Indonesian Journal of Digital Business

Journal homepage: https://ejournal.upi.edu/index.php/IJDB

Integrasi OCR dan Model Bahasa Besar untuk Transliterasi dan Interpretasi Manuskrip Jawi Minangkabau

Farastika Allistio Putri[1], Yeka Hendriyani[2], Khairi Budayawan[3], Yulia Fatmi[4]

1,2,3,4Program Studi Informatika, Departemen Teknik Elektronika, Fakultas Teknik, Universitas Negeri Padang, Indonesia

Korespondensi: E-mail: farastika22@gmail.com

ABSTRAK

Manuskrip beraksara Jawi koleksi UPTD Museum Adityawarman sulit dimanfaatkan karena museum belum memiliki tenaga yang mampu membacanya. Penelitian ini membangun sistem *web mobile* (ReactJS, FastAPI, PostgreSQL; metode *Prototyping*) yang memadukan Tesseract-OCR, korpus 1.500 kosakata Jawi, dan model bahasa besar Qwen melalui Groq API untuk menghasilkan transliterasi Latin serta terjemahan dan interpretasi makna dari Bahasa Minangkabau klasik ke Bahasa Indonesia, dengan penyaringan berbasis ambang dan penanda keraguan pada keluaran. Seluruh 26 skenario *black box* sesuai harapan. Evaluasi OCR memakai 1.080 kata dari empat halaman naskah "Cerita Umbuik Mudo" (koleksi 07/58) yang juga menjadi salah satu sumber korpus, sehingga evaluasi belum independen dan hasilnya tidak dapat digeneralisasi. Diperoleh *word error rate* (WER) 31,34% dan *character error rate* (CER) 16,75%; akurasi kata berlabel 84,17% (ukuran pelengkap) memenuhi ambang 70% yang ditetapkan peneliti, sedangkan 1 − WER (68,66%) belum memenuhinya. Pencocokan korpus menurunkan WER 1,25 poin dibandingkan Tesseract mentah. Satu validator ahli yang juga menyusun transliterasi acuan menilai transliterasi (4,0) dan makna (3,50) berkategori Baik; kesalahan makna dikaitkan dengan transliterasi yang keliru, tetapi kontribusi model belum dipisahkan. *User acceptance testing* tiga petugas museum menghasilkan skor SUS 71,67 yang mengukur penerimaan, bukan ketepatan isi. Keluaran sistem hanya layak sebagai draf awal yang wajib diverifikasi ahli sebelum dipublikasikan.

INFO ARTIKEL

Riwayat Artikel:
Diterima 20 November 2026
Direvisi 25 November 2026
Disetujui 26 November 2026
Tersedia daring 26 November 2026
Tanggal publikasi 26 November 2026

Kata Kunci:
aksara Jawi,
Groq API,
manuskrip Minangkabau,
model bahasa besar,
OCR,
Tesseract.

1. PENDAHULUAN

Manuskrip adalah dokumen tulisan tangan beraksara tertentu yang berusia lebih dari 50 tahun (Nurrahman dkk., 2023). Manuskrip Minangkabau di Sumatera Barat ditulis dengan aksara Jawi (Arab-Melayu) dan memuat sejarah, hukum adat, keagamaan, jimat, politik, pengobatan, serta sastra (Taufiqurrahman dkk., 2021). Hingga Mei 2024, program DREAMSEA mencatat 1.235 manuskrip kuno yang ditemukan di Sumatera Barat (Pramono, 2024). UPTD Museum Adityawarman menyimpan sekitar 40–50 naskah beraksara Jawi. Berdasarkan wawancara dengan konservator naskah pada 21 Juli 2026, seluruh koleksi budaya hanya ditangani tiga staf dan museum belum memiliki tenaga internal yang mampu membaca aksara Jawi Minangkabau.

Pembacaan dan penerjemahan naskah bergantung pada penelitian eksternal yang bersifat insidental dan tidak diarsipkan utuh di museum. Digitalisasi yang ada sebatas restorasi fisik dan katalog *barcode* berisi ringkasan satu atau dua paragraf tanpa transliterasi atau terjemahan lengkap, sehingga petugas, masyarakat, dan akademisi sulit memahami isi naskah secara mandiri.

Optical Character Recognition (OCR) relevan untuk mengubah citra naskah menjadi teks. Tesseract-OCR banyak dipakai di industri digital Indonesia dengan akurasi cukup tinggi (Elmizan dkk., 2021), tetapi menurun pada tulisan tangan: Ramadhani dan Pangestu (2026) melaporkan akurasi sekitar 70% untuk huruf dan angka, sedangkan Umar dan Yunus (2025) melaporkan 68–91% pada manuskrip hadis Arab klasik, dengan mesin berbantuan *machine learning* lebih unggul pada citra berkualitas rendah. Untuk aksara Jawi, penelitian umumnya baru mencapai tingkat karakter. Fitriyah dkk. (2026) memisahkan karakter naskah "Syair Perahu" dengan akurasi 0,801 dan mencatat bahwa data satu naskah membatasi generalisasi. Rahmatika dkk. (2026) memadukan Tesseract-OCR, transliterasi berbasis kamus, dan penerjemahan mesin untuk aksara Arab Pegon, tetapi tidak melaporkan akurasi kuantitatif dan keluaran OCR-nya masih perlu dikoreksi. Pada naskah Jawi Minangkabau, kajian yang ada berupa alih aksara dan alih bahasa manual oleh filolog (Feblinda & Nurizzati, 2020). Perbandingan penelitian terdahulu dan celah yang ditangani diringkas pada Tabel 1.

Tabel 1. Perbandingan penelitian terdahulu dengan penelitian ini

Penelitian	Komponen	Data	Metrik dan hasil	Celah yang ditangani
Ramadhani & Pangestu (2026)	Konversi tulisan tangan menjadi teks dan suara	Huruf dan angka tulisan tangan	Akurasi ±70%	Tidak menyasar manuskrip beraksara Jawi
Umar & Yunus (2025)	OCR; Tesseract dan integrasi *machine learning*	Manuskrip hadis Arab klasik	Akurasi 68–91%; WER Tesseract 18,2%	Bukan aksara Jawi Minangkabau
Fitriyah dkk. (2026)	Segmentasi karakter	Naskah Jawi "Syair Perahu"	Akurasi 0,801	Tingkat karakter; satu naskah
Rahmatika dkk. (2026)	Tesseract-OCR, transliterasi berbasis kamus, penerjemahan mesin (Android)	Aksara Arab Pegon	Tanpa akurasi kuantitatif	Keluaran OCR masih perlu dikoreksi; belum terukur
Feblinda & Nurizzati (2020)	Alih aksara dan alih bahasa manual oleh filolog	Naskah Paparan Adat Limbago Alam Minangkabau	–	Manual; belum ada alur otomatis

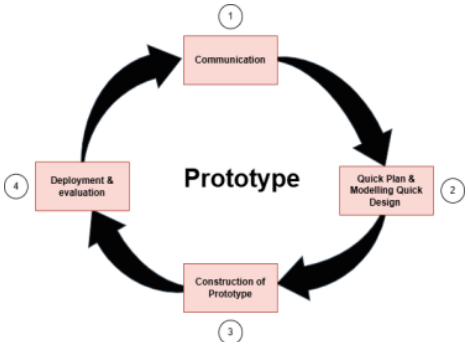
Penelitian	Komponen	Data	Metrik dan hasil	Celah yang ditangani
Penelitian ini	Tesseract-OCR, pemilihan PSM, korpus 1.500 kosakata, ambang *confidence*, Qwen (Groq API)	Empat halaman (1.080 kata) naskah Jawi Minangkabau koleksi Museum Adityawarman	WER, CER, akurasi berlabel, penilaian ahli, SUS	Mengukur alur terpadu terhadap GT ahli beserta analisis kesalahan; evaluasi belum independen

Penelusuran literatur tidak menemukan sistem yang memadukan OCR, pencocokan korpus kosakata Jawi, dan interpretasi makna oleh AI untuk naskah Jawi Minangkabau yang diukur terhadap transliterasi acuan ahli. Penelitian ini mengisi celah tersebut dengan sistem *web mobile* berisi alur terpadu Tesseract-OCR, korpus kosakata Jawi, dan model Qwen melalui Groq API, disertai pemilihan *page segmentation mode* (PSM) berdasarkan skor gabungan *confidence* dan jumlah karakter Arab serta penyaringan hasil OCR berbasis ambang sebelum interpretasi AI. Keluaran OCR hanya berupa transliterasi Latin yang sulit dipahami awam karena bahasa Minangkabau klasik sarat makna tersirat, sehingga interpretasi makna ke Bahasa Indonesia diserahkan pada model bahasa besar (*large language model*, LLM) Qwen melalui layanan Groq berbasis *Application Programming Interface* (API). Kebaruan penelitian ini terutama berupa rancangan alur dan evaluasi; bukti terukur yang disajikan terbatas pada perbandingan Tesseract mentah dengan pencocokan korpus serta analisis kesalahan yang lolos tanpa penanda. Tujuan penelitian ini adalah membangun sistem tersebut, mengevaluasi keluarannya terhadap transliterasi acuan seorang ahli, dan menilai penerimaannya oleh petugas museum. Kinerja sistem dievaluasi pada empat halaman naskah nyata melalui akurasi OCR, validasi ahli, dan uji penerimaan petugas museum. Sistem juga dirancang untuk mempermudah pengujian transliterasi dan meningkatkan aksesibilitas informasi warisan budaya Minangkabau.

2. METODE

2.1. Model Pengembangan Sistem

Sistem dikembangkan dengan metode *Prototyping*, yaitu pengembangan iteratif yang mengevaluasi prototipe bersama pengguna (Mranani dkk., 2025). Metode ini dipilih karena OCR dan integrasi API AI memerlukan pengujian berulang agar akurasinya memadai. Tahapannya ialah *Communication* (kebutuhan dikumpulkan melalui observasi dan wawancara dengan konservator), *Quick Design & Modeling* (*use case*, *activity diagram*, *sequence diagram*, dan *entity relationship diagram*), *Construction of Prototype* (kode program, integrasi OCR-AI, dan pra-pemrosesan citra), serta *Deployment & Evaluation* (uji akurasi OCR dan umpan balik pengguna) (Rohmadi & Yasin, 2020), seperti ditunjukkan pada Gambar 1.



Gambar 1. Tahapan *Prototyping*

2.2. Arsitektur Sistem Tiga Lapis

Arsitektur *client-server* terdiri atas tiga lapis: *Presentation Layer* (React JS, dengan akses kamera melalui Media Capture API), *Communication Layer* (REST API FastAPI berformat JSON), dan *Intelligence Layer* (Python) yang menjalankan pra-pemrosesan OpenCV, Tesseract-OCR, transliterasi dan pencocokan korpus, serta interpretasi makna oleh model Qwen melalui Groq API sebelum hasil disimpan ke basis data.

2.3. Alur Pengolahan Citra dan Data

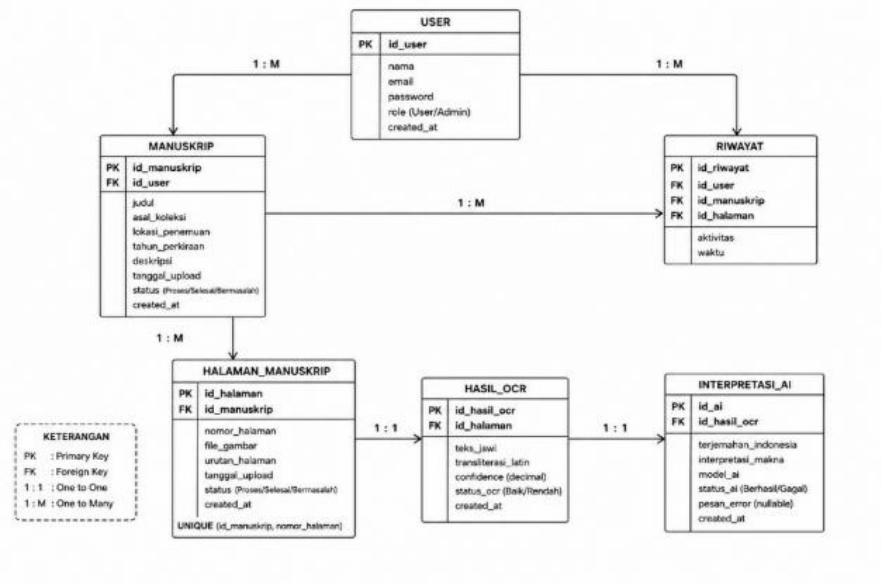
Digitalisasi manuskrip termasuk ranah *computer vision* (Sutisna dkk., 2024; Asvin dkk., 2023) yang memanfaatkan pengolahan citra untuk memperbaiki kualitas citra sebelum dikenali mesin (Jumadi & Sartika, 2021). Alur terdiri atas lima tahap. (1) Pra-pemrosesan: *grayscale*, *median filter* untuk mengurangi derau, binarisasi *adaptive Gaussian thresholding*, *deskewing*, dan normalisasi ukuran karakter (Ramadhani & Pangestu, 2026; Garai & Paul, 2022). (2) Pengenalan teks memakai Tesseract 5.5.0 dengan OEM 1 (LSTM) dan bahasa *Arabic* (ara) tanpa pelatihan lanjut. Setiap variasi citra dikenali dengan PSM 6, 11, dan 12 melalui fungsi image_to_data() pada Pytesseract, lalu konfigurasi dengan skor gabungan tertinggi dipilih; skor gabungan dihitung sebagai $\bar{c} + 0{,}5 \times \min(N, 30)$, dengan $\bar{c}$ rata-rata *confidence* kata Tesseract (0–100) dan N jumlah karakter beraksara Arab pada hasilnya. Teks beraksara Arab dipetakan ke Latin dan dikoreksi dengan kamus Jawi-Minangkabau; model Qwen tidak dipakai untuk transliterasi. (3) Setiap token Latin dicocokkan dengan korpus 1.500 kosakata Jawi pada basis data (sepuluh kandidat terdekat), dengan skor kecocokan berupa rasio kemiripan string (RapidFuzz fuzz.ratio, skala 0–100; 100 bila identik setelah dinormalisasi huruf kecil dan penghapusan tanda baca), margin sebagai selisih skor kandidat pertama dan kedua, dan skor minimal 65 karena ambang yang lebih tinggi membuat hasil kurang akurat akibat sebagian kata berskor rendah. Token diberi status “Diterapkan” (skor 100, kata diganti entri korpus), “Saran” (perlu ditinjau), atau “Tidak Cocok”.

(4) Penyaringan ambang: hasil diteruskan bila *confidence* OCR mencapai 70. Bila tidak, sistem menampilkan peringatan, menahan interpretasi AI, dan meminta pemindaian ulang. Nilai ambang minimum ini ditetapkan peneliti dengan merujuk pada batas bawah akurasi OCR manuskrip historis hadis (68–91%) oleh Umar dan Yunus (2025) dan akurasi tulisan tangan sekitar 70% oleh Ramadhani dan Pangestu (2026). Kedua acuan hanya melaporkan capaian sehingga ambang ini bukan standar baku, dan hasil yang lolos tetap diperiksa manusia. Pada pengujian akurasi, fungsi penyaring ini dijalankan oleh skor kecocokan korpus karena nilai *confidence* per kata tidak tercatat pada data uji. Ambang operasional ini berbeda dari kriteria kelayakan 70% pada pengujian. (5) Interpretasi makna: teks yang lolos dikirim ke model Qwen melalui Groq API dengan *prompt* yang meminta terjemahan alami ke Bahasa Indonesia, penjelasan makna tersirat dan nilai budaya bila didukung teks, serta mempertahankan bagian yang belum pasti tanpa menambah informasi. Parameter model adalah temperatur 0,1, batas keluaran 2.500 token, dan tiga kali percobaan ulang; teks *prompt* lengkap disajikan pada Lampiran 1 di subbab 3.4. Sistem dikembangkan dan diuji pada laptop HP 14s-dq5002TU (Intel Core i7-1260P, RAM 16 GB) dengan Python 3.11, OpenCV 5.0.0 (opencv-python 5.0.0.93), pytesseract 0.3.13, RapidFuzz 3.14.5, Groq SDK 1.7.0, dan model qwen/qwen3.8.27b. Halaman 29–32 yang menjadi data uji juga menjadi salah satu sumber korpus sehingga data uji tidak sepenuhnya terpisah, dan hal ini dicatat sebagai keterbatasan; korpus hanya dipakai pada tahap pencocokan.

2.4. Skema Basis Data

Basis data memakai PostgreSQL dengan enam entitas: users, manuskrip, halaman_manuskrip, hasil_ocr, interpretasi_ai, dan riwayat. PostgreSQL dipilih karena

menjaga integritas relasi melalui *primary key* dan *foreign key* (Indra dkk., 2024), penting karena satu manuskrip memiliki banyak halaman yang harus tetap terhubung dengan hasil OCR dan interpretasinya. Peran Admin dan pengguna dibedakan oleh atribut *role* pada users. Relasi *one-to-many* berlaku pada users–manuskrip dan manuskrip–halaman_manuskrip, sedangkan halaman_manuskrip–hasil_ocr dan hasil_ocr–interpretasi_ai berelasi *one-to-one* (Gambar 2).



Gambar 2. *Entity Relationship Diagram*

2.5. Rencana Pengujian

Pengujian terdiri atas *black box testing*, pengujian akurasi OCR terhadap teks acuan (*ground truth*, GT), validasi ahli aksara Jawi, dan UAT oleh petugas museum sebagai pengujian penerimaan sistem untuk pengguna. Pengujian akurasi memakai 1.080 kata pada empat halaman berurutan (halaman 29–32; 111 baris) naskah "Cerita Umbuik Mudo" (koleksi 07/58). Halaman yang sama juga menjadi salah satu sumber korpus pencocokan sehingga evaluasi belum independen dari korpus; hasilnya dibaca sebagai kinerja pada naskah yang sama, bukan pada naskah baru. Pemisahan halaman uji dari korpus tidak dilakukan karena memerlukan korpus yang lebih besar dan penyusunan GT oleh ahli yang memerlukan waktu lebih banyak dalam penelitian; hal ini diakui sebagai keterbatasan sistem untuk dikembangkan lebih lanjut di masa depan. GT disusun dan divalidasi oleh Kasman, Wali Jorong V Sungai Jariang, Kabupaten Agam, yang mendalami aksara Arab-Melayu sekitar 10 tahun, pada 18 Agustus–10 September 2026; setelah normalisasi GT berisi 1.120 kata dan 7.174 karakter. Hasil utama memakai kata akhir setelah pencocokan korpus, termasuk kandidat berstatus Saran yang belum diterapkan otomatis sehingga angkanya mengandaikan admin menyetujui kandidat tersebut; keluaran Tesseract mentah dihitung sebagai pembanding.

Ukuran utama kesalahan teks adalah *character error rate* (CER) dan *word error rate* (WER) yang dihitung dari jarak edit (*Levenshtein distance*) antara teks hasil sistem dan GT (Vidal dkk., 2023), sebagaimana pada Persamaan (1) dan (2):

$$CER = (S_c + D_c + I_c) / N_c \quad (1)$$

$$WER = (S_w + D_w + I_w) / N_w \quad (2)$$

Pada Persamaan (1) dan (2), S, D, dan I berturut-turut menyatakan jumlah substitusi, penghapusan, dan penyisipan, sedangkan N adalah jumlah karakter (c) atau kata (w) pada GT. Karena nomor baris tidak valid pada 54 dari 1.080 baris (5,00%), penyelarasan dilakukan per halaman dengan algoritma jarak *Levenshtein* berbobot sama untuk substitusi, penghapusan,

dan penyisipan, tanpa verifikasi manual, setelah normalisasi: huruf dikecilkan, tanda baca (termasuk 15 token pada keluaran sistem) dan penanda nomor dibuang, serta tanda hubung diganti spasi. Token *placeholder* "[tidak terbaca]" terpecah menjadi dua token dan bentuk "[tidak" dinormalisasi menjadi "tidak"; token ini tetap dihitung pada hasil utama dan dampaknya dilaporkan sebagai analisis sensitivitas pada subbab 3.3. Ukuran pelengkap adalah akurasi kata berlabel, yaitu proporsi kata yang oleh peneliti dilabeli "sesuai" terhadap seluruh kata yang dievaluasi (Tharwat, 2021), sebagaimana pada Persamaan (3):

$$A_{kata} = N_{sesuai} / N_{total} \times 100\% \quad (3)$$

Label diberikan per kata pada lembar kerja Excel hasil ekspor sistem dengan mengacu pada GT oleh satu penilai tanpa pemeriksaan penilai kedua. Kolom label hanya memuat "sesuai" (909 kata) dan sel kosong (171 kata) yang ditafsirkan "tidak sesuai"; penafsiran ini didukung GT karena 168 dari 170 kata bukan tanda baca bersel kosong (98,82%) memang tidak identik dengan kata GT. Kriteria tertulis label "sesuai" tidak terdokumentasi dan lebih longgar daripada kesamaan identik dengan GT (subbab 3.3); contoh pasangan keluaran dan GT disajikan pada Tabel 6. Interval kepercayaan 95% untuk akurasi kata dihitung dengan metode *Wilson score interval* pada tingkat kata dengan asumsi setiap kata merupakan pengamatan independen; karena kata-kata berasal dari empat halaman berurutan dalam satu naskah, interval ini dilaporkan sebagai estimasi deskriptif. Ambang kelayakan 70% (dasar pada subbab 2.3) dikenakan pada 1 − WER dan akurasi berlabel.

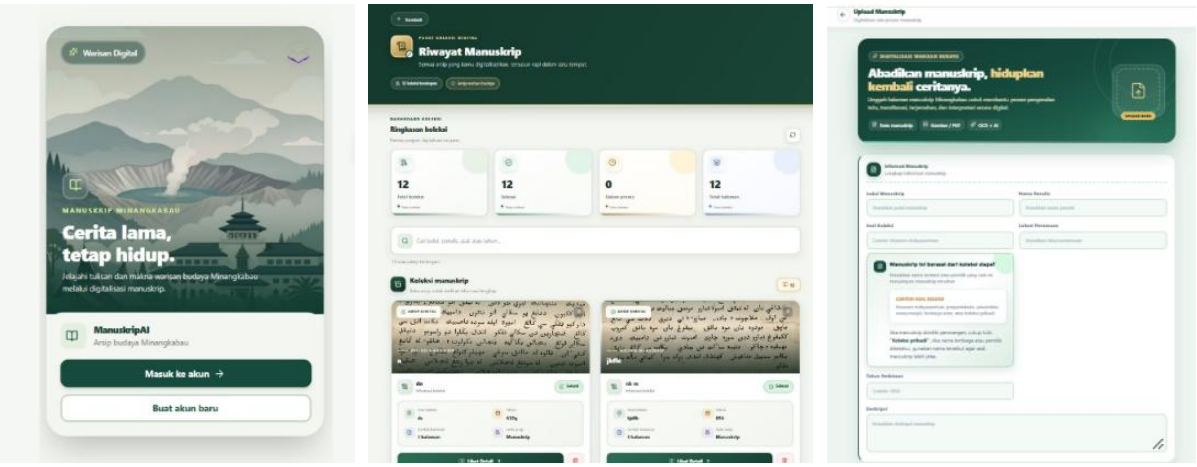
Validasi ahli dilakukan oleh satu validator (V1; praktisi dengan pengalaman 10 tahun membaca aksara Jawi dan 7 tahun naskah Minangkabau) pada 28 Oktober 2026 terhadap halaman uji yang sama melalui skor langsung skala 1–5 beserta kode kesalahan, yaitu skor T untuk transliterasi (1: tidak dapat dipakai; 2: banyak kata salah, makna sulit ditelusuri; 3: beberapa kata salah, makna masih dapat ditelusuri; 4: satu kata salah kecil yang tidak mengubah makna; 5: semua kata benar) dan skor M untuk makna. Penilaian dilengkapi kuesioner Likert (20 butir transliterasi dan 40 butir makna; butir negatif dibalik dengan skor akhir = 6 − skor mentah untuk menyeimbangkan butir positif, mengikuti Sauro dan Lewis, 2011) dan wawancara tujuh pertanyaan, dengan indikator mengacu pada kerangka evaluasi manusia terhadap model bahasa besar (Chang dkk., 2024; Awasthi dkk., 2025; Yang dkk., 2026). Kategori skor: 4,21–5,00 Sangat Baik, 3,41–4,20 Baik, 2,61–3,40 Cukup, 1,81–2,60 Kurang, dan 1,00–1,80 Sangat Kurang, dengan kriteria keputusan rerata minimal 3,41 sebagai usulan peneliti.

UAT mengukur penerimaan dan pengalaman pengguna oleh petugas, bukan ketepatan isi transliterasi atau interpretasi. UAT dilakukan pada 4 November 2026 oleh tiga petugas UPTD Museum Adityawarman (edukator, konservator, dan kurator) memakai *System Usability Scale* (SUS) 10 butir dan 36 butir penilaian enam aspek (fungsionalitas; kemudahan dan antarmuka; kinerja dan keandalan; kualitas keluaran dan transparansi; keamanan dan hak akses; kesesuaian, manfaat, dan penerimaan) yang selaras dengan karakteristik mutu ISO/IEC 25010 (ISO/IEC, 2023), berskala 1–5 dengan opsi N/A, ditambah tiga pertanyaan terbuka yang dianalisis secara tematik. Skor SUS dihitung dengan aturan baku (butir ganjil skor − 1, butir genap 5 − skor, jumlah dikali 2,5) dan dibandingkan dengan rata-rata acuan 68 (Sauro & Lewis, 2016; Hyzy dkk., 2022); karena acuan ini lintas domain dan bukan standar baku, ia dipakai sebagai batas minimal peneliti, dan skor dilaporkan per responden karena sampel kecil (Clark dkk., 2021).

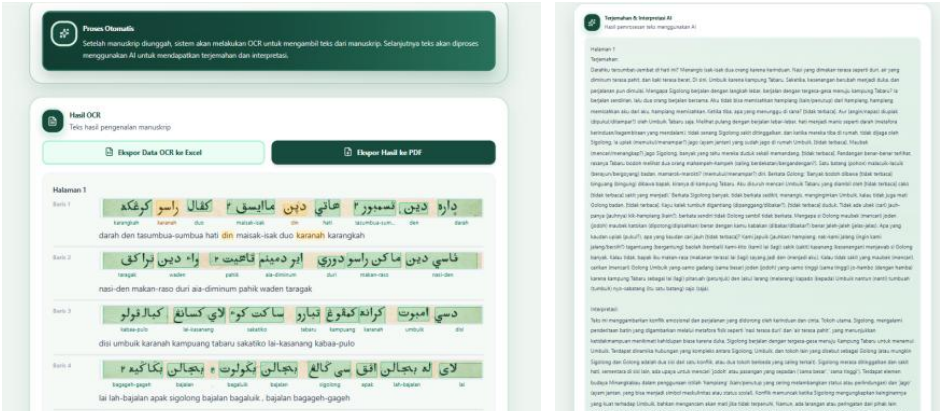
3. HASIL DAN PEMBAHASAN

3.1. Implementasi Sistem

Sistem berjalan sebagai aplikasi web responsif *mobile-first* berpola *single page application*. Pengguna memotret naskah, memilih dari galeri, atau mengunggah PDF, lalu membaca citra asli berdampingan dengan hasil OCR, transliterasi, dan interpretasi makna (Gambar 3 dan 4). Setiap citra diproses berurutan di *backend* (pra-pemrosesan, OCR, pencocokan korpus, penilaian ambang, interpretasi AI, dan penyimpanan dalam satu transaksi); bila Groq API tidak terjangkau, hasil OCR dan transliterasi tetap tersimpan dengan status “Interpretasi AI Gagal”. Admin memperoleh fitur kelola data, monitoring, dan pelatihan baris untuk koreksi label per baris, yang hanya menyimpan data berlabel dan tidak melatih ulang model. Kata sandi disimpan sebagai *hash* bcrypt, identitas dibawa JWT, dan hak akses diperiksa di *backend* melalui *role* pada token.



Gambar 3. Antarmuka pengguna: halaman awal, formulir informasi manuskrip, dan riwayat manuskrip



Gambar 4. Tampilan transliterasi hasil proses OCR dan terjemahan serta interpretasi AI

3.2. Hasil Black Box Testing

Seluruh 26 skenario pada 7 modul menghasilkan keluaran sesuai harapan (Tabel 2), mencakup jalur normal dan jalur gagal seperti kata sandi salah, nama pengguna ganda, formulir tidak lengkap, hasil OCR di bawah ambang, dan kegagalan layanan AI. Hasil OCR di bawah ambang menahan interpretasi AI disertai peringatan, kegagalan layanan AI tidak menghapus hasil OCR yang sudah tersimpan, dan fitur Admin tidak dapat dibuka akun biasa.

Pengujian ini hanya memastikan fungsi berjalan sesuai rancangan dan tidak menilai mutu keluaran OCR.

Tabel 2. Hasil *black box testing*

Modul	Skenario yang diuji	Hasil
Autentikasi dan profil	Login, registrasi, proteksi halaman, profil, logout	Seluruh fungsi berhasil
Dashboard dan riwayat	Pencarian, filter, detail, dan penghapusan manuskrip	Seluruh fungsi berhasil
Unggah manuskrip	Berkas valid dan formulir tidak lengkap	Valid diterima, tidak lengkap ditolak
Proses OCR dan interpretasi AI	Pengenalan teks, ambang skor, transliterasi, interpretasi, gangguan API	Sesuai kondisi uji
Administrasi	Dashboard admin, pengelolaan data pengguna dan manuskrip	Seluruh fungsi berhasil
Monitoring dan basis data	Statistik, pembatasan perubahan, pengelolaan basis data	Sesuai hak akses
Pelatihan baris	Pemotongan baris, koreksi label, penyimpanan data	Koreksi tersimpan

3.3. Akurasi OCR

WER 31,34% (229 substitusi, 27 penghapusan, 95 penyisipan dari 1.120 kata GT) dan CER 16,75% (1.202 operasi edit dari 7.174 karakter) menjadi indikator utama; penyisipan jauh lebih banyak daripada penghapusan, artinya sistem lebih sering menambah token daripada melewatkan kata. Halaman 31 terburuk pada kedua ukuran (WER 41,81%; CER 24,58%) dan memuat 45 dari 95 penyisipan. Akurasi berlabel sebagai ukuran pelengkap adalah 84,17% (909 dari 1.080 kata; interval kepercayaan 95%: 81,87–86,22%, estimasi deskriptif karena kata-kata tidak sepenuhnya independen), dan seluruh halaman berada di atas ambang 70% (Tabel 3).

Tabel 3. Akurasi kata, WER, dan CER per halaman (keluaran setelah pencocokan korpus)

Halaman	Kata uji	Sesuai	Akurasi kata (%)	WER (%)	CER (%)	Kata GT
29	248	205	82,66	26,64	15,44	274
30	267	227	85,02	26,06	11,15	284
31	300	243	81,00	41,81	24,58	299
32	265	234	88,30	30,04	16,03	263
Total	**1.080**	**909**	**84,17**	**31,34**	**16,75**	**1.120**

Akurasi antarhalaman berkisar 81,00–88,30% tanpa perbedaan nyata (χ^2, db = 3: 6,23; p = 0,10). Kontras halaman 31 dan 32 (p = 0,020, uji eksak Fisher tanpa koreksi perbandingan ganda) hanya eksploratif karena kondisi fisik halaman tidak dicatat. Akurasi berlabel (84,17%) lebih tinggi daripada 1 − WER (68,66%) karena penilaian berlabel lebih longgar daripada penyelarasan ketat WER: dari 895 kata bukan tanda baca yang dinilai sesuai, 142 (15,87%) tidak identik dengan GT pada penyelarasan. Terhadap ambang 70%, akurasi berlabel memenuhi, sedangkan 1 − WER kurang 1,34 poin; karena itu WER dan CER dijadikan indikator utama. Karena penyelarasan dilakukan per halaman penuh, urutan baca yang bergeser dapat membuat WER tampak lebih tinggi walaupun kata-katanya sudah dikenali (Vidal dkk., 2023); WER yang mengabaikan urutan kata (*bag-of-words* WER) yang diusulkan Vidal dkk. (2023) dapat dihitung pada penelitian lanjutan. CER 16,75% berada dalam kisaran 15–25% yang dilaporkan untuk naskah Arab kuno, dan WER 31,34% lebih tinggi daripada WER Tesseract 18,2% pada manuskrip hadis oleh Umar dan Yunus (2025); perbandingan ini hanya indikatif karena aksara, korpus, dan ukuran sampel berbeda.

Indikator bawaan sistem tidak cukup menyaring kesalahan (Tabel 4). Skor kecocokan 100 tidak menjamin ketepatan: 125 dari 961 kata berstatus Diterapkan (13,01%) tidak sesuai dengan GT, dan 42,72% kandidat Saran keliru. Namun 93 dari 171 ketidaksesuaian (54,39%) berasal dari kata tanpa penanda sehingga lolos tanpa peringatan.

Tabel 4. Ketidaksesuaian menurut status dan penanda sistem

Kelompok kata	Jumlah	Proporsi (%)	Tidak sesuai	% dalam kelompok
Status Diterapkan	961	88,98	125	13,01
Status Saran	103	9,54	44	42,72
Status Tidak Cocok	16	1,48	2	12,50
Ditandai sistem (“?”, Saran/Tidak Cocok, atau skor < 100)	183	16,94	78	42,62
Tidak ditandai	897	83,06	93	10,37

Kesalahan terpusat pada empat pola yang saling lepas (24 + 68 + 17 + 62 = 171 ketidaksesuaian): 24 (14,04%) dari token *placeholder* “[tidak terbaca]” yang terpecah dua token, 68 (39,77%) dari kata bertanda “?”, 17 (9,94%) dari kata “andak” tanpa padanan pada GT, dan 62 (36,26%) dari kata utuh lain yang ada di korpus sehingga lolos dari skor. Secara keseluruhan 18 dari 21 kemunculan “andak” tidak sesuai; satu kemunculan bertanda “?” tercatat pada kategori kata bertanda “?”. Menurut penyelarasan, 171 ketidaksesuaian terdiri atas 58 sisipan, 109 substitusi, dan 4 lainnya. Token *placeholder* merupakan artefak proses, bukan kesalahan pengenalan aksara, sedangkan “andak” (di GT hanya muncul sekali) diduga berasal dari bagian citra yang bukan kata, dugaan yang perlu dipastikan lewat citra potongan kata. Kata paling banyak dua karakter lebih rawan (ketidaksesuaian 35,85% dibanding 14,60% pada kata empat karakter atau lebih), dan hanya 13 substitusi (12%) berjarak paling banyak dua karakter dari kata GT, sehingga kekeliruan dominan adalah kata yang berbeda atau segmentasi yang bergeser, bukan salah ketik. Bila 26 token *placeholder* dikeluarkan, akurasi kata menjadi 86,05% sehingga kesimpulan terhadap ambang tidak bergantung pada perlakuan token tersebut.

Kontribusi korpus diukur dengan membandingkan Tesseract mentah dan Tesseract dengan pencocokan korpus pada empat halaman, normalisasi, dan penyelarasan yang sama (Tabel 5): WER 32,59% dan CER 17,76% turun menjadi 31,34% dan 16,75% setelah pencocokan (1,25 dan 1,00 poin persentase), karena 944 dari 961 kata berstatus Diterapkan sudah identik dengan entri korpus. Bila hanya kata berstatus Diterapkan yang dianggap berubah, WER 32,23%; sisa penurunan berasal dari kandidat Saran yang baru berlaku setelah disetujui admin.

Tabel 5. Perbandingan Tesseract mentah dan pencocokan korpus

Konfigurasi	WER (%)	CER (%)
Tesseract mentah	32,59	17,76
Tesseract + kata berstatus Diterapkan saja	32,23	–
Tesseract + pencocokan korpus (Diterapkan dan Saran)	31,34	16,75
Selisih terhadap Tesseract mentah (poin persentase)	1,25	1,00

Kontribusi ini kecil karena pada 944 dari 961 kata Diterapkan (98,23%) string OCR sudah identik dengan entri korpus; dari 17 kata yang berubah, 16 adalah normalisasi *placeholder* dan satu adalah “b?” menjadi “,b” (tidak sesuai), sehingga tidak ditemukan koreksi leksikal

otomatis atas kata bermakna. Kesesuaian terhadap korpus dan terhadap GT adalah dua ukuran berbeda: korpus hanya memastikan bahwa suatu kata ada dalam leksikon, sehingga 125 dari 961 kata Diterapkan tetap tidak sesuai dengan GT. Pada 103 kata berstatus Saran, kandidat berskor 100 dinilai sesuai pada 47 dari 49 kata (95,92%), sedangkan kandidat berskor di bawah 100 hanya pada 12 dari 54 kata (22,22%), yang mendukung rancangan menahan kandidat berskor rendah untuk diperiksa admin. Contoh pasangan keluaran OCR mentah, hasil akhir sistem, GT, dan label disajikan pada Tabel 6. Karena halaman uji ikut menyusun korpus (keterbatasan yang diakui pada subbab 2.5), pencocokan ini dapat tampak lebih menguntungkan daripada pada naskah baru. Kontribusi pemilihan PSM belum diuji terpisah dan nilai *confidence* per kata tidak tercatat, sehingga tidak diklaim; fitur pelatihan baris menyediakan jalur koreksi manusia dan dapat menjadi tempat menambah korpus jika diperlukan.

Tabel 6. Contoh pasangan keluaran OCR mentah, hasil akhir sistem, GT, dan label

Keluaran OCR mentah	Hasil akhir sistem	GT	Label
smbil?	sambil (kandidat Saran)	sambil	Sesuai
qad?	padi (kandidat Saran)	padi	Sesuai
nmpqq?	nampak (kandidat Saran)	nampak	Sesuai
[tidak (pecahan *placeholder*)	tidak (skor 100)	Umumnya tanpa padanan	Tidak sesuai
andak	andak (skor 100)	Tanpa padanan (18 dari 21 kemunculan)	Tidak sesuai

3.4. Interpretasi AI

Interpretasi makna oleh model Qwen melalui layanan Groq API merupakan tahap akhir alur. Gambar 4 memperlihatkan keluaran sistem untuk halaman 29 pada naskah Umbuik Mudo; panel Terjemahan dan Interpretasi AI menampilkan dua keluaran terpisah, yaitu terjemahan ke Bahasa Indonesia dan interpretasi makna. Pada contoh tersebut, bagian yang tidak dapat dipastikan (“tidak terbaca”) dan pilihan makna yang belum pasti ditulis dalam tanda kurung disertai tanda tanya, sesuai arahan *prompt* agar model tidak menambah informasi yang tidak didukung teks. Keluaran memuat penjelasan istilah seperti “hampiang” dan “jago” serta norma perjodohan dan status sosial. Teks *prompt* yang digunakan disajikan pada Lampiran 1.

Kinerja model bahasa besar pada bahasa daerah masih terbatas: model sangat besar seperti GPT-4 tertinggal dari sistem terjemahan mesin terbaik pada bahasa berdaya sumber rendah, dan model yang disesuaikan khusus untuk bahasa Bali dan Minangkabau dapat memperbaiki mutu terjemahan (Tan & Zhu, 2024), sementara sistem terjemahan mesin saraf untuk bahasa daerah Indonesia dapat bersaing dengan terjemahan *zero-shot* GPT-3.5-turbo (Susanto dkk., 2023). Selain itu, keluaran model bahasa yang lancar tetapi tidak dapat diverifikasi dari sumber tergolong halusinasi ekstrinsik (Ji dkk., 2023). Karena itu keluaran dibaca sebagai bantuan awal bagi petugas, bukan terjemahan yang sudah terverifikasi, dan ketepatannya dinilai oleh validator ahli pada subbab 3.5.

Lampiran 1. Teks *prompt* terjemahan dan interpretasi makna AI

Anda bertugas menerjemahkan dan menginterpretasikan manuskrip Minangkabau klasik/kuno ke dalam Bahasa Indonesia. Pahami bahasa sumber sebagai bahasa Minangkabau lama yang memiliki kemungkinan kosakata, struktur kalimat, ungkapan, dan kiasan berbeda dari bahasa Minangkabau modern. Hasil terjemahan harus lengkap, akurat, mudah dipahami, sesuai konteks cerita, tidak menghilangkan kosakata yang dapat diterjemahkan, tidak mengubah nama tokoh atau tempat, serta tidak mengarang informasi di luar sumber. Gunakan referensi kamus sebagai sumber pendukung, bukan satu-satunya dasar penentuan makna. Jika suatu kata tidak ditemukan dalam kamus, analisis

berdasarkan bentuk kata, imbuhan, fungsi dalam kalimat, kata sebelum dan sesudahnya, serta penggunaannya pada bagian lain. Perhatikan kosakata Minangkabau lama, Melayu klasik, variasi dialek, kata berimbuhan, kata ulang, ungkapan tradisional, peribahasa, pantun, dan bahasa kiasan. Identifikasi nama orang, tokoh, tempat, gelar, panggilan, dan nama diri berdasarkan konteks, serta jangan menerjemahkannya secara harfiah.

Analisis makna harfiah dan kiasan berdasarkan konteks kalimat dan keseluruhan cerita. Terjemahkan seluruh kata yang dapat dipahami, pertahankan maksud asli, urutan peristiwa, dan hubungan antartokoh, serta gunakan Bahasa Indonesia yang alami dalam bentuk paragraf. Jika makna kata belum dapat dipastikan, pertahankan kata aslinya dengan keterangan [makna belum dapat dipastikan: kata asli]. Gunakan [tidak terbaca] hanya apabila bentuk kata memang tidak dapat dikenali. Interpretasikan isi utama cerita, tokoh, konflik, maksud ucapan dan tindakan, makna kiasan, serta nilai sosial dan budaya Minangkabau apabila didukung teks. Bedakan informasi yang tertulis langsung dengan hasil penafsiran dan jangan mengarang identitas tokoh, motif tindakan, lokasi, latar sejarah, atau pesan moral. Tampilkan kembali transliterasi Latin sesuai sumber tanpa mengubahnya menjadi Bahasa Indonesia. Tentukan tingkat keyakinan Tinggi, Sedang, atau Rendah berdasarkan kejelasan teks dan kekuatan interpretasi, bukan kelancaran hasil terjemahan. Sebelum memberikan hasil, periksa kelengkapan terjemahan, ketepatan nama diri, kesesuaian makna, alur cerita, dan dukungan teks terhadap interpretasi. Keluarkan JSON valid saja tanpa Markdown, code fence, kalimat pembuka, atau penutup, dengan tepat empat properti berupa string: transliterasi_latin, terjemahan_indonesia, interpretasi_makna, dan tingkat_keyakinan. Terjemahan dan interpretasi harus berbentuk paragraf yang runtut dan tidak boleh menambahkan properti lain.

3.5. Validasi Ahli

V1 menyusun GT dan memeriksa kesesuaiannya dengan citra naskah, dan koreksi penyalinan diterapkan sebelum seluruh angka pada subbab 3.3 dihitung; karena V1 juga menjadi validator, penilaiannya tidak sepenuhnya independen. Rerata skor T transliterasi adalah 4,0 (median 4; simpangan baku 0,94; Baik) dengan 8 dari 10 baris (80%) berskor ≥ 4, dan skor M makna 3,50 (Baik), hanya tipis di atas ambang 3,41 (Tabel 7). Dari 91 kata sampel, 18 (19,8%) salah, sekitar 12 poin lebih rendah daripada WER 31,34%; selisih ini dapat berasal dari sampel kecil dan perbedaan cara menghitung. Kode kesalahan dominan ialah T1 (salah huruf atau ejaan) dan T3 (kata hilang atau tambahan), terutama pada kata berbentuk mirip (“dapek”, “diupiak”) dan bagian kosong yang terbaca sebagai kata. Karena empat baris berisi dua kata salah tetap diberi skor 4, angka 80% dibaca sebagai batas atas (rerata menjadi 3,6 bila baris itu diberi 3).

Tabel 7. Ringkasan penilaian validator ahli terhadap transliterasi OCR dan makna AI

Ukuran	Transliterasi OCR	Makna AI
Skor langsung (skala 1–5)	4,0 (Baik)	3,50 (Baik)
Baris atau halaman berskor ≥ 4	8 dari 10 baris	2 dari 4 halaman
Kuesioner, aspek kriteria	4,00 (Baik)	3,68 (Baik)
Kuesioner, seluruh butir	4,25 (Sangat Baik)	3,58 (Baik)
Aspek tertinggi	Tampilan dan fitur koreksi (5,00)	Kejelasan bahasa dan keterbukaan (4,00)
Aspek terendah	Ketepatan karakter (3,67)	Ketahanan terhadap salah transliterasi (3,17)

Penanda sistem belum menangkap seluruh kesalahan. Dari tujuh baris berkesalahan, hanya dua yang seluruh kata salahnya ditandai, sehingga 6–8 dari 18 kata salah (33–44%) tidak bertanda, searah dengan Tabel 4 (54,39%), sedangkan tidak ada kata benar yang diberi “?”. Pada kuesioner transliterasi, validator menilai ketepatan karakter paling rendah karena huruf Jawi yang mirip sering salah dikenali, dan kelayakan 3,75 karena belum yakin memeriksa hasil sistem lebih cepat daripada menransliterasi sendiri. Validator sangat setuju bahwa penanda membantu (Q8 = 5) dan tidak setuju bahwa banyak kesalahan luput dari penanda (Q9), padahal 33–44% kata salah tidak bertanda, sehingga skor kuesioner tidak boleh dibaca sebagai bukti kesalahan senyap rendah.

Pada interpretasi makna, halaman 31 dan 32 berskor 4 sedangkan halaman 29 dan 30 berskor 3. Kode M4 (salah makna akibat transliterasi keliru) muncul pada tiga halaman dan M2 (menambah makna yang tidak ada) pada halaman 31. Sistem mengenali nama tokoh dan tempat, suasana emosional, pantun, dan tata cara adat, tetapi alur inti sering tidak tertangkap: kiasan kesepadanan pasangan hidup tidak teridentifikasi, “umbuik” ditafsirkan sebagai buah, pelaku pengacungan rancong tertukar, dan Si Umbuik Mudo ditafsirkan sebagai perempuan; “kbaa-pulo” bahkan diterjemahkan menjadi “kabur ke pulau”. Pada kuesioner makna, ketepatan makna (3,20), ketahanan terhadap kesalahan transliterasi (3,17), dan kelayakan sebagai bahan awal (3,38) berkategori Cukup. Butir Q30 (nada interpretasi terlalu yakin sehingga awam mengira pasti benar) berskor akhir 2 dan menjadi satu-satunya butir fokus di bawah 3,00. Validator bersedia memakai interpretasi sebagai bahan awal setelah diverifikasi (Q33 dan Q36 = 4), tetapi menilai verifikasi lebih lama daripada menafsirkan sendiri (Q34, skor akhir 2). Wawancara memperkuat hal ini: pada halaman 32 sistem menyatakan teks sulit diinterpretasikan tetapi tetap menyusun narasi, sehingga pembaca non-ahli mudah terarah; validator menyarankan penyempurnaan *prompt* dan pembatasan interpretasi bila banyak kata tidak terbaca, serta mengeluhkan lamanya pemrosesan saat banyak halaman diunggah sekaligus.

Sistem bekerja berantai sehingga mutu interpretasi dibatasi mutu transliterasi: 18 dari 91 kata salah dan seluruh halaman interpretasi ditandai berakibat transliterasi, sehingga dugaan perambatan kesalahan (*error propagation*) mendapat dukungan penilaian ahli. Temuan ini sejalan dengan hasil van Strien dkk. (2020) bahwa kesalahan OCR berdampak konsisten pada tugas hilir dan membuat sebagian tugas rusak parah. Dukungan ini bersifat kualitatif karena kontribusi model belum dipisahkan dari kesalahan transliterasi. Kata yang keliru cenderung ditafsirkan sebagai kata bermakna lain tanpa tanda keraguan, dugaan yang perlu diuji, misalnya dengan menelusuri hubungan kata bertanda dan kesalahan makna. Bagi pembaca non-ahli, tafsiran yang tampak meyakinkan berisiko menimbulkan salah pengertian, sebagaimana risiko halusinasi ekstrinsik (Ji dkk., 2023), sehingga keluaran layak sebagai bahan awal yang diverifikasi ahli tetapi tidak mendukung penayangan langsung tanpa verifikasi dan label belum terverifikasi.

Keterbatasan evaluasi ini ialah sebagai berikut. Kedua pengujian dinilai satu validator yang juga menyusun GT sehingga tidak sepenuhnya independen dan kesepakatan antarpenilai tidak dapat dihitung. Sampel kecil dari satu naskah (10 baris, empat halaman) hanya mendukung penilaian ahli, bukan uji statistik. Interval kepercayaan akurasi kata mengasumsikan independensi antarkata, padahal kata-kata berasal dari empat halaman berurutan dalam satu naskah dan kesalahan OCR dapat mengelompok menurut kondisi halaman, sehingga interval kemungkinan lebih sempit daripada ketidakpastian sebenarnya dan hasil tidak dapat digeneralisasi ke seluruh manuskrip Minangkabau. Qwen belum diuji dengan GT yang telah diperiksa ahli, sehingga kesalahan akibat OCR dan keterbatasan model belum dapat dibedakan secara kuantitatif. Kuesioner transliterasi (20 butir) dan makna (40 butir) tidak setara sehingga tidak dibandingkan langsung, butirnya belum diuji reliabilitasnya, dan kategori serta kriteria keputusan merupakan usulan peneliti.

3.6. User Acceptance Testing oleh Petugas Museum

UAT mengukur penerimaan dan pengalaman pengguna, bukan ketepatan isi. Responden adalah tiga petugas UPTD Museum Adityawarman, yaitu edukator (R1), konservator (R2), dan kurator (R3) dengan masa kerja satu sampai lebih dari lima tahun; dua memakai laptop, satu memakai *smartphone*, dan seluruhnya menguji peran Pengguna dan Admin. Ketiganya tidak

dapat membaca aksara Jawi sehingga representatif bagi pengguna sasaran, tetapi tidak dapat memverifikasi ketepatan transliterasi dan penilaian hasil dibaca sebagai persepsi.

Skor SUS R1–R3 adalah 70,0; 75,0; dan 70,0 (rata-rata 71,67; simpangan baku 2,89), sedikit di atas acuan 68 dan tidak ada yang di bawah 51 (Tabel 8). Interval kepercayaan 95% (64,5–78,8) masih memuat 68, sehingga hasil hanya mendukung kebergunaan yang sebanding dengan rata-rata sistem lain, bukan lebih baik; angka ini berada pada kisaran aplikasi museum sejenis (SUS 72,10 pada versi VR dan 68,10 pada versi non-VR; Othman dkk., 2022). Subskala kebergunaan bernilai 77,08 sedangkan kemudahan dipelajari 50,00, karena dua responden perlu mempelajari banyak hal terlebih dahulu (butir 10) dan satu membutuhkan bantuan teknis (butir 4), sehingga pendampingan awal masih diperlukan.

Tabel 8. Hasil UAT petugas museum: SUS dan penilaian sistem per aspek (skala 1–5)

Ukuran	Nilai	Keterangan
SUS R1, R2, R3	70,0; 75,0; 70,0	Seluruhnya ≥ 68
SUS rata-rata (SD); IK 95%	71,67 (2,89); 64,5–78,8	Acuan 68; IK memuat 68
Subskala kebergunaan; kemudahan dipelajari	77,08; 50,00	Pendampingan awal diperlukan
A. Fungsionalitas (5 butir)	4,33	Sangat Baik
B. Kemudahan dan antarmuka (7)	4,43	Sangat Baik
C. Kinerja dan keandalan (5)	4,47	Sangat Baik
D. Kualitas keluaran dan transparansi (8)	4,55	Sangat Baik
E. Keamanan dan hak akses (4)	4,58	Sangat Baik
F. Kesesuaian, manfaat, penerimaan (7)	4,48	Sangat Baik
Seluruh butir (36; 106 jawaban valid)	4,47	98,1% bernilai 4–5

Keenam aspek penilaian sistem berkategori Sangat Baik, tanpa jawaban di bawah 3. Selisih aspek tertinggi (keamanan, 4,58) dan terendah (fungsionalitas, 4,33) hanya 0,25, sedangkan perbedaan antarresponden lebih besar (R1 4,92; R2 4,39; R3 4,09). Karena seluruh 36 butir bernada positif, skor dapat terdongkrak kecenderungan setuju sehingga SUS yang memuat butir negatif dijadikan patokan utama kebergunaan. Butir kesesuaian OCR, transliterasi, dan terjemahan yang menuntut kemampuan membaca Jawi dijawab N/A oleh dua dari tiga responden; hanya R2 yang memberi skor (4) dan keyakinan 3 (cukup yakin), serta menulis bahwa sebagian terjemahan belum sesuai. Jawaban terbuka ketiga responden menyebut kebutuhan verifikasi ahli.

UAT menunjukkan petugas dapat menerima sistem sebagai alat bantu pemahaman dan penyusunan draf. Namun nilai tinggi pada aspek kualitas keluaran (4,55) mencerminkan kejelasan tampilan, bukan ketepatan isi, karena responden tidak dapat membaca Jawi dan 54,39% kesalahan tidak ditandai (Tabel 4) sehingga tidak dapat ditemukan petugas sendiri. Ketepatan hasil karenanya bertumpu pada pengujian akurasi dan validasi ahli, dan permintaan verifikasi ahli di dalam sistem (status verifikasi, pencatatan pemverifikasi, penyimpanan teks terkoreksi) menjadi arah pengembangan utama. Keterbatasan UAT: hanya tiga responden pada satu hari sehingga hasil deskriptif dan interval SUS lebar; pengukuran berupa persepsi tanpa observasi tugas, waktu, atau galat; responden tidak dapat menilai ketepatan hasil; seluruh butir penilaian sistem bernada positif; dan acuan SUS 68 yang lintas domain serta ambang penerimaan merupakan usulan peneliti. Status penerimaan adalah diterima dengan catatan, bukan validasi mutu keluaran.

4. KESIMPULAN

Sistem digitalisasi manuskrip Minangkabau beraksara Jawi berbasis *web mobile* berhasil dibangun dan seluruh 26 skenario *black box* sesuai harapan. Pada empat halaman uji, akurasi kata 84,17% (WER 31,34%; CER 16,75%) memenuhi ambang 70% menurut akurasi kata tetapi belum menurut 1 − WER (68,66%), dan lebih dari separuh kesalahan lolos tanpa penanda. Pencocokan korpus menurunkan WER sebanyak 1,25 poin, dan kontribusi pemilihan PSM belum diuji terpisah. Validasi ahli menempatkan transliterasi (skor 4,0) dan makna (3,50) pada kategori Baik, sehingga keluaran layak sebagai bahan awal yang diverifikasi ahli, tetapi belum layak ditayangkan tanpa verifikasi karena kesalahan makna berasal dari transliterasi yang keliru dan nada interpretasi terlalu yakin. UAT tiga petugas museum menunjukkan sistem dapat diterima sebagai alat bantu (SUS 71,67; seluruh aspek Sangat Baik) tanpa menilai ketepatan hasil karena responden tidak dapat membaca Jawi, dan seluruhnya meminta verifikasi ahli di dalam sistem. Seluruh hasil berasal dari sampel kecil pada satu naskah dan tidak dapat digeneralisasi.

Prioritas pengembangan adalah mekanisme *human-in-the-loop*: status "belum diverifikasi" pada setiap keluaran, riwayat koreksi, dan persetujuan ahli sebelum teks atau interpretasi ditayangkan kepada masyarakat, disertai panduan singkat dan penempatan sistem pada server museum. Evaluasi berikutnya dapat diperluas pada halaman dan naskah yang tidak dipakai menyusun korpus, disertai pencatatan *confidence* tingkat kata, identifikasi token *placeholder*, pengujian terpisah pemilihan PSM, pengujian Qwen dengan GT yang telah diperiksa ahli untuk memisahkan kesalahan OCR dan kesalahan model, serta evaluasi sebelum dan sesudah koreksi pelatihan baris. Perbaikan hulu perlu diprioritaskan: memperkaya korpus terutama untuk huruf yang bentuknya mirip, meneruskan penanda ketidakpastian ke tahap interpretasi, dan menyempurnakan *prompt* serta glosarium istilah Minangkabau. Validasi dan UAT berikutnya sebaiknya melibatkan lebih banyak validator yang independen dari penyusun GT serta pengguna dari masyarakat umum, akademisi, dan peneliti, dengan observasi tugas, waktu, dan galat.

5. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Ibu Dr. Yeka Hendriyani, S.Kom. Selaku dosen pembimbing, serta Bapak Khairi Budayawan, S.Pd., M.Kom. dan Ibu Yulia Fatmi, S.Kom., M.Kom. selaku penelaah yang telaah memberikan arahan, masukan, dan saran dalam penyusunan penelitian ini. Penulis juga menyampaikan terimakasih kepada UPTD Museum Adityawarman Sumatera Barat atas akses terhadap koleksi naskah dan partisipasi petugas dalam *user acceptance testing*, serta kepada Bapak Kasman atas penyusunan dan validasi transliterasi acuan.

6. CATATAN PENULIS

Penulis menyatakan bahwa tidak terdapat konflik kepentingan terkait publikasi artikel ini. Penulis menegaskan bahwa naskah ini bebas dari plagiarisme.

7. DAFTAR PUSTAKA

Asvin, A., Suradi, M., & Alam, S. (2023). Sistem deteksi kantuk pengemudi mobil berdasarkan analisis rasio mata menggunakan computer vision. JUKI: Jurnal computer dan Informatika *5*, 222–230.

Awasthi, R., et al. (2025). Human evaluation of large language models in healthcare: Gaps, challenges, and the need for standardization. *npj Health Systems, 2*(1), 40. https://doi.org/10.1038/s44401-025-00043-2

Chang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology, 15*(3), 1–45. https://doi.org/10.1145/3641289

Clark, N., Dabkowski, M., Driscoll, P. J., Kennedy, D., Kloo, I., & Shi, H. (2021). Empirical decision rules for improving the uncertainty reporting of small sample System Usability Scale scores. *International Journal of Human–Computer Interaction, 37*(13), 1191–1206. https://doi.org/10.1080/10447318.2020.1870831

Elmizan, Sari, & Rahman. (2021). Optical Character Recognition (OCR) menggunakan Tesseract dan penerapannya pada industri digital di Indonesia. *Jurnal Teknologi Informasi dan Komunikasi, 8*(2), 125–138.

Feblinda, S., & Nurizzati. (2020). Alih aksara dan alih bahasa teks Paparan Adat Limbago Alam Minangkabau. *Bahasa dan Sastra, 8*(3). *Jurnal Bahasa dan sastra, 8(3),* 165-172.

Fitriyah, F., Andriana, D., Aliansyah, M. Z., Hakim, L., & Amrulloh, F. (2026). Identification of paleographic curvature using skeletonization and key point detection. *TELKOMNIKA, 24*(2), 620–634. https://doi.org/10.12928/TELKOMNIKA.v24i2.27502

Garai, S. K., & Paul, O. (2022). A novel method for image to text extraction using Tesseract-OCR **,** *American Journal of electronics & Communication*, 3(2), 8–11.

Hyzy, M., et al. (2022). System usability scale benchmarking for digital health apps: Meta-analysis. *JMIR mHealth and uHealth, 10*(8), e37290. https://doi.org/10.2196/37290

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Jumadi, J., & Sartika, D. (2021). Pengolahan citra digital untuk identifikasi objek menggunakan metode hierarchical agglomerative clustering. *Jurnal Sains dan Teknologi, 10*(2), 148–156.

Mranani, M., Tri, Y., & Agung, S. (2025). Penerapan metode prototyping dalam perancangan user interface (UI) website registrasi online pendaki. *Journal of Information Technology and Computer Science, 8*(5), 1–15.

Nurrahman, I. I., M. I., & Lawanda. (2023). Proses digitisasi naskah kuno berdasarkan NSPK Perpustakaan Nasional RI: Studi kasus Pusat Dokumentasi Sastra X. *Ilmu Informasi Perpustakaan dan Kearsipan, 12*(1), 83–93. https://doi.org/10.24036/jiipk.v11i2.124498

Othman, M. K., Nogoibaeva, A., Leong, L. S., & Barawi, M. H. (2022). Usability evaluation of a virtual reality smartphone app for a living museum. *Universal Access in the Information Society, 21*(4), 995–1012. https://doi.org/10.1007/s10209-021-00820-4

Rahmatika, L., Listyorini, T., & Supriyati, E. (2026). Pengembangan aplikasi translasi teks Arab Pegon ke Bahasa Indonesia menggunakan integrasi OCR dan API AI berbasis Android. *JATI (Jurnal Mahasiswa Teknik Informatika), 10*(2), 2131–2137.

Ramadhani, V., & Pangestu, D. A. (2026). Sistem konversi tulisan tangan huruf atau angka yang ditampilkan menjadi teks dan suara secara jelas. *Prosiding Seminar Nasional Teknologi dan Sains, 5*, 864–871.

Rohmadi, A., & Yasin, V. (2020). Desain dan penerapan website tata kelola percetakan pada CV Apicdesign Kreasindo Jakarta dengan metode prototyping. *Journal of Information System, Informatics and Computing, 4*(1), 70–85.

Sutisna, T., Raharja, A. R., Hariyadi, E., Hafizh, V., & Putra, C. (2024). Penggunaan computer vision untuk menghitung jumlah kendaraan dengan menggunakan metode SSD (Single Shoot Detector). *INNOVATIVE: Journal of Social Science Research*, *4*, 6060–6067.

Taufiqurrahman, Hidayat, A. T., Efrinaldi, Sudarman, & Lukmanulhakim. (2021). The existence of the manuscript in Minangkabau Indonesia and its field in Islamic studies *Journal of Al-Tamaddan*, *16*(1), 125–138.

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Umar, M., & Yunus, Z. A. (2025). Optimalisasi teknologi OCR dalam digitalisasi manuskrip hadis: Studi akurasi dan tantangan linguistik Arab klasik. *Hamidah: Jurnal Ilmu Hadis, 1*(1), 43–54.

Vidal, E., Toselli, A. H., Ríos-Vila, A., & Calvo-Zaragoza, J. (2023). End-to-end page-level assessment of handwritten text recognition. *Pattern Recognition, 142*, 109695. https://doi.org/10.1016/j.patcog.2023.109695

Yang, E., Ko, S., & Woo, H. (2026). The reliability of human evaluation of large language models in health care settings: Scoping review. *Journal of Medical Internet Research, 28*, e98184. https://doi.org/10.2196/98184