dan motivasi selama proses penyusunan skripsi ini, serta pihak-pihak yang telah memberikan inspirasi kepada penulis untuk mendalami bidang *Artificial Intelligence* sehingga penelitian ini dapat terselesaikan dengan baik.

Penulis menyadari bahwa di dalam penulisan dan penyusunan skripsi ini masih terdapat banyak kekurangan karena keterbatasan ilmu dan pengalaman. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun dari semua pihak.

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat yang nyata bagi pengembangan ilmu pengetahuan, pelestarian naskah budaya aksara Jawi, serta dapat menjadi referensi yang berguna bagi penelitian-penelitian selanjutnya di masa yang akan datang.

Banda Aceh, 24 Juli 2026
Penulis

Zidan Mubarak

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang

Aksara Jawi adalah bagian penting dari warisan budaya Nusantara karena selama ratusan tahun digunakan sebagai alat komunikasi dan sarana utama penyebaran ilmu pengetahuan di kawasan Asia Tenggara. Di era sekarang, digitalisasi naskah lama menjadi perhatian global sebagai cara untuk menjaga nilai intelektual agar tetap dapat dimanfaatkan pada masa modern (Bania & Akob, 2025). Sayangnya, kemampuan membaca aksara Jawi di masyarakat saat ini terus menurun, sehingga banyak naskah bersejarah tidak lagi dipahami oleh generasi muda. Situasi ini menuntut adanya teknologi yang mampu mengekstraksi teks secara otomatis agar warisan masa lalu dapat dihubungkan dengan kebutuhan zaman sekarang.

Upaya ekstraksi teks otomatis tersebut menghadapi kendala teknis yang serius akibat karakteristik unik dari aksara Jawi itu sendiri. Tantangan terbesar dalam proses digitalisasi naskah Jawi terletak pada bentuk tulisannya yang bersifat kursif atau saling menyambung. Setiap huruf Jawi dapat memiliki bentuk berbeda tergantung pada posisinya dalam kata, baik di awal, tengah, maupun akhir, ditambah dengan penggunaan titik dan sambungan antar-huruf yang sangat rapat. Hal ini membuat sistem komputer biasa sering sulit membedakan huruf-huruf yang bentuknya mirip (Faizullah et al., 2023). Kesalahan kecil dalam membaca titik atau sambungan huruf dapat berdampak besar karena dapat mengubah arti kata saat dilakukan transliterasi ke huruf Latin.

Untuk mengatasi kompleksitas visual tersebut, sejumlah penelitian sebelumnya telah mencoba menerapkan berbagai teknik visi komputer. Metode standar seperti *Convolutional Neural Network* (CNN) sering digunakan untuk mengenali karakter secara individual. Hasilnya cukup baik ketika karakter tampil terpisah, tetapi performanya menurun tajam pada naskah cetak yang rapat dan saling terhubung (Handoko et al., 2024). Pendekatan ini sangat bergantung pada proses pemisahan karakter satu per satu yang sulit diterapkan pada dokumen lama.

Akibatnya, tingkat kesalahan kata yang dihasilkan masih tinggi dan membutuhkan banyak perbaikan manual setelah proses pengenalan selesai (Syahputra Bania et al., 2025).

Kelemahan pada metode berbasis visi saja menunjukkan bahwa pengenalan aksara Jawi tidak bisa hanya mengandalkan fitur piksel gambar, melainkan juga membutuhkan pemahaman konteks bahasa. Walaupun teknologi pembelajaran mendalam terus berkembang, masih terdapat keterbatasan dalam penelitian yang memanfaatkan model multimodal berbasis visi dan Bahasa (*vision-language*) yang mampu memahami konteks visual dan bahasa secara bersamaan. Penelitian terdahulu umumnya hanya menitikberatkan pada ciri visual gambar tanpa memperhatikan hubungan makna antar kata dalam satu kalimat. Selain itu, banyak sistem yang belum mampu menangani tata letak dokumen yang rumit serta kondisi kertas yang sudah rusak atau menua (Adilazuarda et al., 2025). Padahal, proses ekstraksi teks membutuhkan tidak hanya ketelitian visual, tetapi juga pemahaman bahasa yang baik.

Guna menutup celah antara pengenalan visual dan pemahaman bahasa tersebut, penelitian ini mengusulkan penggunaan teknologi kecerdasan buatan generasi terbaru. Penelitian ini bertujuan untuk menjawab keterbatasan tersebut dengan mengimplementasikan Multimodal *Large Language Model*, khususnya Qwen2.5-VL-3B-Instruct, melalui proses *fine-tuning*. Model ini dipilih karena kemampuannya dalam membaca gambar secara menyeluruh dan memahami bahasa untuk menghasilkan transliterasi yang lebih tepat (Wasfy et al., 2025). Dengan menerapkan teknik *Parameter-Efficient Fine-Tuning* (PEFT), model yang dihasilkan tetap efisien dari sisi sumber daya namun memiliki kemampuan tinggi dalam mengenali aksara Jawi cetak. Diharapkan, hasil penelitian ini dapat menjadi acuan baru dalam digitalisasi naskah Jawi dan membantu menjaga keberlanjutan warisan intelektual Melayu di masa mendatang.

1.2 Rumusan Masalah

1. Bagaimana mengimplementasikan model Qwen2.5-VL-3B-Instruct untuk tugas ekstraksi dan transliterasi teks Jawi cetak?

2. Bagaimana pengaruh penerapan strategi *fine-tuning* menggunakan *Quantized Low-Rank Adaptation* (QLoRA) terhadap kinerja model dalam tugas ekstraksi dan transliterasi teks aksara Jawi?

1.3 Tujuan Penelitian

1. Mengimplementasikan model Qwen2.5-VL-3B-Instruct untuk ekstraksi dan transliterasi teks Jawi.
2. Menganalisis pengaruh strategi *fine-tuning* menggunakan *Quantized Low-Rank Adaptation* (QLoRA) terhadap kinerja model dalam melakukan ekstraksi dan transliterasi teks aksara Jawi.

1.4 Manfaat Penelitian

1. Menambah kajian ilmiah mengenai penerapan Multimodal *Large Language Model* dalam ekstraksi dan transliterasi aksara Jawi pada dokumen cetak.
2. Menjadi referensi bagi penelitian selanjutnya terkait pengolahan dan digitalisasi dokumen beraksara Jawi menggunakan kecerdasan buatan.
3. Memberikan gambaran kinerja strategi *fine-tuning* dalam meningkatkan hasil ekstraksi dan transliterasi aksara Jawi.

1.5 Batasan Masalah

1. Dokumen yang digunakan dalam penelitian berupa dokumen cetak beraksara Jawi.
2. Penelitian ini tidak membahas penerjemahan makna teks, melainkan terbatas pada ekstraksi dan transliterasi.
3. Evaluasi penelitian dibatasi pada hasil ekstraksi teks dan transliterasi aksara Jawi ke huruf Latin.
4. Model multimodal yang digunakan adalah Qwen2.5-VL-3B-Instruct.
5. Penggunaan kemampuan multimodal dalam penelitian ini dibatasi pada pemrosesan informasi visual berupa gambar naskah Jawi dan informasi bahasa berupa teks. Modalitas lain, seperti audio dan video, tidak digunakan dalam proses pelatihan maupun pengujian model.
6. Citra masukan yang digunakan dalam penelitian dibatasi pada citra yang memuat satu baris teks Jawi, sehingga penelitian tidak mencakup proses

ekstraksi dan transliterasi dari dokumen atau halaman penuh yang memuat beberapa baris teks.

7. Penerapan model dalam penelitian ini hanya mencakup pembuatan antarmuka inferensi sederhana melalui platform *Hugging Face Space*, dan tidak mencakup pengembangan website atau aplikasi secara penuh.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Tabel 2.1 Penelitian Terkait

Penelitian, Tahun	Judul Penelitian	Fokus & Metode	Hasil Penelitian
(Abdiansah et al., 2025)	Penerapan Algoritma Convolutional Neural Networks untuk Pengenalan Tulisan Tangan Aksara Jawa	Tulisan tangan Aksara Jawa; CNN deep learning	Akurasi ~99,83% dalam pengenalan karakter tulisan tangan Jawa menunjukkan CNN sangat baik untuk dataset terbatas dan struktur visual
(Safrizal, 2019)	PENGENALAN KARAKTER JAWI TULISAN TANGAN MENGGUNAKAN FITUR SUDUT	Pengenalan karakter Jawi tulisan tangan; SVM dengan ekstraksi fitur sudut	Akurasi ~78,86% (SVM klasik) menunjukkan metode ML tradisional kurang optimal untuk teks kompleks
Wasfy et al. (2025)	QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large	OCR tulisan Arab (complex script); Fine-tuned Vision-Language	WER 0.160 & CER 0.061 menunjukkan VLM sangat efektif mengatasi kompleksitas

	Language Model Adaptation	Model (derived dari Qwen2-VL)	skrip, dianggap sebagai benchmark pendekatan VLM
(Bhatia et al., 2024)	Qalam: A Multimodal LLM for Arabic Optical Character and Handwriting Recognition	OCR teks cetak & tulisan tangan Arab; Multimodal LLM (SwinV2 + RoBERTa)	WER 0.80% (HWR) & 1.18% (OCR) menunjukkan model multimodal transformer kuat dalam pengenalan skrip kompleks
(Nagasubramanian et al., 2025)	OCRNet a robust deep learning framework for alphanumeric character recognition to assist the visually impaired	OCR teknologi modern; Survey horizontal (CNN vs Transformer-based)	Transformer menunjukkan peningkatan pada pengenalan teks kompleks dan konteks diluar kemampuan CNN tradisional
(AR et al., 2025)	An Applied Computer Mathematics Approach to Transliteration: YOLOv8-Based Detection of Harah Jawoe Script	Deteksi objek YOLOv8 untuk transliterasi Harah Jawoe level kata	mAP50–95 72.4%, akurasi 99,95%; tidak integrasi LLM multimodal atau ekstraksi end-to-end

Berdasarkan penelitian-penelitian terdahulu pada Tabel 2.1, dapat disimpulkan bahwa pendekatan umum seperti SVM dan CNN telah banyak digunakan untuk pengenalan karakter aksara Jawa dan Jawi, terutama pada tulisan tangan dan klasifikasi karakter tunggal, namun masih memiliki keterbatasan dalam menangani teks dokumen yang kompleks serta belum mengintegrasikan proses transliterasi secara *end-to-end*. Penelitian mutakhir berbasis *Vision-Language Model* dan Multimodal *Large Language Model* telah menunjukkan kinerja yang sangat baik pada pengenalan teks Arab standar, tetapi belum secara khusus diterapkan pada aksara Jawi yang memiliki karakter tambahan, perbedaan cara penulisan dan kesulitan dalam memahami makna kalimat. Selain itu, penelitian transliterasi Jawi yang ada masih bertumpu pada pendekatan deteksi objek dan belum memanfaatkan pemahaman bahasa dari model multimodal. Oleh karena itu, masih terdapat celah penelitian dalam pengembangan sistem ekstraksi teks dan transliterasi aksara Jawi dari dokumen cetak secara *end-to-end* menggunakan *Vision-Language Model*, khususnya dengan pendekatan *fine-tuning* efisien berbasis *Quantized Low-Rank Adaptation* (QLoRA), yang dievaluasi menggunakan metrik *Character Error Rate* dan *Word Error Rate*.

2.2 Landasan Teori

2.2.1 Kecerdasan Buatan, *Machine Learning*, dan *Deep Learning*

Kecerdasan Buatan atau *Artificial Intelligence* (AI) adalah bidang ilmu komputer yang berfokus pada pembuatan sistem yang dapat meniru kemampuan manusia, seperti mengenali pola, memahami informasi, dan mengambil keputusan. Dalam beberapa tahun terakhir, AI berkembang sangat cepat karena dukungan perangkat keras yang semakin kuat dan ketersediaan data dalam jumlah besar. Teknologi ini kini banyak digunakan di berbagai bidang, termasuk pendidikan dan pengolahan dokumen digital, karena mampu mempercepat pekerjaan dan meningkatkan ketepatan hasil.

Machine Learning (ML) merupakan bagian dari AI yang memungkinkan sistem belajar langsung dari data tanpa harus diatur dengan aturan yang rinci. Model ML mempelajari pola dari data latihan, lalu menggunakan pola tersebut untuk memproses data baru. Salah satu bagian penting dari ML adalah *Deep Learning*

(DL), yang menggunakan jaringan saraf tiruan dengan banyak lapisan. Pendekatan ini sangat efektif untuk mengolah gambar dan teks, sehingga sering digunakan dalam sistem pengenalan teks berbasis citra (Jumrah Jamil & Suharto Pulukadang, 2025).

2.2.2 Aksara Jawi

Aksara Jawi adalah bentuk tulisan yang diadaptasi dari huruf Arab dan digunakan untuk menulis bahasa Melayu. Penulisannya dilakukan dari kanan ke kiri dan bersifat menyambung antar huruf. Selain huruf Arab standar, Jawi memiliki beberapa tambahan huruf seperti ڬ ,ڠ ,ڤ ,چ, dan ڽ untuk mewakili bunyi khas Melayu. Dalam praktiknya, penulisan Jawi sering tidak menggunakan tanda vokal, sehingga arti kata sangat bergantung pada konteks kalimat. Ragam huruf Arab Jawi beserta huruf tambahan yang digunakan dalam bahasa Melayu ditunjukkan pada Gambar 2.1.

ا (Alif)	ب (Ba)	ت (Ta)	ة (Ta Marbutah)	ث (Sa)
ج (Jim)	چ *(Ca)	ح (Ha)	خ (Kha)	د (Dal)
ذ (Zal)	ر (Ra)	ز (Zai)	س (Sin)	ش (Syin)
ص (Sad)	ض (Dad)	ط (Ta)	ظ (Za)	ع (Ain)
غ (Ghain)	ڠ *(Nga)	ف (Fa)	ڤ *(Pa)	ق (Qaf)
ك (Kaf)	ڬ *(Ga)	ل (Lam)	م (Mim)	ن (Nun)
و (Wau)	ۏ *(Va)	ه (Ha)	ء (Hamzah)	ي (Ya)
ى *(Ye)	ڽ *(Nya)			

Gambar 2.1 Aksara Arab Jawi

Seiring waktu, penggunaan aksara Jawi semakin berkurang karena huruf Latin lebih dominan dalam pendidikan dan administrasi. Hal ini membuat digitalisasi dokumen Jawi menjadi penting sebagai upaya pelestarian bahasa dan budaya Melayu. Dari sisi teknologi, aksara Jawi cukup sulit dikenali oleh komputer karena hurufnya saling menyambung, banyak ligatur, serta sangat bergantung pada posisi titik dan tanda baca. Ditambah lagi, data Jawi yang sudah diberi label masih

terbatas, sehingga menyulitkan pengembangan sistem OCR yang akurat (Coluzzi, 2022).

2.2.3 *Optical Character Recognition* (OCR)

Optical Character Recognition atau OCR adalah teknologi yang berfungsi mengubah tulisan dalam gambar, seperti hasil *scan* atau foto dokumen, menjadi teks digital. Secara umum, proses OCR terdiri dari beberapa tahap, mulai dari perbaikan kualitas gambar, pencarian area teks, pengenalan huruf atau kata, hingga perbaikan hasil akhir. Proses ini sangat bergantung pada kualitas gambar yang diproses, karena faktor seperti pencahayaan dan resolusi dapat memengaruhi akurasi pengenalan karakter (Darmi et al., 2025)

Perkembangan OCR saat ini mulai meninggalkan metode lama yang memisahkan huruf satu per satu. Pendekatan terbaru menggunakan *deep learning end-to-end* yang langsung mengubah fitur visual menjadi teks tanpa proses pemisahan karakter. Metode ini lebih tahan terhadap perbedaan bentuk huruf, tata letak, dan kualitas gambar. Namun, untuk aksara Arab dan turunannya seperti Jawi, tantangannya lebih besar karena huruf saling terhubung dan kesalahan kecil pada titik dapat mengubah arti kata. Oleh sebab itu, penyesuaian model besar dan penggunaan data khusus Jawi sangat dibutuhkan untuk meningkatkan hasil pengenalan.

2.2.4 *Vision-Language Models* (VLM)

Vision-Language Model (VLM) adalah jenis model AI yang mampu memproses gambar dan teks secara bersamaan. Model ini memungkinkan sistem tidak hanya melihat bentuk visual, tetapi juga memahami makna bahasa yang terkait dengan gambar tersebut. Karena kemampuan ini, VLM sangat cocok digunakan untuk tugas OCR berbasis konteks, analisis dokumen, dan pengambilan informasi dari gambar. Dengan demikian, penerapan VLM dalam analisis dokumen dapat memberikan hasil yang lebih baik dengan efisiensi yang lebih tinggi, seperti yang ditunjukkan oleh metode DocVLM (Nacson et al., 2024).

Secara umum, VLM terdiri dari dua bagian utama, yaitu bagian pemroses gambar dan bagian pemroses bahasa. Informasi visual dari gambar diubah menjadi

representasi yang dapat dipahami oleh model bahasa. Dengan cara ini, teks yang dihasilkan tidak hanya bergantung pada bentuk huruf, tetapi juga pada konteks kalimat. Istilah "multimodal" pada penelitian ini dipahami secara spesifik sebagai kemampuan model dalam memproses dua modalitas, yaitu visual dan tekstual, bukan cakupan modalitas yang lebih luas seperti suara maupun video. Dalam penelitian ini, VLM menjadi dasar utama untuk membangun sistem pengambilan teks dan transliterasi aksara Jawi dari dokumen cetak.

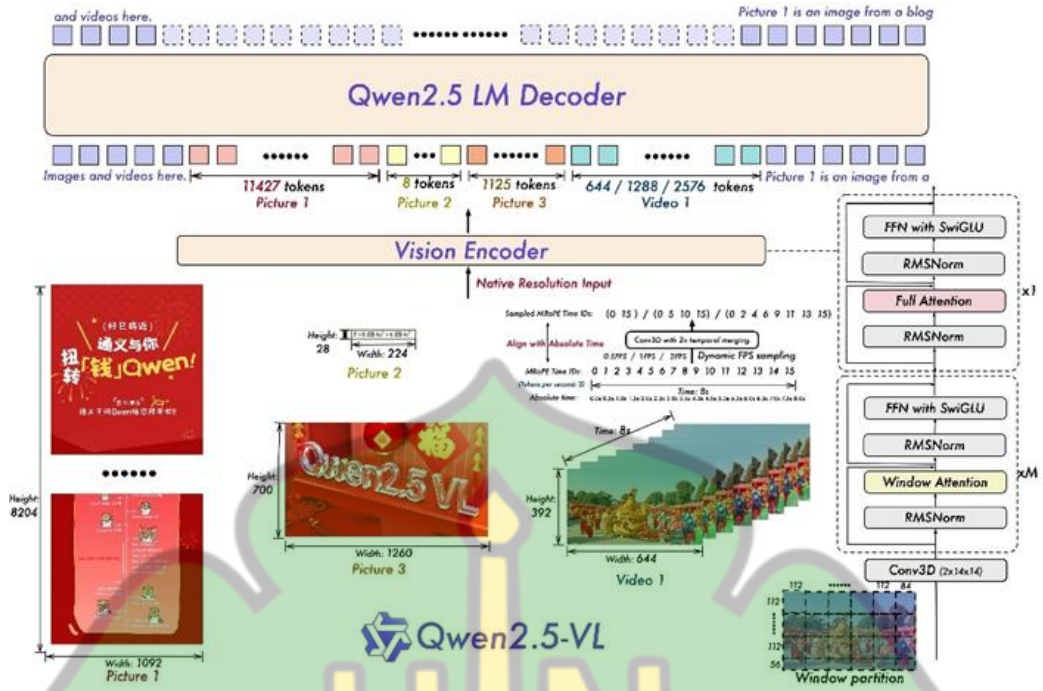
2.2.5 *Transformer* dan *Vision Transformer*

Transformer adalah arsitektur model yang digunakan untuk memproses data berurutan dengan memanfaatkan mekanisme perhatian atau *self-attention*. Berbeda dengan model lama yang bekerja secara bertahap, *Transformer* memproses seluruh data sekaligus sehingga lebih cepat dan mampu memahami hubungan antar bagian data yang berjauhan. Arsitektur ini menjadi fondasi utama bagi banyak model bahasa modern.

Vision Transformer (ViT) menerapkan konsep *Transformer* pada data gambar dengan cara membagi gambar menjadi bagian-bagian kecil atau patch. Setiap bagian diperlakukan seperti token pada teks dan diproses menggunakan *self-attention*. Cara ini memungkinkan model memahami hubungan antar bagian gambar secara menyeluruh, yang sangat penting dalam membaca dokumen dengan tata letak rumit. Gabungan *Transformer* dan ViT menjadi dasar bagi banyak VLM yang digunakan dalam OCR dan analisis dokumen (Dosovitskiy et al., 2021).

2.2.6 Arsitektur Qwen2.5-VL-3B-Instruct

Qwen2.5-VL-3B-Instruct adalah model multimodal yang menggabungkan pemrosesan gambar berbasis *Vision Transformer* dengan model bahasa dari keluarga Qwen2.5. Model ini dirancang untuk menerima input berupa gambar dan teks, lalu menghasilkan keluaran teks dengan mempertimbangkan konteks visual.



Gambar 2.2 Arsitektur Qwen2.5-VL-3B-Instruct

Gambar 2.2 menunjukkan alur kerja Qwen2.5-VL-3B-Instruct mulai dari beberapa masukan visual berupa gambar dan video hingga menghasilkan teks. Gambar dan video terlebih dahulu masuk ke *Vision Encoder*, di mana setiap gambar dapat menghasilkan jumlah token yang berbeda tergantung ukuran dan tingkat detailnya, misalnya gambar beresolusi tinggi akan menghasilkan token yang lebih banyak. Di dalam *Vision Encoder*, data visual diproses melalui beberapa lapisan yang bekerja berulang, yaitu lapisan untuk mengolah informasi dasar gambar, lapisan yang melihat gambar secara keseluruhan, serta lapisan yang fokus pada bagian-bagian kecil gambar. Untuk video, terdapat lapisan khusus yang membaca perubahan antar frame sehingga isi video tetap bisa dipahami. Pada bagian bawah diagram terlihat bahwa gambar dan video dipecah menjadi potongan-potongan kecil sebelum diproses, sehingga sistem lebih ringan dan detail visual seperti huruf dapat dibaca dengan lebih baik. Setelah semua informasi visual diproses, hasilnya dikirim ke Qwen2.5 *Language Model Decoder* dalam bentuk token visual, lalu *decoder* ini menyusun teks berdasarkan isi gambar dan konteksnya. Diagram ini juga menunjukkan bahwa model mampu menerima gambar dengan ukuran asli dan memilih frame video secara fleksibel, sehingga dapat menangani berbagai jenis dokumen dan media dengan ukuran yang berbeda-beda.

Versi dengan 3 miliar parameter dipilih karena memiliki keseimbangan antara kemampuan dan kebutuhan komputasi. Model ini cukup kuat untuk melakukan ekstraksi teks dan transliterasi, tetapi masih bisa dilatih ulang dengan sumber daya yang terbatas. Selain itu, Qwen2.5-VL mampu menangani berbagai ukuran gambar dan tata letak dokumen yang rumit, sehingga cocok digunakan pada dokumen cetak beraksara Jawi (Bai et al., 2025).

Pertimbangan pemilihan ukuran model ini juga didasarkan pada perbandingan dengan varian Qwen2.5-VL lain yang tersedia, yaitu 7B dan 72B. Varian 72B tidak dipertimbangkan karena kebutuhan VRAM yang jauh melampaui kapasitas perangkat keras yang tersedia untuk penelitian ini, bahkan dengan skema kuantisasi 4-bit sekalipun. Dibandingkan dengan varian 7B, varian 3B menawarkan kemampuan *vision-language* yang tetap memadai untuk memahami tata letak dokumen serta relasi antar karakter, sementara kebutuhan VRAM untuk proses *fine-tuning* menggunakan QLoRA jauh lebih ringan sehingga sesuai dengan kapasitas GPU tunggal yang digunakan pada penelitian ini. Varian 7B berpotensi memberikan kemampuan pemahaman bahasa yang lebih kuat, tetapi membutuhkan kapasitas VRAM yang jauh lebih besar sehingga meningkatkan waktu dan biaya pelatihan tanpa jaminan peningkatan performa yang sepadan untuk tugas ekstraksi dan transliterasi aksara Jawi yang relatif spesifik. Dengan demikian, pemilihan model 3B merupakan hasil pertimbangan antara kapasitas model, kebutuhan komputasi, dan sumber daya perangkat keras yang tersedia, bukan semata-mata memilih ukuran yang berada di tengah.

2.2.7 *Fine-Tuning* dan *Quantized Low-Rank Adaptation* (QLoRA)

Fine-tuning adalah proses melatih ulang model yang sudah dilatih sebelumnya dengan data khusus agar lebih sesuai untuk tugas tertentu. Pada model berukuran besar, *fine-tuning* penuh membutuhkan banyak sumber daya dan berisiko merusak kemampuan umum model.

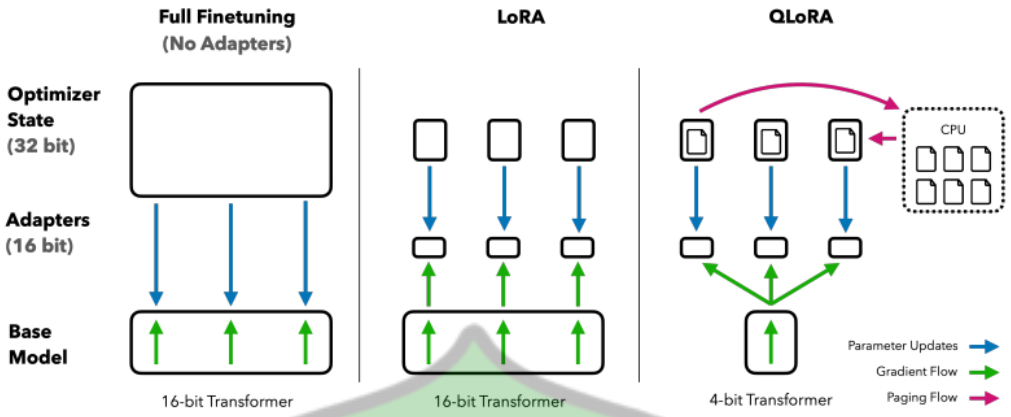
Parameter-Efficient Fine-Tuning (PEFT) adalah pendekatan umum yang mengelompokkan berbagai teknik fine-tuning efisien, di mana hanya sebagian kecil parameter model yang diperbarui selama pelatihan. Pendekatan ini memungkinkan adaptasi model besar dengan kebutuhan memori dan komputasi yang jauh lebih

rendah dibanding fine-tuning penuh. Dalam kerangka PEFT, terdapat beberapa teknik turunan, di antaranya LoRA dan QLoRA yang digunakan dalam penelitian ini.

Low-Rank Adaptation (LoRA) merupakan metode *fine-tuning* yang lebih hemat sumber daya dengan menambahkan sejumlah kecil parameter baru pada bagian tertentu model, sementara parameter utama tetap tidak berubah (Hu et al., 2021).

Quantized Low-Rank Adaptation (QLoRA) adalah pengembangan dari LoRA yang mengintegrasikan teknik kuantisasi 4-bit pada model dasar. Secara bertahap, cara kerja QLoRA dapat dijelaskan sebagai berikut: (1) seluruh bobot model dasar hasil *pre-training* dikompresi dari presisi 16-bit menjadi format NormalFloat4 (NF4) sehingga ukurannya menyusut jauh lebih kecil; (2) bobot model dasar yang telah dikuantisasi tersebut dibekukan, artinya nilainya tidak diperbarui sama sekali selama pelatihan; (3) di beberapa titik pada lapisan *attention* dan *feed-forward*, disisipkan modul adapter LoRA berukuran kecil dengan presisi 16-bit; (4) hanya parameter pada adapter inilah yang dihitung gradiennya dan diperbarui pada setiap langkah pelatihan, sementara model dasar hanya dilalui saat *forward pass* tanpa perlu disimpan gradien maupun status optimizer untuknya. Pendekatan ini mengurangi kebutuhan memori hingga 75% tanpa mengorbankan performa secara signifikan, sehingga memungkinkan *fine-tuning* model besar dengan sumber daya komputasi terbatas (Dettmers et al., 2023).

Perbedaan mendasar antara ketiga pendekatan tersebut, khususnya dari sisi kebutuhan memori dan bagian model yang diperbarui selama pelatihan, dapat dilihat pada Gambar 2.3.



Gambar 2.3 Perbandingan Metode *Full Finetuning*, LoRA, dan QLoRA (Dettmers et al., 2023)

Gambar 2.3 memperlihatkan tiga skema *fine-tuning* secara berdampingan. Pada *full finetuning*, seluruh bobot model dasar (16-bit) diperbarui secara langsung, sehingga status optimizer yang menyertainya juga harus disimpan dalam presisi 32-bit untuk seluruh parameter, membuat kebutuhan memorinya paling besar di antara ketiga pendekatan. Pada LoRA, bobot model dasar tetap dibekukan dan tidak diperbarui, sedangkan pembaruan hanya dilakukan pada adapter berukuran kecil dengan presisi 16-bit yang disisipkan di beberapa bagian model, sehingga kebutuhan memori jauh berkurang dibanding *full finetuning*. Pada QLoRA, model dasar dikuantisasi lebih lanjut menjadi presisi 4-bit dan tetap dibekukan, sementara adapter yang dilatih tetap berada pada presisi 16-bit. Selain itu, QLoRA menerapkan mekanisme *paged optimizer*, yaitu status optimizer yang jarang diakses akan dipindahkan sementara ke memori CPU dan dikembalikan ke GPU hanya saat dibutuhkan, sehingga lonjakan kebutuhan memori pada GPU dapat dihindari. Kombinasi kuantisasi 4-bit dan *paged optimizer* inilah yang membuat QLoRA menjadi pendekatan paling hemat memori di antara ketiganya, sekaligus menjadi alasan utama pemilihan QLoRA pada penelitian ini.

2.2.8 Transliterasi

Transliterasi adalah proses mengubah tulisan dari satu jenis aksara ke aksara lain tanpa mengubah makna kata. Dalam penelitian ini, transliterasi difokuskan pada pengubahan teks Jawi menjadi huruf Latin. Proses ini cukup menantang

karena tulisan Jawi tidak selalu memiliki tanda vokal, memiliki variasi penulisan, dan aturan ejaan yang berbeda.

Istilah "Rumi" yang dipakai pada penelitian ini bukan istilah buatan penulis, melainkan istilah baku yang lazim digunakan dalam kajian linguistik dan pengolahan bahasa Melayu untuk merujuk pada sistem tulisan Latin bahasa Melayu, sebagai lawan dari sistem tulisan Jawi yang berbasis huruf Arab. Dikotomi Jawi-Rumi ini digunakan secara konsisten dalam literatur digitalisasi naskah Melayu, termasuk pada penelitian mengenai pelestarian aksara Jawi yang menjadi rujukan Bab I dan Bab II skripsi ini (Coluzzi, 2022), sehingga penggunaan istilah "Rumi" pada skripsi ini konsisten dengan konvensi akademik yang berlaku, bukan sekadar istilah umum "romanisasi" yang dapat merujuk pada bahasa apa pun.

Pendekatan modern untuk transliterasi memanfaatkan model pembelajaran mesin yang mempelajari pola dari pasangan teks Jawi dan Latin. Dengan menggabungkan hasil OCR dan model transliterasi berbasis data, sistem diharapkan mampu menghasilkan teks Latin yang lebih tepat dan konsisten saat membaca dokumen Jawi cetak (Razak et al., 2019).

2.2.9 Hugging Face Space

Hungging Face Space adalah platform hosting gratis yang disediakan oleh Hungging Face untuk men-*deploy machine learning* sebagai aplikasi interaktif berbasis web. Platform ini memungkinkan peneliti dan pengembang untuk mempublikasikan model mereka sehingga dapat diakses dan dicoba langsung oleh publik tanpa memerlukan infrastruktur server sendiri. Hungging Face Space mendukung berbagai framework antarmuka seperti Gradio dan Streamlit, yang memudahkan pembuatan tampilan inferensi sederhana hanya dengan beberapa baris kode Python.

Dalam penelitian ini, Hungging Face Space digunakan sebagai platform *deployment* model hasil *fine-tuning*. Pengguna dapat mengunggah gambar teks aksara Jawi dan mendapatkan hasil ekstraksi sekaligus transliterasi dalam format teks Latin secara langsung melalui antarmuka Gradio yang dibangun diatas platform tersebut. Pendekatan deployment ini dipilih karena integrasi langsung

dengan ekosistem Hungging Face untuk pengelolaan model dan dataset, serta memungkinkan peneliti lain untuk mereproduksi dan menguji sistem secara langsung tanpa perlu membangun infrastruktur web yang kompleks

2.2.10 Metrik Evaluasi Pengenalan Teks

Untuk menilai kinerja sistem, digunakan dua ukuran utama, yaitu *Character Error Rate* (CER) dan *Word Error Rate* (WER). CER menghitung tingkat kesalahan pada tingkat huruf, sehingga sangat peka terhadap kesalahan karakter atau tanda baca. Sementara itu, WER mengukur kesalahan pada tingkat kata dan menunjukkan seberapa mudah teks hasil sistem dapat dibaca.

Character Error Rate (CER) didefinisikan sebagai:

$$CER = \frac{S_C + D_C + I_C}{N_C} \tag{1}$$

dengan S_c adalah substitusi karakter, D_c penghapusan, I_c penyisipan, dan N_c jumlah karakter referensi.

Word Error Rate (WER) dirumuskan sebagai:

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{2}$$

dengan notasi yang sama tetapi pada tingkat kata.

Penggunaan CER dan WER secara bersamaan memberikan gambaran yang lebih lengkap tentang kualitas sistem. CER membantu menemukan kesalahan teknis pada pengenalan huruf, sedangkan WER menunjukkan pengaruh kesalahan tersebut terhadap pemahaman teks secara keseluruhan (Neudecker et al., 2021).

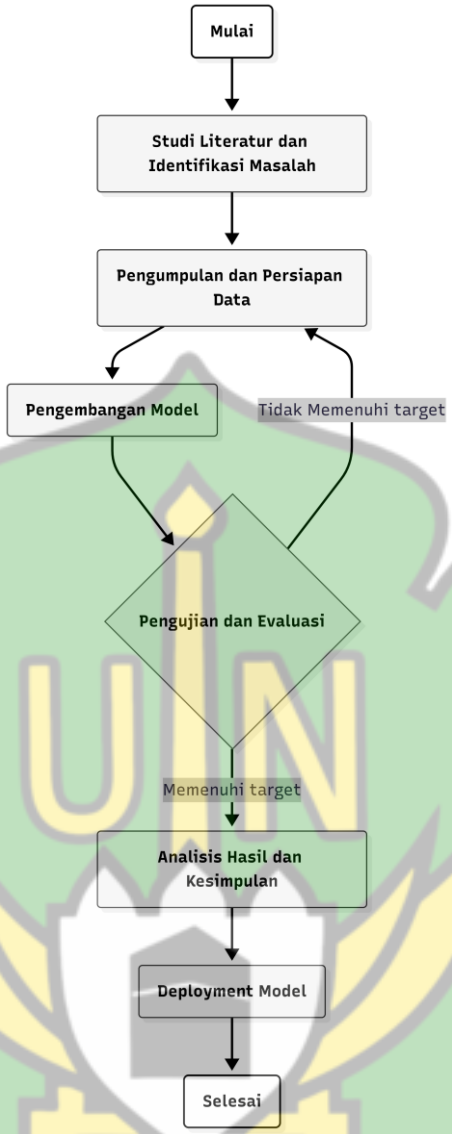
BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan penelitian eksperimental kuantitatif berbasis rekayasa sistem. Fokus utama penelitian adalah mengimplementasikan dan mengoptimalkan model *Vision-Language Model* (VLM) Qwen2.5-VL-3B-Instruct untuk tugas ekstraksi dan transliterasi aksara Jawi. Pendekatan eksperimental dilakukan melalui proses *fine-tuning* menggunakan metode *Quantized Low-Rank Adaptation* (QLoRA), sedangkan pendekatan kuantitatif digunakan untuk mengukur kinerja model secara objektif melalui metrik *Character Error Rate* (CER) dan *Word Error Rate* (WER).

3.2 Tahapan Penelitian

Penelitian dilaksanakan melalui rangkaian tahapan sistematis yang dimulai dari identifikasi masalah hingga analisis hasil akhir untuk memastikan model yang dikembangkan memiliki akurasi yang optimal. Alur penelitian ini secara visual dapat dilihat pada Gambar 3.1.



Gambar 3.1 Alur Penelitian

3.2.1 Studi Literatur dan Identifikasi Masalah

Tahap awal penelitian dimulai dengan kajian mendalam terhadap literatur ilmiah terkini dari tahun 2020 hingga 2025 mengenai Vision-Language Models, teknologi OCR untuk aksara non-Latin, dan metode *Parameter-Efficient Fine-Tuning*. Studi literatur mencakup analisis terhadap penelitian-penelitian terdahulu tentang digitalisasi aksara Jawi untuk mengidentifikasi gap penelitian yang ada.

Masalah diidentifikasi melalui tantangan dalam membaca dan mengubah aksara Jawi, seperti bentuk tulisan sambung yang rumit, beragamnya jenis cetakan, kerusakan pada dokumen lama, serta kurangnya ketersediaan data penelitian. Dari

proses ini, dirumuskan masalah penelitian yang spesifik mengenai bagaimana memanfaatkan teknologi *Vision-Language Model* untuk mengatasi tantangan-tantangan tersebut secara efektif dan efisien.

3.2.2 Pengumpulan dan Persiapan Data

Keberhasilan proses *fine-tuning* sangat dipengaruhi oleh kualitas data yang digunakan. Oleh karena itu, penelitian ini menerapkan pendekatan gabungan dengan memanfaatkan data asli dan data buatan agar model dapat mengenali aksara Jawi dalam berbagai kondisi.

3.2.2.1 Sumber dan Klasifikasi Data

Data Autentik, data ini dikumpulkan dari hasil pemotretan langsung buku Jawi cetak dan tangkapan layar dokumen PDF yang tersedia secara digital. Data autentik mencerminkan kondisi nyata seperti adanya gangguan visual (*noise*), tekstur kertas lama, dan kualitas cetak yang beragam.

Data Sintetik, data ini dibuat secara otomatis melalui proses *web crawling* dari situs-situs Islami untuk memperoleh pasangan teks Jawi dan Rumi dalam jumlah besar, sehingga data latih menjadi lebih banyak.

3.2.2.2 Standarisasi Aksara Jawi

Penyusunan dataset dalam penelitian ini mengacu pada standarisasi aksara Jawi yang ditetapkan oleh Dewan Bahasa dan Pustaka Malaysia (DBP) melalui panduan *Pedoman Ejaan Jawi Yang Disempurnakan* (PEJYD). Standarisasi ini mencakup penetapan karakter dasar 28 huruf Arab beserta 5 huruf tambahan khas Melayu, yaitu Ca (چ), Nga (ڠ), Pa (ڤ), Ga (گ), dan Nya (ڽ), yang tidak terdapat dalam alfabet Arab standar. Selain itu, aturan penulisan vokal, penggunaan tanda baca, dan konvensi penyambungan huruf dalam kata juga mengikuti pedoman tersebut. Penetapan standarisasi ini penting untuk memastikan konsistensi label pada dataset, sehingga model dapat mempelajari pola karakter secara terstruktur dan hasil transliterasi dapat direproduksi secara ilmiah oleh peneliti lain.

3.2.2.3 Alur Pra-Pemrosesan Data Autentik

Data autentik diproses melalui beberapa tahap agar menghasilkan pasangan gambar dan teks yang akurat:

1. Segmentasi Layout, baris teks dideteksi secara otomatis menggunakan *library Kraken*. Setelah itu, penyesuaian manual dilakukan dengan bantuan *VGG Image Annotator* (VIA) agar posisi *bounding box* benar-benar tepat.
2. Pemotongan Gambar, setiap *bounding box* dipotong menjadi satu gambar baris teks yang dijadikan satu data pelatihan.
3. Pemeriksaan dan Pelabelan, gambar dan teks dicek ulang menggunakan antarmuka *Gradio*. Jika ditemukan kesalahan huruf atau titik, perbaikan dilakukan secara manual dengan *keyboard* Jawi khusus untuk menghasilkan *ground truth* yang valid.

3.2.2.4 Alur Pra-Pemrosesan Data Sintetik

Data sintetik dibuat melalui tahapan yang terkontrol untuk membantu proses awal pelatihan model:

1. *Crawling dan Cleaning*, kalimat dikumpulkan dari sekitar 5 *URL* situs web keagamaan, lalu dibersihkan dari duplikasi dan simbol yang tidak diperlukan.
2. Konversi Rumi ke Jawi, teks Latin (Rumi) diubah secara otomatis ke aksara Jawi digital sesuai aturan ejaan yang berlaku.
3. *Text-to-Image Synthesis*, teks Jawi digital diubah menjadi gambar dengan berbagai jenis font seperti *Amiri*, *Harmattan*, *Lateef*, *Scheherazade New*, dan *Traditional Arabic* untuk meniru variasi tulisan pada dokumen asli.

3.2.2.5 Penggabungan dan Pembagian Dataset

Seluruh data yang telah diproses kemudian disatukan dalam satu sistem data terpadu:

1. Penggabungan Data, semua baris teks Jawi dipasangkan dengan teks Latin (Rumi) yang sesuai serta lokasi file gambar ke dalam satu file CSV terstruktur.
2. Format Multimodal, setiap data disusun dalam bentuk perintah yang berisi gambar baris teks Jawi dan instruksi sistem untuk mengekstraksi serta mentransliterasikan teks ke huruf Latin.
3. Pembagian Data, dataset dirancang untuk dibagi menjadi tiga bagian khusus agar evaluasi hasil tetap objektif. Data latih (*training set*) dibentuk menggunakan keseluruhan data sintetik atau 100% data buatan, yang kemudian digabungkan dengan 90% data autentik. Data validasi (*validation set*) diambil sebesar 5% dari total gabungan data latih tersebut. Sementara itu, data uji (*testing set*) disiapkan murni menggunakan 10% sisa data autentik, tanpa ada campuran data sintetik sama sekali. Data uji murni ini dipastikan belum pernah dilihat oleh model selama masa pelatihan agar evaluasi benar-benar mencerminkan performa pada naskah asli.

   Secara rinci, mekanisme pembagian data dilakukan melalui tiga langkah berikut. Pertama, sebanyak 10% dari data autentik dipisahkan sebagai data uji murni, sedangkan sisa digabungkan ke dalam data latih. Kedua, seluruh data sintetik digabungkan dengan data autentik. Ketiga, dari kumpulan gabungan ini diambil 5% secara acak sebagai data validasi, dan sisanya digunakan sebagai data latih akhir. Pengambilan sampel pada setiap tahap dilakukan secara acak (*random sampling*) dengan proporsi tetap menggunakan fungsi pengacakan pandas, yaitu DataFrame.sample(frac=1, random_state=42) untuk mengacak urutan baris, dilanjutkan pemotongan posisi (slicing) sesuai proporsi yang ditentukan, bukan berdasarkan pemilihan manual atau urutan berkas, sehingga setiap *subset* memiliki peluang yang sama untuk memuat variasi font, kualitas citra, dan panjang kalimat yang representatif, sekaligus dapat direproduksi karena nilai random_state (*seed*) yang tetap. Pemilihan proporsi 10% untuk data uji didasarkan pada pertimbangan bahwa data autentik jumlahnya jauh lebih terbatas

dibandingkan data sintetik, sehingga porsi tersebut perlu dijaga cukup besar agar hasil evaluasi tetap stabil secara statistik, tanpa mengorbankan terlalu banyak data autentik yang justru dibutuhkan untuk melatih model mengenali karakteristik naskah nyata.

Penggunaan gabungan data autentik dan data sintetik pada validation set, alih-alih data autentik murni seperti pada testing set, didasarkan pada perbedaan fungsi kedua *subset* tersebut dalam penelitian ini. Validation set berfungsi untuk memantau proses pelatihan secara berkala, yaitu untuk memilih *checkpoint epoch* terbaik dan mendeteksi gejala *overfitting* melalui pergerakan *validation loss* (sebagaimana ditunjukkan pada Tabel 4.4), sehingga *validation set* perlu merepresentasikan kedua karakteristik data yang dipelajari model selama pelatihan, yaitu data sintetik yang dominan secara jumlah maupun data autentik yang merepresentasikan kondisi nyata. Apabila *validation set* dibuat murni dari data autentik seperti *testing set*, jumlah datanya akan sangat terbatas karena harus diambil dari sisa yang sebagian besar sudah dialokasikan untuk *training*, sehingga pemantauan *loss* berisiko kurang stabil akibat ukuran sampel yang kecil. Sebaliknya, *testing set* sengaja dibuat murni dari data autentik karena fungsinya berbeda, yaitu sebagai tolok ukur akhir yang mencerminkan performa model pada kondisi penggunaan yang senyatanya, bukan untuk memandu proses pelatihan. Dengan kata lain, *validation set* menjawab pertanyaan "kapan model berhenti belajar dengan baik", sedangkan *testing set* menjawab pertanyaan "seberapa baik model bekerja pada naskah nyata yang belum pernah dilihat sebelumnya".

4. Proses Pengacakan (*Shuffle*), untuk mencegah masalah model yang melupakan pola sebelumnya akibat melihat data secara berurutan (*catastrophic forgetting*), proses pengacakan atau *shuffle* diterapkan secara ketat dalam tiga tahap. Tahap pertama saat memuat data mentah, tahap kedua saat membuat bagian untuk eksperimen, dan tahap ketiga saat menggabungkan semua data untuk pelatihan utama. Langkah ini

memastikan bahwa setiap tahapan pelatihan berisi campuran seimbang antara gambar buatan dan gambar asli.

3.2.2.6 Normalisasi dan Spesifikasi Citra

Agar model Qwen2.5-VL-3B-Instruct bekerja optimal, dilakukan penyesuaian teknis pada gambar:

1. Resolusi Dinamis, gambar tidak dipaksa ke ukuran tertentu karena model mampu menerima berbagai resolusi tanpa menghilangkan detail penting seperti titik pada aksara Jawi.
2. Format Warna, semua gambar diubah ke format RGB, sementara normalisasi nilai piksel dilakukan secara otomatis oleh sistem pemroses gambar bawaan model.

3.2.3 Pengembangan Model

Tahap ini berfokus pada penyesuaian model Qwen2.5-VL-3B-Instruct agar lebih terampil mengenali aksara Jawi melalui proses *fine-tuning* yang hemat sumber daya. Model ini dipilih karena memiliki kemampuan memahami dokumen yang baik dengan jumlah parameter yang masih memungkinkan untuk dilatih pada perangkat terbatas.

3.2.3.1 *Fine-Tuning* Menggunakan QLoRA

Untuk meningkatkan efisiensi pelatihan model, penelitian ini menggunakan pendekatan *Quantized Low-Rank Adaptation* atau QLoRA. Model dasar terlebih dahulu dikompresi ke dalam format 4-bit dengan metode *NormalFloat4* (NF4) sehingga kebutuhan memori GPU dapat ditekan hingga sekitar 75% tanpa penurunan akurasi yang signifikan. Selanjutnya, seluruh bobot utama model dibekukan agar pengetahuan umum yang telah dipelajari sebelumnya tetap terjaga dan tidak berubah selama proses pelatihan lanjutan. Penyesuaian terhadap karakter Jawi dilakukan dengan menambahkan modul adapter LoRA yang memiliki tingkat ketelitian angka paling tinggi pada lapisan pemrosesan data (*attention* dan *feed-forward*), sehingga model dapat mempelajari pola khusus tanpa harus memperbarui seluruh parameter.

3.2.3.2 Konfigurasi *Hyperparameter*

Keberhasilan proses *fine-tuning* sangat dipengaruhi oleh pengaturan *hyperparameter* yang tepat. Dalam penelitian ini, eksperimen *hyperparameter* dirancang untuk menguji beberapa konfigurasi agar didapatkan performa paling optimal sebelum pelatihan utama dijalankan. Opsi yang akan dieksperimenkan mencakup variasi nilai LoRA Rank sebesar 32 dan 64, yang dipadukan dengan laju pembelajaran (*learning rate*) pada rentang 1e-4 hingga 5e-5. Parameter *dropout* LoRA diatur pada nilai nol agar model dapat fokus menyerap informasi secara maksimal. Untuk mengatasi keterbatasan memori pada GPU, penelitian ini menerapkan ukuran kelompok data (*batch size*) yang kecil dipadukan dengan teknik akumulasi gradien (*gradient accumulation*), sehingga efek pembelajarannya setara dengan menggunakan ukuran kelompok data yang besar.

Selain pengaturan pelatihan umum, konfigurasi parameter khusus QLoRA juga ditetapkan secara eksplisit. Modul target yang diadaptasi tidak hanya terbatas pada mekanisme *attention*, melainkan mencakup seluruh lapisan kritis pada arsitektur *Vision-Language Model*, yaitu lapisan visi, lapisan bahasa, serta modul pemrosesan lanjutan (*MLP modules*). Selain itu, panjang urutan maksimal (*max sequence length*) dirancang pada batasan 1024 token. Pemilihan batasan ini didasarkan pada analisis statistik distribusi token pada dataset, yang menunjukkan bahwa nilai 1024 mampu menaungi mayoritas sampel secara penuh tanpa memotong teks, sekaligus menghindari pemborosan alokasi memori yang sering terjadi jika menggunakan batas yang jauh lebih tinggi.

Ringkasan seluruh konfigurasi *hyperparameter* yang digunakan dalam penelitian ini disajikan pada Tabel 3.1.

Tabel 3.1 Konfigurasi *Hyperparameter* QLoRA

No	Parameter	Nilai	Keterangan
1	*Quantization Bits*	4-bit	Kuantisasi model dasar dengan NF4 untuk efisiensi memori GPU

No	Parameter	Nilai	Keterangan
2	*LoRA Rank* (r)	32 dan 64	Dimensi matriks adaptor yang akan dieksperimenkan
3	*LoRA Alpha*	2× *rank*	Faktor skala pembaruan bobot adaptor LoRA
4	*LoRA Dropout*	0.0	Menonaktifkan *dropout* untuk fokus penyerapan informasi
5	*Target Modules*	*Vision, Language, Attention, MLP*	Menargetkan seluruh lapisan kritis model untuk adaptasi penuh
6	*Learning Rate*	1e-4 hingga 5e-5	Laju pembelajaran untuk dieksperimenkan mencari stabilitas terbaik
7	*Batch Size & Accumulation*	*Batch* 4, *Accumulation* 4	Pengaturan memori untuk menyamai *batch size* efektif sebesar 16
8	*Max Sequence Length*	1024	Panjang token maksimal berdasarkan hasil analisis batas persentil distribusi data
9	*Save Strategy*	per *epoch*	Model akan disimpan pada setiap putaran *epoch* untuk evaluasi rinci

Sebagai gambaran konkret mengenai tingkat efesiensi pendekatan QLoRA yang diterapkan, dari keseluruhan 3.836.792.832 parameter yang dimiliki model Qwen2.5-VL-3B-Instruct, hanya 82.169.856 parameter atau setara 2,14% yang benar-benar diperbarui selama proses pelatihan melalui adapter LoRA. Sisanya, yaitu sekitar 97,86% parameter model dasar, tetap dibekukan dan tidak mengalami perubahan sama sekali selama pelatihan berlangsung. Angka ini dihasilkan secara

otomatis oleh *library* Unsloth setelah konfigurasi target modul (*vision, language, attention,* dan MLP) diterapkan pada model. Sebagai pembanding, pendekatan *fine-tuning* penuh (*full fine-tuning*) akan memperbarui seluruh atau 100% parameter model, sehingga jauh lebih boros dari sisi kebutuhan memori GPU maupun komputasi dibandingkan dengan QLoRA.

3.2.3.3 *Prompt Engineering*

Prompt engineering dilakukan dengan merancang instruksi sistem yang jelas dan konsisten agar model menghasilkan keluaran transliterasi sesuai dengan kaidah yang diharapkan. Instruksi dirancang untuk menegaskan bahwa tugas model terbatas pada ekstraksi teks dan transliterasi aksara Jawi ke huruf Latin, tanpa melakukan penerjemahan makna. *Prompt* juga mencakup format keluaran yang seragam untuk memudahkan proses evaluasi.

3.2.4 Pengujian dan Evaluasi

Evaluasi dilakukan secara komparatif untuk mengukur efektivitas intervensi *fine-tuning* terhadap kemampuan model dalam menangani aksara Jawi.

3.2.4.1 Pengujian Model *Baseline*

Sebelum masuk ke tahap *fine-tuning*, model diuji dalam kondisi asli (*pre-trained*) untuk mengukur kemampuan dasar model dalam mengenali skrip Jawi. Hasil dari tahap ini menjadi titik acuan untuk melihat peningkatan performa setelah teknik QLoRA diterapkan.

3.2.4.2 Pengujian Model *Fine-Tuned*

Setelah proses *fine-tuning* selesai, model kembali diuji menggunakan dataset uji yang bersifat independen dari data pelatihan. Model menerima input multimodal berupa citra teks aksara Jawi beserta instruksi yang sama seperti pada pengujian *baseline*. Output transliterasi yang dihasilkan kemudian dibandingkan dengan data *ground truth* untuk menghitung nilai kesalahan secara kuantitatif, sehingga dampak penerapan fine-tuning dapat dievaluasi secara langsung.

3.2.4.3 Metrik Evaluasi Pengenalan Teks

Evaluasi kinerja model dilakukan secara terpisah untuk dua tugas utama sistem, yaitu ekstraksi aksara asli dan transliterasi ke huruf Latin. Metrik yang digunakan adalah *Character Error Rate* (CER) dan *Word Error Rate* (WER). Untuk menilai ketepatan model dalam membaca gambar teks asli, sistem akan mengukur nilai CER Ekstraksi Jawi dan WER Ekstraksi Jawi. Sementara itu, untuk menilai keakuratan pengubahan ke dalam bahasa Latin, sistem akan mengukur nilai CER Transliterasi Rumi dan WER Transliterasi Rumi. Penggunaan metrik secara spesifik ini sangat penting karena kesalahan kecil seperti pembacaan titik pada aksara Jawi dapat berdampak besar pada struktur kata Latin yang dihasilkan.

Indikator keberhasilan model dalam alur penelitian (Gambar 3.1) ditentukan secara tegas berdasarkan ambang batas nilai kesalahan. Hasil pengujian dikategorikan "memenuhi target" apabila nilai metrik kesalahan, terutama *Character Error Rate*, berhasil ditekan hingga di bawah angka 15%. Angka ini ditetapkan sebagai standar kelayakan yang menunjukkan bahwa model mampu mengenali sebagian besar karakter Jawi dengan tepat. Sebaliknya, hasil pengujian akan dikategorikan "tidak memenuhi target" apabila nilai kesalahan masih berada pada atau melampaui batas 15%. Sesuai dengan bagan alur penelitian, apabila target belum terpenuhi, siklus penelitian akan berputar kembali ke tahap pengumpulan dan persiapan data. Pada tahap perulangan tersebut, akan dilakukan evaluasi ulang terhadap kualitas dataset serta penyesuaian ulang parameter pelatihan hingga model akhirnya berhasil mencapai standar akurasi yang telah ditetapkan.

Perlu dicatat bahwa ambang batas 15% ditetapkan sebagai target kerja internal penelitian, bukan sebagai standar baku yang berlaku umum, mengingat nilai CER yang dianggap baik sangat bergantung pada konteks penggunaannya, seperti jenis dokumen, kondisi cetak, dan kompleksitas aksara yang diproses. Penetapan ambang batas ini digunakan semata-mata sebagai kriteria praktis untuk menghentikan siklus perulangan pada alur penelitian (Gambar 3.1), bukan sebagai klaim bahwa nilai tersebut merupakan tolok ukur kualitas OCR secara ilmiah.

3.2.4.4 Desain Studi Ablasi Komposisi Data

Selain membandingkan model *baseline* dan model hasil *fine-tuning*, penelitian ini juga merancang studi ablasi (*ablation study*) untuk menguji pengaruh komposisi data latih terhadap kemampuan model. Studi ablasi ini dirancang untuk menjawab pertanyaan apakah model cukup dilatih menggunakan satu jenis data saja, yaitu data sintetik saja atau data autentik saja, ataukah kombinasi keduanya benar-benar diperlukan untuk mencapai kemampuan membaca aksara Jawi yang optimal. Tiga skenario pelatihan dirancang untuk menjawab pertanyaan tersebut: Skenario A menggunakan seluruh data sintetik sebagai data latih, Skenario B menggunakan porsi data latih dari data autentik saja, dan Skenario C menggunakan kombinasi keduanya sebagaimana model akhir penelitian ini (lihat sub-bab 3.2.2.5). Agar pengaruh komposisi data latih dapat diukur secara adil, ketiga skenario dirancang untuk dievaluasi pada satu set data uji yang identik, yaitu data uji autentik yang sama dengan yang digunakan pada pengujian model *baseline* dan model final (sub-bab 3.2.4.1), sehingga yang bervariasi hanya komposisi data latih, bukan data ujinya. Hasil dan pembahasan lengkap studi ablasi ini disajikan pada sub-bab 4.7.

3.2.5 Analisis Hasil dan Kesimpulan

Tahap akhir penelitian ini difokuskan pada analisis hasil eksperimen dengan membandingkan nilai *Character Error Rate* (CER) dan *Word Error Rate* (WER) antara model *baseline* dan model hasil *fine-tuning*. Perbandingan ini bertujuan untuk melihat pengaruh penerapan strategi QLoRA terhadap peningkatan kinerja model dalam tugas ekstraksi dan transliterasi aksara Jawi. Selain itu, dilakukan analisis kesalahan untuk mengidentifikasi karakter atau kondisi citra tertentu yang masih sulit dikenali oleh model.

Berdasarkan hasil analisis tersebut, disusun kesimpulan penelitian yang merangkum temuan utama untuk menjawab rumusan masalah. Kesimpulan ini menegaskan tingkat efektivitas implementasi Qwen2.5-VL-3B-Instruct dalam melakukan ekstraksi dan transliterasi aksara Jawi, sekaligus menggambarkan batasan sistem yang dapat menjadi dasar pengembangan penelitian selanjutnya.

3.2.6 *Deployment* Model

Sebagai tahap akhir, model terbaik hasil *fine-tuning* akan di-*deploy* ke platform *Hugging Face Spaces* sebagai antarmuka demonstrasi yang dapat diakses publik. Platform ini dipilih karena menyediakan infrastruktur gratis untuk *hosting* model *machine learning*, terintegrasi langsung dengan ekosistem *Hugging Face* untuk pengelolaan model dan dataset, serta memungkinkan peneliti lain untuk mencoba sistem secara langsung. Melalui antarmuka ini, pengguna dapat mengunggah gambar teks aksara Jawi dan mendapatkan hasil ekstraksi sekaligus transliterasi dalam format teks Latin secara langsung.

3.3 Waktu dan Lokasi Penelitian

3.3.1 Waktu Penelitian

Penelitian ini dilaksanakan dalam periode 6 bulan dengan rincian jadwal sebagai berikut:

Tabel 3.2 Tahapan Kegiatan Penelitian

No	Tahapan Kegiatan	Bulan 1	Bulan 2	Bulan 3	Bulan 4	Bulan 5	Bulan 6
1	Studi Literatur & Identifikasi Masalah						
2	Pengumpulan & Persiapan Data						
3	Pengembangan Model						
4	Evaluasi Model & Pengujian						
5	Analisis Hasil dan Kesimpulan						

3.3.2 Lokasi Penelitian

Penelitian ini dilaksanakan di Laboratorium Komputasi Program Studi Teknologi Informasi Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh untuk pengembangan dan pelatihan model.

Remote computing dilakukan dengan penggunaan cloud computing platforms seperti Google Colab Pro, Kaggle, atau layanan cloud lainnya untuk akselerasi komputasi GPU yang diperlukan dalam fine-tuning model deep learning.

3.4 Alat dan Bahan

3.4.1 Perangkat Keras

Tabel 3.3 Perangkat Keras

No	Komponen	Spesifikasi
1	Komputer Pengembangan	Processor Intel Core i5 Gen 11
2	RAM	24 GB DDR4
3	Storage Lokal	SSD 512 GB

3.4.2 Perangkat Lunak

Tabel 3.4 Perangkat Lunak

No	Kategori	Tools/Library	Versi
1	Programming Language	Python	3.9+
2	Deep Learning Framework	PyTorch	2.0+
		CUDA Toolkit	11.8+
3	Model & Training	Transformers (Hugging Face)	4.35+
		PEFT	0.7+
		bitsandbytes	Lates
		Unsloth	Lates
4	Evaluation & Metrics	JiWER	3.0+
		Python-Levenshtein	0.21+
5	Development Tools	Google Colab	-
		Antigravity	-

BAB IV
HASIL DAN PEMBAHASAN

Bab ini membahas hasil pengujian model Qwen2.5-VL-3B-Instruct setelah melalui proses *fine-tuning* untuk mengekstrak aksara Jawi dan mentransliterasikan ke huruf Latin (Rumi). Pembahasan disusun mulai dari gambaran umum dataset, hasil pengujian model dasar, eksperimen pencarian *hyperparameter*, hasil pelatihan akhir, studi ablasi (*ablation study*) pengaruh komposisi data, sampai analisis kesalahan dan demonstrasi aplikasi yang dibangun.

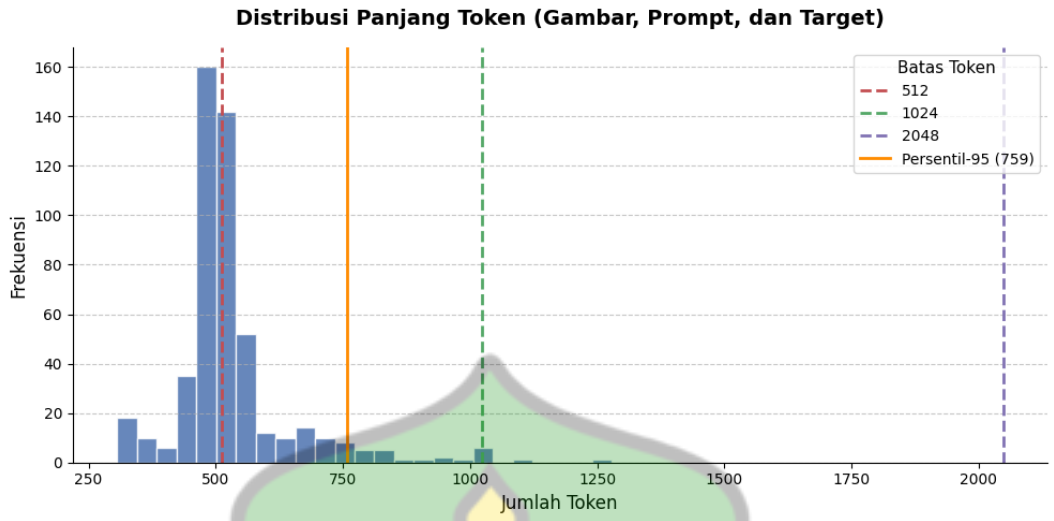
4.1 Gambaran Umum Hasil Penelitian

Penelitian ini berhasil menerapkan model multimodal Qwen2.5-VL-3B-Instruct untuk mengekstraksi dan mentransliterasikan aksara Jawi cetak. Setelah melalui proses *fine-tuning* dengan QLoRA, model menunjukkan peningkatan kemampuan yang cukup besar dibanding kondisi awalnya sebelum dilatih. Pada bagian-bagian berikut akan dijelaskan secara berurutan bagaimana data disiapkan, bagaimana model diuji pada tiap tahap, dan apa saja yang ditemukan dari proses tersebut, termasuk batasan yang masih dimiliki sistem ini.

4.2 Hasil Pengolahan Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari dua jenis data, yaitu data sintetik yang dibuat dari *font* komputer dan data autentik yang diambil dari naskah Jawi asli.

Sebelum data digunakan untuk melatih model, dilakukan analisis terhadap panjang token pada teks supaya batas panjang input model (*max sequence length*) bisa ditentukan dengan tepat. Hasil analisis menunjukkan rata-rata panjang token sekitar 530, dengan nilai persentil ke-95 berada di angka 759 token. Berdasarkan temuan ini, batas panjang token ditetapkan pada 1024, karena nilai tersebut sudah cukup menampung hampir seluruh data tanpa harus memotong teks, sekaligus tidak terlalu besar sehingga membuang-buang memori GPU. Sebagai gambaran, jika batas yang dipakai hanya 512 token (nilai default yang umum dipakai), lebih dari separuh data justru akan terpotong karena di atas rata-rata. Hal ini dapat dilihat jelas pada Gambar 4.1.



Gambar 4.1 Grafik Distribusi Panjang Token Dataset Jawi

Pembagian dataset dilakukan dengan memisahkan 10% data autentik sebagai data uji akhir, tanpa campuran data sintetik sama sekali. Sisa data autentik digabung dengan seluruh data sintetik untuk menjadi data latih dan data validasi. Pemisahan ini sengaja dibuat murni dari data autentik karena tujuan akhir sistem adalah membaca naskah Jawi yang nyata, bukan teks buatan yang relatif lebih mudah dibaca. Jika data sintetik ikut dimasukkan ke dalam data uji, hasil evaluasi berisiko terlihat lebih bagus dari kenyataan.

4.3 Hasil Pengujian Model Dasar (*Baseline*)

Sebelum dilatih, model terlebih dahulu diuji dalam kondisi aslinya (*zero-shot*) untuk mengetahui sejauh mana kemampuan dasarnya dalam mengekstraksi dan mentransliterasikan aksara Jawi. Hasil pengujian ini menjadi acuan pembanding untuk melihat seberapa besar peningkatan yang dihasilkan oleh proses *fine-tuning*. Tabel 4.2 menunjukkan hasil pengujian model dasar pada data uji autentik.

Tabel 4.1 Hasil Pengujian Model Dasar (*Baseline*)

Tugas	**CER**	**WER**
Ekstraksi Jawi	0.825	0.961
Transliterasi Rumi	0.906	1.071

Dari tabel di atas terlihat bahwa model dasar menghasilkan tingkat kesalahan yang sangat tinggi pada keempat metrik. Nilai WER Transliterasi Rumi yang mencapai 1,071 menunjukkan bahwa keluaran model umumnya jauh berbeda dari teks aslinya, baik dari segi jumlah kata maupun susunannya, bukan sekadar salah menebak sebagian kata. Kondisi ini cukup wajar terjadi karena model dasar dilatih untuk memahami bahasa-bahasa umum seperti Inggris dan belum pernah diberi contoh khusus aksara Jawi, sehingga adanya proses *fine-tuning* memang diperlukan agar model bisa menyesuaikan diri dengan karakteristik aksara ini.

4.4 Eksperimen Konfigurasi *Hyperparameter*

Sebelum pelatihan utama dijalankan, dilakukan eksperimen pendahuluan untuk mencari kombinasi *hyperparameter* yang paling sesuai. Tiga konfigurasi dicoba dengan variasi pada nilai LoRA *rank* dan *learning rate*, masing-masing dilatih pada subset data kecil (500 sampel, 1 *epoch*) supaya proses pencarian ini tidak memakan waktu dan biaya komputasi yang besar. Hasil dari ketiga konfigurasi tersebut dapat dilihat pada Tabel 4.3.

Tabel 4.2 Hasil Eksperimen Konfigurasi *Hyperparameter*

Konfigurasi Eksperimen	***Rank LoRA***	***Learning Rate***	**CER Jawi**	**CER Rumi**
Konfigurasi 1	32	5e-5	0,271	0,447
Konfigurasi 2	64	2e-5	0,402	0,575
Konfigurasi 3	32	1e-4	0,246	0,412

Pemilihan konfigurasi terbaik pada tahap ini difokuskan pada nilai CER, bukan pada *training loss* atau *validation loss*. Alasannya, CER mengukur langsung seberapa tepat model membaca bentuk huruf, sedangkan nilai *loss* hanya menggambarkan perhitungan matematis yang tidak selalu sejalan dengan kualitas hasil bacaan yang sebenarnya. Konfigurasi 3, dengan *rank* 32 dan *learning rate* 1e-4, terbukti memberikan hasil CER paling rendah pada kedua metrik. Konfigurasi inilah yang kemudian dipakai untuk seluruh proses pelatihan selanjutnya, termasuk pelatihan final dan studi ablasi.

4.5 Konfigurasi *Prompt* yang Digunakan dalam Pelatihan dan Pengujian

Keberhasilan model dalam menghasilkan keluaran yang konsisten dan mudah diproses secara otomatis sangat bergantung pada rancangan instruksi (*prompt*) yang diberikan kepada model. Prompt yang digunakan pada penelitian ini terdiri atas dua bagian, yaitu *system prompt* yang mendefinisikan peran dan aturan keluaran model, serta *user prompt* yang menjadi instruksi spesifik yang menyertai setiap gambar. Kedua prompt ini digunakan secara identik dan konsisten pada seluruh tahapan penelitian, baik saat menyusun data pelatihan, saat proses *fine-tuning* berlangsung, maupun saat pengujian model *baseline* dan model hasil *fine-tuning*, sehingga perbandingan performa antar tahap tetap adil karena tidak ada perbedaan instruksi yang memengaruhi hasil.

System prompt di atas dirancang dengan dua instruksi berurutan yang tegas. Pertama, model diminta membaca dan mengekstrak teks aksara Jawi dari gambar seakurat mungkin, persis seperti yang tertulis pada naskah. Kedua, hasil ekstraksi tersebut ditransliterasikan ke huruf Latin (Rumi) dengan mengikuti pedoman Ejaan Jawi yang Disempurnakan (PEJYD) dari Dewan Bahasa dan Pustaka Malaysia. Format keluaran juga ditentukan secara eksplisit dalam bentuk dua baris berlabel "Jawi:" dan "Rumi:", tanpa penjelasan atau komentar tambahan. Ketegasan format ini penting agar keluaran model dapat langsung diuraikan (*parsing*) secara otomatis oleh sistem evaluasi tanpa memerlukan penyaringan manual, sekaligus mencegah model menyisipkan tafsiran atau penjelasan makna yang tidak diminta. Sementara itu, *user prompt* berfungsi sebagai instruksi singkat yang menyertai setiap gambar yang diunggah, menegaskan kembali bahwa tugas yang diminta mencakup ekstraksi sekaligus transliterasi dalam satu kali proses inferensi.

4.6 Hasil Pelatihan Final

Model dilatih menggunakan konfigurasi terbaik di atas selama 5 *epoch*, dengan data latih gabungan dari seluruh data sintetik dan 90% data autentik. Pergerakan nilai *loss* pada setiap *epoch* direkam pada Tabel 4.4.

Tabel 4.3 Riwayat Nilai *Loss* Pelatihan Final (5 *Epoch*)

Epoch	***Training Loss***	***Validation Loss***	**Indikasi Status**
1	0,0277	0,0306	Normal
2	0,0197	0,0256	Normal
3	0,0094	0,0252	Terbaik
4	0,0038	0,0271	*Overfitting*
5	0,0013	0,0288	*Overfitting*

Berdasarkan Tabel 4.4, nilai *validation loss* menurun secara konsisten dari *epoch* 1 sampai *epoch* 3, lalu mulai naik lagi pada *epoch* 4 dan 5 meskipun *training loss*-nya terus turun. Pola ini menjadi tanda *overfitting*: model semakin menghafal data latih, tetapi justru semakin buruk saat menghadapi data yang belum pernah dilihat. Karena alasan ini, sistem secara otomatis memilih *checkpoint epoch* 3 sebagai model akhir melalui mekanisme *load_best_model_at_end*, dan *checkpoint* inilah yang dipakai pada seluruh pengujian berikutnya di bab ini.

Sebelum masuk ke pembahasan studi ablasi yang membandingkan beberapa skenario komposisi data, perlu ditampilkan terlebih dahulu hasil evaluasi dari model akhir yang sesungguhnya, yaitu *checkpoint epoch* 3 pada Tabel 4.4. Model inilah yang menjadi luaran utama penelitian ini dan yang akan digunakan pada aplikasi inferensi. Tabel 4.5 menunjukkan hasil evaluasinya pada data uji autentik yang sama dengan yang digunakan pada pengujian model dasar.

Tabel 4.4 Hasil Evaluasi Model Final (*Epoch* 3) pada Data Uji

Tugas	**CER**	**WER**
Ekstraksi Jawi	0,174	0,500
Transliterasi Rumi	0,165	0,394

Nilai- nilai pada Tabel 4.5 inilah yang menjadi acuan utama untuk menilai keberhasilan penelitian, sekaligus menjadi rujukan pembanding bagi Skenario C pada studi ablasi di bagian selanjutnya. Dengan hasil ini sebagai titik tolak, studi ablasi berikutnya diarahkan untuk menjawab pertanyaan mengapa komposisi data

gabungan sintetik dan autentik dipilih, dibandingkan dengan alternatif menggunakan salah satu jenis data saja.

4.7 Studi Ablasi Pengaruh Komposisi Data

Studi ablasi ini dirancang untuk menjawab satu pertanyaan mendasar, yaitu apakah model cukup dilatih menggunakan satu jenis data saja, baik data sintetik saja maupun data autentik saja, ataukah kombinasi keduanya benar-benar diperlukan untuk mencapai kemampuan membaca aksara Jawi yang optimal. Untuk menjawab pertanyaan tersebut, dilakukan tiga skenario pelatihan, yaitu Skenario A yang hanya menggunakan data sintetik, Skenario B yang hanya menggunakan data autentik, dan Skenario C yang menggunakan kombinasi kedua jenis data sebagai model akhir. Agar ketiga skenario benar-benar dapat menjawab pertanyaan tersebut secara adil, prinsip yang diterapkan pada studi ablasi ini adalah menjaga data uji tetap konstan sementara komposisi data latih yang divariasikan. Dengan kata lain, Skenario A, B, dan C sama-sama dievaluasi pada satu set data uji yang identik, yaitu citra data autentik yang telah ditetapkan pada sub-bab 3.2.2.5 dan yang sama dengan data uji yang dipakai pada pengujian model *baseline* (Tabel 4.2) dan model final Skenario C (Tabel 4.5). Data uji sengaja dipertahankan berupa data autentik, bukan data sintetik, karena tujuan akhir sistem adalah membaca naskah Jawi yang nyata (lihat juga sub-bab 3.2.2.1), sehingga kemampuan generalisasi ke data autentik inilah yang sebenarnya hendak diukur dari ketiga skenario. Dengan menjaga data uji tetap sama, perbedaan nilai CER/WER antar skenario dapat dimurnikan sebagai akibat dari perbedaan komposisi data latih semata, yaitu satu jenis data versus kombinasi keduanya, bukan akibat perbedaan tingkat kesulitan data uji.

4.7.1 Skenario A (Hanya Data Sintetik)

Pada skenario ini, model dilatih selama satu *epoch* menggunakan seluruh data sintetik sebagai data pelatihan, kemudian dievaluasi menggunakan citra data autentik yang sama dengan data uji pada Tabel 4.2 dan Tabel 4.5, yang tidak pernah digunakan selama proses pelatihan. Pengujian ini bertujuan untuk mengetahui sejauh mana model yang hanya mempelajari karakteristik data sintetik mampu

mengenali teks Jawi pada naskah autentik. Hasil evaluasi ditampilkan pada Tabel 4.6.

Tabel 4.5 Hasil Evaluasi Skenario A (Data Sintetik)

Tugas	**CER**	**WER**
Ekstraksi Jawi	0,311	0,744
Transliterasi Rumi	0,338	0,718

Berdasarkan hasil evaluasi, model yang hanya dilatih menggunakan data sintetik masih menghasilkan tingkat kesalahan yang relatif tinggi ketika diuji pada data autentik. Hal ini menunjukkan bahwa meskipun data sintetik mampu membantu model mempelajari bentuk dasar karakter Jawi, model masih mengalami kesulitan ketika dihadapkan pada karakteristik naskah asli yang memiliki variasi kualitas citra, seperti bagian yang memudar, maupun perbedaan bentuk penulisan. Dengan demikian, data sintetik saja belum cukup untuk menghasilkan kemampuan generalisasi yang baik terhadap data autentik.

4.7.2 Skenario B (Hanya Data Autentik)

Pada skenario ini, model dilatih selama satu epoch menggunakan data autentik pada porsi data latih sebagai data pelatihan tanpa memanfaatkan data sintetik. Selanjutnya, model dievaluasi menggunakan citra data uji autentik yang sama dengan data uji *Baseline*, Skenario A, dan Skenario C, sehingga performa Skenario B dapat dibandingkan secara adil dengan skenario lain. Pengujian ini bertujuan untuk mengetahui kemampuan model dalam mengenali karakter Jawi pada naskah autentik setelah mempelajari data autentik semata. Hasil evaluasi ditampilkan pada Tabel 4.7.

Tabel 4.6 Hasil Evaluasi Skenario B (Data Autentik)

Tugas	**CER**	**WER**
Ekstraksi Jawi	0,208	0,570
Transliterasi Rumi	0,229	0,508

Hasil evaluasi menunjukkan bahwa model yang dilatih menggunakan data autentik memperoleh tingkat kesalahan yang lebih rendah dibandingkan Skenario A. Meskipun jumlah data pelatihannya jauh lebih sedikit dibanding data sintetik, data autentik memiliki karakteristik yang lebih mendekati kondisi nyata data uji sehingga mampu membantu model mempelajari pola penulisan aksara Jawi dengan lebih baik untuk kasus pengujian pada naskah autentik. Hasil ini menunjukkan bahwa kualitas dan kesesuaian karakteristik data pelatihan dengan kondisi penggunaan sebenarnya memberikan pengaruh yang besar terhadap performa model, bahkan lebih besar dibanding sekadar kuantitas data. Karena Skenario A dan Skenario B kini sama-sama dievaluasi pada citra data uji autentik yang identik, kedua hasil ini dapat dibandingkan secara adil: performa Skenario B yang lebih baik daripada Skenario A menunjukkan bahwa, ketika hanya satu jenis data yang tersedia, data autentik memberi bekal generalisasi yang lebih relevan untuk tugas ini dibanding data sintetik. Namun demikian, kedua skenario satu-sumber ini tetap kalah dibanding Skenario C pada sub-bab berikutnya, yang menegaskan bahwa kombinasi keduanya adalah pilihan terbaik.

4.7.3 Skenario C (Gabungan Data Sintetik dan Autentik)

Skenario C adalah model final yang hasil evaluasinya sudah ditampilkan pada Tabel 4.5, yaitu model yang dilatih dengan gabungan seluruh data sintetik dan 90% data autentik selama 5 *epoch*. Untuk melihat bagaimana performanya berkembang di setiap *epoch*, dilakukan evaluasi ulang pada *checkpoint* tiap *epoch* menggunakan data uji yang sama. Hasilnya direkap pada Tabel 4.8.

Tabel 4.7 Hasil Evaluasi Skenario C per *Epoch*

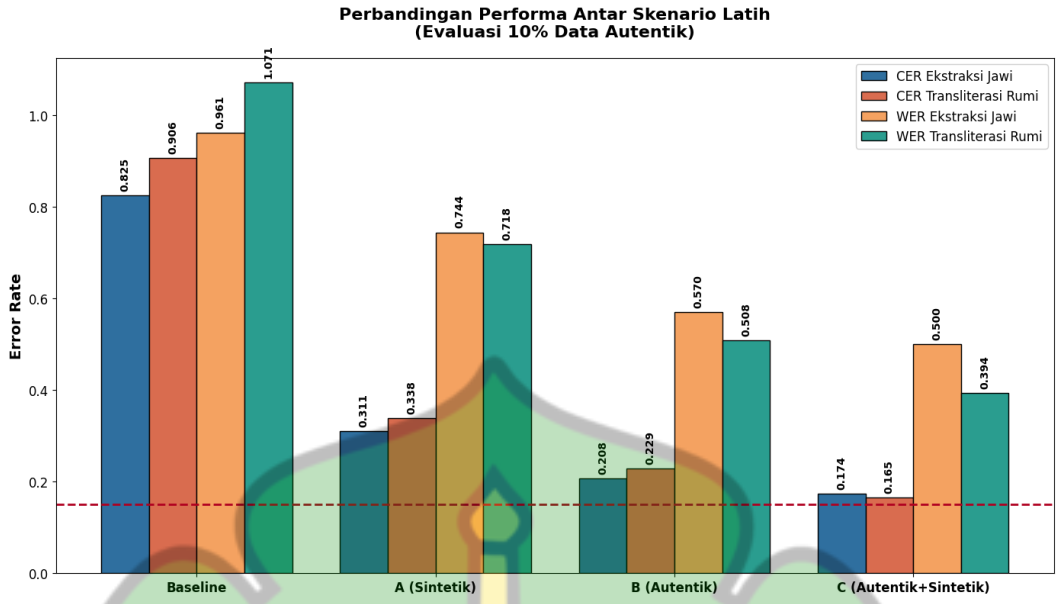
Epoch	**Ekstraksi Jawi**		**Transliterasi Rumi**	
	CER	**WER**	**CER**	**WER**
1	0,201	0,535	0,201	0,445
2	0,183	0,525	0,171	0,413
3	**0,174**	**0,500**	**0,165**	**0,394**
4	0,173	0,496	0,166	0,399
5	0,174	0,498	0,171	0,406

Pada Tabel 4.8, seluruh metrik menunjukkan penurunan yang cukup signifikan dari *epoch* pertama hingga *epoch* ketiga. Hal ini menunjukkan bahwa model semakin mampu mengenali karakter aksara Jawi dan menghasilkan transliterasi yang lebih mendekati *ground truth*. Setelah *epoch* ketiga, perubahan nilai evaluasi cenderung relatif kecil. Meskipun pada *epoch* keempat nilai CER dan WER untuk tugas ekstraksi Jawi sedikit lebih rendah dibandingkan *epoch* ketiga, perbedaannya sangat kecil, yaitu hanya sebesar 0,001 pada CER dan 0,004 pada WER. Sebaliknya, pada tugas transliterasi Rumi, nilai CER dan WER justru mengalami sedikit peningkatan dibandingkan *epoch* ketiga.

Dengan demikian, tidak terdapat peningkatan performa yang konsisten pada seluruh metrik setelah *epoch* ketiga. Selain itu, berdasarkan hasil pemantauan selama pelatihan, nilai *validation loss* mencapai titik terendah pada *epoch* ketiga sebelum mulai meningkat pada *epoch* berikutnya. Kondisi tersebut menunjukkan bahwa model mulai memasuki fase *overfitting*, yaitu ketika model semakin menyesuaikan diri terhadap data pelatihan, tetapi kemampuan generalisasinya terhadap data yang belum pernah dilihat tidak lagi meningkat secara konsisten. Oleh karena itu, *checkpoint* pada *epoch* ketiga dipilih sebagai model final karena memberikan keseimbangan terbaik antara hasil evaluasi menggunakan CER dan WER serta kemampuan generalisasi model yang ditunjukkan oleh nilai *validation loss*.

4.7.4 Evaluasi Metrik Antar Skenario

Untuk melihat gambaran besarnya, hasil ketiga skenario di atas dibandingkan langsung dengan model dasar (baseline) pada Gambar 4.2. Karena keempat kelompok data yang ditampilkan (Baseline, Skenario A, Skenario B, dan Skenario C) seluruhnya dievaluasi pada data uji autentik yang sama, keempatnya dapat dibandingkan secara adil dalam satu grafik.



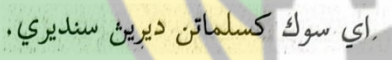
Gambar 4.2 Grafik Perbandingan Evaluasi *Baseline*, Skenario A, B, dan C

Dari keempat hasil tersebut, yang seluruhnya diuji pada citra autentik yang sama, urutan performa dari yang paling lemah ke paling kuat adalah *Baseline*, kemudian Skenario A, disusul Skenario B, dan yang terbaik adalah Skenario C. Urutan ini menguatkan dua hal. Pertama, *fine-tuning* dengan data apa pun jelas lebih baik daripada tidak sama sekali, karena ketiga skenario yang dilatih (A, B, maupun C) seluruhnya jauh mengungguli *Baseline*. Kedua, kombinasi data sintetik dan autentik pada Skenario C memberi hasil paling baik dibandingkan penggunaan salah satu jenis data saja (Skenario A atau B), karena keduanya saling melengkapi: data sintetik membekali model dengan pemahaman bentuk dasar huruf secara bersih dan terstruktur dalam jumlah besar, sementara data autentik melatih model untuk tahan terhadap gangguan visual yang nyata meskipun jumlahnya lebih terbatas. Dengan kata lain, menggunakan data secara terpisah (Skenario A atau B saja) tetap memberi peningkatan dibanding tidak *fine-tuning* sama sekali, tetapi menggabungkan keduanya (Skenario C) terbukti menjadi strategi komposisi data yang paling optimal untuk tugas ekstraksi dan transliterasi aksara Jawi pada penelitian ini.

4.7.5 Analisis Prediksi Terbaik dan Terburuk

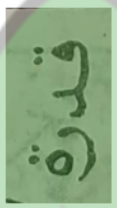
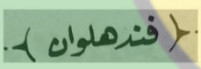
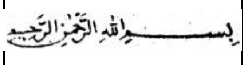
Untuk memahami lebih jauh kapan model bekerja baik dan kapan gagal, dilakukan pemeriksaan terhadap sampel-sampel dengan nilai kesalahan paling rendah dan paling tinggi dari model final (Skenario C, *epoch* 3).

Tabel 4.8 Contoh Prediksi dengan Akurasi Tinggi

No	Gambar Naskah	Teks Asli	Prediksi Model
1		**Jawi:** والله اعلم . **Rumi:** wallahu a'lam.	**Jawi:** والله اعلم . **Rumi:** wallahu a'lam. *(CER Jawi: 0,000; WER Jawi: 0,000; CER Rumi: 0,000; WER Rumi: 0,000)*
2		**Jawi:** اي سوك كسلماتن ديريڽ سنديري . **Rumi:** ia suka keselamatan dirinya sendiri.	**Jawi:** اي سوك كسلماتن ديريڽ سنديري . **Rumi:** ia suka keselamatan dirinya sendiri. *(CER Jawi: 0,000; WER Jawi: 0,000; CER Rumi: 0,000; WER Rumi: 0,000)*
3		**Jawi:** دسوسن اوله **Rumi:** disusun oleh	**Jawi:** دسوسن اوله **Rumi:** disusun oleh *(CER Jawi: 0,000; WER Jawi: 0,000; CER Rumi: 0,000; WER Rumi: 0,000)*

Sampel-sampel pada Tabel 4.9 menunjukkan bahwa model bekerja sangat baik pada gambar dengan pencahayaan jelas, potongan baris yang rapi, dan panjang teks yang wajar (tidak terlalu pendek). Pada kondisi ini, model bisa membaca dan mentransliterasikan teks tanpa kesalahan sama sekali.

Tabel 4.9 Contoh Prediksi dengan Akurasi Rendah

No	Gambar Naskah	Teks Asli	Prediksi Model
1		**Jawi:** قدرة **Rumi:** qudrah	**Jawi:** اف باپق **Rumi:** apa banyak *(CER Jawi: 1.750; WER Jawi: 2,000; CER Rumi: 1,500; WER Rumi: 2,000)*
2		**Jawi:** (فندهلوان) **Rumi:** (pendahuluan)	**Jawi:** (فندر هلوا ن) **Rumi:** (pender halwan) *(CER Jawi: 0,364; WER Jawi: 2,000; CER Rumi: 0,385; WER Rumi: 2,000)*
3		**Jawi:** بِسمِ اللهِ الرَّحْمَنِ الرَّحِيْمِ **Rumi:** bismillahirrahmanir rahim	**Jawi:** بسم الله وَالْحَمْدُ لِلّٰهِ رَبِّ الْعَالَمِيْنَ **Rumi:** bismillahirrahmani rrah-imualhamdulilla-hrabbal alamin *(CER Jawi: 0.811; WER Jawi: 1.500; CER Rumi: 1.125; WER Rumi: 2,000)*

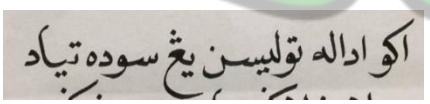
Sebaliknya, Tabel 4.10 memperlihatkan tiga kasus dengan penyebab kesalahan yang berbeda-beda jika ditelusuri dari gambar sumbernya masing-masing. Pada gambar pertama, gambar sumber ternyata memiliki orientasi vertikal, berbeda dari mayoritas data latih yang berupa potongan baris horizontal. Pada gambar kedua dan ketiga, model masih kesulitan saat menghadapi teks yang memuat tanda baca atau simbol tambahan, karena mayoritas data latih berupa teks tanpa harakat dan simbol sehingga model belum terbiasa dengan variasi semacam itu.

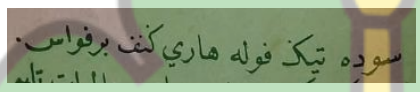
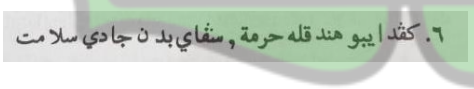
Satu hal penting yang juga perlu disampaikan sebagai batasan sistem, model ini paling optimal untuk membaca potongan gambar berupa satu baris, sesuai dengan karakteristik mayoritas data latihnya. Untuk gambar berupa paragraf panjang, model cenderung mengulang format "Rumi:" di setiap baris karena belum banyak contoh paragraf utuh selama pelatihan. Ini menjadi salah satu catatan yang perlu ditindaklanjuti pada penelitian selanjutnya.

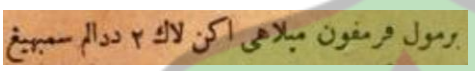
4.8 Inferensi Perbandingan Model *Baseline* dan *Fine-Tuned*

Selain melihat angka metrik, perbedaan kemampuan model sebelum dan sesudah *fine-tuning* juga dapat diamati langsung dari hasil pembacaan pada contoh gambar baru. Tabel 4.11 menunjukkan perbandingan keluaran antara model dasar (*baseline*) dan model yang sudah dilatih (*fine-tuned*) pada beberapa potongan naskah uji.

Tabel 4.10 Perbandingan Prediksi Inferensi Teks Jawi

No	Potongan Naskah Uji	*Baseline*	*Fine-Tuned*
1		**Jawi**: اکو اداله تولیسن یثُ سوده تیااد **Rumi**: أکو أداله تولیسن یثُ سوده تیااد *(CER Jawi: 0,137; WER Jawi: 0,500;*	**Jawi**: اکو اداله تولیسن یڠ سوده تیاد **Rumi**: aku adalah tulisan yang sudah tiada

		CER Rumi: 0,857; WER Rumi: 1,000)	*(CER Jawi:0,000; WER Jawi: 0,000; CER Rumi: 0,000; WER Rumi: 0,000)*
2		**Jawi**: سوده تيک فوله هاري کنف برفواس **Rumi**: Sode teik foulah hari kent burfaras *(CER Jawi: 0,133; WER Jawi: 0,500; CER Rumi: 0,486; WER Rumi: 0,833)*	**Jawi**: سوده تيڬ فوله هاري کفد برفواس . **Rumi**: sudah tiga puluh hari kepada berpuasa. *(CER Jawi: 0,133; WER Jawi: 0,500; CER Rumi: 0,108; WER Rumi: 0,166)*
3		**Jawi**: كَفَدَا اِبْوُ هَنْدَ قَلْهُ حَرْمَةٌ ، سَتَفَيْا بَدْنَ جَادِي سَلاَمَت **Rumi**: Kafada ibnu Hindah qalhuh harma, satfaya bdn jadi salamat *(CER Jawi: 0,760; WER*	**Jawi**: ٦ . كڤد ايبو هند قله حرمة , ستيا فى بدن جادي سلامت **Rumi**: 6. kepada ibu hendaklah hormat, setiap benda jadi selamat

		Jawi: 0,909; CER Rumi: 0,403; WER Rumi: 0,888)	*(CER Jawi: 0,152; WER Jawi: 0,454; CER Rumi: 0,140; WER Rumi: 0,222)*
4		**Jawi**: رمول فرمفون مبلاهي أكن لاك ٢ ددام سمهيج **Rumi**: Ramul Farumfon Mbalahi Aken Lak 2 Dam Samahyeg *(CER Jawi: 0,190; WER Jawi: 0,625; CER Rumi: 0.541; WER Rumi: 1.142)*	**Jawi**: برمول فرمفون مبلاهى اكن لاک ٢ ددالم سمبهيغ . **Rumi**: Bermula perampun melampaui akan laka-laka didalam sembahyang. *(CER Jawi: 0,071; WER Jawi: 0,250; CER Rumi: 0.180; WER Rumi: 0.714)*

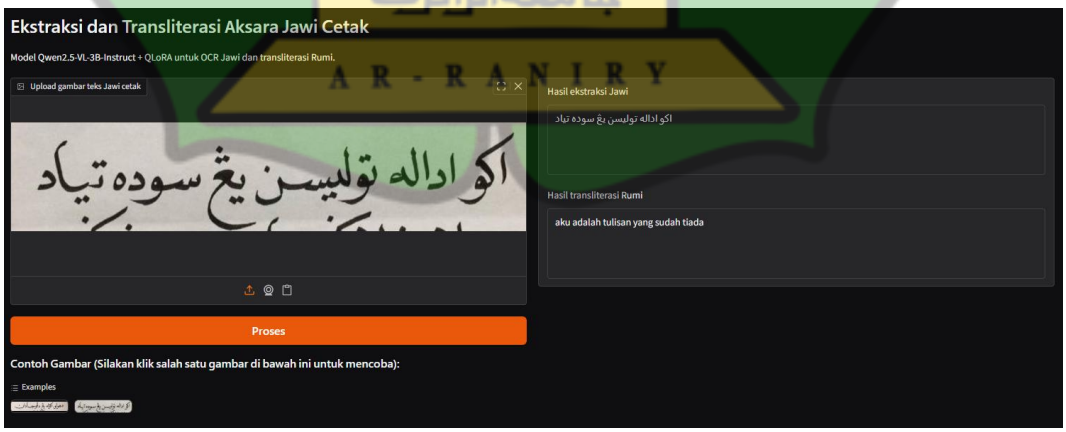
Berdasarkan Tabel 4.11, kolom "Potongan Naskah Uji" berisi citra masukan yang diberikan ke model. Kolom "Baseline" berisi dua baris keluaran mentah model sebelum *fine-tuning*, yaitu baris "Jawi:" berupa hasil ekstraksi aksara Arab Jawi dan baris "Rumi:" yang seharusnya berupa transliterasi Latin. Kolom "Fine-Tuned" memuat dua baris yang sama untuk model setelah *fine-tuning*. Pada baris pertama Tabel 4.11, kolom "Rumi" milik model *Baseline* masih menampilkan aksara Arab, bukan huruf Latin. Ini bukan kesalahan penyalinan tabel, melainkan justru contoh kegagalan model dasar yang paling khas, karena model dasar belum pernah dilatih dengan pasangan Jawi-Rumi, model tidak memahami instruksi

"transliterasikan ke huruf Latin" dan hanya menyalin ulang (atau sedikit mengubah) teks Jawi yang sama pada bagian yang seharusnya berisi hasil transliterasi. Kegagalan mengikuti format keluaran inilah yang justru menjadi salah satu bukti kuantitatif rendahnya kemampuan model *baseline*, dan sejalan dengan nilai WER Transliterasi Rumi baseline yang tinggi pada Tabel 4.2. Setelah *fine-tuning*, sebagaimana terlihat pada kolom "Fine-Tuned" baris pertama, model sudah mampu menghasilkan baris "Rumi:" dalam huruf Latin yang sesuai.

Meskipun masih terdapat beberapa kesalahan kecil, seperti karakter tertentu yang belum dikenali secara sempurna maupun penggunaan tanda baca yang kurang tepat, hasil inferensi menunjukkan bahwa proses *fine-tuning* berhasil meningkatkan kemampuan model dalam melakukan ekstraksi dan transliterasi naskah Jawi cetak. Temuan ini juga konsisten dengan hasil evaluasi kuantitatif pada bagian sebelumnya yang menunjukkan penurunan nilai *Character Error Rate* (CER) dan *Word Error Rate* (WER), sehingga model *fine-tuned* terbukti memberikan performa yang lebih baik dibandingkan model *baseline*.

4.9 Demonstrasi Aplikasi Inferensi

Sebagai bagian akhir dari penelitian ini, model final (Skenario C, *epoch* 3) dibungkus ke dalam aplikasi inferensi sederhana berbasis Gradio. Aplikasi ini disebarluaskan melalui platform Hugging Face Spaces agar dapat diakses dan diuji oleh siapa saja tanpa perlu melakukan instalasi atau pengaturan teknis apa pun. Tampilan antarmuka aplikasi tersebut ditunjukkan pada Gambar 4.3.



Gambar 4.3 Tampilan Antarmuka Aplikasi Inferensi Gradio

Sebagaimana terlihat pada Gambar 4.3, pengguna cukup mengunggah gambar potongan naskah Jawi, dan aplikasi akan menampilkan dua kotak keluaran secara terpisah, satu untuk teks Jawi hasil ekstraksi dan satu untuk teks Rumi hasil transliterasi. Pemisahan dua kotak ini sengaja dibuat agar pengguna bisa dengan mudah memeriksa hasil dari masing-masing tugas tanpa harus mengurai format teks secara manual.

4.10 Pembahasan

Secara keseluruhan, hasil penelitian menunjukkan bahwa proses *fine-tuning* pada model multimodal menggunakan metode QLoRA mampu meningkatkan kemampuan model dalam melakukan ekstraksi dan transliterasi aksara Jawi. Peningkatan ini terlihat dari penurunan nilai *Character Error Rate* (CER) dan *Word Error Rate* (WER), serta hasil inferensi yang lebih akurat dibandingkan model *baseline*. Model tidak hanya mampu mengenali bentuk karakter Jawi, tetapi juga menghasilkan transliterasi yang lebih sesuai dengan konteks kata dan kalimat.

Penggunaan kombinasi data sintetik dan data autentik pada proses pelatihan memberikan hasil yang lebih baik dibandingkan penggunaan salah satu jenis data saja. Data sintetik membantu model mempelajari pola penulisan karakter Jawi secara konsisten, sedangkan data autentik memberikan variasi kondisi yang umum ditemukan pada naskah asli, seperti kualitas citra yang menurun, tekstur kertas, dan perbedaan bentuk penulisan. Meskipun nilai CER akhir untuk ekstraksi teks Jawi sebesar 0,174 belum mencapai target penelitian sebesar 0,15, hasil tersebut menunjukkan peningkatan yang signifikan dibandingkan model sebelum *fine-tuning* yang memiliki nilai CER sebesar 0,825. Penurunan tingkat kesalahan sekitar 79% menunjukkan bahwa proses pelatihan berhasil meningkatkan performa model secara nyata.

Apabila dibandingkan dengan penelitian terdahulu pada Tabel 2.1, seperti QARI-OCR yang mencapai CER 0,061 dan WER 0,160 pada teks Arab (Wasfy et al., 2025), maupun Qalam yang mencapai WER 0,80% pada tulisan tangan Arab dan 1,18% pada teks cetak Arab (Bhatia et al., 2024), nilai CER dan WER penelitian ini pada aksara Jawi memang masih lebih tinggi. Selisih ini tidak serta-merta menunjukkan bahwa pendekatan pada penelitian ini kurang tepat, melainkan lebih

mencerminkan perbedaan kondisi eksperimen yang mendasar antara kedua kelompok penelitian tersebut. Pertama, dari sisi bahasa dan aksara, penelitian terdahulu umumnya berfokus pada bahasa Arab standar (Modern Standard Arabic) yang memiliki data latih berlabel jauh lebih besar dan tersedia luas di internet, sedangkan aksara Jawi merupakan aksara dengan ketersediaan data digital yang jauh lebih terbatas, sebagaimana disinggung pula pada sub-bab 2.2.2. Kedua, aksara Jawi memiliki huruf tambahan khas Melayu di luar 28 huruf Arab standar serta pola ejaan yang tidak selalu konsisten karena minimnya penggunaan tanda vokal, sehingga tugas pengenalannya secara inheren lebih kompleks dibanding Arab standar. Ketiga, dari sisi karakteristik data, penelitian ini secara khusus menguji model pada data autentik, yaitu foto naskah cetak dengan variasi kualitas cetak, noda, dan tekstur kertas lama, yang secara umum lebih menantang dibanding data uji yang lebih bersih. Keempat, penelitian ini melakukan tugas ganda sekaligus dalam satu kali proses inferensi, yaitu ekstraksi aksara Arab Jawi sekaligus transliterasi ke huruf Latin, sehingga kesalahan pada tahap ekstraksi berpotensi merambat dan memperbesar kesalahan pada tahap transliterasi, berbeda dengan penelitian pembanding yang umumnya hanya mengukur satu tugas OCR saja. Kelima, dari sisi ukuran model dan sumber daya, penelitian ini menggunakan model 3B parameter dengan *fine-tuning* QLoRA pada perangkat keras terbatas, sedangkan sebagian penelitian pembanding memanfaatkan model maupun sumber daya komputasi yang lebih besar. Dengan mempertimbangkan kelima faktor tersebut, angka CER dan WER penelitian ini sebaiknya dimaknai dalam konteks kondisi eksperimennya sendiri, dan bukan sebagai perbandingan langsung yang menyatakan penelitian ini gagal hanya karena nilai numeriknya lebih tinggi.

Perbedaan antara hasil yang diperoleh dan target penelitian dipengaruhi oleh beberapa faktor. Sebagian besar data latih masih berupa potongan teks dengan panjang yang relatif pendek sehingga variasi pola kalimat yang dipelajari model masih terbatas. Selain itu, beberapa citra naskah autentik memiliki kualitas yang kurang baik akibat kondisi fisik dokumen, seperti noda, bagian yang memudar, atau karakter yang tidak tercetak dengan jelas. Model juga masih mengalami kesulitan ketika memproses teks yang mengandung banyak tanda baca atau harakat, karena variasi tersebut belum banyak terdapat pada data pelatihan.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa pendekatan *Vision-Language Model* yang dipadukan dengan *fine-tuning* menggunakan QLoRA mampu memberikan kinerja yang baik untuk ekstraksi dan transliterasi naskah Jawi cetak. Meskipun masih terdapat beberapa keterbatasan, hasil penelitian ini menunjukkan bahwa pendekatan yang digunakan memiliki potensi untuk mendukung proses digitalisasi naskah Jawi dan dapat dikembangkan lebih lanjut melalui penambahan variasi data maupun penyempurnaan proses pelatihan.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan seluruh tahapan penelitian yang telah dilakukan, mulai dari persiapan dataset, pencarian konfigurasi *hyperparameter*, pelatihan model, hingga pengujian dan studi ablasi, dapat ditarik beberapa kesimpulan berikut.

1. Penerapan model multimodal Qwen2.5-VL-3B-Instruct untuk tugas ekstraksi teks Jawi sekaligus transliterasi ke huruf Latin berhasil dilakukan secara menyeluruh, mulai dari tahap persiapan data sampai model siap digunakan melalui aplikasi inferensi. Keberhasilan ini didukung oleh penyusunan dataset yang menggabungkan data sintetik dan data autentik, serta penerapan pengacakan data pada beberapa tahap untuk mencegah model terlalu condong mempelajari satu jenis data secara berurutan. Penentuan batas panjang token sebesar 1024, yang diambil berdasarkan analisis distribusi panjang teks pada dataset, juga memastikan hampir seluruh data dapat diproses model tanpa terpotong.
2. Proses *fine-tuning* dengan metode QLoRA terbukti efektif meningkatkan kemampuan model dalam membaca aksara Jawi. Menggunakan konfigurasi rank LoRA sebesar 32 dan learning rate 1e-4 yang dipilih melalui eksperimen pendahuluan, model berhasil menurunkan CER Ekstraksi Jawi dari 0,825 pada kondisi awal sebelum dilatih menjadi 0,174 setelah dilatih, atau setara penurunan kesalahan sekitar 79%. Pola serupa juga terlihat pada metrik CER Transliterasi Rumi yang turun dari 0,906 menjadi 0,165. Hasil ini memang masih berada sedikit di atas target awal sebesar 0,15, namun penurunan kesalahan yang dicapai tetap menunjukkan bahwa proses penyesuaian model berjalan sesuai arah yang diharapkan. Model terbaik diperoleh pada *epoch* ke-3 dari total 5 *epoch* pelatihan, sebelum gejala *overfitting* mulai terlihat pada epoch berikutnya.
3. Studi ablasi yang membandingkan tiga skenario komposisi data latih (sintetik saja, autentik saja, dan gabungan keduanya), yang seluruhnya

dievaluasi pada citra data uji autentik yang sama, menunjukkan bahwa kombinasi data sintetik dan autentik (Skenario C) menghasilkan performa terbaik, disusul model yang hanya dilatih dengan data autentik saja (Skenario B), kemudian model yang hanya dilatih dengan data sintetik saja (Skenario A), dan yang terlemah adalah model baseline tanpa *fine-tuning*. Temuan ini juga memperlihatkan bahwa data autentik, meskipun jumlahnya jauh lebih sedikit dibanding data sintetik, memberi kontribusi yang lebih besar terhadap penurunan tingkat kesalahan per skenario tunggal, karena data ini lebih mencerminkan kondisi naskah Jawi yang sesungguhnya. Meskipun demikian, menggunakan data autentik saja (Skenario B) tetap belum sebaik menggabungkan kedua jenis data (Skenario C), sehingga kombinasi keduanya tetap menjadi strategi komposisi data yang paling disarankan.

4. Hasil analisis kesalahan menunjukkan bahwa model bekerja sangat baik pada gambar dengan kualitas baik, potongan teks yang rapi, dan panjang baris yang wajar, tetapi masih mengalami kesulitan pada beberapa kondisi tertentu, yaitu kata-kata pendek yang minim konteks, gambar dengan gangguan visual seperti noda atau tinta tembus, serta teks yang memuat tanda baca atau harakat yang padat. Model juga belum diuji secara memadai pada gambar berupa paragraf panjang karena mayoritas data latih berupa potongan satu sampai dua baris, sehingga aspek ini menjadi salah satu batasan yang perlu diperhatikan sebelum sistem digunakan pada naskah dengan format yang lebih kompleks.

5.2 Saran

Berdasarkan temuan dan keterbatasan yang ditemukan selama penelitian, berikut beberapa saran yang dapat dipertimbangkan untuk pengembangan lebih lanjut.

1. Penelitian selanjutnya disarankan untuk memperkaya dataset dengan menambahkan gambar naskah dalam format paragraf penuh atau halaman utuh, tidak hanya potongan baris pendek seperti pada penelitian ini. Penambahan variasi ini diharapkan dapat mengurangi kecenderungan

model mengulang format keluaran ketika dihadapkan pada teks yang lebih panjang.

2. Pengumpulan data tambahan yang secara khusus memuat teks Jawi dengan harakat (tanda vokal) yang lebih lengkap juga perlu dipertimbangkan, mengingat keterbatasan paparan terhadap jenis teks ini menjadi salah satu sumber kesalahan yang ditemukan pada tahap analisis kesalahan.
3. Pemeriksaan ulang terhadap kesesuaian antara label teks dan potongan gambar pada dataset autentik disarankan untuk dilakukan secara lebih cermat, khususnya untuk menyaring kemungkinan ketidaksesuaian antara gambar dan label yang dapat memengaruhi akurasi evaluasi.
4. Pada sisi teknis pelatihan, penelitian selanjutnya dapat mempertimbangkan eksperimen dengan jumlah *epoch* yang lebih bervariasi disertai strategi tambahan, mengingat gejala *overfitting* pada penelitian ini sudah mulai terlihat sejak *epoch* keempat. Eksperimen dengan rentang nilai rank LoRA atau *learning rate* yang lebih luas dari yang diuji pada penelitian ini juga berpotensi memberikan hasil yang lebih baik, mengingat keterbatasan sumber daya komputasi membuat eksperimen *hyperparameter* pada penelitian ini masih terbatas pada tiga konfigurasi saja.
5. Terakhir, mengingat sistem ini baru diuji pada data uji berskala terbatas, pengujian lanjutan pada kumpulan naskah Jawi yang lebih beragam, baik dari segi sumber, kondisi fisik, maupun gaya tulisan, akan membantu memastikan sejauh mana model ini dapat diandalkan ketika diterapkan pada kasus penggunaan yang lebih luas.

DAFTAR PUSTAKA

Abdiansah, L., Sumarno, S., Eviyanti, A., & Azizah, N. L. (2025). Penerapan Algoritma Convolutional Neural Networks untuk Pengenalan Tulisan Tangan Aksara Jawa. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 496–504. https://doi.org/10.57152/malcom.v5i2.1814

Adilazuarda, M. F., Wijanarko, M. I., Susanto, L., Nur'aini, K., Wijaya, D. T., & Aji, A. F. (2025). NusaAksara: A Multimodal and Multilingual Benchmark for Preserving Indonesian Indigenous Scripts. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 28371–28401. https://doi.org/10.18653/v1/2025.acl-long.1377

AR, K., Ahmadian, H., Nasihah, A. D., & Hamdan, A. M. (2025). An Applied Computer Mathematics Approach to Transliteration: YOLOv8-Based Detection of Harah Jawoe Script. *ZERO: Jurnal Sains, Matematika Dan Terapan*, *9*(1), 66. https://doi.org/10.30829/zero.v9i1.24174

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., … Lin, J. (2025). Qwen2.5-VL Technical Report. *ArXiv*.

Bania, A. S., & Akob, B. (2025). Preserving the Jawi script in Aceh: Assessing literacy, cultural heritage, and modern paradigm challenges. *Studies in English Language and Education*, *12*(1), 457–470. https://doi.org/10.24815/siele.v12i1.36629

Bhatia, G., Nagoudi, E. M. B., Alwajih, F., & Abdul-Mageed, M. (2024). Qalam: A Multimodal LLM for Arabic Optical Character and Handwriting Recognition. *Proceedings of The Second Arabic Natural Language Processing Conference*, 210–224. https://doi.org/10.18653/v1/2024.arabicnlp-1.19

Coluzzi, P. (2022). Jawi, an endangered orthography in the Malaysian linguistic landscape. *International Journal of Multilingualism*, *19*(4), 630–646. https://doi.org/10.1080/14790718.2020.1784178

Darmi, Y., Fajri Sepriansyah, M., Darnita, Y., & Pahrizal, P. (2025). Penerapan Metode Optical Character Recognition (OCR) Untuk Mengidentifikasi Teks Pada Identitas Dokumen Surat Izin Mengemudi (SIM). *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(4), 5992–5998. https://doi.org/10.36040/jati.v9i4.13987

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *ArXiv*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*.

Faizullah, S., Ayub, M. S., Hussain, S., & Khan, M. A. (2023). A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges. *Applied Sciences*, *13*(7), 4584. https://doi.org/10.3390/app13074584

Handoko, A. A., Rosid, M. A., & Indahyanti, U. (2024). Implementasi Convolutional Neural Network (CNN) Untuk Pengenalan Tulisan Tangan Aksara Bima. *SMATIKA JURNAL*, *14*(01), 96–110. https://doi.org/10.32664/smatika.v14i01.1196

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*.

Jumrah Jamil, & Suharto Pulukadang. (2025). Application of Deep Learning Method in Learning. *Formosa Journal of Sustainable Research*, *4*(6), 1011–1028. https://doi.org/10.55927/fjsr.v4i6.308

Nacson, M. S., Aberdam, A., Ganz, R., Avraham, E. Ben, Golts, A., Kittenplon, Y., Mazor, S., & Litman, R. (2024). DocVLM: Make Your VLM an Efficient Reader. *ArXiv*.

Nagasubramanian, A., Almazyad, A. S., Balakrishnan, S. A., MM, B. R., Potti, D., Zakariah, M., & AlSekait, D. M. (2025). OCRNet a robust deep learning framework for alphanumeric character recognition to assist the visually

impaired. *Scientific Reports*, *15*(1), 41344. https://doi.org/10.1038/s41598-025-25278-9

Neudecker, C., Baierer, K., Gerber, M., Christian, C., Apostolos, A., & Stefan, P. (2021). A survey of OCR evaluation tools and metrics. *ACM International Conference Proceeding Series*, 13–18. https://doi.org/10.1145/3476887.3476888

Razak, S. M. A., Seman, M. S. A., Ali, W., Wan, W. Y., Nizan, N. H., & Noor, M. (2019). Malay Manuscripts Transliteration Using Statistical Machine Translation (SMT). *2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 137–141. https://doi.org/10.1109/AiDAS47888.2019.8970867

Safrizal, S. (2019). Pengenalan Karakter Jawi Tulisan Tangan Menggunakan Fitur Sudut. *VOCATECH: Vocational Education and Technology Journal*, *1*. https://doi.org/10.38038/vocatech.v1i0.1

Syahputra Bania, A., Faridy, N., & Akob, B. (2025). Evaluating The Readability of Jawi To Latin Transliteration Via AI-Based Text Photography Applications. *INJECT (Interdisciplinary Journal of Communication)*, *10*(1), 115–132. https://doi.org/10.18326/inject.v10i1.4383

Wasfy, A., Nacar, O., Elkhateb, A., Reda, M., Elshehy, O., Ammar, A., & Boulila, W. (2025). QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation. *ArXiv*.