GlotOCR Bench: OCR Models Still Struggle Beyond a Handful of Unicode Scripts

Amir Hossein Kargaran [1 2] **Nafiseh Nikeghbal** [3 2] **Jana Diesner** [3 2] **François Yvon** [4] **Hinrich Schütze** [1 2]

Abstract

OCR has improved quickly with vision-language models, but evaluation still focuses on a small set of high- and mid-resource scripts. We introduce GlotOCR Bench, a benchmark for OCR generalization across 100+ Unicode scripts, using clean and degraded images rendered from real multilingual text with Google Fonts, HarfBuzz, and FreeType. Evaluating both open and proprietary models, we find that most work well on fewer than ten scripts, and even the best models generalize to fewer than thirty. OCR performance relies heavily on script coverage in pretraining and visual recognition, with unfamiliar scripts often yielding noise or hallucinated lookalikes. We release the benchmark and rendering pipeline for reproducibility.

🤗 hf.co/datasets/cis-lmu/glotocr-bench
github.com/cisnlp/glotocr-bench

1. Introduction

Optical character recognition (OCR) is one of the oldest problems in pattern recognition, but its evaluation has narrowed over time. Leading benchmarks, including OCRBench (Liu et al., 2024b), OCRBench v2 (Fu et al., 2025), CC-OCR (Yang et al., 2025), and OmniDocBench (Ouyang et al., 2025), focus mainly on Latin, CJK, and a few mid-resource scripts. Even multilingual OCR work such as (Li et al., 2025) emphasizes *languages* rather than *scripts*, leaving underlying script diversity limited, while minority-script efforts (Liu et al., 2026) still cover only a small number of writing systems. No existing benchmark evaluates OCR across the full Unicode script space.

[1]Ludwig Maximilian University of Munich, Munich, Germany [2]Munich Center for Machine Learning, Munich, Germany [3]Technical University of Munich, Munich, Germany [4]Sorbonne Université & CNRS, ISIR, Paris, France. Correspondence to: Amir Hossein Kargaran <amir@cis.lmu.de>.

Published at ICML 2026 GlobalSouthML.

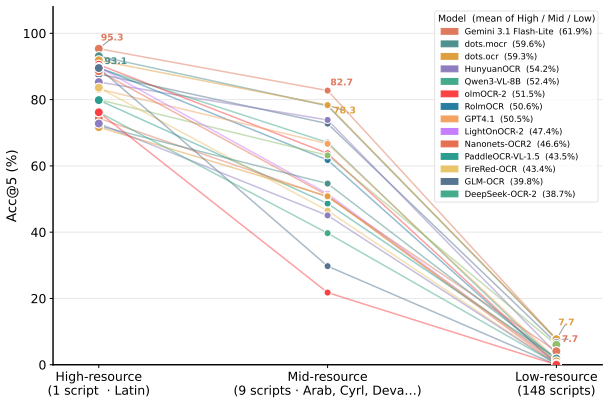


Figure 1. Acc@5 (see §4) by script resource tier (high, mid, low). Performance drops sharply on low-resource scripts. Gemini 3.1 leads (95.3%, 82.7%) but falls to 7.7%; others score <1%.

This gap matters because Unicode 17.0 encodes 172 scripts spanning thousands of years of writing. Many remain in active use, while others are crucial for historical linguistics, archaeology, and cultural preservation (Diao et al., 2025). OCR is essential for digitizing documents in these scripts, yet we still lack a systematic picture of where current models succeed or fail. It is also key to unlocking scanned books and documents as training data for low-resource and historical languages, an opportunity recent work has begun to explore (Kydlíček et al., 2025).

We address this gap with GlotOCR Bench, a benchmark covering 158 Unicode scripts with clean and degraded image variants, carefully curated text, and controlled rendering. Evaluating a broad set of open-weight and API-based vision-language models, we find a simple but striking pattern, summarized in Figure 1.

All models, from frontier APIs to specialized open OCR systems, perform well on Latin and drop substantially on mid-resource scripts such as Arabic, Cyrillic, and Devanagari. On the remaining 148 scripts—94% of our benchmark—every model scores below 10% Acc@5 (see §4); the best result is 7.7%, achieved by Gemini 3.1 Flash-Lite, dots.ocr, and dots.mocr, while most models remain below 1%. These scripts are not marginal cases: they include widely used scripts such as Ethiopic, Khmer, and Sinhala; major literary scripts such as Armenian, Tibetan, and Myanmar; and

scripts vital for cultural preservation, including Linear B, N'Ko, Vai, and many historical systems (Daniels & Bright, 1996). Failure is also not silent: instead of refusing, models often generate fluent-looking text in a familiar script, such as Devanagari for Gujarati or Arabic for Thaana. This suggests that OCR performance depends heavily on script coverage in pretraining as well as visual recognition. We make the following contributions:

- GlotOCR Bench, a **benchmark** covering 158 Unicode scripts with clean and degraded image variants from real multilingual texts, rendered with script-aware font selection and proper bidirectional handling.
- A **comprehensive evaluation** of open-weight and frontier OCR models, reporting character error rate and acceptance rate (Acc@0 and Acc@5) by script and resource level.
- **Core findings**: (a) OCR generalization remains limited to a small set of scripts; (b) performance broadly follows script-level pretraining coverage, with a sharp drop from mid- to low-resource scripts; and (c) models faced with unfamiliar scripts hallucinate known ones rather than failing silently.
- A **public release** of the benchmark dataset, rendering pipeline, evaluation code, and per-model results.

2. Related Work

OCR benchmarks. OCR benchmarks have historically concentrated on a small set of high- and mid-resource scripts. Recent multimodal benchmarks have expanded the range of evaluated tasks but remain narrow in script coverage. OCRBench (Liu et al., 2024b) and its successor OCRBench v2 (Fu et al., 2025) evaluate large multimodal models across document parsing, key information extraction, and multilingual recognition, but coverage remains largely limited to Latin and Chinese scripts. CC-OCR (Yang et al., 2025) covers eleven languages including Arabic, Japanese, Korean, and Vietnamese. Reasoning-OCR (He et al., 2025) probes logical reasoning from OCR cues. OmniDocBench (Ouyang et al., 2025) benchmarks end-to-end document parsing in English and Chinese. olmOCR-Bench (Poznanski et al., 2025a) focuses on English-language PDF parsing. KITAB-Bench (Heakl et al., 2025) targets Arabic OCR and document understanding specifically. OCRTurk (Yılmaz et al., 2026) addresses Turkish specifically. Sohail et al. (2024) benchmark LLM-based OCR for low-resource scripts including Urdu and Tajik, finding that performance degrades with text length. Dasanaike (2026) introduce socOCRbench, a private benchmark targeting social science documents across multiple world regions and scripts; the script coverage is broader than most benchmarks and their model rankings are broadly consistent with ours, but the dataset is not publicly available. These last two works are the closest in spirit to ours, yet none of the benchmarks discussed above treats script coverage as the primary axis of evaluation.

OCR datasets and low-resource adaptation. Agarwal & Anastasopoulos (2024) survey OCR techniques for low-resource languages with a focus on indigenous languages of the Americas, identifying data scarcity and script support as key open challenges. For Indic scripts, Saini et al. (2022) introduce a synthetic dataset across 23 Indic languages, and Kolavi et al. (2025) propose a LoRA-based adaptation framework for ten Indic languages using synthetic data. Sarkar et al. (2024) address printed OCR for extremely low-resource Indic languages, introducing synthetic and real word-level datasets for nine Indian languages. CAMIO (Arrigo et al., 2022) covers 35 languages across 24 scripts as a data resource with transcriptions for only 13 languages, and is available only through the LDC catalog at cost. OmniOCR (Liu et al., 2026) introduces dynamic LoRA adaptation for minority scripts, applied to only four writing systems: Tibetan, Shui, Ancient Yi, and Dongba, though limited to single-character classification datasets (Yuan et al., 2018; Liu et al., 2024a; Luo et al., 2023). dots.ocr (Li et al., 2025) and its XDocParse benchmark span 126 languages, yet script diversity is not the primary axis. Historical document OCR has received attention from Greif et al. (2025), but only for Latin-script historical documents. These efforts demonstrate growing awareness of the problem but address individual script families rather than Unicode coverage as a whole. GlotOCR Bench is the first benchmark to evaluate OCR generalization across most of the Unicode script inventory.

3. GlotOCR Bench

GlotOCR Bench covers 158 Unicode scripts, with two image variants per sentence: a clean rendering on a white background and a degraded rendering simulating aged documents. As in prior OCR dataset construction work (Yim et al., 2021; Malik et al., 2026), we render real multilingual text into images for transcription, extending this approach to a much broader set of scripts; unlike Malik et al. (2026), who generate word-level synthetic OCR data for Kashmiri, we use sentence-level text where available and fall back to words for low-resource scripts.

We sample up to 100 sentences per script, except for Latin with 4,000 sentences and a small set of mid-resource scripts with 400 sentences each for per-language analysis, yielding 16,375 sentences in total. For 68 scripts, fewer than 100 examples are available; although this limits statistical strength, nearly all models fail even at script identification (ScriptAcc; Table 4), suggesting genuine weakness on low-resource scripts rather than evaluation artifact. We cap evaluation at 100 sentences per script to balance cost and

Table 1. Font availability per resource tier.

Tier	Scripts	Total	Median	Min	Max
High	1	1907	1907	1907	1907
Mid	9	814	59	29	323
Low	148	380	1	1	29
All	158	3101	1	1	1907

coverage. Scripts are grouped by web prevalence into three resource tiers:[1] High (Latin), Mid (Arabic, Cyrillic, Devanagari, Han, Japanese, Hangul, Greek, Hebrew, and Thai), and Low (the remaining 148 scripts).

3.1. Text

Text data was compiled from multiple sources to maximize Unicode script coverage. Our primary source is GlotLID v3 (Kargaran et al., 2023; 2024a), which covers over 2,102 language–script pairs; from it, we prioritize publicly shareable sentences of 30–100 characters. For scripts with limited GlotLID coverage, we add data from Wiktionary (Wiktionary Contributors, 2026), WikiSource (Wikimedia Foundation, 2026), Omniglot (Ager, 2026), Google Fonts language data (Google Fonts Team, 2026), and script-converted texts for scripts with little native digital data (Rajan, 2024; Karamolegkou et al., 2025). We also use Common Crawl–derived data from GlotCC (Kargaran et al., 2024a) and FineWeb2 (Penedo et al., 2025) for `und_*` script labels, filtering out mismatched scripts and mapping entries to primary languages where possible, following work such as Sefat et al. (2026). For all newly collected sources, we verify each sentence's Unicode script with GlotScript (Kargaran et al., 2024b). We exclude randomly generated character sequences, since they may help augmentation but are not linguistically valid for evaluation.

3.2. Image

Font. All fonts were sourced from Google Fonts Files (Google Fonts Team, 2026) under the SIL Open Font License v1.1.[2] We grouped fonts by supported Unicode scripts using repository metadata. For each sentence, we selected a font by filtering for script match, full codepoint coverage, and successful glyph rendering, then choosing randomly from the remaining candidates. These checks were necessary because metadata alone did not always guarantee correct rendering. We also manually inspected ten rendered images per script at different sizes, verifying common scripts with external editors and rare scripts against Unicode character charts.

Rendering. Images are rendered using HarfBuzz (Esfahbod et al., 2026) for text shaping and FreeType (Turner et al., 2024) for glyph rasterization. Sentences with mixed bidirectional content are excluded; all rendered text is uniformly LTR or RTL. For each sentence we produce two image variants. The *clean* variant renders text on a plain white canvas with slight random rotation. The *degraded* variant applies a pipeline of augmentations simulating an aged physical document: textured paper backgrounds, ink spread and wear, geometric distortions, resolution downsampling, and JPEG compression artifacts. These operations reflect common artifacts in document capture and scanning pipelines (Groleau et al., 2023) and standard augmentation practices in OCR literature (Gupta et al., 2016; Yim et al., 2021). Full rendering parameters are given in Appendix A. Appendix Figure 2 displays representative examples for Greek and Aghwan scripts. Note that vertically written scripts such as Mongolian are treated as horizontal text, as vertical rendering is not supported by our pipeline.

4. Evaluation Setup

Evaluation pipeline. All models are evaluated in zero-shot mode using the `uv-scripts/ocr` inference suite (van Strien, 2026). Where a chat template is available it is applied; otherwise the prompt is passed directly. The prompt simply asks the model to transcribe the text in the image, return it wrapped in tags, and provide no commentary or explanation. Images are provided at their native rendered resolution without further preprocessing.

Models. We evaluate the following open-weight OCR models: dots.ocr (Li et al., 2025), dots.mocr (dots.ocr-1.5) (Zheng et al., 2026), olmOCR-2 (Poznanski et al., 2025b), RolmOCR (Reducto AI, 2025), LightOnOCR-2 (Taghadouini et al., 2026), Nanonets-OCR2 (Mandal et al., 2025), PaddleOCR-VL-1.5 (Cui et al., 2026), FireRed-OCR (Wu et al., 2026), GLM-OCR (Duan et al., 2026), DeepSeek-OCR-2 (Wei et al., 2026), HunyuanOCR (Hunyuan Vision Team et al., 2025), and Qwen3-VL-8B (Bai et al., 2025). We additionally evaluate two proprietary models via their respective APIs: Gemini 3.1 Flash-Lite (Google DeepMind, 2026; Comanici et al., 2025) and GPT-4.1 (OpenAI et al., 2024).

Metrics. We use three metrics throughout the paper. Our primary metric is CER, defined as the normalized Levenshtein edit distance at the character level with whitespace ignored: CER $= \min\left(1, \frac{S+D+I}{N}\right)$, where S, D, and I are respectively the substitution, deletion, and insertion counts, and N is the ground-truth length. To account for minor expected variations in model output, we report the best CER across four configurations: the original output, the reversed output string, the lowercased output, and the version with Unicode combining marks removed. We additionally report Acc@k (A@k for short): the fraction of sentences for which CER $\leq k/100$, with $k \in \{0, 5\}$. Acc@5 is our primary accuracy metric, measuring near-perfect transcription

[1] `wikipedia:Languages_used_on_the_Internet`
[2] `github.com/google/fonts/tree/main/ofl`

Table 2. GlotOCR Bench benchmark results by resource tier. Each tier result is the macro average over its scripts. A@0 and A@5 denote the fraction of predictions with CER ≤ 0 and ≤ 0.05, respectively. Bold = best; underline = 2nd best

Model	**High** (1 script)			**Mid** (9 scripts)			**Low** (148 scripts)			**Mean** (across tiers)		
	CER ↓	A@0 ↑	A@5 ↑	CER ↓	A@0 ↑	A@5 ↑	CER ↓	A@0 ↑	A@5 ↑	CER ↓	A@0 ↑	A@5 ↑
Gemini 3.1 Flash-Lite	**0.9**	**86.0**	**95.3**	**3.0**	**66.1**	**82.7**	**79.0**	5.0	**7.7**	**27.7**	**52.4**	**61.9**
dots.mocr	1.5	82.5	93.1	6.0	57.0	78.1	84.1	5.1	**7.7**	30.5	48.2	59.6
dots.ocr	1.6	80.4	91.8	5.0	55.4	78.3	82.6	**5.2**	**7.7**	29.7	47.0	59.3
HunyuanOCR	3.4	56.0	85.3	6.3	52.5	73.9	87.3	1.7	3.4	32.3	36.8	54.2
Qwen3-VL-8B	2.0	72.8	89.5	7.2	47.6	67.1	89.8	0.3	0.8	33.0	40.3	52.4
olmOCR-2	2.0	75.0	90.5	8.3	45.3	63.8	90.2	0.2	0.3	33.5	40.2	51.5
RolmOCR	2.0	72.7	89.6	10.0	44.6	61.8	92.0	0.1	0.2	34.7	39.1	50.6
GPT4.1	2.7	58.7	83.2	6.3	45.6	66.7	85.9	0.6	1.6	31.7	35.0	50.5
LightOnOCR-2	2.2	75.6	89.8	13.0	28.4	51.6	91.6	0.2	0.7	35.6	34.7	47.4
Nanonets-OCR2	2.3	70.7	88.6	12.2	34.1	51.1	91.9	0.1	0.2	35.5	35.0	46.6
PaddleOCR-VL-1.5	4.6	57.0	79.8	26.9	33.9	48.6	91.8	1.5	2.0	41.1	30.8	43.5
FireRed-OCR	3.4	59.2	83.6	19.3	27.9	46.5	91.9	0.1	0.2	38.2	29.0	43.4
GLM-OCR	2.1	70.9	89.5	31.1	17.8	29.8	91.2	0.0	0.0	41.5	29.5	39.8
DeepSeek-OCR-2	5.5	50.1	76.2	24.4	22.2	39.7	92.0	0.1	0.3	40.6	24.1	38.7

and mapping naturally onto the binary question of whether a model can operate in a given script. Finally, Script Accuracy (ScriptAcc) measures whether the model responds in the correct script regardless of transcription accuracy, as determined by GlotScript (Kargaran et al., 2024b). Throughout this paper, scripts are identified by their ISO 15924 four-letter codes (e.g., `Arab` for Arabic). This metric disentangles script identification from transcription quality and serves as a diagnostic for cross-script hallucination.

5. Results

Table 2 presents results for fourteen OCR systems across three resource tiers: High (Latin), Mid (9 scripts), and Low (148 scripts), along with the overall mean.

> **Finding 1.** OCR performance is largely solved for high-resource Latin script but degrades severely in mid- and low-resource tiers, where even the best models fail on over 92% of low-resource sentences. The transition from mid- to low-resource is not a smooth degradation but a sharp discontinuity, suggesting a threshold phenomenon: models either have sufficient training exposure to a script or they do not. The gap between top proprietary and best open-weight models is small — thanks to dots.ocr and dots.mocr — but most open-weight approaches still lag substantially.

5.1. Appendix Results

The appendix provides a more detailed view of the same pattern shown in Table 2. Appendix G confirms that strong OCR performance is confined to a small set of high- and mid-resource scripts, with performance collapsing on most low-resource scripts. Appendix B shows that this pattern also holds within scripts: even where aggregate scores are strong, performance can vary substantially across languages, with Arabic showing the largest within-script degradation.

Appendix C shows that script recognition is only a partial predictor of OCR accuracy. Some scripts, such as Arabic, achieve relatively high ScriptAcc but remain difficult to transcribe, while others such as Japanese perform better than expected. Appendix D further shows that script-aware hinting yields only limited overall gains, helping a small number of scripts but leaving most low-resource scripts unsolved.

Appendix E shows that all models worsen on degraded images, with sensitivity increasing outside the high-resource tier. Finally, Appendix F shows that these failures are systematic rather than random: when models fail, they usually hallucinate a small set of familiar higher-resource scripts instead of remaining silent.

6. Conclusion

We introduce GlotOCR Bench, a benchmark for OCR generalization across 158 Unicode scripts using clean and degraded images rendered from real multilingual text. Evaluating 14 open and proprietary vision-language models, we find strong results on Latin, weaker performance on mid-resource scripts, and near-universal failure on the other 148 low-resource scripts; even the best model correctly transcribes fewer than 7.7% of sentences in this tier at under 5% character error rate. These failures are often non-silent: models hallucinate fluent text in familiar scripts, while script-aware hints offer only limited gains. Performance drops sharply rather than gradually beyond mid-resource scripts, pointing to training coverage as the main bottleneck. We release the benchmark, pipeline, and code to support reproducible research and to encourage OCR development beyond the small set of scripts that currently receive most attention.

References

Agarwal, M. and Anastasopoulos, A. A concise survey of OCR for low-resource languages. In Mager, M., Ebrahimi, A., Rijhwani, S., Oncevay, A., Chiruzzo, L., Pugh, R., and von der Wense, K. (eds.), *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pp. 88–102, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.americasnlp-1.10. URL https://aclanthology.org/2024.americasnlp-1.10/.

Ager, S. Omniglot: Writing systems and languages of the world. https://www.omniglot.com, 2026.

Arrigo, M., Strassel, S., King, N., Tran, T., and Mason, L. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1209–1216, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.129/.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., Zhang, Y., Zhang, Y., Liu, Y., Yu, D., and Ma, Y. Paddleocr-vl-1.5: Towards a multi-task 0.9b vlm for robust in-the-wild document parsing, 2026. URL https://arxiv.org/abs/2601.21957.

Daniels, P. T. and Bright, W. *The world's writing systems*. Oxford University Press, 1996.

Dasanaike, N. sococrbench: An OCR benchmark for social science documents. Working paper, 2026. URL https://noahdasanaike.github.io/posts/sococrbench.html.

Diao, X., Bo, R., Xiao, Y., Shi, L., Zhou, Z., Xu, H., Li, C., Tang, X., Poesio, M., John, C. M., and Shi, D. Ancient script image recognition and processing: A review, 2025. URL https://arxiv.org/abs/2506.19208.

Duan, S., Xue, Y., Wang, W., Su, Z., Liu, H., Yang, S., Gan, G., Wang, G., Wang, Z., Yan, S., Jin, D., Zhang, Y., Wen, G., Wang, Y., Zhang, Y., Zhang, X., Hong, W., Cen, Y., Yin, D., Chen, B., Yu, W., Gu, X., and Tang, J. GLM-OCR Technical Report, 2026. URL https://arxiv.org/abs/2603.10910.

Esfahbod, B. et al. HarfBuzz: A text shaping engine, 2026. URL https://github.com/harfbuzz/harfbuzz.

Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., and Bai, X. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=Vb6i3Dp24N.

Google DeepMind. Gemini 3.1 flash-lite: Built for intelligence at scale, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/.

Google Fonts Team. Google fonts. GitHub, 2026. URL https://github.com/google/fonts.

Greif, G., Griesshaber, N., and Greif, R. Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents, 2025. URL https://arxiv.org/abs/2504.00414.

Groleau, A., Chee, K. W., Larson, S., Maini, S., and Boarman, J. Augraphy: A data augmentation library for document images, 2023. URL https://arxiv.org/abs/2208.14558.

Gupta, A., Vedaldi, A., and Zisserman, A. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Gupta_Synthetic_Data_for_CVPR_2016_paper.html.

He, H., Ye, M., Zhang, J., Cai, X., Liu, J., Du, B., and Tao, D. Reasoning-OCR: Can large multimodal models solve complex logical reasoning problems from ocr cues?, 2025. URL https://arxiv.org/abs/2505.12766.

Heakl, A., Sohail, M. A., Ranjan, M., Elbadry, R., Ahmad, G. S., El-Geish, M., Maher, O., Shen, Z., Khan, F. S., and Khan, S. KITAB-bench: A comprehensive multi-domain benchmark for Arabic OCR and document understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22006–22024, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1135. URL https://aclanthology.org/2025.findings-acl.1135/.

Hunyuan Vision Team, Lyu, P., Wan, X., Li, G., Peng, S., Wang, W., Wu, L., Shen, H., Zhou, Y., Tang, C., Yang, Q., Peng, Q., Luo, B., Yang, H., Zhang, X., Zhang, J., Peng, H., Yang, H., Xie, S., Zhou, L., Pei, G., Wu, B., Yan, R., Wu, K., Yang, J., Wang, B., Liu, K., Zhu, J., Jiang, J., Linus, Hu, H., and Zhang, C. HunyuanOCR Technical Report, 2025. URL https://arxiv.org/abs/2511.19575.

Karamolegkou, A., Nikandrou, M., Pantazopoulos, G., Sanchez Villegas, D., Rust, P., Dhar, R., Hershcovich, D., and Søgaard, A. Evaluating multimodal language models as visual assistants for visually impaired users. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25949–25982, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1260. URL https://aclanthology.org/2025.acl-long.1260/.

Kargaran, A. H., Imani, A., Yvon, F., and Schuetze, H. GlotLID: Language identification for low-resource languages. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6155–6218, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.410. URL https://aclanthology.org/2023.findings-emnlp.410/.

Kargaran, A. H., Yvon, F., and Schuetze, H. GlotCC: An open broad-coverage commoncrawl corpus and pipeline for minority languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=aJ1yse8GEr.

Kargaran, A. H., Yvon, F., and Schütze, H. GlotScript: A resource and tool for low resource writing system identification. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7774–7784, Torino, Italia, May 2024b. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.687/.

Kolavi, A., P, S., and Jain, V. Nayana OCR: A scalable framework for document OCR in low-resource languages. In Truong, S., Putri, R. A., Nguyen, D., Wang, A., Ho, D., Oh, A., and Koyejo, S. (eds.), *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pp. 86–103, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-242-8. doi: 10.18653/v1/2025.lm4uc-1.11. URL https://aclanthology.org/2025.lm4uc-1.11/.

Kydlíček, H., Penedo, G., and von Werra, L. FinePDFs. https://huggingface.co/datasets/HuggingFaceFW/finepdfs, 2025.

Li, Y., Yang, G., Liu, H., Wang, B., and Zhang, C. dots.ocr: Multilingual document layout parsing in a single vision-language model, 2025. URL https://arxiv.org/abs/2512.02498.

Liu, B., Zhang, Z., Meng, B., Wang, H., Zhang, H., Wang, C., Ergu, D., and Cai, Y. Omniocr: Generalist ocr for ethnic minority languages, 2026. URL https://arxiv.org/abs/2602.21042.

Liu, X., Han, X., Chen, S., Dai, W., and Ruan, Q. Ancient yi script handwriting sample repository. *Scientific Data*, 11(1):1183, 2024a. URL https://doi.org/10.1038/s41597-024-03918-5.

Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.-C., Liu, C.-L., Jin, L., and Bai, X. OCRBench: On the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b. doi: 10.1007/s11432-024-4235-6. URL https://doi.org/10.1007/s11432-024-4235-6.

Luo, Y., Sun, Y., and Bi, X. Multiple attentional aggregation network for handwritten Dongba character recognition. *Expert Systems with Applications*, 213:118865, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.118865. URL https://www.sciencedirect.com/science/article/pii/S0957417422018838.

Malik, H. N., Shafi, K. M., and Reshi, T. A. synthocr-gen: A synthetic OCR dataset generator for low-resource languages- breaking the data barrier, 2026. URL https://arxiv.org/abs/2601.16113.

Mandal, S., Talewar, A., Thakuria, S., Ahuja, P., and Juvatkar, P. Nanonets-OCR2: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging, 2025. URL https://huggingface.co/nanonets/Nanonets-OCR2-3B.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., Shi, J., Wu, F., Chu, P., Liu, M., Li, Z., Xu, C., Zhang, B., Shi, B., Tu, Z., and He, C. OmniDocBench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848, June 2025.

Penedo, G., Kydlíček, H., Sabolčec, V., Messmer, B., Foroutan, N., Kargaran, A. H., Raffel, C., Jaggi, M., Werra, L. V., and Wolf, T. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=jnRBe6zatP.

Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., and Soldaini, L. olmOCR: Unlocking trillions of tokens in pdfs with vision language models, 2025a. URL https://arxiv.org/abs/2502.18443.

Poznanski, J., Soldaini, L., and Lo, K. olmOCR 2: Unit test rewards for document ocr, 2025b. URL https://arxiv.org/abs/2510.19817.

Rajan, V. Aksharamukha: Script conversion web tool. https://www.aksharamukha.com/converter, 2024.

Reducto AI. Rolmocr: A faster, lighter open-source ocr model, 2025. URL https://reducto.ai/blog.

Saini, N., Pinto, P., Bheemaraj, A., Kumar, D., Daga, D., Yadav, S., and Nagaraj, S. Ocr synthetic benchmark dataset for indic languages, 2022. URL https://arxiv.org/abs/2205.02543.

Sarkar, A., Mondal, A., Lehal, G. S., and Jawahar, C. Printed ocr for extremely low-resource indic languages. In *International Conference on Computer Vision and Image Processing*, pp. 108–122. Springer, 2024. URL https://ilocr.iiit.ac.in/dataset/static/assets/img/publication/printed/printed_ocr.pdf.

Sefat, A. A., Kargaran, A. H., Yvon, F., and Schütze, H. GlotWeb: Web indexing for minority languages. In *Proceedings of the ACM Web Conference 2026*, pp. 8469–8472, 2026. URL https://dl.acm.org/doi/abs/10.1145/3774904.3792887.

Sohail, M. A., Masood, S., and Iqbal, H. Deciphering the underserved: Benchmarking llm ocr for low-resource scripts, 2024. URL https://arxiv.org/abs/2412.16119.

Taghadouini, S., Cavaillès, A., and Aubertin, B. LightOnOCR: A 1b end-to-end multilingual vision-language model for state-of-the-art ocr, 2026. URL https://arxiv.org/abs/2601.14251.

Turner, D., Wilhelm, R., and Lemberg, W. FreeType: A free, high-quality and portable font engine, 2024. URL https://freetype.org.

van Strien, D. Ocr uv scripts. Hugging Face, 2026. URL https://huggingface.co/datasets/uv-scripts/ocr.

Wei, H., Sun, Y., and Li, Y. DeepSeek-OCR 2: Visual causal flow, 2026. URL https://arxiv.org/abs/2601.20552.

Wikimedia Foundation. Wikisource: The free online library. https://wikisource.org, 2026.

Wiktionary Contributors. Wiktionary, the free dictionary, 2026. URL https://www.wiktionary.org/.

Wu, H., Lou, H., Li, X., Zhong, Z., Sun, Z., Chen, P., Zhou, X., Zuo, K., Chen, Y., Tang, X., Hu, Y., Zhou, B., Wu, J., Wu, Y., Yu, W., Liu, Y., Huang, Y., Xu, M., Liu, G., Ma, Y., Sun, Z., and Qiao, C. Firered-ocr technical report, 2026. URL https://arxiv.org/abs/2603.01840.

Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., Jin, L., and Lin, J. CC-OCR: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21744–21754, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_CC-OCR_A_Comprehensive_and_Challenging_OCR_Benchmark_for_Evaluating_Large_ICCV_2025_paper.html.

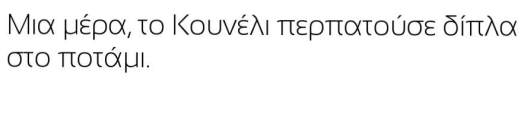
Yılmaz, D., Munis, E. A., Toraman, Ç., Köse, S. K., Aktaş, B., Baytekin, M. C., and Görür, B. K. Ocrturk: A

comprehensive ocr benchmark for turkish, 2026. URL https://arxiv.org/abs/2602.03693.

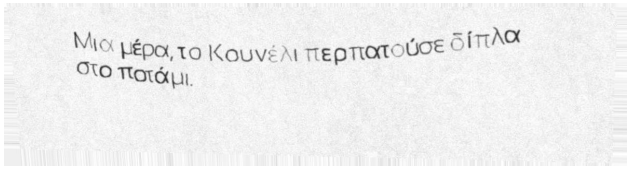
Yim, M., Kim, Y., Cho, H.-C., and Park, S. Synthtiger: Synthetic text image generator towards better text recognition models. In *Document Analysis and Recognition – ICDAR 2021*, pp. 109–124, Cham, 2021. Springer International Publishing.

Yuan, M., Xianmu, C., Tang, J., et al. Tibetanmnist: Tibetan handwritten digit dataset, 2018. URL https://www.heywhale.com/mw/dataset/5bfe734a954d6e0010683839.

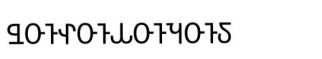
Zheng, H., Li, Y., Zhang, K., Xin, L., Zhao, G., Liu, H., Chen, J., Lou, J., Qiu, J., Fu, Q., Yang, R., Jiang, S., Luo, W., Su, W., Zhang, W., Zhu, X., Li, Y., ma, Y., Chen, Y., Yu, Z., Yang, G., Zhang, C., Zhang, L., Liu, Y., and Bai, X. Multimodal OCR: Parse anything from documents, 2026. URL https://arxiv.org/abs/2603.13032.



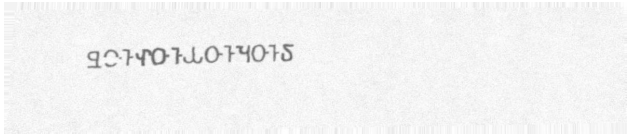
(a) Greek — clean



(b) Greek — degraded



(c) Aghwan — clean



(d) Aghwan — degraded

Figure 2. Example images from GlotOCR Bench for Greek (Grek, Mid tier resource) and Aghwan (Aghb, Low tier resource), each shown in clean and degraded variants.

A. Rendering Pipeline Details

Images are rendered using HarfBuzz (Esfahbod et al., 2026) for text shaping and FreeType (Turner et al., 2024) for glyph rasterization. Sentences with mixed bidirectional content are excluded; all rendered text is uniformly LTR or RTL. For each sentence we produce two image variants as described below and illustrated in Figure 2.

Clean variant. Text is rendered at 48px on a plain white 1000px-wide canvas with 40px padding, with a negligible random rotation of up to $\pm 1^\circ$ to simulate minor page tilt.

Degraded variant. A sequential pipeline simulates an aged physical document:

1. **Paper background and rotation.** The image is placed onto a randomly cropped scanned paper texture, then rotated by up to $\pm 2^\circ$ to simulate page tilt.
2. **Elastic deformation and Gaussian noise.** A smooth displacement field (17×17 Gaussian kernel, ± 8px amplitude) warps the image; independent Gaussian noise ($\sigma = 8$) is then added.
3. **Ink effects.** Between 10 and 30 white rectangular patches ($\leq 40 \times 15$px) simulate ink dropout; pixel intensities are then scaled to 50–85% with texture noise ($\sigma = 10$) to simulate ink fading.
4. **Resolution and compression.** Images are downsampled to 40–70% of original resolution and upscaled back (area/bilinear interpolation), then JPEG-compressed at quality 30–80.
5. **Perspective distortion.** The four corners are independently warped by up to 10% of the image dimensions.

Additionally, at the glyph level during rendering, character spacing is perturbed by -2 to $+4$ pixels, each glyph is independently dilated (prob. 0.4) or eroded (prob. 0.25) with a 2×2 kernel, each line is vertically jittered by up to ± 3 pixels, and glyphs are displaced vertically by a parabolic curvature to simulate page curl.

These degradation operations reflect common artifacts in document capture, scanning, and photocopying pipelines (Groleau et al., 2023), as well as standard augmentation practices in OCR literature (Gupta et al., 2016; Yim et al., 2021): geometric distortions model page misalignment, photometric noise simulates low-quality digitization, and morphological perturbations reflect ink spread and aging. All transformations are applied within the parameter ranges listed above, validated by human inspection across Latin, Greek, Cyrillic, and Arabic scripts, to ensure legibility while still providing sufficient challenge for model robustness evaluation.

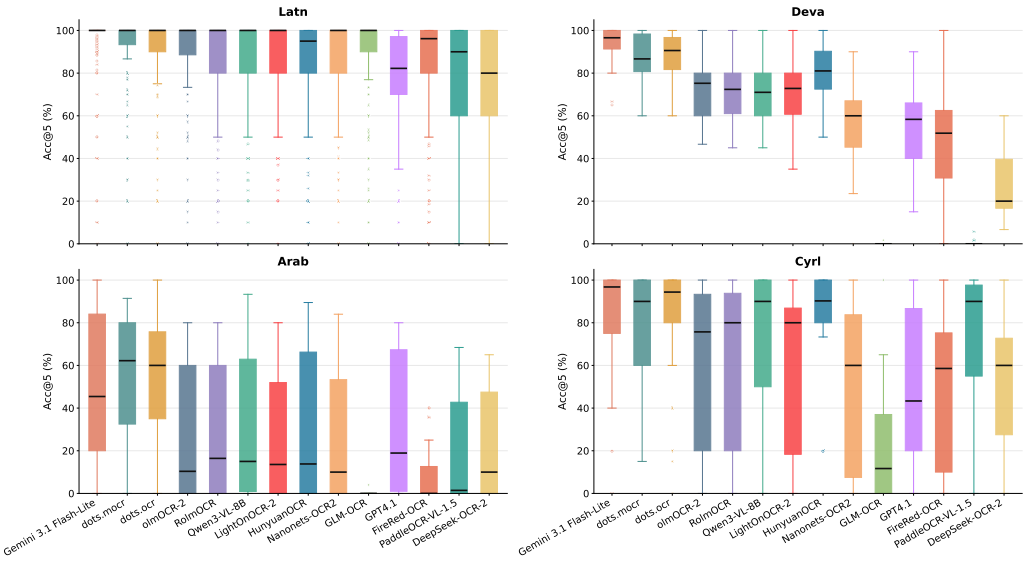


Figure 3. Acc@5 distributions for four scripts (Latin, Devanagari, Arabic, Cyrillic). Boxes correspond to models; each point is the score for one language within the script.

B. Per-Language Variance Within Scripts

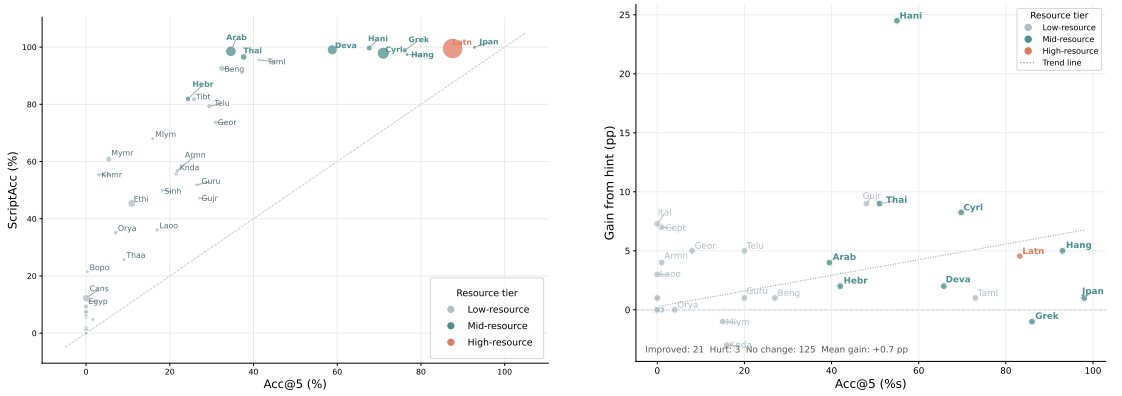
Figure 3 presents the per-language Acc@5 distribution across models for four scripts — Latin, Devanagari, Arabic, and Cyrillic — selected as the four scripts with the most languages in the benchmark. For Latin-script languages, models generally achieve high median performance (typically above 90%), but with notable variability across languages and low-scoring outliers, revealing that strong aggregate performance does not imply uniform coverage across all Latin-script languages.

Performance on non-Latin scripts is generally more variable and degraded. For Devanagari, median accuracies are lower than Latin with moderate spread, though most models maintain moderate performance. Remaining errors are often attributable to conjunct characters — where multiple characters merge into a single glyph — though models generally handle them well as they are a core feature of the script. Arabic shows the most severe degradation: most models exhibit low medians, wide interquartile ranges, and strong downward skew — reflecting Arabic's orthographic complexity, where visually similar characters, optional diacritics, and numerous script variations make generalization across its many languages particularly challenging. Cyrillic presents a comparatively stronger pattern, with several models achieving medians comparable to Latin, though variability remains substantial and some models perform poorly. Across all scripts, models such as Gemini 3.1 Flash-Lite, dots.ocr, and dots.mocr demonstrate tighter distributions and higher medians, indicating more stable cross-language performance — yet even these models exhibit failures on specific languages.

> **Finding 2.** Per-language performance varies substantially within and across scripts. Among the four scripts evaluated, Arabic exhibits the steepest degradation, reflecting its orthographic complexity and the diversity of languages it encodes.

C. Script Accuracy vs. OCR Accuracy

Figure 4a presents the relationship between script recognition accuracy (ScriptAcc) and OCR accuracy (Acc@5), averaged across models. ScriptAcc serves as a prerequisite for Acc@5: models that fail to produce correct script characters cannot achieve high OCR accuracy. This is reflected in the strong diagonal correlation, where resource tier largely determines placement — high- and mid-resource scripts (Latin, Japanese, Greek, Han) cluster in the upper-right, low-resource scripts occupy the middle band and bottom-left. Notable deviations reveal additional factors at play. Arabic achieves high ScriptAcc yet lags in Acc@5, suggesting that script-level confusions are not the bottleneck; rather, intra-script variation and visually similar characters drive OCR errors. Hebrew presents a different failure mode: its ScriptAcc is comparatively low due to frequent confusion with Thai (see Table 5), pulling its OCR performance below scripts like Tamil that suffer less from such



(a) Script-level recognition accuracy (ScriptAcc) vs. OCR accuracy (Acc@5), averaged across all models. Bubble size $\propto$ log number of languages using the script. Resource tier is indicated by color.

(b) Gain in Acc@5 from script-identity hint vs. baseline Acc@5, per script for GPT4.1. Most scripts show modest gains (+0.7 pp on average), with Han (Hani) benefiting most.

Figure 4. Script Recognition and Hint-Guided OCR Analysis.

cross-script confusion. Japanese is a notable positive outlier, achieving higher Acc@5 than even Latin despite combining three writing systems — Hiragana, Katakana, and Kanji. This suggests that OCR models can handle mixed-script sentences well, though we do not investigate code-switching further in this paper.

> **Finding 3.** ScriptAcc is a weak but early indicator of Acc@5. Deviations reveal distinct failure modes: Arabic suffers from intra-script variation despite high ScriptAcc, Hebrew is hurt by cross-script confusion with Thai, and Japanese exceeds Latin in OCR accuracy despite combining three writing systems.

D. Effect of Script-Aware Hinting on OCR Performance

Figure 4b illustrates the relationship between baseline OCR accuracy (Acc@5) and the gain obtained from providing GPT-4.1 with an explicit hint — informing the model of the language, script, and the exact characters present in the image, deduplicated and sorted by Unicode code point. Since this provides an unfair advantage for short texts, we exclude samples with fewer than 10 characters, leaving 149 scripts. Out of all evaluated scripts, 125 show no change, 21 improve, and only 3 are negatively affected, yielding a mean gain of +0.7 percentage points — indicating that script-identity hinting provides selective but limited benefits overall. Several mid-resource scripts show pronounced improvements: Hani exhibits a gain exceeding 20 percentage points, despite GPT-4.1 already having good ScriptAcc for it. The model tends to produce common tokens rather than visually similar rare characters; providing the exact character set corrects these substitutions. This is expected given Han's large character inventory — constraining the candidate set has a greater impact when thousands of characters are possible. Cyrillic and Thai also benefit notably, suggesting that character ambiguity is a non-negligible factor for these scripts. Some low-resource scripts (e.g., Ital, Copt) show modest gains from near-zero baselines, improving to between 5–10%, which is encouraging but still far from usable performance. Overall, hinting is not the primary bottleneck for most scripts — particularly in low-resource tier where the model lacks both the underlying visual recognition capability and sufficient pretraining exposure to those characters entirely.

> **Finding 4.** Script-aware hinting yields only marginal overall gains (+0.7 pp mean), with 125 of 149 scripts showing no change. Benefits are selective: Hani, Cyrillic, and Thai improve notably due to character ambiguity, while low-resource scripts remain largely unsolved — indicating that the bottleneck is insufficient visual recognition and pretraining exposure, not script identity.

E. Robustness to Image Degradation Across Resource Tiers

We evaluate the robustness of the six best-performing models from the previous experiments by comparing their performance on clean images versus degraded "old document" variants across three resource tiers. Figure 5 reports Acc@5 for both conditions, along with the performance drop. Across all models and tiers, performance consistently drops under image

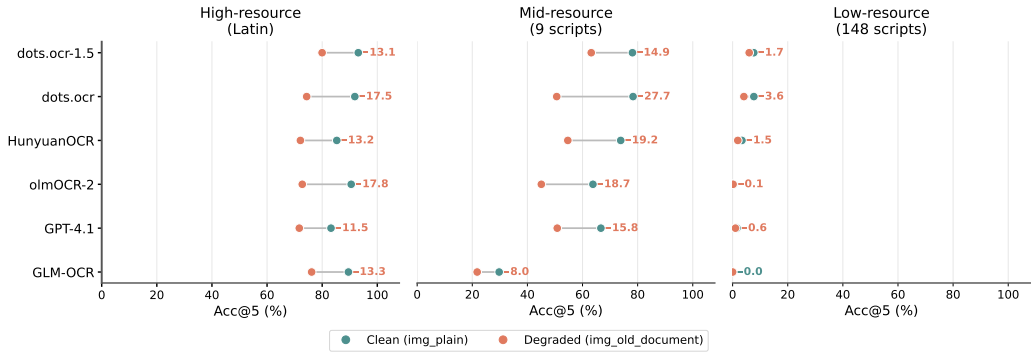


Figure 5. Acc@5 on clean vs. degraded images for the six best-performing models across high-, mid-, and low-resource tiers, with corresponding performance drops.

degradation, though the magnitude of the loss varies substantially by resource level. In the high-resource tier, all models exhibit noticeable performance drops, indicating that even well-represented scripts remain sensitive to image degradation. GPT-4.1 displays relatively better robustness (13.8% relative drop), while olmOCR-2 shows a larger decline (19.7% relative drop). The mid-resource tier reveals greater sensitivity than high-resource scripts — lower clean performance and larger absolute drops — suggesting that models struggle more to recover correct text for lower-resource scripts. This likely reflects reduced familiarity with script-specific visual patterns and weaker generalization under degraded conditions. For low-resource scripts, although absolute degradation is small, relative degradation is greater, given that baseline performance is already near zero.

> **Finding 5.** All models suffer performance loss under image degradation across all resource tiers. Clean images represent an upper bound on OCR performance. Latin scripts show consistent but moderate drops, with GPT-4.1 being the most robust among the evaluated models. Sensitivity increases as resource levels decrease.

F. Cross-Script Hallucination

For each prediction we detect the dominant script of the output and compare it to the expected script. We distinguish three failure modes: *cross-script hallucination*, where the model produces text in a different recognizable script; *silence*, where the model returns an empty or whitespace-only response; and *artifact*, where the output contains characters that GlotScript cannot assign to any real script—typically repetitive digit strings, punctuation loops, or model-specific wrapper tokens left over from structured-output formats.

Table 3 reports the hallucination, silence, and artifact rates for each model, computed as a macro-average over scripts. Across all models, only 12.5% of predictions are on average assigned to the correct script. Cross-script hallucination accounts for 68.4% on average, artifacts for 13.1%, and silence for only 6.0%, showing that models overwhelmingly prefer to confabulate in a wrong script rather than abstain. The artifact rate is highly model-dependent: DeepSeek-OCR2 (26.2%) produce far more artifacts than dots.ocr (3.8%), consistent with models trained to always emit output—even from blank images. dots.ocr, by contrast, mostly chooses to remain silent when it cannot recognise a script (42.1%).

Appendix Tables 4–6 report, for each target script, the two scripts most frequently observed in model outputs. These show that hallucinated outputs are not random: hallucination targets concentrate on a small set of high- and mid-resource attractor scripts, with Latin, Arabic, and Devanagari collectively accounting for the majority of cross-script substitutions, reflecting their dominance in OCR training corpora. Several substitution pairs reflect genuine script-family proximity—Syriac → Arabic, Grantha → Tamil, Coptic → Greek, Newa → Devanagari, Tangut → Han, Lisu → Latin—pairing each lower-resource script with its closest higher-resource relative. Other substitutions are purely distributional: Old Uyghur and Mongolian are most often predicted as Arabic, likely because both are rendered horizontally in our benchmark despite being traditionally vertical scripts, and their horizontal rendering may share superficial visual features with Arabic's connected cursive strokes. Ogham is rendered almost exclusively as Latin. This suggests models conflate visual similarity with statistical co-occurrence in training data, defaulting to whichever script is most compatible with the image features they extract

Table 3. Script-level error rates (%) per model, macro-averaged over scripts, ranked by cross-script hallucination (Hall.) rate. The four categories are mutually exclusive and sum to 100%.

Model	Correct ↑	Hall. ↓	Silent	Artifact
dots.ocr	15.8	38.3	42.1	3.8
dots-mocr	16.1	50.2	16.3	17.4
FireRed-OCR	8.3	62.4	12.7	16.6
DeepSeek-OCR2	6.5	63.7	3.6	26.2
Hunyuan-OCR	11.5	68.2	0.0	20.3
PaddleOCR-VL-1.5	7.6	69.7	0.0	22.7
Qwen3-VL-8B	11.8	70.1	3.6	14.5
Gemini-Flash-Lite	22.6	70.2	0.5	6.7
GPT-4.1	17.6	72.1	0.9	9.4
GLM-OCR	10.8	74.2	3.0	12.0
Nanonets-OCR2	11.1	78.6	0.0	10.3
RolmOCR	13.3	78.9	0.0	7.8
LightOn-OCR-2	8.3	79.2	0.9	11.6
olmOCR-2	14.4	81.7	0.0	3.9
Average	12.5	68.4	6.0	13.1

Finding 6. Cross-script hallucination is the dominant failure mode: models overwhelmingly confabulate in a wrong script rather than abstain. Hallucination targets concentrate on high- and mid-resource attractor scripts, with some substitutions reflecting genuine script-family proximity and visual resemblance, while others are purely distributional, driven by the dominance of certain scripts in OCR training corpora.

G. Per-Script Results

Table 4 shows the scripts for which all models obtain zero in the ScriptAcc metric; for these, we report the sentence count in the benchmark (n) and the two scripts most frequently observed in model outputs (as a diagnostic for hallucination). Tables 5 and 6 report per-script Acc@5 and ScriptAcc for all evaluated models, along with the sentence count and the most frequently observed output scripts.

Table 4. Scripts for which all models obtain zero in both Acc@5 and ScriptAcc. These scripts are not identifiable by any model. Top-2 Out.Script shows which scripts appear most frequently in model outputs for these samples (avg. over all models).

Script	n	**Top-2 Out.Script**	**Script**	n	**Top-2 Out.Script**	**Script**	n	**Top-2 Out.Script**
Lepc	100	Latn, Arab	Limb	100	Mtei, Tibt	Mand	100	Arab, Syrc
Modi	100	Deva, Thai	Mong	100	Hani, Shaw	Newa	100	Deva, Beng
Nkoo	100	Arab, Latn	Ogam	100	Latn, Hani	Olck	100	Latn, Sinh
Hung	100	Latn, Cyrl	Ital	100	Latn, Cyrl	Ougr	100	Arab, Mong
Phli	100	Hebr, Latn	Prti	100	Hebr, Tibt	Sarb	100	Latn, Tibt
Shrd	100	Deva, Guru	Dsrt	100	Shaw, Latn	Glag	100	Latn, Beng
Gran	100	Taml, Thai	Ugar	100	Latn, Xsux	Wara	100	Latn, Geor
Bugi	100	Latn, Cans	Cakm	100	Mlym, Mymr	Cham	100	Mymr, Latn
Sund	100	Latn, Hebr	Tale	100	Latn, Hebr	Talu	100	Geor, Mymr
Tavt	100	Thai, Latn	Kali	100	Khmr, Thai	Khar	100	Hebr, Cans
Kthi	100	Deva, Gujr	Lana	100	Mymr, Tibt	Adlm	100	Thai, Latn
Ahom	100	Thai, Latn	Avst	100	Arab, Latn	Bhks	100	Latn, Olck
Sgnw	90	Latn, Arab	Soyo	86	Deva, Tibt	Saur	85	Latn, Beng
Mahj	82	Latn, Geor	Diak	79	Mlym, Beng	Zanb	76	Mtei, Latn
Tirh	75	Beng, Thai	Nand	68	Deva, Thai	Sunu	67	Beng, Latn
Mani	65	Arab, Latn	Hatr	63	Hebr, Arab	Palm	63	Hebr, Phnx
Gonm	62	Latn, Ethi	Nbat	61	Hebr, Ethi	Kits	60	Hani, Latn
Samr	59	Latn, Phnx	Elym	57	Hebr, Latn	Osge	48	Grek, Latn
Armi	45	Hebr, Latn	Gong	43	Knda, Tibt	Berf	41	Ethi, Latn
Osma	40	Latn, Armn	Mult	38	Latn, Geor	Lydi	35	Latn, Runr
Sind	34	Latn, Geor	Mroo	32	Latn, Cans	Plrd	31	Latn, Grek
Rjng	31	Latn, Cans	Takr	30	Guru, Gujr	Rohg	28	Arab, Latn
Dogr	27	Deva, Guru	Cprt	27	Latn, Tibt	Nagm	26	Latn, Geor
Kawi	26	Khmr, Thai	Batk	26	Latn, Cans	Hano	25	Latn, Runr
Lyci	23	Latn, Grek	Lina	21	Latn, Hani	Medf	21	Taml, Latn
Hluw	21	Latn, Egyp	Sogd	18	Arab, Latn	Khoj	17	Gujr, Beng
Narb	16	Latn, Arab	Hmnp	16	Thai, Latn	Toto	16	Geor, Grek
Maka	15	Ethi, Shaw	Phag	15	Latn, Beng	Wcho	15	Arab, Geor
Tnsa	14	Latn, Armn	Hmng	13	Thai, Khmr	Perm	13	Latn, Tibt
Yezi	13	Grek, Latn	Todr	12	Latn, Thai	Vith	12	Latn, Grek
Krai	12	Thai, Taml	Elba	11	Latn, Geor	Mend	9	Latn, Ethi
Cari	9	Latn, Grek	Buhd	8	Latn, Tibt	Tagb	8	Latn, Grek
Nshu	5	Hani, Latn	Bass	5	Cans, Geor	Sogo	4	Latn, Hani
Sora	3	Latn, Mymr	Chrs	2	Latn, Arab	Pauc	1	Thai, Cyrl
Phlp	1	Arab, Latn						

Table 5. A@5 (Acc@5)/SA (ScriptAcc) (%) per script. In order: PaddleOCR-VL-1.5; olmOCR-2; Gemini 3.1 Flash-Lite; dots.mocr; dots.ocr; HunyuanOCR; GPT-4.1.

Script	n	**Top-2 Out.Script**	**Paddle-1.5** A@5/SA↑	**olmOCR2** A@5/SA↑	**Gemini-FL** A@5/SA↑	**dots.mocr** A@5/SA↑	**dots.ocr** A@5/SA↑	**Hunyuan** A@5/SA↑	**GPT-4.1** A@5/SA↑
Latn	4000	Latn	79.8/99.2	90.5/99.6	95.3/99.7	93.1/99.8	91.8/99.5	85.3/99.5	83.2/99.7
Cyrl	400	Cyrl, Latn	78.0/97.8	71.2/98.2	88.8/99.2	83.5/99.2	86.0/99.2	86.8/98.5	69.8/98.8
Hani	400	Hani	60.8/99.2	78.2/100.0	85.5/99.8	86.5/100.0	81.5/99.2	83.0/99.5	55.0/99.5
Deva	400	Deva	1.0/99.8	71.5/100.0	93.5/100.0	86.0/98.5	90.2/100.0	79.8/100.0	65.8/100.0
Arab	400	Arab, Latn	21.0/99.8	31.0/99.8	58.8/99.8	62.0/97.5	63.2/98.2	38.2/97.5	39.5/99.8
Jpan	100	Jpan	96.0/100.0	98.0/100.0	98.0/99.0	92.0/100.0	88.0/100.0	99.0/100.0	98.0/100.0
Hang	100	Hang, Latn	96.0/100.0	98.0/100.0	100.0/100.0	76.0/97.0	76.0/95.0	94.0/100.0	93.0/100.0
Grek	100	Grek, Cyrl	68.0/98.0	83.0/100.0	91.0/100.0	89.0/100.0	84.0/100.0	77.0/99.0	86.0/100.0
Taml	100	Taml, Telu	98.0/100.0	4.0/100.0	99.0/100.0	86.0/99.0	88.0/100.0	65.0/95.0	73.0/100.0
Telu	100	Telu, Beng	81.0/100.0	0.0/100.0	92.0/100.0	90.0/97.0	93.0/100.0	35.0/72.0	20.0/100.0
Thai	100	Thai, Latn	17.0/100.0	34.0/100.0	68.0/100.0	79.0/99.0	81.0/100.0	65.0/100.0	51.0/100.0
Geor	100	Geor, Thai	0.0/0.0	0.0/100.0	89.0/97.0	97.0/97.0	97.0/97.0	88.0/90.0	8.0/96.0
Gujr	100	Gujr, Deva	0.0/0.0	13.0/100.0	98.0/100.0	90.0/98.0	95.0/100.0	6.0/9.0	48.0/100.0
Guru	100	Guru, Deva	0.0/0.0	0.0/71.0	85.0/100.0	79.0/94.0	89.0/100.0	71.0/77.0	20.0/100.0
Beng	100	Beng, Deva	29.0/99.0	27.0/100.0	63.0/100.0	58.0/99.0	73.0/100.0	53.0/98.0	27.0/99.0
Tibt	100	Tibt, Deva	81.0/100.0	0.0/86.0	19.0/100.0	78.0/100.0	76.0/100.0	71.0/100.0	0.0/98.0
Armn	100	Armn, Thai	0.0/0.0	0.0/52.0	97.0/100.0	95.0/100.0	93.0/100.0	17.0/73.0	1.0/100.0
Knda	100	Knda, Telu	0.0/0.0	0.0/89.0	85.0/100.0	87.0/89.0	80.0/90.0	33.0/76.0	16.0/94.0
Hebr	100	Hebr, Thai	0.0/0.0	9.0/97.0	61.0/100.0	49.0/93.0	55.0/98.0	42.0/91.0	42.0/99.0
Sinh	100	Telu, Sinh	0.0/0.0	0.0/99.0	74.0/100.0	89.0/100.0	89.0/100.0	2.0/4.0	1.0/100.0
Laoo	100	Laoo, Thai	0.0/0.0	0.0/15.0	72.0/100.0	74.0/97.0	79.0/99.0	0.0/0.0	0.0/65.0
Mlym	100	Mlym, Telu	0.0/0.0	0.0/99.0	81.0/100.0	69.0/99.0	35.0/100.0	23.0/74.0	15.0/100.0
Ethi	100	Ethi, Tibt	0.0/0.0	0.0/97.0	19.0/100.0	51.0/96.0	65.0/100.0	18.0/29.0	0.0/100.0
Thaa	100	Thaa, Arab	0.0/0.0	0.0/0.0	0.0/100.0	58.0/98.0	69.0/100.0	0.0/0.0	0.0/58.0
Orya	100	Orya, Thai	0.0/0.0	0.0/91.0	82.0/100.0	11.0/93.0	0.0/4.0	2.0/3.0	4.0/100.0
Mymr	100	Mymr, Thai	0.0/0.0	0.0/75.0	32.0/99.0	15.0/95.0	13.0/94.0	16.0/62.0	0.0/99.0
Khmr	100	Khmr, Thai	0.0/0.0	0.0/96.0	28.0/100.0	8.0/95.0	6.0/89.0	1.0/43.0	0.0/100.0
Copt	100	Grek, Copt	0.0/0.0	0.0/0.0	22.0/36.0	0.0/0.0	0.0/0.0	0.0/0.0	1.0/31.0
Bopo	100	Bopo, Hani	3.0/20.0	0.0/13.0	1.0/97.0	0.0/19.0	0.0/41.0	0.0/24.0	0.0/0.0
Cans	100	Cans, Latn	0.0/0.0	0.0/0.0	0.0/74.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/97.0
Egyp	100	Egyp, Latn	0.0/0.0	0.0/0.0	0.0/99.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/56.0
Xsux	100	Xsux, Hani	0.0/0.0	0.0/0.0	0.0/71.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/54.0
Syrc	100	Arab, Syrc	0.0/0.0	0.0/0.0	0.0/84.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/20.0
Runr	100	Runr, Latn	0.0/0.0	0.0/0.0	0.0/84.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/16.0
Mtei	100	Mtei, Tibt	0.0/0.0	0.0/0.0	0.0/97.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Brai	100	Brai, Latn	0.0/0.0	0.0/0.0	0.0/88.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Cher	100	Cher, Latn	0.0/0.0	0.0/0.0	0.0/88.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Xpeo	100	Xpeo, Latn	0.0/0.0	0.0/0.0	0.0/77.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tglg	100	Geor, Shaw	0.0/0.0	0.0/0.0	0.0/25.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Java	100	Khmr, Mlym	0.0/0.0	0.0/0.0	0.0/15.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Bali	100	Khmr, Laoo	0.0/0.0	0.0/0.0	0.0/14.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Phnx	100	Latn, Tibt	0.0/0.0	0.0/0.0	0.0/8.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Linb	100	Latn, Hani	0.0/0.0	0.0/0.0	0.0/7.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Vaii	100	Shaw, Ethi	0.0/0.0	0.0/0.0	0.0/4.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tang	100	Hani, Latn	0.0/0.0	0.0/0.0	0.0/3.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Orkh	100	Runr, Latn	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/2.0
Brah	100	Cans, Ethi	0.0/0.0	0.0/0.0	0.0/2.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Lisu	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Shaw	100	Latn, Arab	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Goth	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tfng	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Yiii	100	Hani, Hang	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/1.0
Sylo	100	Deva, Beng	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Aghb	100	Latn, Armn	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Sidd	75	Deva, Tibt	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/1.3

Table 6. A@5 (Acc@5)/SA (ScriptAcc) (%) per script. In order: Qwen3-VL-8B; GLM-OCR; RolmOCR; LightOnOCR-2; DeepSeek-OCR-2; FireRed-OCR; Nanonets-OCR2.

Script	n	**Top-2 Out.Script**	**Qwen3-8B** A@5/SA↑	**GLM** A@5/SA↑	**Rolm** A@5/SA↑	**LightOn-2** A@5/SA↑	**DeepSeek-2** A@5/SA↑	**FireRed** A@5/SA↑	**Nanonets-2** A@5/SA↑
Latn	4000	Latn	89.5/99.7	89.5/99.6	89.6/99.8	89.8/99.8	76.2/98.2	83.6/99.5	88.6/99.8
Cyrl	400	Cyrl, Latn	79.0/99.2	26.2/98.2	75.2/98.8	68.8/95.5	57.8/94.2	59.2/95.0	63.7/98.2
Hani	400	Hani, Latn	78.0/100.0	82.0/100.0	68.8/99.8	26.2/98.8	48.5/100.0	66.8/99.5	46.5/100.0
Deva	400	Deva, Guru	70.8/100.0	0.2/100.0	70.2/100.0	65.2/96.0	28.5/93.8	46.5/99.0	54.2/99.8
Arab	400	Arab, Latn	36.8/99.8	0.2/99.8	32.8/99.8	30.0/99.2	28.5/99.0	13.0/89.8	29.2/99.8
Jpan	100	Jpan	95.0/100.0	92.0/100.0	93.0/100.0	83.0/100.0	80.0/100.0	95.0/100.0	92.0/100.0
Hang	100	Hang, Latn	97.0/100.0	17.0/100.0	97.0/100.0	53.0/93.0	31.0/84.0	65.0/95.0	81.0/100.0
Grek	100	Grek, Latn	83.0/100.0	50.0/94.0	82.0/100.0	75.0/100.0	71.0/100.0	57.0/93.0	70.0/100.0
Taml	100	Taml, Thai	7.0/100.0	0.0/100.0	3.0/100.0	29.0/88.0	27.0/62.0	1.0/93.0	0.0/100.0
Telu	100	Telu, Knda	0.0/98.0	0.0/1.0	0.0/100.0	0.0/31.0	1.0/20.0	0.0/94.0	0.0/98.0
Thai	100	Thai, Tibt	31.0/100.0	0.0/94.0	33.0/100.0	25.0/98.0	7.0/74.0	14.0/87.0	22.0/100.0
Geor	100	Geor, Thai	17.0/97.0	0.0/99.0	0.0/97.0	36.0/46.0	1.0/3.0	0.0/16.0	0.0/97.0
Gujr	100	Deva, Gujr	13.0/46.0	0.0/0.0	4.0/71.0	6.0/8.0	6.0/11.0	0.0/0.0	0.0/18.0
Guru	100	Deva, Guru	22.0/85.0	0.0/0.0	0.0/28.0	2.0/27.0	0.0/0.0	2.0/43.0	0.0/1.0
Beng	100	Beng, Deva	33.0/100.0	0.0/96.0	30.0/100.0	15.0/83.0	0.0/22.0	20.0/100.0	26.0/100.0
Tibt	100	Tibt, Beng	12.0/98.0	0.0/97.0	0.0/44.0	21.0/82.0	1.0/70.0	2.0/60.0	0.0/10.0
Armn	100	Armn, Latn	0.0/48.0	0.0/9.0	0.0/98.0	0.0/13.0	2.0/11.0	0.0/3.0	0.0/87.0
Knda	100	Knda, Telu	0.0/5.0	0.0/79.0	0.0/74.0	0.0/7.0	0.0/0.0	0.0/0.0	0.0/77.0
Hebr	100	Hebr, Thai	33.0/99.0	0.0/95.0	4.0/93.0	38.0/95.0	5.0/31.0	2.0/64.0	1.0/92.0
Sinh	100	Sinh, Tibt	0.0/26.0	0.0/63.0	0.0/59.0	0.0/4.0	0.0/1.0	0.0/0.0	0.0/43.0
Laoo	100	Thai, Laoo	11.0/49.0	0.0/0.0	0.0/0.0	0.0/6.0	1.0/23.0	1.0/51.0	0.0/0.0
Mlym	100	Mlym, Thai	0.0/80.0	0.0/100.0	0.0/99.0	0.0/9.0	0.0/0.0	0.0/0.0	0.0/92.0
Ethi	100	Mymr, Latn	0.0/16.0	0.0/0.0	0.0/89.0	0.0/1.0	0.0/3.0	0.0/0.0	0.0/4.0
Thaa	100	Arab, Latn	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/4.0	0.0/0.0	0.0/0.0
Orya	100	Beng, Orya	0.0/17.0	0.0/1.0	0.0/74.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/9.0
Mymr	100	Mymr, Latn	0.0/62.0	0.0/85.0	0.0/85.0	0.0/14.0	0.0/15.0	0.0/2.0	0.0/64.0
Khmr	100	Khmr, Thai	0.0/21.0	0.0/89.0	0.0/79.0	0.0/18.0	0.0/10.0	0.0/0.0	0.0/36.0
Copt	100	Grek, Cyrl	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Bopo	100	Kana, Hani	0.0/11.0	0.0/5.0	0.0/16.0	0.0/0.0	0.0/0.0	0.0/34.0	0.0/20.0
Cans	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Egyp	100	Latn, Avst	0.0/2.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/1.0
Xsux	100	Latn, Hani	0.0/0.0	0.0/0.0	0.0/5.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Syrc	100	Arab, Hebr	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Runr	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/1.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/1.0
Mtei	100	Geor, Tibt	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Brai	100	Latn, Cyrl	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Cher	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Xpeo	100	Latn, Tibt	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tglg	100	Latn, Beng	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Java	100	Khmr, Thai	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Bali	100	Thai, Khmr	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Phnx	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Linb	100	Latn, Hang	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Vaii	100	Latn, Mymr	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tang	100	Hani, Latn	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Orkh	100	Latn, Hebr	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Brah	100	Latn, Hang	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Lisu	100	Latn, Cyrl	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Shaw	100	Latn, Hebr	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Goth	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tfng	100	Latn, Grek	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Yiii	100	Latn, Hani	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Sylo	100	Deva, Guru	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Aghb	100	Latn, Beng	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Sidd	75	Beng, Deva	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0