

When OCR Changes the Script, Does It Preserve the Number?

A Controlled Evaluation of Numeral Recognition

Syamjith NK

Independent Researcher

Abu Dhabi, United Arab Emirates

syamjithnk.com

Preprint — not peer reviewed

Abstract

We audit five released open-weight OCR-capable models on 4,000 core and 3,000 robustness trials of synthetic numeral strings, with typeface-matched Western controls. Value accuracy and script fidelity are distinct outcomes. A Unicode scoring correction changes 22 trial categories; the neutral and preservation prompts achieve joint accuracy on 41/1,500 and 33/1,500 non-Western core trials. The Arabic-Indic H1 comparison contains 344 neutral outputs classified as preserved or all-Western. Its observed-cluster standardized difference is -17.3 percentage points (converted minus preserved). The H2 adjusted joint contrast is -0.53 points. All H1 95% sensitivity intervals exclude zero, but observed-target coverage in a limited diagnostic was 55-65% for percentile and 85-88% for basic intervals at nominal 95%. We therefore withhold a calibrated H1 significance claim and formal H2 equivalence; the percentile-based claim beyond -10 points is not reliable enough to assert. Arabic-Indic neutral joint scores range from 0 to 30 out of 100; paired tests distinguish the handwritten model from each other system after Holm adjustment. These clean synthetic probes are not production failure rates.

1 Motivation and scope

Normalizing digit characters before string comparison can hide changes in numeral script. Character/edit-distance metrics do not inherently require such normalization; the evaluation preprocessing must be specified. We report numeric value, script fidelity and joint success separately, including outputs that emit no decimal digit. We do not claim that output conversion causes value errors, that fine-tuning causes failures, or that these results rank frontier systems or production OCR pipelines. ArabicNum-Bench studies Arabic number reading (Alhumud

et al., 2026). Normalization-sensitive ASR evaluation (Manohar and Pillai, 2024), script-normalized scoring (Pattnayak, 2026), and script-collapse measurement (Rahman, 2026) motivate separating value from script. Related visual benchmarks examine financial fact accuracy (He et al., 2025), Unicode-script coverage (Kargaran et al., 2026), and transcription faithfulness (Lee et al., 2026). Multi-font Arabic word recognition provides related rendering context (Tabzaoui and Chakir, 2026). These tasks differ from the present isolated-numeral probe; this work contributes a controlled empirical comparison rather than a new metric.

2 Data and scoring

The supplied protocol bears a local freeze record dated 8 October 2026. Its original bytes and digest are preserved, but the archive alone is not an independent timestamp attestation. The 100 seeded values reproduce byte-for-byte, are disjoint from pilot values, and contain 33 short, 33 medium and 34 long strings. Each core value is rendered in Western, Arabic-Indic, Persian and Devanagari digits at one typeface/size under neutral and preservation prompts. Five systems were run, rather than the planned six; the frontier API was withheld according to the historical record. No new trials were collected. All core Western controls pass. All GeezaPro robustness cells fail, excluding 1,000 trials. The robustness subset contains only 25 short values.

Value accuracy extracts all Unicode decimal digits, normalizes their decimal values and compares the resulting string to the intended value. Script fidelity requires every emitted decimal digit to belong to the input block, with digit-emitting trials as denominator. Joint accuracy requires both properties and uses all trials as denominator. Any decimal digit outside the input and ASCII Western blocks triggers third-script substitution. Full-

width, Odia, Thai and other decimal digits cannot be silently ignored. The original outputs remain unchanged; rescored fields and a complete old/new change log are derived. For non-Western inputs, the mutually exclusive categories are A (all emitted decimal digits in the input block), B (all ASCII Western), C (a mixture of input-block and ASCII Western digits only), D (at least one decimal digit outside both blocks), and E (no decimal digit). Category A describes script preservation, irrespective of numeric correctness.

The evaluated models are Qari-OCR-v0.3-VL-2B-Instruct, Qwen2-VL-2B-Instruct, Qwen2.5-VL-3B-Instruct, arabic-legal-documents-ocr-1.0, and Arabic-English-handwritten-OCR-v3. These identifiers define the order used for the model counts below. A Western control cell must reach at least 90% value accuracy to pass.

3 Analysis and amendments

H1 uses Arabic-Indic, neutral-prompt outputs classified as preserved or all-Western; three-script pooling is explicitly exploratory. H2 joint and value contrasts use all retained non-Western core trials. The script-fidelity contrast conditions on emitting digits and is subject to prompt-dependent selection. Mixed Bernoulli-logit models include random intercepts for test value and model. Variational-Bayes fitting, priors, deterministic starts and optimizer diagnostics are documented in METHOD_DECISIONS.txt. The fit uses statsmodels 0.15.0 mean-field variational Bayes, fixed-effect prior SD 2, and log random-effect SD prior SD 1. Starting means are zero and posterior SDs are 0.6; BFGS uses a 1,000-iteration limit and gradient tolerance 1e-5. Odds-ratio credible intervals are not called confidence intervals.

The frozen text did not resolve conditioning on realized random effects versus integrating them when combining observed-population averaging with cluster resampling. We therefore expose two distinct targets: (i) average predictions using fitted observed random effects, with conditional outcome simulation; (ii) a new population-integrated Gaussian-random-effect target, with cluster effects redrawn during simulation and the same integrated statistic recomputed after refitting. Neither target is silently substituted for the other. Both basic and percentile intervals from 2,000 draws are available as sensitivity analyses. All methodological changes are post-analysis amendments.

4 Corrected results

Core non-Western categories A/B/C/D/E are 112/749/2/213/424 for the neutral prompt and 93/898/4/200/305 for preservation. Joint accuracy is 2.73% versus 2.20%. Conditional script fidelity is 112/1,076 (10.41%) versus 93/1,195 (7.78%). H1 contains 71 preserved and 273 all-Western outputs across 94 values. Its observed-cluster point estimate is -17.25 pp; its 95% basic interval is [-22.0, -1.9] and its 95% percentile interval is [-32.6, -12.5]. The respective 90% intervals are [-21.0, -3.3] and [-31.2, -13.5]. These are sensitivity calculations, not a claim that nominal coverage has been established. Only the percentile 90% interval has its upper bound below -10 pp; the basic interval crosses that threshold. The percentile procedure has substantially poorer diagnostic coverage (Section 5), so this threshold claim is not supported robustly.

For the population-integrated sensitivity target, H1 is -17.33 pp. Its 95% basic and percentile intervals are [-29.8, -4.3] and [-30.4, -4.9]; its 90% intervals are [-28.1, -6.5] and [-28.2, -6.6]. Both 90% intervals cross -10 pp, so neither satisfies the directional one-sided 95% threshold rule. All four H1 95% intervals exclude zero, supporting a consistent negative direction across these sensitivity calculations. We withhold a calibrated significance claim because even observed-target basic intervals undercovered and the limited population diagnostic does not validate that alternative target and procedure. Agreement in interval sign alone does not resolve coverage.

The H2 observed-cluster adjusted joint contrast is -0.53 pp; basic and percentile 95% intervals are [-1.4, +0.3] and [-1.4, +0.3]. For the population-integrated target, H2 joint is -1.33 pp, with 95% basic [-3.0, +1.7] and percentile [-4.4, +0.3] intervals. This magnitude is about 2.5 times the observed-target estimate, illustrating target sensitivity; neither 95% interval excludes zero. The previously asserted formal equivalence conclusion is withheld because the uncertainty procedure lacks adequate calibration; this is not evidence of harm.

Secondary observed-target H2 script fidelity also depends on interval choice: the 95% percentile interval [-4.2, -0.1] excludes zero, while the 95% basic interval [-4.0, +0.1] includes zero. The raw rates remain descriptive; no confirmed secondary significance claim is made, particularly given the output-selected denominator.

Arabic-Indic neutral joint counts by model are 0, 0, 1, 6 and 30 of 100; corresponding value counts are 2, 4, 7, 7 and 51. Exact paired McNemar tests with Holm adjustment over ten pairs distinguish the handwritten model from each other model; no other pair meets the adjusted 0.05 threshold. The test choice and reduced family size are explicit post-registration specifications. Full per-stratum, per-model, robustness and exploratory empty-output results are generated by the release workflow and not described as unplanned confirmatory discoveries.

5 Calibration and interval sensitivity

The observed-target H1 bootstrap mean is -22.03 pp, against a point estimate of -17.25 pp: a displacement of -4.78 pp. This bootstrap bias estimate helps explain why percentile intervals lie farther in the negative direction and basic intervals, which reflect the bootstrap quantiles about the estimate, are shifted toward zero. It does not establish the true bias or prove that shrinkage alone caused it.

A known-truth diagnostic at one fitted H1 setting used 40 outer datasets per target and 99 inner bootstrap draws. Observed-target nominal 95% coverage was 22/40 (55%) for percentile and 34/40 (85%) for basic intervals. Reusing the same 40 outer datasets with 399 inner draws gave 26/40 (65%) and 35/40 (87.5%), respectively. These are not two independent calibration samples. The 95% Monte Carlo Wilson ranges were 39.8-69.3% and 70.9-92.9% initially, and 49.5-77.9% and 73.9-94.5% in the refinement. Basic performs better in this diagnostic but still falls below nominal coverage. The evidence weighs against relying on the stronger percentile-based threshold claim; it does not certify basic intervals as a replacement. Coverage at 2,000 inner draws was not measured by this diagnostic.

The population-target initial diagnostic covered 33/40 (82.5%) for basic and 38/40 (95%) for percentile 95% intervals, with Monte Carlo Wilson ranges 68.1-91.3% and 83.5-98.6%. This one fitted setting does not establish coverage near zero or the -10 pp threshold, under misspecification, or under alternative variance/prior settings. A more favorable result for a different target does not justify changing the scientific question after observing the results.

6 Reproducibility

The supplement contains unchanged raw records and frozen protocol, corrected derived scores, a complete change log, active code, dependency lock, bootstrap draw diagnostics, paired tests, calibration evidence, and a SHA-256 manifest. Historical code and pilot records are retained separately in the author audit archive; the reviewer candidate includes readable correction narratives. Local hashes establish content integrity, not independent chronology. All post-analysis amendments are labelled.

7 Conclusion

Value accuracy and numeral-script fidelity answer different questions. The corrected synthetic probe shows low joint success and a consistently negative adjusted H1 direction across the reported sensitivity calculations. Poor observed-target coverage prevents a calibrated significance or 10-point threshold claim. The H2 equivalence claim is also withheld. The contribution is the controlled probe, auditable scoring corrections and explicit sensitivity evidence; it does not establish production performance or a causal effect of script conversion.

Limitations

H1 is observational and selects on output script; sparse within-model overlap limits what its adjusted contrast can support. Five model levels weakly identify the random-model variance. The raw/adjusted gap is not proof of a single confounding mechanism.

The old ad hoc ICC and universal design effect are withdrawn. Latent-scale correlations for different shared-cluster cases are reported separately in the supplement.

Clean isolated renders are not production documents. Short-string robustness cannot establish generalization to longer strings, handwriting, surrounding text, degraded scans or full document layouts. Target-script glyph support and exact historical rendering cannot be reconstructed solely from a passing Western control. Original model revisions, font hashes and rendered stimuli were not supplied; exact original evaluation reproduction is qualified.

Expected input script is required for a runtime Unicode mismatch check; the image itself has no Unicode encoding. Metadata, corpus assumptions or a validated script-identification step may supply

that expectation. Such a check cannot establish that the numeric value is correct.

Ethical considerations

The evaluation uses synthetic numeral strings and archived outputs from released models. The supplied study record does not describe recruited participants or participant-derived test inputs. The intended use is research on numeral transcription, rather than deployment certification. Misreading these findings as production error rates or treating script-preserving output as proof of numeric correctness could encourage unreliable downstream use. The five selected models do not represent all systems or all writing systems. Code and data licences are supplied separately; third-party model and font rights are not transferred by those licences. No model weights or proprietary font files are redistributed. Missing historical provenance and the limits of the inferential procedures are disclosed.

Acknowledgements: Human contribution and AI assistance

Syamjith NK led and directed this research, using author-operated AI agents, Claude, and ChatGPT (identified by the author as “Astra”) as assistive tools. AI assistance supported code, analysis and manuscript drafting. In the revision workflow, Codex assisted with inspecting archived materials, rescoring and analysis code, executing checks, bibliography verification, text revision and package preparation. Exact model versions and the allocation of original-study tasks among the reported systems have not yet been specified. Automated checks establish neither nominal interval coverage nor human author approval.

References

- Anas Alhumud, Abdulaziz Alhammadi, and Muhammad Badruddin Khan. 2026. [ArabicNumBench: Evaluating Arabic Number Reading in Large Language Models](#). arXiv:2602.18776.
- Yueru He, Xueqing Peng, Yupeng Cao, Yan Wang, Lingfei Qian, Haohang Li, Yi Han, Shuyao Wang, Ruoyu Xiang, Fan Zhang, Zhuohan Xie, Mingquan Lin, Prayag Tiwari, Jimin Huang, Guojun Xiong, and Sophia Ananiadou. 2025. [FinCriticalED: A Visual Benchmark for Financial Fact-Level OCR](#). arXiv:2511.14998.
- Amir Hossein Kargaran, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. 2026. [Glo-tOCR Bench: OCR Models Still Struggle Beyond a Handful of Unicode Scripts](#). arXiv:2604.12978.
- Gwang Gook Lee, Kenan Emir Ak, Jay Mohta, Yan Xu, and Dimitrios Dimitriadis. 2026. [Do VLMs Read or Rewrite? On Transcription Faithfulness in Vision-Language Models](#). arXiv:2607.21617.
- Kavya Manohar and Leena G Pillai. 2024. [What is lost in Normalization? Exploring Pitfalls in Multilingual ASR Model Evaluations](#). Proceedings of EMNLP 2024, pages 10864-10869. DOI: 10.18653/v1/2024.emnlp-main.607.
- Priyaranjan Pattnayak. 2026. [SN-WER: Script-Normalized WER for Multi-Script Indic ASR Evaluation](#). arXiv:2606.02548.
- Hanif Rahman. 2026. [Script collapse in multilingual ASR: A reference-free metric and 100-pair benchmark](#). arXiv:2604.08786.
- Abderrahime Tabzaoui and Loqman Chakir. 2026. [A Statistically Validated and Decoding-Aware CNN-Transformer-CTC Framework for Multi-Font Printed Arabic Word Recognition](#). Applied Sciences, 16(9):4071. DOI: 10.3390/app16094071.