Beyond Character Error Rate: Evaluating Diacritic Fidelity and Post-Processing Effects in Classical Arabic OCR

Mohammed Fazalullah Qudrath
Independent Researcher
ORCID: 0009-0003-3684-3130

Preprint, October 26, 2026

Abstract

Classical Arabic books print diacritics selectively, and the marks can change meaning. Faithful OCR of these books must copy the marks the page shows, adding none and dropping none. Arabic OCR is still commonly scored by character error rate (CER) computed after diacritics are removed. We fine-tune the vision-language OCR model dots.mocr with LoRA on synthetic and human-verified classical pages and evaluate it on 124 human-verified test pages, 38 human-verified unvowelled scans and further unverified scans.

We report three findings. First, letters-only CER misses most of what changes. On the test split, mean CER is unchanged (0.051 vs. 0.051), while the diacritic error rate (DER) falls from 0.204 to 0.177 and invented marks per 1,000 letters fall from 36.9 to 21.6. Second, gains are large on typeset print (DER 0.053 → 0.008) and on several scanned sets, but absent on the only human-verified low-resolution scan book. Third, replacing recognised Qur'anic quotations with canonical text, a common post-processing step, changes the conclusions: on the test split the DER improvement becomes non-significant, and on the low-resolution book a null result becomes a significant regression. We also disclose a release gate revised after a failed evaluation, bootstrap-seed sensitivity and run-to-run variance.

1 Introduction

Arabic script writes consonants and long vowels as letters. Short vowels, gemination (*shadda*), the absence of a vowel (*sukun*) and case endings (*tanwin*) are written as optional marks above or below the letter. Modern prose usually omits them. Classical Islamic texts mark them, and editions differ in how much they mark: a fully vowelled *matn*, a commentary vowelled only at ambiguous words, and an unvowelled modern commentary that marks only Qur'anic quotations. The marks are not decoration. In Figure 5(a), كَفَّرَهُمْ ("he declared them unbelievers") and كَفَرَهُم ("he disbelieved in them") differ only by a shadda.

We call OCR *faithful* when it reproduces the letters and marks that the page prints: no marks added to unvowelled text, none dropped from vowelled text, and Qur'anic quotations copied as printed. This is the requirement for digital editions, for language-learning material built on a specific printing, and for any downstream use that must cite the page. It differs from producing correct or fully vowelled Arabic. A model can produce fluent, correctly vowelled text that is not on the page, and a diacritic-free score will rate it highly.

Evaluating faithfulness raises three practical problems, which this paper studies on a concrete system.

1. **Metric choice.** Arabic OCR results are often reported as CER after removing diacritics, or after other normalisations. That number cannot tell a model that copies marks from one that invents or drops them.
2. **Document conditions.** Gains measured on clean typeset pages may not hold on degraded scans, and human-verified pages of the latter are scarce.
3. **Post-processing.** Production pipelines often replace recognised Qur'anic quotations with canonical text from a reference edition. That changes what is being measured.

We fine-tune the open vision-language OCR model dots.mocr with low-rank adapters (LoRA) [17] on synthetic renders and human-verified classical pages. We evaluate it against references that transcribe the printed source, with and without a deterministic Qur'an quotation layer. All numbers in this paper were recomputed from stored predictions and references during a reproducibility audit of the project. We report the issues that audit found alongside the results.

Contributions.

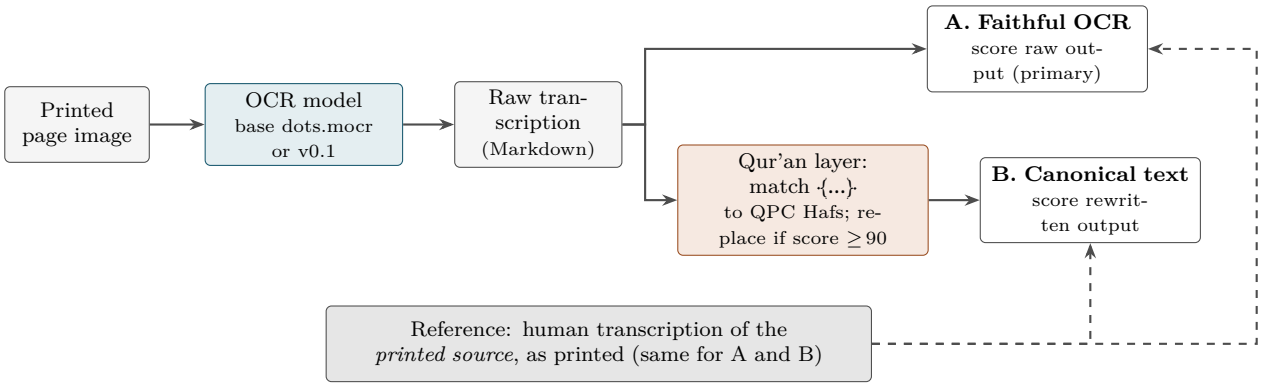


Figure 1: Two processing paths evaluated in this paper. Path A (primary) scores the model's raw transcription against a human transcription of the printed page. Path B first passes the output through a deterministic quotation layer that replaces recognised Qur'anic spans with canonical King Fahd Complex (QPC) Hafs text. The reference is identical in both paths; only the hypothesis changes. Path B therefore mixes recognition quality with the behaviour of the matcher and with differences between the printed edition and the canonical text.

- An evaluation protocol for faithful Arabic OCR. It defines letters-only CER, a diacritic error rate (DER) over aligned letters, and pooled rates of invented and omitted marks. It reports page-paired bootstrap intervals and shows where each metric is blind (Section 3).
- Evidence that letters-only CER is insufficient. On the same test pages, CER does not move while DER and invented-mark rates change substantially (Section 5).
- Results for a fine-tuned dots.mocr across document conditions, reported separately for test, selection and synthetic pages and for verified and unverified references. Large gains on typeset print do not appear on the one human-verified low-resolution scan book (Section 5).
- An analysis of canonical-verse post-processing. It shows that the layer narrows or reverses model comparisons and changes release-gate verdicts (Section 6).
- A transparent account of evaluation-process issues: a revised release gate, bootstrap-seed and run-to-run sensitivity, and reference errors. Also an error analysis on real page crops (Sections 7–8).

We make no claim that the fine-tuned model is better on scanned documents in general. We also make no causal claim about which synthetic-data changes produced the differences between our two training runs: the runs differ in several variables at once (Table 2).

2 Related Work

Arabic OCR. Printed Arabic OCR must handle cursive joining, context-dependent letter shapes, dots that separate letters, and optional diacritics; surveys document these challenges and the methods used against them [3]. Systems and benchmarks for printed Arabic include:

- Tesseract [36], the open engine;
- the APTI benchmark of synthetic printed text images [35];
- OCR for large historical Arabic-script corpora, as in the OpenITI/KITAB work with Kraken [20, 34].

Recent work applies vision-language models to Arabic: a broad benchmark covering text recognition and document understanding [15], and models fine-tuned for Arabic OCR and document parsing [16, 31, 40]. Several of these report diacritic-sensitive results, but standard reporting still centres on CER and WER under some normalisation. We do not compare against these systems. We study evaluation choices for one system on a narrower target: copying exactly the diacritics a classical edition prints.

Historical documents and OCR evaluation. Work on historical printings stresses ground-truth quality and transcription guidelines [32]. Character and word error rates derive from edit distance [24], with alignment procedures for OCR accuracy formalised early [33]. Evaluation tools and metrics, and how their normalisation choices affect results, have been surveyed [27] and implemented in open tools [7]. Human judgements of OCR quality have also been collected as a dataset [43]. Our contribution is specific to Arabic diacritics: we make the normalisation explicit, separate letters from marks, and show the effect of a canonical-text substitution step.

Arabic diacritics. Automatic diacritisation (restoring marks to undiacritised text) is long studied [1, 2, 8, 12, 13, 45]. It is usually evaluated with a diacritic error rate over letters, often with and without case

endings. Tashkeela [42], a corpus of mostly classical vowelled text, is a standard training source; we use it as text for synthetic pages. Diacritisation aims at linguistically correct, typically full vowelling. Faithful OCR aims at the vowelling of a particular printing. A diacritisation-style model is therefore a source of *invented* marks under our metric. Our DER borrows the diacritisation definition but restricts it to letters the OCR output reproduces.

Vision-language models for documents. OCR-free document models such as Donut [21], TrOCR [25] and Nougat [6] read text directly from images with transformer decoders. Later systems scale this to general document parsing:

- GOT-OCR 2.0 [41] and olmOCR [29];
- general vision-language models such as Qwen2-VL and Qwen2.5-VL [4, 39];
- the dots.ocr and dots.mocr models [9, 26, 44], the base of this work.

Because such models generate text with a language model, they can produce fluent text that is not on the page. Unrequested vowelling is the Arabic instance of this.

Synthetic data, adaptation and evaluation statistics. Synthetic rendered text is an established way to train text recognisers [14, 18], and synthetic-to-real gaps are a known risk. We fine-tune with LoRA [17], serve with vLLM [23], and use a CTRL-style repetition penalty [19] in a retry. For uncertainty we use the bootstrap [10, 11] in its paired form for system comparison [22]. Significance tests in NLP are known to be fragile and sensitive to sample and setup [5]; our seed and rerun analyses illustrate this for a release gate. chrF [28] is computed by the scorer but not used here. Qur'anic text comes from the King Fahd Complex (QPC) Hafs edition via the quran.com API [30], with Tanzil text [37] for pause marks. The honorific ligatures we discuss were added in Unicode 14.0 [38].

Positioning. CER, WER and diacritic error rates are established metrics; we do not propose them, and we do not present the quotation layer as a contribution. The contribution of this paper is empirical. On a single system, with human-verified references, we show three things. Letters-only CER and diacritic metrics lead to different conclusions about the same model. Gains on typeset print do not transfer to the one human-verified low-resolution scan book, while they do appear on verified unvowelled scans. And a deterministic canonical-verse post-processing step changes both measured error and model-comparison conclusions. We also document how evaluation-process choices, such as gate revision, bootstrap seed and repeated inference, affect a release decision.

3 Task Definition and Evaluation Metrics

Task. Given a page image, the system outputs the page body in reading order: headings, text and footnotes, but not running headers or page numbers. Light Markdown is allowed for headings and emphasis. The reference is a person's transcription of the *printed source*. It keeps every printed diacritic and adds none, and Qur'anic quotations are transcribed as printed, in King Fahd Complex (QPC) Uthmani encoding where the book prints Uthmani text. Where a scan is illegible but a printed copy was consulted, the reference follows the printed copy and the page is flagged; these flags are counted in Section 4.

Text views. All metrics compare two normalised views of the hypothesis h and reference r. They are computed by the project's scorer (a Go port of an earlier Python harness). The *plain* view strips Markdown and HTML markup, collapses runs of three or more leader characters to an ellipsis, applies Unicode NFC (not NFKC), and collapses whitespace. The *letters* view is the plain view with three changes:

- it removes the eight harakat (U+064B–0652) and the superscript alif (U+0670), the Qur'anic annotation marks (U+06D6–06ED, U+08D3–08FF) and tatweel;
- it maps Arabic-Indic and Persian digits to ASCII;
- it turns every other non-word character into a space.

No letter folding is applied: ا أ إ آ, ى/ي and ه/ة remain distinct. Appendix A lists further edge cases. One of them affects this paper: the honorific ligature ﷺ is a letter in this view, but ﷺ (U+FD41) is not and disappears.

Character error rate. For a page, $\mathrm{CER}(h, r) = \mathrm{Lev}(h, r)/\max(1, |r|)$, with Levenshtein distance [24] over code points and unit costs. We report *letters-only CER*, computed on the letters view, because it is the conventional diacritic-free score and the one this project used in its summaries. CER is unbounded above 1, because a looping output can be many times longer than the reference.

Diacritic error rate. Both plain views are split into clusters: each base character carries the harakat that follow it, sorted so that their order does not matter. The two base sequences are aligned with an optimal Levenshtein alignment, and only positions where the bases are equal and the reference base is an Arabic letter are kept. Let A be this set of aligned letters, and $M \subseteq A$ the letters marked on at least one side. Then

$$\mathrm{DER} = \frac{|\{a \in M : \mathrm{marks}_h(a) \neq \mathrm{marks}_r(a)\}|}{|M|}, \tag{1}$$

which is undefined when $M = \emptyset$. Each $a \in M$ with differing marks is classified as *invented* (reference bare, hypothesis marked), *omitted* (reference marked, hypothesis bare) or *changed* (both marked, differently). This DER follows the usual diacritisation definition, the proportion of letters with wrong diacritics [13, 45]. It differs in two ways. The denominator holds only letters marked on at least one side. And letters that the alignment does not match (substituted, inserted or deleted letters) are excluded from both the numerator and the denominator, so DER measures marks conditional on correct letters.

Consequences of these definitions. Three of them matter for interpreting results:

- On a reference with no marks, page DER is 1.0 if the system invents any mark, and undefined otherwise. It carries almost no information there, so we also report *invented* and *omitted marks per 1,000 aligned letters*, pooled over a set.
- An empty or catastrophically wrong answer has few or no aligned letters, so its DER is undefined or computed on very few letters. Such pages leave the DER mean and appear only in CER and in the broken-page count.
- CER and DER are complementary. A model can lower DER while raising CER (by dropping text) and the reverse.

Page-level failure counts. A page is *broken* if it loops (a 30-character window of the letters view occurs more than four times), is truncated at the 4,096-token generation limit, is nearly empty (under 20 % of the reference length), or has letters-only CER above 0.5. An *omitted run* is a maximal sequence of at least five consecutive reference words with no counterpart in a word-level alignment.

Aggregation and uncertainty. Set-level CER and DER are unweighted means of page values (DER over pages where it is defined); invented and omitted marks are pooled rates. Model comparisons use a paired bootstrap over pages [10, 22]: 4,000 resamples, seed 1, percentile 95% interval of the mean per-page difference, over pages where the metric is defined for both systems. We apply no multiple-comparison correction and treat the intervals as descriptive. Because pages are resampled within the evaluated sets, an interval describes variation across these particular pages. It is *not* an estimate of uncertainty about other books, editions, typefaces or publishers. With one typeset edition series and one verified low-resolution book, between-book variation is unmeasured.

Two scoring conditions. In the *raw* condition (layer off), the model's output is scored directly. This is our primary evaluation. In the *post-processed* condition (layer on), the output first passes through a deterministic Qur'an layer (Figure 1). Every span between the ornate brackets ﴿...﴾ with at least eight letters is matched against the Qur'an. Matching uses a normalised text that folds hamza and alef variants and strips marks. If a window of one to four verses scores at least 90 and the word counts agree to within two, the span is replaced by the canonical QPC Hafs wording of the quoted part. Inline verse numbers are also rewritten as ۝n. The reference is the same in both conditions. Only the hypothesis changes.

4 Dataset and Experimental Setup

Human-verified gold set. The gold set has 350 pages from seven books (Table 1, Figure 2):

- six classical *mutoon* (core texts) in a modern typeset edition, rendered from vector PDFs at 2,600 px on the long side;
- one scanned modern commentary, *Sharh Ma'na at-Taghut*.

Each draft transcription came from an LLM-based OCR pipeline and was corrected by a person:

- the *mutoon* against the printed text and, the author reports, against recitation audio;
- the commentary against the scan and photographs of a printed copy.

Splits are by book. The test split contains Kashf ash-Shubuhat (80 typeset pages) and the commentary (44 low-resolution scan pages). On 24 commentary pages the reference follows the printed copy where the scan is illegible or the edition differs. Every gold page with partial or no vowelling belongs to this one scanned book, so the training and validation material is entirely fully vowelled.

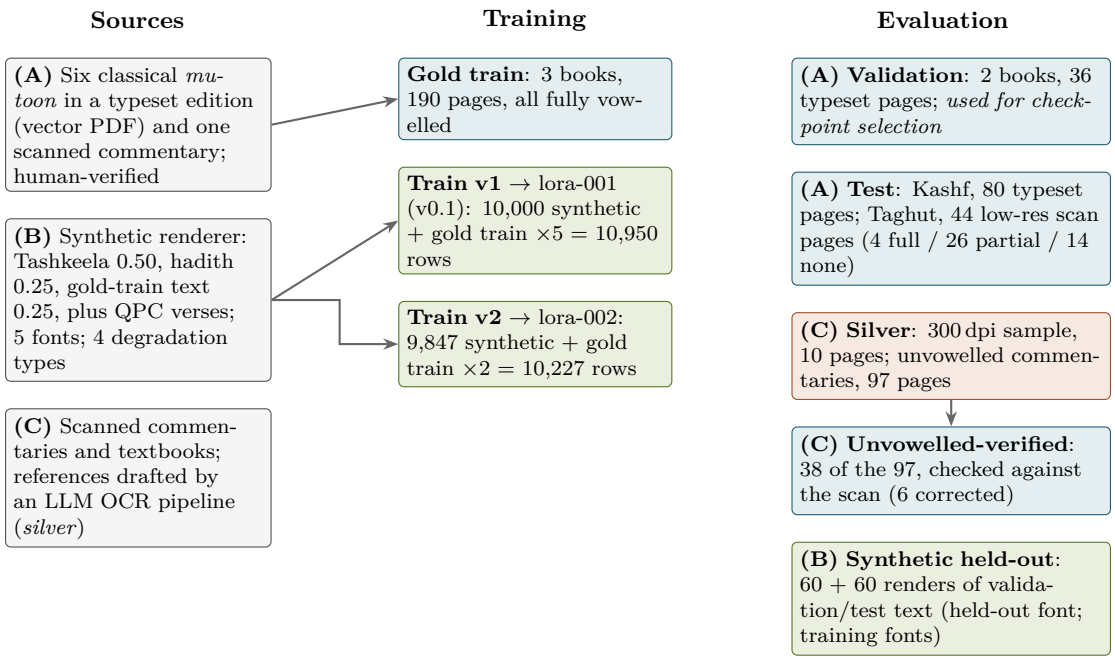


Figure 2: Data used in this study. Letters A–C give each evaluation set's source. Books are separated by split: no book appears in more than one of train, validation and test. All human-verified pages with partial or no vowelling come from a single scanned book (Taghut) in the test split, so the training and validation material is entirely fully vowelled. The silver references were produced by a commercial LLM-based OCR pipeline and were not checked by a person, except for the 38-page verified subset. Counts are given in Table 1.

Silver and verified scan sets. Two further sets have references drafted by an LLM-based OCR pipeline (a commercial Gemini model with automatic checks) and no human check; we call these *silver*:

- the *300 dpi sample*: ten 300 dpi pages, five from each of two books;
- 97 pages from two mostly unvowelled commentaries: *Inna Kunna fi Jahiliyya* (300 dpi, 49 pages) and *al-Jami' li-'Ibadat Allah Wahdah* (low resolution, 48 pages).

The results also pool all pages scanned at 300 dpi into one 59-page aggregate, *all 300 dpi pages*. It is the 10-page 300 dpi sample plus the 49 *Inna Kunna* pages, so it overlaps the unvowelled set.

From the 97, 38 pages were checked against the scan to form the *unvowelled-verified* set:

- the 21 pages where base and fine-tuned outputs disagreed most;
- a 20-page sample of body pages spread through both books, of which 3 pages needed corrections;
- three pages the reviewer marked as uncertain were excluded.

Six pages were corrected, for a missing verse opening, a wrong word, verse numbers and a garbled honorific. Before the check, the honorific "may God be pleased with him/them" was normalised across all 97 references to the single printed symbol (U+FD41/FD43/FD44), replacing spelled-out words and a font-mapping artefact. That change altered 24 references and 230 reference letters. Because the verified pages were partly chosen by model disagreement, the set is not a random sample.

Synthetic data. Synthetic pages are rendered in a headless browser and then degraded.

- **Text.** Mixed from the Tashkeela corpus [42] (weight 0.5), hadith collections (0.25) and gold-*training* text (0.25), with occasional QPC verse blocks and footnotes.
- **Fonts.** Five open Naskh-style fonts, with Kufi headings.
- **Vowelling.** Each page keeps full vowelling, strips it, or keeps a partial subset.
- **Degradation.** Clean, scanned (tint, skew, blur, noise, JPEG), low-resolution (down- and up-sampling) or bilevel.

Two synthetic test sets render validation and test *text* (60 pages each): one in a held-out font, one in the training fonts. They measure in-distribution behaviour only. The raw renders of the training sets were later deleted. Their per-page settings were recovered by replaying the deterministic generator, which reproduces the surviving synthetic test sets on 60 of 60 pages; the training labels and resized images survive.

Overlap between training and evaluation. We measured shared 8-word shingles between all training text (gold train and every synthetic label) and each evaluation set, outside Qur'anic brackets, on letters-only

Table 1: Dataset composition and reference status. Letters are Arabic letters in the reference. Vowelling levels use the repository's marks-per-letter thresholds: full ≥ 0.5, partial ≥ 0.05, none below that. "Verified" means a person checked the text against the page. "Silver" means an LLM-based OCR pipeline drafted the text and no person checked it. The synthetic labels are exact by construction.

Set (use)	Books	Print	Pages	Letters	Full / partial / none	Reference
Gold train (training)	Tawhid, Shurut, Qawa'id	typeset	190	43,300	190 / 0 / 0	verified
Gold validation (selection)	Usul ath-Thalatha, Nawaqid	typeset	36	9,819	36 / 0 / 0	verified
Gold test	Kashf ash-Shubuhat	typeset	80	19,682	80 / 0 / 0	verified
	Sharh Ma'na at-Taghut	low-res scan	44	21,215	4 / 26 / 14	verified
Unvowelled (silver)	Inna Kunna, al-Jami'	scans	97	52,008	8 / 45 / 44	silver
of which verified	Inna Kunna 17, al-Jami' 21	scans	38	18,914	7 / 20 / 11	verified
300 dpi sample (silver)	two books, 5 pages each	300 dpi scans	10	8,525	4 / 6 / 0	silver
Synthetic held-out font	val./test text, Harmattan	rendered	60	53,877	23 / 31 / 6	synthetic
Synthetic held-out text	val./test text, training fonts	rendered	60	55,851	14 / 40 / 6	synthetic
Train v1 (lora-001)	synthetic + gold train ×5	rendered	10,950 rows	—	synth. 4,012 / 4,397 / 1,591	synthetic
Train v2 (lora-002)	synthetic + gold train ×2	rendered	10,227 rows	—	synth. 2,568 / 4,058 / 3,221	synthetic

folded text. The shared fraction is 0.17 % for the gold test split, 2.4 % for validation, 0.18 % for unvowelled-verified and 0.40 % for the full unvowelled set. It consists mostly of hadith quoted in several books. Validation's 2.4 % is a single passage: the same author's opening formula, which also appears in a training book. No evaluation book's text was used to render training pages. A leakage filter removed Tashkeela and hadith passages that overlapped validation and test text, but it did not cover gold-train text or the later scan sets. Appendix C gives the full table.

Models and training. The base model is dots.mocr (`dots-studio/dots.mocr` on Hugging Face), a vision-language model for document parsing [9, 44]. We train LoRA adapters [17] on the language model only, with the vision encoder frozen, using the serving prompt and loss on answer tokens (Table 2).

- **lora-001** trained on synthetic set v1 plus five copies of gold train. It was released as *v0.1*.
- **lora-002** trained on synthetic set v2 plus two copies of gold train. v2 changes the vowelling mix, the partial-vowelling rule and the share of low-resolution renders. The run also differs in steps, time limit, data loading and trainer version.

Each run kept the adapter with the lowest validation loss, computed on the 36 validation pages. Those pages are therefore not independent evidence, and we report them separately. Inference uses vLLM [23] with greedy decoding and a 4,096-token limit. *v0.1 + retry* re-sends a page that hit the limit once with repetition penalty [19] 1.1.

Evaluation protocol. The primary condition is raw transcription (layer off), reported separately for:

- the test split (never used for training or selection);
- additional real scans: verified and silver, never used for training or selection;
- the selection pages;
- synthetic renders.

The layer-on condition is reported separately (Section 6). Images are sent at 2,048 px on the long side. Every number was recomputed from the stored predictions with the project's scorer (Appendix D).

5 Results

Table 3 and Figure 3 give the primary, raw-transcription results for base dots.mocr and v0.1. Appendix Table 7 adds lora-002 and the retry variant.

5.1 Letters-only CER misses most of what changes

On the 124 test pages, mean letters-only CER is the same for both models: 0.051 vs. 0.051, paired difference −0.000 [−0.026, +0.040]. The diacritic metrics on the same pages all move:

- DER falls from 0.204 to 0.177 (−0.027 [−0.041, −0.010]);
- invented marks fall from 36.9 to 21.6 per 1,000 aligned letters;
- omitted marks fall from 9.2 to 4.7.

The two metric families disagree in both directions:

Table 2: Systems and training differences. Settings shared by both runs: LoRA rank 16, $\alpha = 32$, dropout 0.05 on the language model's attention and MLP projections, vision encoder frozen; AdamW, learning rate 10^{-4} with cosine schedule; batch 1 with gradient accumulation 8; bf16; 2048 px long side; loss on answer tokens only, with the serving prompt. Rows marked † differ between lora-001 and lora-002. Because several of them change together, the comparison cannot attribute an effect to any one of them.

	lora-001 (published as v0.1)	lora-002 (not released)
Training rows†	10,950: synthetic v1 10,000 + gold train ×5	10,227: synthetic v2 9,847 + gold train ×2
Synthetic vowelling mix†	full 40 / partial 35 / none 25 %	full 25 / partial 40 / none 35 %
Partial-vowelling rule†	per word (whole 12 %, shadda 33 %, ...)	per page density 0.03–0.15
Synthetic degradation mix†	clean 35 / scan 40 / low-res 15 / bilevel 10 %	15 / 35 / 35 / 15 %
Steps (share of one pass)†	1,062 of 1,369 (78 %), 170-min limit	1,149 of 1,279 (90 %), 105-min limit
Checkpoint	best validation loss (step 1,000)	best validation loss (step inferred)
Data loading / code version†	inline; trainer before commit `0262cfa`	6 workers; later trainer
Hardware	H100 (stated, not logged)	80 GB GPU (memory logged; H100 inferred)
Seed	not logged (code default 1)	not logged (code default 1)

Inference (all systems): vLLM 0.30, temperature 0, at most 4,096 generated tokens, prompt "Extract the text content from this image." *Base:* `dots-studio/dots.mocr`. The revision is not pinned in code; HF `main` = `e539fbb` is inferred. *v0.1 + retry:* the same weights; a page that hits the token limit is re-sent once with repetition penalty 1.1. *Released v0.1:* HF `pandanlabs/dots-mocr-arabic-classical`, revision `037b472`. Identity with the evaluated checkpoint rests on the storage path, timestamps and shard sizes; no byte-level hash comparison was made.

- **Unvowelled-verified (38 pages).** v0.1 lowers DER by 0.158 and invented marks from 58.3 to 20.7, yet its mean CER rises from 0.139 to 0.251. The rise comes from one page where v0.1 repeats digits to the token limit (CER 6.51). With the retry, mean CER is 0.083 and DER is unchanged.
- **Silver low-resolution book.** The same CER and DER pattern appears (Appendix Table 7).

On typeset text, CER understates the change. On Kashf, the diacritic-free error is already low for base (0.019), while base invents 30.3 marks per 1,000 letters and its DER is 0.053, six times v0.1's 0.008. A single diacritic-free number cannot separate these behaviours: copying marks, inventing them, dropping them, and losing text. Figure 5 shows each on real pages.

5.2 Gains depend on document condition

Typeset print. On the 80 typeset test pages, v0.1 reduces DER from 0.053 to 0.008 (-0.045 [-0.053, -0.037]) and CER from 0.019 to 0.005. Invented marks drop from 30.3 to 3.0 per 1,000 letters. The 36 selection pages show a similar pattern (0.091 → 0.010). Because they chose the checkpoint, we do not count them as independent evidence. All typeset pages come from one edition series, and the same publisher's typesetting appears in training.

Human-verified low-resolution scan. On the 44-page scanned commentary there is no DER improvement: 0.492 vs. 0.499, difference +0.007 [-0.031, +0.050]. Invented marks barely change (43.4 → 39.9) and omitted marks fall (13.1 → 7.3). v0.1 has fewer broken pages (3 → 1) but more omitted five-word runs (2 → 5), and its CER is higher without the retry (0.110 → 0.133) and lower with it (0.090). This book carries all of the test split's partially and unvowelled pages, so its result cannot be separated from the vowelling-density shift. Some of its references may also disagree with the scan (Section 7).

Other real scans.

- **Unvowelled-verified (human-checked).** DER improves by 0.158 [-0.222, -0.096], mainly because v0.1 adds far fewer marks to unvowelled text.
- **Silver sets.** Improvements are similar: all 300 dpi pages (59) -0.129; low-resolution al-Jami' -0.126. Their references were never checked by a person. The 59-page aggregate is the 10-page 300 dpi sample plus the 49 *Inna Kunna* pages, and the verified set overlaps both commentaries, so these sets are not independent of each other.

Synthetic renders. DER falls from 0.218 to 0.035 and from 0.261 to 0.038. These pages share the generator and text sources with training, so they show only in-distribution learning.

In short, v0.1 is clearly better on the evaluated typeset pages and on the unvowelled commentary scans, and not better on the one human-verified low-resolution scan book. We do not describe it as better on scanned documents in general.

Table 3: Primary results: raw transcription (Qur'an layer *off*), base dots.mocr vs. v0.1 (lora-001). DER and CER are page means; ΔDER is the paired mean difference over pages where DER is defined for both models, with a 95% percentile bootstrap interval (4,000 resamples, seed 1); bold intervals exclude zero. CER is letters-only; the second v0.1 CER value uses the one-shot retry for pages that hit the token limit (DER is unchanged by the retry on the test split). Invented/omitted marks are pooled per 1,000 aligned letters. n is the number of pages scored for CER; the DER columns are computed over the subset of pages where DER is defined for both systems, whose count is the n of Table 4 (e.g. 122 of the 124 gold test pages).

			DER			Letters-only CER		Invented	Omitted
Evaluation set	Ref.	n	base	v0.1	Δ [95% CI]	base	v0.1 / retry	/1k	/1k
Test split (never used for training or checkpoint selection)									
Kashf (typeset)	verified	80	0.053	0.008	**−0.045 [−0.053, −0.037]**	0.019	0.005 / 0.005	30.3 → 3.0	5.2 → 2.0
Taghut (low-res scan)	verified	44	0.492	0.499	+0.007 [−0.031, +0.050]	0.110	0.133 / 0.090	43.4 → 39.9	13.1 → 7.3
Gold test, all	verified	124	0.204	0.177	**−0.027 [−0.041, −0.010]**	0.051	0.051 / 0.035	36.9 → 21.6	9.2 → 4.7
Additional real scans (not used for training or selection)									
Unvowelled scans	verified	38	0.455	0.297	**−0.158 [−0.222, −0.096]**	0.139	0.251 / 0.083	58.3 → 20.7	14.6 → 6.3
All 300 dpi pages (59)	silver	59	0.453	0.324	**−0.129 [−0.172, −0.086]**	0.078	0.054 / 0.054	35.7 → 11.9	17.2 → 9.8
al-Jami' (low-res scan)	silver	48	0.493	0.367	**−0.126 [−0.193, −0.061]**	0.140	0.246 / 0.061	72.5 → 31.6	16.6 → 5.6
Checkpoint-selection pages (not independent evidence)									
Validation (selection)	verified	36	0.091	0.010	**−0.080 [−0.095, −0.066]**	0.033	0.011 / 0.011	55.7 → 2.7	7.9 → 2.5
Synthetic held-out renders (in-distribution)									
Synthetic, held-out font	synthetic	60	0.218	0.035	**−0.183 [−0.207, −0.159]**	0.026	0.007 / 0.007	39.0 → 2.4	15.5 → 3.2
Synthetic, held-out text	synthetic	60	0.261	0.038	**−0.223 [−0.263, −0.188]**	0.023	0.006 / 0.006	61.6 → 2.4	15.1 → 1.8

5.3 Second training run (negative result)

lora-002 matches lora-001 closely on typeset DER (Kashf 0.011 vs. 0.008). On the 116 typeset pages it is slightly worse, +0.002 [+0.000, +0.004]. It produces more catastrophic outputs: test CER is 0.160 against 0.051, with 3 broken pages against 1 (Table 7). The CER mean is dominated by one looping title page of the low-resolution book, with page CER 12.4, which illustrates how unbounded page CERs distort set means. It was not released. The two runs differ in synthetic vowelling mix, partial-vowelling rule, share of degraded renders, number of gold copies, steps, time limit and trainer version (Table 2). So this comparison supports no claim about which change caused the difference. As a descriptive aside: v2's partially vowelled synthetic pages (0.047 marks per letter outside verses) are closer to the real partially vowelled scans (0.025–0.056) than v1's (0.115), and that did not translate into better scan results here.

5.4 Variability and the release decision

Run-to-run variability. The base model was run twice on the gold pages with the same weights and greedy decoding. 28 of 157 answers differ between the runs; the cause, such as serving non-determinism or an unlogged setting, is not recorded. The DER and CER differences are small, but on the 113 typeset pages with the layer on, the paired DER interval excludes zero (−0.0005 [−0.0009, −0.0001]). Under the project's release gate the model therefore "passes" against itself with the layer on, and fails with it off (Table 4). Intervals of this width should not be read as evidence of improvement.

Release-gate chronology. The project used an automatic gate to decide whether to release fine-tuned weights. All times are 25 October 2026, UTC+3:

- 15:56: the original gate (commit `0262cfa`) required DER to improve over all 160 gold pages, no slice to worsen by more than 0.005, and no increase in broken pages, invented marks or omitted runs.
- 18:39: lora-001 *failed* it. Overall DER went 0.124 → 0.131 ([−0.008, +0.022]). The scanned book worsened (+0.090 [+0.045, +0.138]). Partially vowelled pages worsened. Omitted runs rose from 2 to 5.
- 22:30: a tiered gate was committed (`57ad5b7`). It applies the strict rules only to non-low-resolution pages and fails low-resolution pages only if the lower bound of the DER increase exceeds 0.05.
- On the same stored outputs, lora-001 *passed* the tiered gate. Its low-resolution lower bound is 0.0448.
- 23:13: the weights were published.

The record does not show whether the rule was chosen after seeing the result; we report the order of events.

On the gold set, the strict tier contains only typeset, fully vowelled pages. It never sees a scan or a partially vowelled page. The gate verdicts also depend on the layer and on the bootstrap seed:

- lora-002 failed the tiered gate at seed 1 (lower bound 0.0501).

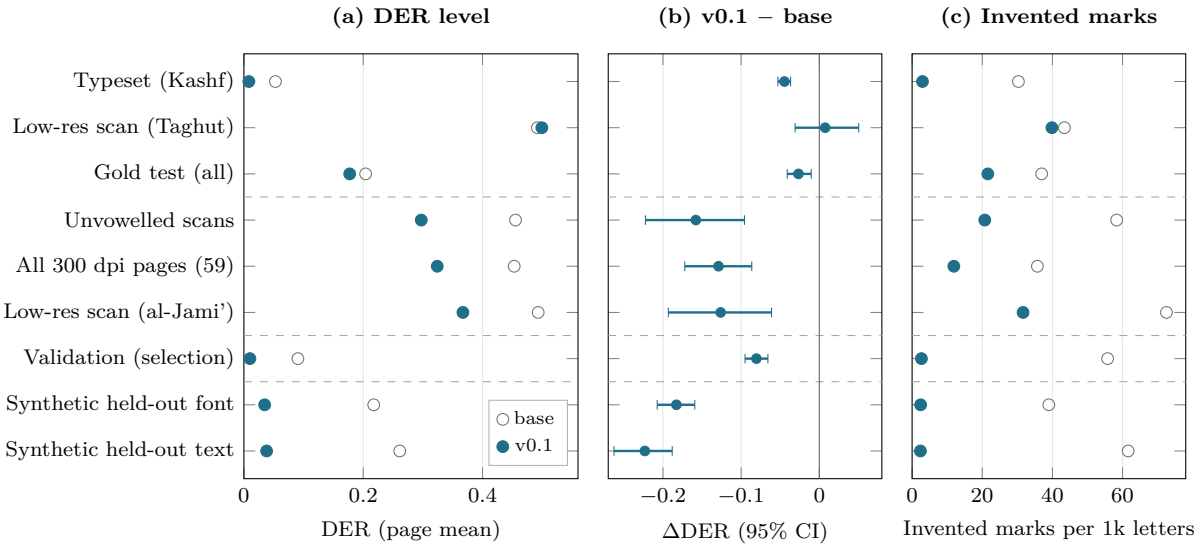


Figure 3: Raw transcription (Qur'an layer off) by document condition, base dots.mocr (open) vs. v0.1 (filled). Rows above the first dashed line are the test split, all human-verified. The next block holds additional real scans: the 38-page unvowelled set is human-verified, and the other two are silver. Below that come the 36 checkpoint-selection pages and the synthetic held-out renders. (b) Paired bootstrap 95% intervals for the change in page-mean DER. (c) Invented marks are pooled over aligned letters, so no interval is available. Sample sizes are in Table 3. Data: `figure_data/fig3_conditions_layer_off.csv`.

- Over seeds 1–200 the same stored outputs pass the low-resolution rule in 125 of 200 cases (bounds 0.047–0.052); lora-001 passes in 200 of 200 (bounds 0.041–0.046).

The release distinction between the two runs is therefore within bootstrap noise. All gold-set numbers in this paper are reported with intervals rather than gate verdicts.

6 Analysis of Post-Processing Effects

Classical texts quote the Qur'an constantly, and pipelines for such texts often swap a recognised quotation for canonical text from a reference edition. Our layer does this deterministically (Section 3). The reference is still the printed page, so the layer can help in two ways: by correcting a misread verse, and by supplying the exact Uthmani marks and spelling the print uses. It can also hurt in three:

- it replaces a near-miss with the wrong verse window;
- it truncates a quotation that runs past the page end;
- it imposes the canonical edition where the printed edition differs.

The layer changes absolute error. With the layer on, base DER falls substantially on every set that contains verses: Kashf 0.053 → 0.030, the scanned test book 0.492 → 0.386, and the selection pages 0.091 → 0.030. v0.1's DER falls much less: Kashf 0.008 → 0.006, scanned book 0.499 → 0.476 (Figure 4b). Most of base's typeset diacritic error therefore lies inside Qur'anic quotations, which the layer overwrites with canonical text.

The layer changes the comparison. Because the layer helps base more, the measured benefit of fine-tuning shrinks everywhere and changes sign or significance on three sets (Table 4, Figure 4a):

- **Gold test.** A significant improvement (−0.027 [−0.041, −0.010]) becomes a non-significant difference (+0.016 [−0.002, +0.036]).
- **Low-resolution scanned book.** A null result (+0.007 [−0.031, +0.050]) becomes a significant regression (+0.090 [+0.045, +0.138]).
- **Silver low-resolution book.** A significant improvement (−0.126) becomes non-significant (−0.054 [−0.129, +0.015]).

Release verdicts flip too. lora-002 passes the gold release gate with the layer off and fails it with the layer on, and the base model passes against itself only with the layer on. Every release decision in the project was taken with the layer on. The regression on the scanned book that drove the original gate failure (Section 5.4) is therefore largely a property of the post-processed condition, not of raw transcription.

Table 4: Effect of the Qur'an quotation-replacement layer on measured DER. Δ = v0.1 − base (paired, 95% bootstrap CI; bold excludes zero). The reference is always the printed source; the layer rewrites only model output. Lower panel: release-gate verdicts on the 160 gold pages computed by the repository's gate code with the layer off and on. Here n is the number of *paired* pages on which DER is defined for both systems (the same with the layer off and on). It can be smaller than the n of Table 3, which counts all pages scored for CER, because DER is undefined on a page with no marked letter on either side (Section 3); e.g. 122 vs. 124 on the gold test split.

		Layer off (primary)		Layer on	
Evaluation set	n	base → v0.1	ΔDER [95% CI]	base → v0.1	ΔDER [95% CI]
Kashf (typeset)	80	0.053 → 0.008	**−0.045 [−0.053, −0.037]**	0.030 → 0.006	**−0.023 [−0.028, −0.019]**
Taghut (low-res scan)	42	0.492 → 0.499	+0.007 [−0.031, +0.050]	0.386 → 0.476	**+0.090 [+0.045, +0.138]**
Gold test, all	122	0.204 → 0.177	**−0.027 [−0.041, −0.010]**	0.152 → 0.168	+0.016 [−0.002, +0.036]
Unvowelled scans	35	0.455 → 0.297	**−0.158 [−0.222, −0.096]**	0.376 → 0.274	**−0.101 [−0.175, −0.031]**
All 300 dpi pages (59)	58	0.453 → 0.324	**−0.129 [−0.172, −0.086]**	0.421 → 0.316	**−0.105 [−0.149, −0.061]**
al-Jami' (low-res scan)	45	0.493 → 0.367	**−0.126 [−0.193, −0.061]**	0.389 → 0.335	−0.054 [−0.129, +0.015]
Validation (selection)	36	0.091 → 0.010	**−0.080 [−0.095, −0.066]**	0.030 → 0.005	**−0.024 [−0.033, −0.017]**
Synthetic, held-out font	59	0.218 → 0.035	**−0.183 [−0.207, −0.159]**	0.097 → 0.027	**−0.070 [−0.095, −0.047]**
Synthetic, held-out text	59	0.261 → 0.038	**−0.223 [−0.263, −0.188]**	0.154 → 0.031	**−0.123 [−0.169, −0.082]**
Release-gate verdict on gold v0 (tiered gate, current code)					
Comparison		Layer off		Layer on	
base → v0.1 (lora-001)		PASS		PASS	
base → lora-002		PASS		FAIL	
lora-001 → lora-002		FAIL		FAIL	
base run 1 → base run 2 (same model)		FAIL		PASS	

Why the layer favours one model. The data support a partial mechanism.

1. **Errors count against the verse span.** The layer only replaces a quotation it recognises. A misreading of two words in a short verse is enough to block the match, and then every error inside the span counts (Figure 5c): base misreads one word of ﴿لَآ إِكۡرَاهَ فِي ٱلدِّينِ﴾ and is fully corrected; v0.1 misreads two and is left unchanged.
2. **v0.1 is matched less often.** The share of a system's quotation spans that the layer can identify as verses is lower for v0.1 than for base: 0.864 → 0.830 on the unvowelled scans, 0.825 → 0.750 on the verified subset, and 0.907 → 0.901 on the gold pages.
3. **Hypothesis: less to gain.** v0.1, trained on QPC verse blocks, may already reproduce more of the canonical form in the spans it does get right. If so, the layer has less to gain there.

We did not measure how often the layer introduces errors where the printed edition differs from QPC Hafs. We observed one case where it deletes words from a quotation cut off at the page end.

Implication. A canonical-text layer measures a different task: verse identification plus canonical substitution, not transcription. Reporting only post-processed scores would have hidden a significant raw-transcription gain on the test split and attributed a large regression to the model on the scanned book. We recommend reporting raw transcription as the primary result and post-processing as a separately labelled condition. The quotation-identification rate should be reported with it, and it should not be read as a transcription accuracy.

7 Error Analysis and Discussion

Table 5 groups the failure modes found in the stored outputs, and Figure 5 shows three on real crops.

Invented and omitted marks. The base model's dominant failure on typeset print is mark-level, not letter-level. It drops shaddas that carry meaning (Figure 5a) and adds case endings the edition does not print. On unvowelled commentary it fully vowels the text and, in doing so, sometimes changes words: in Figure 5b إنا سنرجع becomes an unrelated vowelled phrase. The fine-tuned model mostly copies what is printed. Letters-only CER scores the first case as nearly perfect and the second as a few substitutions, so it misses the main behavioural difference between the systems.

Degraded scans. On the low-resolution scans, errors are dominated by dotted-letter confusions (ب/ي, ف/ق), garbled calligraphic title pages, and loops on short pages. Eight of the ten worst v0.1 test pages belong to the low-resolution book. Fine-tuning does not fix these; the retry removes most loops.

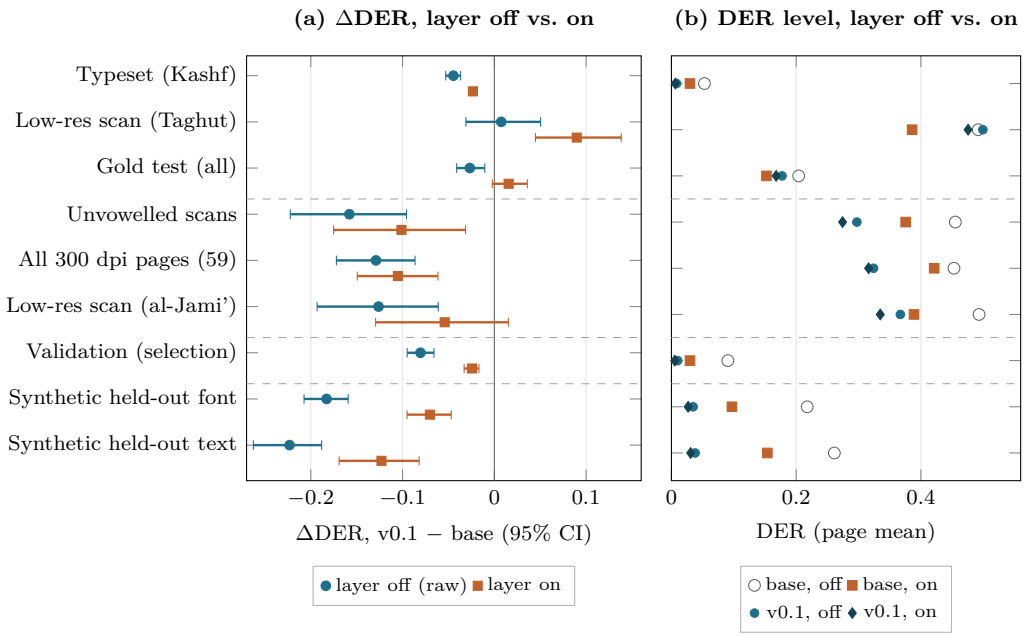


Figure 4: The quotation layer changes the measured comparison. (a) Paired change in DER with the layer off (circles) and on (squares). On the gold test split the layer turns a significant improvement (−0.027, CI excludes zero) into a non-significant difference (+0.016). On the verified low-resolution scan (Taghut) it turns a null result into a significant regression (+0.090). (b) The layer lowers the base model's DER much more than v0.1's (e.g. typeset Kashf: base 0.053 → 0.030, v0.1 0.008 → 0.006). Data: `figure_data/fig4_layer_effect.csv`.

Metric pathologies seen in practice. Five recur in the stored outputs:

- **Unbounded CER.** A single looping page (page CER 6.5 or 12.4) can dominate a set mean.
- **DER on unvowelled pages.** On references without vowels, page DER takes only the values 0, 1 or undefined.
- **Text loss hidden from DER.** DER ignores unaligned letters, so a model can lose text without any DER penalty.
- **Missed loops.** The loop detector reads letters only, so repeated verse-number markers (◯) are missed and caught only as truncation.
- **Inconsistent honorifics.** The two honorific ligatures are treated inconsistently by the letters view (Section 3).

None of these changes the direction of the main results, but each shows why we report several metrics together with page-level failure counts.

Reference quality. References are the weakest link in evaluating faithful OCR, because an error in the reference is scored as a model error:

- **Silver errors.** One silver key contains a stray CJK character.
- **Corrections on review.** Review of 41 silver pages found errors on 6.
- **Honorific change.** Normalising the honorific changed 24 of 97 silver references.
- **The verified scanned test book.** A visual check of crops suggests that the references omit marks the scan prints on two pages, where v0.1 is charged with "invented" marks it read correctly. On another page they add marks the scan does not show, and v0.1 is charged with "omissions" for following the print. On the title and colophon pages, names may not match.

These observations have not been re-adjudicated by a second annotator. Because this book is the only human-verified low-resolution source, its DER of about 0.5 for both systems contains an unknown contribution from reference disagreement. The direction of that bias is unknown. Marks missing from the reference inflate a system's invented-mark count, and marks the reference adds inflate its omitted count. Errors can also cancel within a page, so the true DER may be lower or higher than measured, and the bias may differ between the two systems.

Why do typeset gains not carry over? The evidence does not settle this, and we offer hypotheses only:

1. **Training distribution.** Real training pages are all typeset and fully vowelled, so the model never sees a real degraded or partially vowelled page in training.

(a) Typeset test page, human-verified (Kashf ash-Shubuhat, p. 73): *omitted shaddas that change the meaning.*

Scan excerpt omitted from this web edition pending rights clearance; the transcriptions below are unchanged.

	first excerpt (line 1–2 of the crop)	what differs
Reference	أَمَا سَمِعْتَ اللَّهَ كَفَّرَهُمْ بِكَلِمَةٍ	
Base	أَمَا سَمِعْتَ اللَّهَ كَفَرَهُم بِكَلِمَةٍ	shadda omitted: *kaffarahum* "declared them unbelievers" → *kafarahum* "disbelieved in them"
v0.1	أَمَا سَمِعْتَ اللَّهَ كَفَّرَهُمْ بِكَلِمَةٍ	identical to the reference
	second excerpt (last line of the crop)	
Reference	وَيَحُجُّونَ، وَيُوَحِّدُونَ -؟	
Base	وَيَحُجُّونَ، وَيُوحَدُونَ - ؟	shadda omitted (and a vowel changed): *yuwaḥḥidūna* "affirm God's oneness" → *yūḥadūna*
v0.1	وَيَحُجُّونَ، وَيُوَحِّدُونَ -؟	identical to the reference

Page DER (layer off): base 0.119 (17 invented, 7 omitted and 7 changed marks); v0.1 0.000.

(b) Unvowelled scan, human-verified (Sharh al-Jami', p. 33): *invented vowels and changed words.* The print has no vowels.

Scan excerpt omitted from this web edition pending rights clearance; the transcriptions below are unchanged.

Reference	وأصحابه بعد وقعة أحد وقالوا: إنا سنرجع	as printed, without vowels
Base	وَأَصْحَابُهُ بَعْدَ وَقْعَةِ أَحَدٍ وَقَالُوا: إِنَّا سُنِجَ عَلَيْنَآ	vowels added to unvowelled text; the final words are replaced by an unrelated phrase
v0.1	وأصحابه بعد وقعة أحد وقالوا: إنا سنرجع	identical to the reference

Page (layer off, verified reference): base CER 0.077, DER 0.677, 194 invented marks; v0.1 CER 0.023, DER 0.165, 12 invented and 7 omitted.

(c) Low-resolution test scan, human-verified (Sharh Ma'na at-Taghut, p. 8): *the quotation layer.* The verse appears five times on the page.

Scan excerpt omitted from this web edition pending rights clearance; the transcriptions below are unchanged.

	raw output (layer off)	after the Qur'an layer
Reference	﴿لَآ إِكۡرَاهَ فِي ٱلدِّينِ﴾	﴿لَآ إِكۡرَاهَ فِي ٱلدِّينِ﴾
Base	﴿لَا إِكْرَاهَ فِي ٱلَّذِينَ﴾	﴿لَآ إِكۡرَاهَ فِي ٱلدِّينِ﴾
v0.1	﴿لَآ إِكْرَأَ فِي ٱلَّذِينَ﴾	﴿لَآ إِكْرَأَ فِي ٱلَّذِينَ﴾

Base misreads one word, and the layer replaces the span with the canonical verse (page DER 0.362 → 0.048). v0.1 misreads two words, so the span is not matched and is left unchanged (DER 0.576 → 0.576). The reference is the same in both columns.

Figure 5: Real page crops with the reference (the printed source as verified), base dots.mocr and v0.1 output. Words that differ from the reference have a red background; the corresponding correct words have a green one. The prediction strings are reproduced exactly. Crops are short excerpts of the evaluation images (see §8 on rights). Page-level metrics as defined in Section 3.

Table 5: Failure categories, with examples from stored outputs and the limits of what the metrics show about each. All examples come from `paper/evidence/failure_analysis.md`; excerpts were checked byte-for-byte against the predictions.

Category	Example (system, page)	What the metric does and does not show
Invented diacritics	Base fully vowels an unvowelled commentary line and changes words (al-Jami' p. 33: 194 invented marks; v0.1: 12)	Counted per aligned letter. On references without vowels, page DER is 1.0 if any mark is invented, so the invented-marks rate is the informative measure.
Omitted / changed marks	Base drops meaning-bearing shaddas (Kashf p. 73)	DER counts them only where the base letters align; a substituted letter removes its marks from the denominator.
Wrong letters on degraded scans	Dotted-letter confusions ب/ي and ف/ق (Taghut p. 19)	Letters-only CER captures them; DER ignores the affected positions.
Omitted text	Base leaves out a whole verse (Inna Kunna p. 19); lora-002 stops after a title (Inna Kunna p. 1)	Runs of $\geq$5 omitted words are counted; shorter gaps and substitutions are not.
Repetition loops	v0.1 repeats digits to the token limit (al-Jami' p. 39, CER 6.51; retry 0.14)	CER is unbounded and page-averaged, so one loop page can dominate a set mean. The loop detector reads letters only and misses repeated ◌ markers.
Quotation-layer effects	Layer repairs a one-word misreading but not a two-word one (Taghut p. 8); it truncates a quote cut off at the page end (Usul p. 21)	Layer-on scores mix recognition with matcher behaviour (Section 6).
Reference errors	Silver key with a stray CJK character (al-Jami' p. 3); verified Taghut pages where a visual check suggests the key omits printed marks (pp. 36–37) or adds unprinted ones (p. 28)	Errors in the reference are counted as model errors. The Taghut cases are an author-side visual reading and have not been re-adjudicated.

2. **Synthetic degradation.** It may not resemble the low-resolution scan, a scan of a printed book with its own typeface and wear.
3. **Mark density.** Synthetic partially vowelled pages differ in mark density from real ones. v2 matched real density more closely without better scan results.
4. **Narrow and noisy evidence.** The verified scan evidence is one book with imperfect references.

Separating these explanations would need more verified scan books and controlled training variations. The present data support only the observation that the gains are condition-dependent.

8 Limitations and Reproducibility

Scope of evidence.

- **Typeset print.** The typeset evidence is one edition series.
- **Low-resolution scans.** The human-verified low-resolution evidence is one book of 44 pages, which also holds every partially or unvowelled test page.
- **Other scans.** The other scan results come from two commentaries with silver references, plus a 38-page verified subset that was partly selected by model disagreement.
- **Checkpoint selection.** The checkpoint was selected by validation loss on 36 gold pages that the project's own release evaluation also used. We report them separately and do not count them as independent evidence.
- **Training runs.** Each training configuration was run once, with an unlogged seed.
- **Statistics.** No multiple-comparison correction is applied to the many slice intervals. The intervals resample pages within the evaluated sets, so they do not cover variation across books, editions or publishers; conclusions about other books are not supported by them.

Evaluation-process issues.

- **Gate revision.** The release gate was revised after the released model failed its first version (Section 5.4).
- **Seed sensitivity.** The pass/fail distinction between the two runs depends on the bootstrap seed.
- **Run-to-run variability.** Repeated inference of the same model changed 28 of 157 answers.
- **Layer dependence.** All release decisions used the quotation layer, which Section 6 shows changes conclusions.

We report these so that readers can weigh the results. We do not claim that the released model passed a pre-registered criterion.

Audit mismatches. When all reported project numbers were recomputed, 85 of 96 matched exactly. The 11 mismatches have five causes and are not 11 independent errors:

- one is the gate verdict that changed with the revised rule;
- one is a last-digit rounding difference;
- two report the same page with scores from the layer-off condition inside a layer-on table;
- one states that a mean was "unchanged" when it moved by 0.001;
- six are values computed before the 26 October honorific normalisation of the silver references; they reproduce exactly against the earlier references.

Model identity. The base model was served from Hugging Face `dots-studio/dots.mocr` without a pinned revision; the revision in use (`e539fbb`) is inferred from the hub history. The released v0.1 (`pandan labs/dots-mocr-arabic-classical`, revision `037b472`) is consistent with the evaluated checkpoint: same storage path, timestamps before the evaluation, identical shard names and sizes. We did not compare byte-level hashes, and two checkpoints of this architecture have identical shard sizes, so the identity is inferred rather than verified.

Reproducibility.

- **Publicly released:** only the fine-tuned model weights (see Code and Model Availability). The scorer, the evaluation and data-generation code, and the stored predictions behind every table and figure are preserved in a private repository, with the audit scripts and SHA-256 manifests of their inputs, but are not public.
- **Missing:** the exact training and synthesis command lines, the seeds, and the deleted raw synthetic renders. The renders' per-page settings were reconstructed by deterministic replay.
- **Not version-controlled:** raw predictions and logs.

Appendix D lists commits, revisions and file locations.

Rights and data release. The classical *mutoon* are centuries old, but the page images are of a recent typeset edition, whose editorial vowelling and typesetting may carry rights. The scanned commentaries are modern works by a living author; one carries an "all rights reserved" notice. Fonts, the Qur'anic text and the synthetic text corpora have their own terms. We therefore do not redistribute page images here beyond the short crops in Figure 5, and evaluation-data release is subject to rights clearance. This limits independent reproduction of the scan results.

9 Conclusion

Faithful OCR of classical Arabic print has to be evaluated on the marks the page prints, not only on its letters. On human-verified test pages, a LoRA-fine-tuned dots.mocr leaves letters-only CER unchanged while substantially reducing diacritic errors and invented marks. The gains are large on typeset print and on unvowelled commentary scans, and absent on the one human-verified low-resolution scan book. Canonical replacement of Qur'anic quotations, a plausible production step, changes the comparison enough to hide a significant test-set improvement and to manufacture a significant regression.

For this task we recommend reporting:

- DER with invented and omitted-mark rates beside letters-only CER;
- page-paired intervals and page-level failure counts;
- raw transcription as the primary condition, with post-processing as a separate one;
- which references were checked by a person;
- test and selection pages kept apart.

We also recommend fixing release criteria before evaluation, and checking their sensitivity to bootstrap seeds and repeated inference. The evaluated model is public; the code and evaluation data are not (see Code and Model Availability below).

Code and Model Availability

The evaluated fine-tuned model is publicly available on Hugging Face. Evaluation code, audit scripts and associated research artifacts are currently maintained in a private repository. The reproducibility manifest

in Appendix D documents the available artifacts and the remaining limitations.

- **Fine-tuned model (v0.1, public):** `https://huggingface.co/pandanlabs/dots-mocr-arabic-classical`. The evaluated weights correspond to revision `037b472`: `https://huggingface.co/pandanlabs/dots-mocr-arabic-classical/tree/037b472f2a2fd004e02ce012549844af119eea62`. Identity with the evaluated checkpoint is inferred, not hash-verified (Section 8).
- **Base model (public, third party):** `https://huggingface.co/dots-studio/dots.mocr`.
- **Not publicly available at the time of writing:** the evaluation and training code, the scorer and quotation layer, the stored predictions, the audit materials, and the evaluation datasets with their page images. The page images come from editions whose rights have not been cleared for redistribution (Section 8).

References

[1] Gheith A. Abandah, Alex Graves, Balkees Al-Shagoor, Alaa Arabiyat, Fuad Jamour, and Majid Al-Taee. Automatic diacritization of Arabic text using recurrent neural networks. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(2):183–197, 2015. doi: 10.1007/s10032-015-0242-2.

[2] Faris Alasmary, Orjuwan Zaafarani, and Ahmad Ghannam. CATT: Character-based Arabic tashkeel transformer, 2024. URL `https://arxiv.org/abs/2407.03236`.

[3] Salah Alghyaline. Arabic optical character recognition: A review. *Computer Modeling in Engineering & Sciences*, 135(3):1825–1861, 2023. doi: 10.32604/cmes.2022.024555.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

[5] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In Jun'ichi Tsujii, James Henderson, and Marius Paşca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `https://aclanthology.org/D12-1091/`.

[6] Lukas Blecher, Guillem Cucurull Preixens, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations (ICLR)*, pages 37646–37663, 2024. URL `https://proceedings.iclr.cc/paper_files/paper/2024/hash/a39a9aceda771cded859ae7560530e09-Abstract-Conference.html`. arXiv:2308.13418.

[7] Rafael C. Carrasco. An open-source OCR evaluation tool. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014)*, pages 179–184. ACM, 2014. doi: 10.1145/2595188.2595221.

[8] Kareem Darwish, Hamdy Mubarak, and Ahmed Abdelali. Arabic diacritization: Stats, rules, and hacks. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 9–17, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1302. URL `https://aclanthology.org/W17-1302/`.

[9] dots-studio (rednote hi lab). dots.mocr. Hugging Face model card, 2026. URL `https://huggingface.co/dots-studio/dots.mocr`. MIT licence. Code: `https://github.com/rednote-hilab/dots.mocr`.

[10] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

[11] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Springer US, Boston, MA, 1993. doi: 10.1007/978-1-4899-4541-9.

[12] Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. Arabic text diacritization using deep neural networks. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–7. IEEE, 2019. doi: 10.1109/CAIS.2019.8769512.

[13] Ali Fadel, Ibraheem Tuffaha, Bara' Al-Jawarneh, and Mahmoud Al-Ayyoub. Neural Arabic text diacritization: State of the art results and a novel approach for machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 215–225, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5229. URL `https://aclanthology.org/D19-5229/`.

[14] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2315–2324. IEEE, 2016. doi: 10.1109/CVPR.2016.254.

[15] Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. KITAB-bench: A comprehensive multi-domain benchmark for Arabic OCR and document understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22006–22024, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.1135. URL https://aclanthology.org/2025.findings-acl.1135/.

[16] Khalil Hennara, Muhammad Hreden, Mohamed Motasim Hamed, Ahmad Bastati, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. Baseer: A vision-language model for Arabic document-to-markdown OCR, 2025. URL https://arxiv.org/abs/2509.18174.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. arXiv:2106.09685.

[18] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition, 2014. URL https://arxiv.org/abs/1406.2227.

[19] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation, 2019. URL https://arxiv.org/abs/1909.05858.

[20] Benjamin Kiessling. Version 5 of the Kraken ATR engine for the humanities. In *Document Analysis and Recognition – ICDAR 2025*, Lecture Notes in Computer Science, pages 443–458. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04624-6_26.

[21] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 498–517. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19815-1_29.

[22] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*, pages 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

[24] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. English translation of Doklady Akademii Nauk SSSR 163(4):845–848, 1965.

[25] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102, 2023. doi: 10.1609/aaai.v37i11.26538.

[26] Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. dots.ocr: Multilingual document layout parsing in a single vision-language model, 2025. URL https://arxiv.org/abs/2512.02498.

[27] Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. A survey of OCR evaluation tools and metrics. In *The 6th International Workshop on Historical Document Imaging and Processing (HIP '21)*, pages 13–18. ACM, 2021. doi: 10.1145/3476887.3476888.

[28] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

[29] Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking trillions of tokens in PDFs with vision language models, 2025. URL https://arxiv.org/abs/2502.18443.

[30] Quran.com. QPC Hafs text (King Fahd Glorious Qur'an Printing Complex) via the Quran.com content API v4. https://api-docs.quran.com/docs/content_apis_versioned/4.0.0/quran-verses-by-script/, 2026. Script identifier qpc_hafs; field text_qpc_hafs. Accessed 2026.

[31] Mohamed Rashad. Arabic-Nougat: Fine-tuning vision transformers for Arabic OCR and markdown extraction, 2024. URL https://arxiv.org/abs/2411.17835.

[32] Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. OCR4all—an open-source tool providing a (semi-)automatic OCR workflow for historical printings. *Applied Sciences*, 9(22):4853, 2019. doi: 10.3390/app9224853.

[33] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. An algorithm for matching OCR-generated text strings. In *Document Image Analysis*, Series in Machine Perception and Artificial Intelligence, pages 263–272. World Scientific, 1994. doi: 10.1142/9789812797797_0015.

[34] Maxim Romanov, Matthew Thomas Miller, Sarah Bowen Savant, and Benjamin Kiessling. Important new developments in Arabographic optical character recognition (OCR), 2017. URL https://arxiv.org/abs/1703.09550.

[35] Fouad Slimane, Rolf Ingold, Slim Kanoun, Adel M. Alimi, and Jean Hennebert. A new Arabic printed text image database and evaluation protocols. In *2009 10th International Conference on Document Analysis and Recognition*, pages 946–950. IEEE, 2009. doi: 10.1109/ICDAR.2009.155.

[36] R. Smith. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007. doi: 10.1109/ICDAR.2007.4376991.

[37] Tanzil Project. Tanzil Quran text. https://tanzil.net/download/, 2021. Version 1.1, February 2021. Simple, simple-clean and Uthmani text types.

[38] The Unicode Consortium. *The Unicode Standard, Version 14.0.0*. The Unicode Consortium, Mountain View, CA, 2021. ISBN 978-1-936213-29-0. URL http://www.unicode.org/versions/Unicode14.0.0/. Arabic honorific ligatures U+FD40–U+FD4F added in this version.

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

[40] Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. QARI-OCR: High-fidelity Arabic text recognition through multimodal large language model adaptation, 2025. URL https://arxiv.org/abs/2506.02295.

[41] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General OCR theory: Towards OCR-2.0 via a unified end-to-end model, 2024. URL https://arxiv.org/abs/2409.01704.

[42] Taha Zerrouki and Amar Balla. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11:147–151, 2017. doi: 10.1016/j.dib.2017.01.011.

[43] Yulong Zhang. OCR-Quality: A human-annotated dataset for OCR quality assessment, 2025. URL https://arxiv.org/abs/2510.21774.

[44] Handong Zheng, Yumeng Li, Kaile Zhang, Liang Xin, Guangwei Zhao, Hao Liu, Jiayu Chen, Jie Lou, Qi Fu, Rui Yang, Shuo Jiang, Weijian Luo, Weijie Su, Weijun Zhang, Xingyu Zhu, Yabin Li, Yiwei Ma, Yu Chen, Yuqiu Ji, Zhaohui Yu, Guang Yang, Colin Zhang, Lei Zhang, Yuliang Liu, and Xiang Bai. Multimodal OCR: Parse anything from documents, 2026. URL `https://arxiv.org/abs/2603.13032`. Technical report for dots.mocr.

[45] Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. Maximum entropy based restoration of Arabic diacritics. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 577–584, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220248. URL `https://aclanthology.org/P06-1073/`.

A Normalisation Edge Cases

Table 6 lists how the scorer treats characters that matter for classical Arabic. It was established by running the project's scorer on constructed strings; the inputs and outputs are in `paper/evidence/evaluation_methodology.md`, §9.

Table 6: Treatment of selected phenomena by the scorer's letters view (CER) and diacritic clusters (DER).

Phenomenon	Treatment
Alef/hamza variants ا أ إ آ; ى/ي; ة/ه	Not folded; each counts as a distinct letter.
Decomposed hamza/madda	Composed by NFC where possible; otherwise the combining mark becomes a space in the letters view and splits the word.
Presentation forms (U+FB50–FDFF, FE70–FEFF)	Not normalised (no NFKC): ﻻ ≠ لا.
ﷺ (U+FDFA) vs. ﷺ (U+FD41)	The first is a letter in the letters view, the second is not and is removed.
Tatweel	Removed from the letters view and from DER clusters; runs of ≥ 3 become an ellipsis in the plain view.
Harakat set	U+064B–0652 and U+0670. Maddah (U+0653), hamza marks (U+0654/5) and Qur'anic small signs are not harakat.
Qur'anic annotation marks	Removed from the letters view and from DER (U+06D6–06ED, U+08D3–08FF).
Digits	Arabic-Indic and Persian digits map to ASCII in the letters view only.
Mark order on a letter	Sorted before comparison; duplicated marks count as a change.
Marks on spaces or punctuation	Ignored by DER, counted in raw CER.

B All Systems, Variability and Gate Definitions

Same-model rerun. Two inference runs of base dots.mocr on the gold pages (157 pages common to both) give these paired differences (run 2 − run 1):

- **Layer off.** DER −0.0027 [−0.0096, +0.0011]; CER +0.0003 [−0.0015, +0.0020]; typeset DER −0.0004 [−0.0009, +0.0001].
- **Layer on.** Typeset DER −0.0005 [−0.0009, −0.0001].

Data: `figure_data/same_model_rerun.csv`.

Bootstrap-seed sensitivity. For the tiered gate's low-resolution rule (fail if the lower 95% bound of the DER increase exceeds 0.05), we recomputed the bound for seeds 1–200 on the same stored outputs:

- **lora-001:** bounds 0.0406–0.0463, passing in 200/200 seeds;
- **lora-002:** bounds 0.0472–0.0516, passing in 125/200 seeds.

Data: `figure_data/bootstrap_seed_sensitivity.csv`.

Gate definitions. The *original* gate (commit `0262cfa`) fails a candidate unless:

- DER is better over all pages with 95% confidence;
- no slice's DER lower bound exceeds +0.005;
- broken pages, invented marks per 1k (allowing a factor of 1.05 plus 0.1) and omitted five-word runs do not increase;

Table 7: All evaluated systems, layer off. DER page mean (paired Δ vs. base with 95% CI), letters-only CER page mean, broken pages (loop, truncation, empty or CER>0.5) and runs of ≥5 consecutive omitted reference words. Bold intervals exclude zero.

Evaluation set	System	n	DER [Δ 95% CI]	CER	Broken	Omitted runs
Kashf (typeset)	base	80	0.053	0.019	0	0
	v0.1		0.008 [**−0.045 [−0.053, −0.037]**]	0.005	0	0
	v0.1 + retry		0.008 [**−0.045 [−0.053, −0.037]**]	0.005	0	0
	lora-002		0.011 [**−0.042 [−0.051, −0.034]**]	0.005	0	0
Taghut (low-res scan)	base	44	0.492	0.110	3	2
	v0.1		0.499 [+0.007 [−0.031, +0.050]]	0.133	1	5
	v0.1 + retry		0.499 [+0.007 [−0.031, +0.050]]	0.090	1	2
	lora-002		0.517 [+0.021 [−0.022, +0.068]]	0.441	3	3
Gold test, all	base	124	0.204	0.051	3	2
	v0.1		0.177 [**−0.027 [−0.041, −0.010]**]	0.051	1	5
	v0.1 + retry		0.177 [**−0.027 [−0.041, −0.010]**]	0.035	1	2
	lora-002		0.182 [**−0.021 [−0.038, −0.002]**]	0.160	3	3
Unvowelled scans	base	38	0.455	0.139	2	6
	v0.1		0.297 [**−0.158 [−0.222, −0.096]**]	0.251	3	6
	v0.1 + retry		0.300 [**−0.156 [−0.216, −0.095]**]	0.083	2	4
	lora-002		0.324 [**−0.146 [−0.218, −0.079]**]	0.092	1	5
All 300 dpi pages (59)	base	59	0.453	0.078	1	22
	v0.1		0.324 [**−0.129 [−0.172, −0.086]**]	0.054	1	22
	v0.1 + retry		0.324 [**−0.129 [−0.172, −0.086]**]	0.054	1	22
	lora-002		0.363 [**−0.100 [−0.152, −0.050]**]	0.062	2	25
al-Jami' (low-res scan)	base	48	0.493	0.140	2	7
	v0.1		0.367 [**−0.126 [−0.193, −0.061]**]	0.246	3	8
	v0.1 + retry		0.369 [**−0.124 [−0.190, −0.060]**]	0.061	2	6
	lora-002		0.355 [**−0.138 [−0.209, −0.070]**]	0.116	1	6

- the quotation-match rate does not fall by more than 0.02.

The *tiered* gate (commit `57ad5b7`) applies these rules only to pages not labelled low-resolution. For low-resolution pages it fails only if the lower bound of the DER increase exceeds 0.05, and it does not check broken pages, invented marks, omitted runs or quotation matching for them. Neither gate checks CER.

C Text Overlap Between Training and Evaluation

Table 8: Shared 8-word shingles between all training text (gold train and all synthetic labels of both versions) and each evaluation set, counted outside Qur'anic brackets on letters-only, folded text. "Other" is shared text that is neither a known hadith nor Qur'anic. Source: `paper/evidence/dataset_audit.md`, §6.

Evaluation set	Shared / total shingles	Hadith	Other	Pages with "other"
Gold validation	23 / 949 (2.4 %)	5	18	1
Gold test	12 / 7,056 (0.17 %)	10	2	1
Unvowelled (silver)	40 / 9,969 (0.40 %)	24	16	4
Unvowelled-verified	6 / 3,279 (0.18 %)	6	0	0
300 dpi sample (10 pages)	12 / 1,599	12	0	0
Synthetic held-out font	82 / 7,867	28	54	3
Synthetic held-out text	55 / 8,763	14	41	3

D Reproducibility Manifest

- **Code (private repository, not public):** `pandan-factory`, evaluation code at commit `6ce3642`.
  - Scorer: `internal/textmetrics`, `internal/evals`.
  - Quotation layer: `internal/quran`.
  - Synthetic generator: `internal/synth` (v1 at `0262cfa`, v2 at `a492235`).

- **Audit:** evidence files, recomputed results (18,939 rows) and preserved analysis scripts with SHA-256 manifests of 2,343 input files are on branch `research/arabic-ocr-paper`, commits `b348698` and `064e856`, under `paper/evidence/`.
- **Figures and tables:** `scripts/extract_figure_data.py` copies the plotted values from `paper/evidence/results_recomputed.csv` into `figure_data/`; `scripts/make_tables.py` writes Tables 3, 4 and 7; `scripts/make_crops.sh` cuts the crops in Figure 5.
- **Models:** base `dots-studio/dots.mocr` (revision not pinned; `e539fbb` inferred). v0.1: `pandanlabs/dots-mocr-arabic-classical`, revision `037b472`.
- **Inference:** vLLM 0.30, greedy decoding, 4,096-token limit, images at 2,048 px long side.
- **Missing:** training and synthesis command lines, seeds, raw synthetic renders (settings reconstructed by replay), byte-level hashes of the released weights.