CoRSAL-OCR: Evaluating Zero-Shot OCR for Language Archive Materials

Luke Gessler
Indiana University Bloomington
lgessler@iu.edu

Andrew Haynes
The Woodlands College Park High School
drew.naoki@gmail.com

Abstract

Language archives contain valuable linguistic materials that are undigitized and therefore difficult to access. Modern optical character recognition (OCR) systems have great potential to make these collections more accessible, but there are few system evaluations which can assess the quality of an OCR system specifically for language archive materials. We present CoRSAL-OCR, an OCR evaluation dataset of over 200 document pages with gold-standard transcriptions from two South Asian languages: Bodo (written in Devanagari) and Garo (written in Latin script). Using this dataset together with the 8-language AILLA-OCR benchmark, we evaluate four OCR systems: Tesseract, Google Cloud Vision, Gemini 3 Flash, and Qwen3.5-27B (an open-weight model). We find that vision language models (VLMs), when given appropriate prompts, achieve the lowest error rates on these datasets. However, prompt design has a large effect on VLM performance, with a detailed generic prompt reducing CER by up to six-fold compared to a minimal prompt. We release our dataset at https://github.com/larc-iu/corsal-ocr to support further research on OCR for language archives.

1 Introduction

Many of the world's languages are at risk of no longer being spoken by the close of this century (Krauss, 1992). Amid these pressures, many language communities and researchers are engaged in efforts to document and revitalize them, with archival material from previous descriptive work often playing a crucial role. However, the language data in these archival materials are often undigitized, rendering them inaccessible for many humans and machines, impeding efforts to create derivative products such as educational materials or language technologies.

ACHIK SONGBAD
(Weekly)
Editor: N. M. Marak.
Ondimani:
Bilsi gimikna Rs.6/8
Ja dokna Rs. 3/4
Kingsani dam Ana 2
Vol. IV] TURA, MARCH 30, 1956. No. 12.

BRITISH ARO RUSSIA RIPENG DAKNA MANGSONGENGA.

Birmingham, March, 29. Soviet Russiani Electric-ni aro mamanti bosturangko bikotgipani Mantri Mr. Malenkcov Englandona reange British-rangni Atom Bomb aro Hydrogea Bomb dakarangni karkanarangko nienga.

British aro Russia panguan kumonggrikgen ine Britishni Atom Bomb aro Hydrogen Bombrangko dakgipani Director Sir John Cocorofordna aganenga.

Britishrangni General Electric karkanako Mr. George Malenkcov niaha.

AMERICAO BILONGEN SURI ONENGA.

New York, March, 19. Daororo Americao bilongen suri onenga.

Surini gimin New York songjinmao Tramrang aro motorrang rena manjaenga.

Badiaba badiaba biaprango put 15 mang ritchae suri dongenga.

RUSSIANRANG STALIN-NI SASON KAANIKO NAMNIKJAENGAHA.

London, March, 19. London songjinmao Vojsabhao, Soviet Russiani Prime Minister gitcham Mr. George Malenkcov, Stalin sagini sason kaanganiko dao Soviet Russia namnikjaengaha ine aganenga.

Mr. George Malenkcov indakeba aganenga: Anga uni tangmitingo, uni jakasini bil gita angko jakkalaha, gisepo angko sootna mangsongarangba donga, banga jongadarangko, uni ningo kam kagiparangko ua hajalni hajal sootaha, silairangchi goe galaha. Asongsa asonggipinrang baksa kumong nangrimgrikani cholrangko onjaha. Ska gita toromko manigiparangko soote galaha. Stalinni tangmitingo Russia asongo ska gita Kristian toromrangko manina onjaha. Indakarangni gimin Soviet Russia dao Stalinni sason kaanganiko namnikjaengaha.

Dao Soviet Russia gital asongko sason kaanirangko dakna mangsongenga, agilsakni jakgitelgipa asongrang baksa kumong nangrimgrikna cholrangko ongrikna mangsongenga, agilsakni asongtong muan 1500 mang batengnang uko gogipa March/56 tariko muanprako Rs. 34/- kae palataha. Indake ongoba saksa sakgni muan 300 mang gitakode dam chuongmanchakuja ine paljaha. Kiltengko kae donna manderangni kumonganio ian duk chakbatgipa manderangni kumonggrikaniko mesokaha. Achik asongni darang antioba iana bate manderang kiliko palanio kumonggrikaniko dakkujachim.

Dalu antioba joljol anti-brinade kaaha, aro bonkamgipa antio muanprako Rs. 34/8/- kae watataha.

Mikkangchi manderangni kumonggrikaniko dakna manbatbeaniko mesoka ine angajanapatna namnika,

—J. C. Marak,
C. M. Organiser.

HYDROGEN BOMBMARANGCHI DAKGRIKODE AGILSAKNI MANDERANGNA NAMJAWA INE CHANCHIENGA.

London, March, 19. Daororo agilsak sosojengjeng ongenga; indake onganganni kri agilsakni manderang dakgrikani ongnaba donga.

Figure 1: Front page of *Achik Songbad*, a Garo-language weekly newspaper (1956), from the CoRSAL archive. This document exhibits several challenges for OCR for archival materials: multi-column layout, aged paper, and Garo's use of a middle dot character for glottal stops.

Historically, performing optical character recognition (OCR) on such materials has been challenging due to the paucity of data available for supervised learning for traditional OCR systems. Much progress has been made in the past five years on languages which are written in major world scripts, such as the Latin alphabet, by composing commercially available OCR products such as Google Vision and applying supervised post-correction algorithms to the output (Rijhwani et al., 2020, 2021; Agarwal and Anastasopoulos, 2024, 2025, *inter alia*). For endangered languages, this approach has significant advantages, as it alleviates the need for the full amount of data required to train a dedicated OCR system. However, it is limited by the fact that

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 125–135

a considerable annotation effort is still required, as pairs of "raw" and corrected OCR system output are required in order to train the post-correction model.

In the past two years, powerful vision language models (VLMs) have emerged, which augment the textual capabilities of modern large language models (LLMs) with the ability to process images as input. Large VLMs are trained on massive datasets covering hundreds or thousands of languages, and we hypothesize that they may be suitable as zero-shot OCR systems for archival data. While we do not expect the VLM to have been trained on any data from the languages in question, we suspect that given the highly multilingual nature of its training data, strong transfer may be possible regardless for unseen languages written in a major world script.

In this work, we therefore empirically investigate the capabilities of VLMs for zero-shot OCR on archival material, and additionally present a new dataset for zero-shot OCR in this setting. We make the following contributions:

1. We release CoRSAL-OCR, a publicly available dataset of document images and human-transcribed text from 2 languages in CoRSAL (Bodo and Garo), expanding the resources available for evaluating OCR on archival materials.
2. We conduct an evaluation comparing traditional OCR systems and VLMs in a zero-shot setting on real archival documents from CoRSAL-OCR, finding that appropriately prompted VLMs achieve the lowest error rates across all evaluation datasets, with CER as low as 1.9% on Bodo and 4.6% on Garo, outperforming traditional alternatives such as Google Cloud Vision and Tesseract.

2 Related Work

OCR for Endangered Languages For many endangered languages, archives hold valuable materials—dictionaries, field notes, grammars, pedagogical texts, and more—that remain in non-machine-readable formats such as scanned images (Agarwal and Anastasopoulos, 2024). Applying general-purpose OCR to these materials is challenging: these languages may be written in their own scripts, or written in a widely used script but with modifications that render it significantly out-of-distribution for models pre-trained on high-resource languages. Further, materials may be multilingual, with e.g. translations in another language. Endangered languages with materials such as these typically do not have enough annotated image–text pairs to support supervised training of OCR models.

Moreover, there is a lack of publicly available evaluation datasets specifically addressing the genres present in archival materials: while Agarwal and Anastasopoulos (2025) released gold transcriptions for eight Indigenous languages, the corresponding document images are not publicly distributed, limiting ease of use.

Supervised Post-Correction In the past several years, the dominant approach to this problem has been supervised post-correction, in which the output of a general-purpose OCR engine is automatically corrected by a model trained on paired raw-OCR and gold-transcription data. Rijhwani et al. (2020) introduced a neural post-correction method for endangered languages, reducing character error rates in experiments on data from three endangered languages. Subsequent work extended this paradigm with semi-supervised self-training and lexically aware decoding (Rijhwani et al., 2021), and Rijhwani et al. (2023) conducted a user-centric evaluation of the resulting systems for Kwak'wala. Most recently, Agarwal and Anastasopoulos (2025) released the first OCR dataset for eight Indigenous languages of Latin America, enabling further work on post-correction–based OCR for endangered languages.

While effective, these methods all require a non-trivial annotation effort to produce paired training data for each target language—a requirement that limits scalability to the hundreds of endangered languages for which no such data exists.

VLMs for OCR Several recent studies have examined the potential of VLMs for zero-shot OCR. Sohail et al. (2024) benchmarked GPT-4o in a zero-shot setting on low-resource scripts including Urdu, Albanian, and Tajik, but evaluated on synthetic images with controlled variations rather than real documents, and concluded that zero-shot performance remains limited. Haq et al. (2025) similarly evaluated multiple VLMs on a synthetic Pashto dataset. Other work has pursued fine-tuning: Kolavi et al. (2025) adapted VLMs to ten Indic languages using LoRA and synthetic data, while Chung and Choi (2025) fine-tuned VLMs for Manchu OCR on syn-

thetic word images, achieving strong results but requiring language-specific training. To our knowledge, no prior work has evaluated VLMs in a truly zero-shot setting on real documents from language archives.

3 Methods

3.1 Data

Our evaluation data comes from two sources: a new dataset that we create and release from CoRSAL, and an existing benchmark from AILLA.

CoRSAL-OCR The Computational Resource for South Asian Languages (CoRSAL) is a digital archive hosted at the University of North Texas, serving over 30 South Asian languages (Chelliah and Phillips, 2023). Many languages in CoRSAL have high-quality document scans but no corresponding transcriptions, making them ideal candidates for OCR evaluation in a truly zero-resource setting.

We create a new evaluation dataset from two CoRSAL languages:

- **Bodo** (ISO 639-3: `brx`), a Sino-Tibetan language spoken in Assam, India, written in Devanagari script. Our dataset includes 53 pages from 13 documents.
- **Garo** (ISO 639-3: `grt`), a Sino-Tibetan language spoken in Meghalaya, India. While Garo is written in a Latin-based alphabet, it uses a middle dot (·) to represent glottal stops. We hypothesize this may pose an issue for OCR systems expecting standard Latin text. Our dataset includes 154 pages from 20 documents.

We choose these two languages because they represent two major script families (Devanagari and Latin) and because substantial archival material is available for both in CoRSAL. The documents span a variety of genres, including newspapers, religious texts, dictionaries, language learning texts, poems, and literary works. We used these documents with permission from CoRSAL's maintainers. Table 1 summarizes the dataset.

Each page was transcribed by a hired native-speaker annotator (annotator E for Bodo, annotator Q for Garo). To assess transcription quality, 12 pages per language were independently double-annotated by a second, non-native-speaker annotator (annotator M for Bodo, annotator B for Garo).

	Bodo	Garo	AILLA
Pages	53	154	296
Documents	13	20	—
Total chars	83,054	233,136	381,625
Total words	11,898	33,850	60,194
Avg. chars/page	1,567	1,514	1,289
Avg. words/page	225	220	203

Table 1: Dataset statistics. AILLA totals span 8 languages; per-language statistics are in Appendix B.

Inter-annotator CER is 0.8% for Bodo and 1.7% for Garo (see Appendix A for details), indicating high transcription consistency. We publicly release all 207 page images with their gold-standard transcriptions.[1]

AILLA To broaden our evaluation beyond South Asian languages, we additionally use the AILLA-OCR benchmark introduced by Agarwal and Anastasopoulos (2025). This benchmark provides page images with verified ground-truth transcriptions spanning 8 Indigenous languages of Latin America, representing diverse language families and geographic regions.

3.2 Systems

We evaluate four systems spanning three categories: a traditional open-source OCR engine, a commercial OCR API, and two vision language models (VLMs).

Tesseract (TESS) As a baseline, we use Tesseract (Smith, 2007), a widely used open-source OCR engine. Tesseract's LSTM-based recognition models are trained on major world languages, and we use the English (`eng`) model for Latin-script languages and the Hindi (`hin`) model for Devanagari-script languages. Crucially, no language-specific models exist for any of the languages in our evaluation, making this a true zero-shot baseline. We select the closest script-matching models available; no alternative Tesseract configuration would be expected to perform substantially better for these languages.[2]

Google Cloud Vision (GV) Google Cloud Vision is a commercial OCR product whose models

[1] https://github.com/larc-iu/corsal-ocr

[2] One reviewer of this work pointed out that monolingual models are very biased towards producing words in just the language that they were trained on, and suggests that multilingual Tesseract models may exhibit less of this bias and therefore perform better in zero-shot settings. We find this suggestion compelling, but leave investigating it to future work.

are continuously updated on a vast and diverse corpus of images from the web. In recent work on OCR for low-resource and endangered languages, it is common to use GV as a strong baseline (Rijhwani et al., 2020, 2021; Agarwal and Anastasopoulos, 2025, *inter alia*), and we include it to facilitate comparison with prior work. We note that we accessed GV in March 2026, and emphasize that the internal details of its operation may differ without any way for anyone outside of Google to know it at times before and after this one.

Gemini 3 Flash (GEMINI) For our closed-weight VLM, we use Gemini 3 Flash,[3] Google's frontier vision language model accessed via API. Gemini 3 Flash achieves strong performance on multimodal reasoning benchmarks, allowing it to process entire document pages with detailed instructions. We choose Flash instead of the related and larger model, Pro, as preliminary experiments did not indicate a measurable difference between the two.

Qwen3.5-27B (QWEN) For our open-weight VLM, we use Qwen3.5-27B,[4] a natively multimodal model. At 5-bit quantization, the model fits entirely in the VRAM of a single consumer GPU (e.g., an NVIDIA RTX 4090 with 24 GB), requiring no cloud infrastructure. Qwen3.5 is highly multilingual and achieves strong OCR performance, with expanded support for rare characters and scripts. As an open-weight model that can be run locally, it represents a fully reproducible and transparent alternative to commercial APIs—an important consideration for language documentation workflows where data sensitivity or infrastructure constraints may preclude the use of external services.

3.3 Prompting

Unlike traditional OCR systems, VLMs accept natural language instructions that can influence their output. We investigate the effect of prompting by evaluating each VLM with three prompts:

- **Minimal** (MIN): A single sentence ("Transcribe the text in this image."), serving as a zero-effort baseline.
- **Generic detailed** (GEN): Language-agnostic instructions specifying exact transcription, preservation of diacritics, multi-column reading order, and handling of non-text elements.
- **Language-specific** (LANG): The generic detailed prompt augmented with a paragraph describing the target language, its script, and key orthographic features (e.g., the Garo middle dot, Devanagari conjuncts, or AILLA glottal stop conventions).

The traditional OCR systems (TESS and GV) do not accept prompts and are evaluated in a single condition. Full prompt texts are given in Appendix F.

3.4 Evaluation

System outputs are evaluated against gold-standard transcriptions using character error rate (CER) and word error rate (WER). CER is the character-level Levenshtein distance between the system output and reference, normalized by reference length; WER is the analogous word-level metric. Before comparison, both texts undergo a normalization pipeline designed to remove formatting variation that does not reflect transcription quality (e.g., whitespace conventions around punctuation); full details are given in Appendix C. We report micro-averaged results, where metrics are computed over all characters (or words) in a dataset rather than averaged across documents, so that longer documents contribute proportionally more to the aggregate score.

4 Results

Table 2 presents the main results across all evaluation datasets and prompt conditions.

Prompting Effects The choice of prompt has a dramatic effect on VLM performance. On Bodo, GEMINI with the generic detailed prompt (GEN) achieves a CER of 1.9%—a six-fold reduction compared to the minimal prompt (11.6%) and a nearly five-fold reduction compared to GV (9.0%), the best traditional system. The effect is even more pronounced for QWEN, which drops from 51.1% CER with the minimal prompt to 11.2% with the language-specific prompt—a 78% relative reduction. On Garo, both VLMs benefit substantially from prompting, with GEMINI LANG achieving 4.6% CER and QWEN LANG achieving 5.0%, compared to 12.9% for GV.

Gemini vs. Qwen GEMINI generally outperforms QWEN, though QWEN is competitive or

[3] Available via Google's APIs under the identifier `gemini-3-flash-preview`

[4] https://huggingface.co/Qwen/Qwen3.5-27B

System	Prompt	**Garo** (Latin) CER	WER	**Bodo** (Devanagari) CER	WER	**AILLA** (Latin) CER	WER	**Macro Avg.** CER	WER
TESS	—	16.8	43.4	23.7	48.7	23.2	42.5	21.2	44.9
GV	—	12.9	36.9	9.0	20.0	22.7	32.1	14.9	29.7
GEMINI	MIN	10.7	22.5	11.6	22.3	31.6	39.2	18.0	28.0
GEMINI	GEN	6.2	14.7	**1.9**	**10.3**	19.7	31.2	9.3	18.7
GEMINI	LANG	**4.6**	**13.8**	2.2	10.7	**19.6**	**30.7**	**8.8**	**18.4**
QWEN	MIN	16.9	32.0	51.1	71.9	31.1	42.6	33.0	48.8
QWEN	GEN	5.3	21.2	25.5	49.0	21.4	34.1	17.4	34.8
QWEN	LANG	5.0	19.9	11.2	38.5	19.9	32.8	12.0	30.4

Table 2: Micro-averaged CER and WER (%) across all evaluation datasets and prompt conditions, with macro-averaged means across datasets. Bold indicates the best result in each column. TESS uses the eng model for Latin-script data and `hin` for Bodo. Prompt conditions: MIN = minimal, GEN = generic detailed, LANG = language-specific (see §3.3).

slightly better in some conditions (e.g., Garo with the GEN prompt). With language-specific prompts, the gap is modest on Garo (CER 4.6% vs. 5.0%) and AILLA (19.6% vs. 19.9%). On Bodo (Devanagari), the gap is larger (2.2% vs. 11.2%), suggesting that QWEN has weaker Devanagari support. Nevertheless, on Garo and AILLA, the open-weight QWEN model running on a single consumer GPU achieves results competitive with a frontier commercial API.

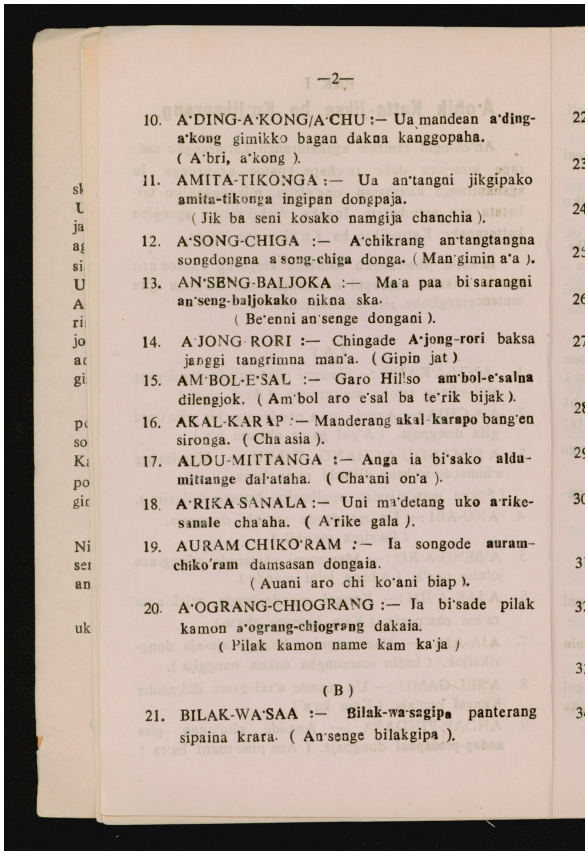
AILLA Results On the AILLA benchmark, the effect of prompting is again evident. With a minimal prompt, both VLMs underperform GV (GEMINI CER 31.6%, QWEN 31.1%, vs. GV 22.7%). With language-specific prompts, both surpass GV: GEMINI achieves 19.6% CER and QWEN 19.9%, compared to GV's 22.7%. Per-language results (Appendix B) reveal substantial variation, with best-system CER ranging from 3.8% (Mixe) to 55.4% (Cusco Quechua). The two Quechua subsets are particularly informative: despite being closely related languages, Cusco Quechua (quch, 17 pages) has over three times the CER of South Bolivian Quechua (quh, 50 pages). This gap is driven by document format rather than language: the quch data consists of three-column vocabulary lists whose tabular layout all systems struggle to linearize, while the quh data is bilingual prose with straightforward paragraph structure.

5 Error Analysis

To understand the qualitative differences between systems, we manually examined predictions on a sample of pages from each dataset. We identify three recurring error patterns.

Special character handling. The most distinctive orthographic feature in our data is the Garo middle dot (·), used for glottal stops in approximately 10% of all words. Table 3 illustrates how each system handles this character. TESS and GV never produce the middle dot, substituting apostrophes, hyphens, or spaces; GV is particularly inconsistent, sometimes dropping the character entirely. Both VLMs correctly produce the middle dot on this page, though on other pages QWEN sometimes substitutes an apostrophe. However, GEMINI also *over-generates* the middle dot, inserting it into words where the source document has none: on one Garo page with zero middle dots in the gold, GEMINI LANG produced 86 spurious instances (e.g., gold *biaprangko* → *biap·rangko*). It also systematically converts periods to middle dots in section headers (e.g., *III.* → *III·*). This hypercorrection appears to be a direct consequence of the language-specific prompt emphasizing the importance of the middle dot character, and illustrates a general risk of language-specific prompting: providing the model with targeted orthographic guidance can cause it to over-apply that guidance. GEMINI also converts periods to middle dots at line-end hyphenation points (*bikot-* → *bikot·-*), with over 130 such substitutions across the Garo dataset. Notably, the generic detailed prompt (GEN) largely avoids this problem while still achieving strong results (6.2% CER vs. 4.6% for LANG), suggesting that language-specific prompts should be validated on a small sample before deployment.

On Bodo, the analogous issue is Devanagari character confusion. QWEN systematically misrecognizes Bodo-specific characters that are rare in Hindi; for example, it renders the title



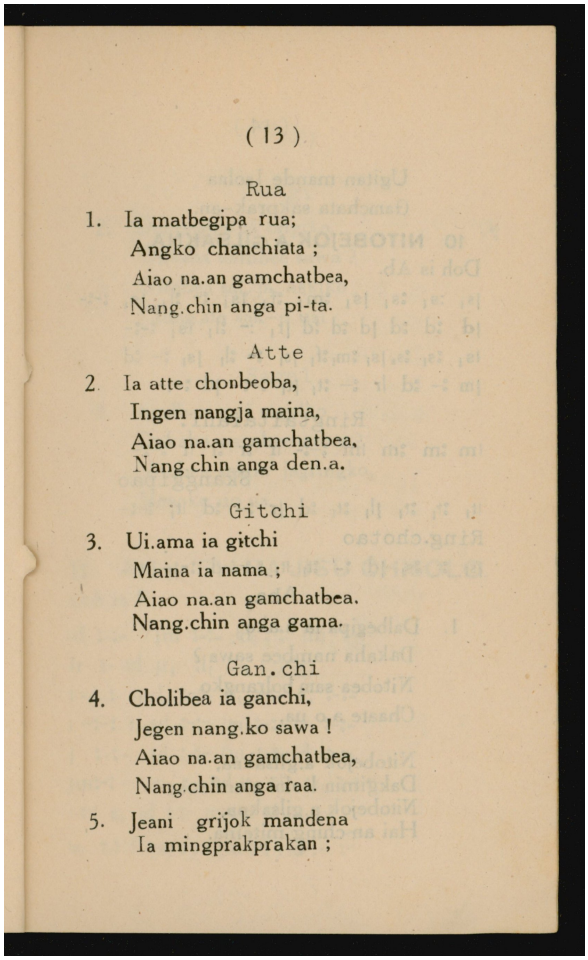
—2—

10. A·DING-A·KONG/A·CHU :— Ua mandean **a'ding-a'kong** gimikko bagan dakna kanggopaha. (A'bri, a'kong).
11. AMITA-TIKONGA :— Ua an'tangni jikgipako **amita-tikonga** ingipan dongpaja. (Jik ba seni kosako namgija chanchia).
12. A·SONG-CHIGA :— A'chikrang an'tangtangna songdongna a'song-chiga donga. (Man'gimin a'a).
13. AN·SENG-BALJOKA :— Ma'a paa bi'sarangni **an'seng-baljokako** nikna ska. (Be'enni an'senge dongani).
14. A·JONG·RORI :— Chingade **A'jong-rori** baksa janggi tangrimna man'a. (Gipin jat)
15. AM·BOL-E'SAL :— Garo Hil'so **am'bol-e'salna** dilengjok. (Am'bol aro e'sal ba te'rik bijak).
16. AKAL-KARAP :— Manderang **akal-karapo** bang'en sironga. (Cha asia).
17. ALDU-MITTANGA :— Anga ia bi'sako **aldu-mittange** dal'ataha. (Cha'ani on'a).
18. A·RIKA·SANALA :— Uni ma'detang uko **a'rike-sanale** cha'aha. (A'rike gala).
19. AURAM CHIKO'RAM :— Ia songode **auram-chiko'ram** damsasan dongaia. (Auani aro chi ko'ani biap).
20. A·OGRANG-CHIOGRANG :— Ia bi'sade pilak kamon **a'ograng-chiograng** dakaia. (Pilak kamon name kam ka'ja)

(B)

21. BILAK-WA·SAA :— **Bilak-wa'sagipa** panterang sipaina krara. (An'senge bilakgipa).

(a) Garo dictionary



(13)

Rua

1. Ia matbegipa rua;
   Angko chanchiata ;
   Aiao na.an gamchatbea,
   Nang.chin anga pi-ta.

Atte

2. Ia atte chonbeoba,
   Ingen nangja maina,
   Aiao na.an gamchatbea,
   Nang chin anga den.a.

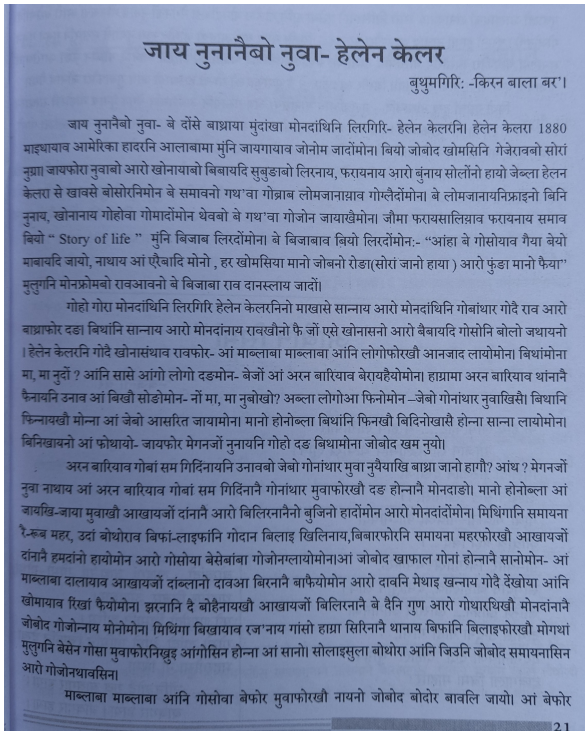
Gitchi

3. Ui.ama ia gitchi
   Maina ia nama ;
   Aiao na.an gamchatbea.
   Nang.chin anga gama.

Gan.chi

4. Cholibea ia ganchi,
   Jegen nang.ko sawa !
   Aiao na.an gamchatbea,
   Nang.chin anga raa.
5. Jeani grijok mandena
   Ia mingprakprakan ;

(b) Garo hymnal



जाय नुनानैबो नुवा- हेलेन केलर

बुथुमगिरि: -किरन बाला बर'।

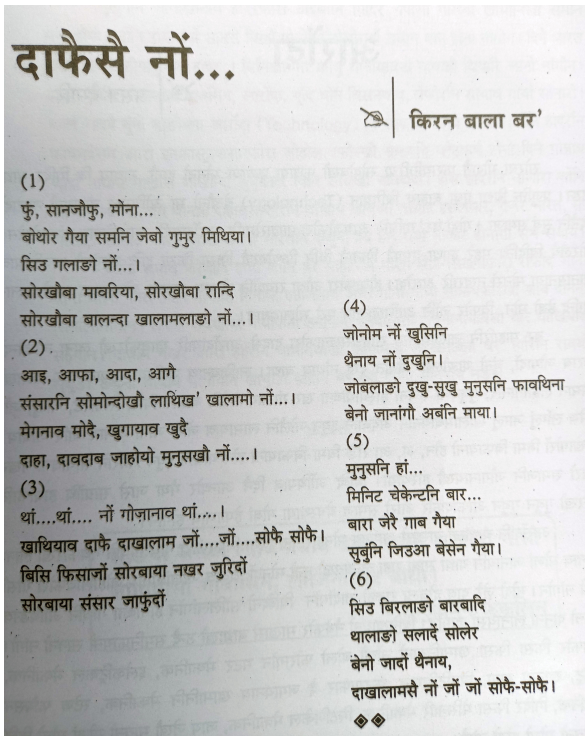
जाय नुनानैबो नुवा- बे दोंसे बाश्राया मुंदांखा मोनदांथिनि लिरगिरि- हेलेन केलरनि। हेलेन केलरा 1880 माइशावाव आमेरिका हादरनि आलाबामा मुंनि जायगायाव जोनोम जादोंमोन। बियो जोबोद खोमसिनि गेजेराबबो सोरां नुग्रा जायफोरा नुवाबो आरो खोनायाबो बिबायदि सुबुडाबो लिरनाय, फरायनाय आरो बुंनाय सोलोंनो हायो जेब्ला हेलन केलरा से खावसे बोसोरनिमोन बे समावनो गथ'वा गोब्राब लोमजानाग्रावं गोल्लेदोंमोना बे लोमजानायनिफ्राइनो बिनि नुनाय, खोनानाय गोहोवा गोमादोंमोन थेवबो बे गथ'वा गोजोन जायाखैमोना। जौमा फरायसालिग्राव फरायनाय समाव बियो " Story of life " मुंनि बिजाब लिरदोंमोना बे बिजाबाव बियो लिरदोंमोन:- "आंहा बे गोसोयाव गैया बेयो माबायदि जायो, नाथाय आं ऐबादि मोनो , हर खोमसिया मानो जोबनो रोडा(सोरां जानो हाया) आरो फुंडा मानो फैया" मुलुगनि मोनफ्रोमबो रावआवनो बे बिजाबा राव दानस्ताय जादों।

गोहो गोरा मोनदांथिनि लिरगिरि हेलेन केलरनिनो माखासे सान्नाय आरो मोनदांथिनि गोबांथार गोदै राव आरो बाश्राफोर दडा बिथांनि सान्नाय आरो मोनदांनाय रावखौनो फै जों एसे खोनासनो आरो बैबायदि गोसोनि बोलो जथायनो । हेलेन केलरनि गोदै खोनासंथाव रावफोर- आं माब्लाबा माब्लाबा आंनि लोगोफोरखौ आनजाद लायोमोना बिथांमोना मा, मा नुदों ? आंनि सासे आंगो लोगो दडमोन- बेजों आं अरन बारियाव बेरायहैयोमोना हाग्रामा अरन बारियाव थांनानै फैनायनि उनाव आं बिखौ सोडोमोन- नों मा, मा नुबोखो? अब्ला लोगोआ फिनोमोन –जेबो गोनांथार नुबाखिसै बिथानि फिनायखौ मोन्ना आं जेबो आसरित जायामोना मानो होनोब्ला बिथांनि फिनखौ बिदिनोखासै होन्ना सान्ना लायोमोना बिनिखायनो आं फोथायो- जायफोर मेगनजों नुनायनि गोहो दडं बिथामोना जोबोद खम नुयो।

अरन बारियाव गोबां सम गिदिंनायनि उनावबो जेबो गोनांथार मुबा नुयैयाखि बाग्रा जानो हागौ? आंथ ? मेगनजों नुबा नाथाय आं अरन बारियाव गोबां सम गिदिंनानै गोनांथार मुबाफोरखौ दडं होन्नानै मोनदाडो। मानो होनोब्ला आं जायखि-जाया मुबाखौ आखायजों दांनानै आरो बिलिरनानैनो बुजिनो हादोंमोन आरो मोनदांदोंमोना मिथिंगानि समायना बै-ब्लब महर, उदां बोथोराव बिफां-लाइफांनि गोदान बिलाइ खिलिनाय,बिबारफोरनि समायना महरफोरखौ आखायजों दांनानै हमदांनो हायोमोन आरो गोसोया बेसेबांबा गोजोनग्लायोमोनाआं जोबोद खाफाल गोनां होन्नानै सानोमोन- आं माब्लाबा दालायाव आखायजों दांब्लानो दावआ बिरनानै बाफैयोमोन आरो दावनि मेथाइ खन्नाय गोदै देंखोया आंनि खोमायाव रिंखां फैयोमोना झरनानि दै बोहैनायखौ आखायजों बिलिरनानै बे दैनि गुण आरो गोथारथिखौ मोनदांनानै जोबोद गोजोन्नाय मोनोमोना मिथिंगा बिखायाव रज'नाय गांसो हाग्रा सिरिनानै थानाय बिफांनि बिलाइफोरखौ मोगथां मुलुगनि बेसेन गोसा मुबाफोरनिखुइ आंगोसिन होन्ना आं सानो। सोलाइसुला बोथोरा आंनि जिउनि जोबोद समायनासिन आरो गोजोनथावसिन।

माब्लाबा माब्लाबा आंनि गोसोबा बेफोर मुबाफोरखौ नायनो जोबोद बोदोर बावलि जायो। आं बेफोर

(c) Bodo prose



दाफैसै नों...

किरन बाला बर'

(1)
फुं, सानजौफु, मोना...
बोथोर गैया समनि जेबो गुमुर मिथिया।
सिउ गलाङो नों...।
सोरखौबा मावरिया, सोरखौबा रान्दि
सोरखौबा बालन्दा खालामलाङो नों...।

(2)
आइ, आफा, आदा, आगै
संसारनि सोमोन्दोखौ लाथिख' खालामो नों...।
मेगनाव मोदै, खुगायाव खुदै
दाहा, दावदाब जाहोयो मुनुसखौ नों....।

(3)
थां....थां.... नों गोज़ानाव थां....।
खाथियाव दाफै दाखालाम जों....जों....सोफै सोफै।
बिसि फिसाजों सोरबाया नखर जुरिदों
सोरबाया संसार जाफुंदों

(4)
जोनोम नों खुसिनि
थैनाय नों दुखुनि।
जोबंलाङो दुखु-सुखु मुनुसनि फावथिना
बेनो जानांगौ अबंनि माया।

(5)
मुनुसनि हां...
मिनिट चेकेन्टनि बार...
बारा जेरै गाब गैया
सुबुंनि जिउआ बेसेन गैया।

(6)
सिउ बिरलाङो बारबादि
थालाङो सलादें सोलेर
बेनो जादों थैनाय,
दाखालामसै नों जों जों सोफै-सोफै।

◆◆

(d) Bodo poem

Figure 2: Sample pages from the CoRSAL dataset illustrating genre and script diversity. (a) A Garo dictionary with frequent middle dots (·) in headwords. (b) A Garo hymnal with numbered verses. (c) Dense Bodo prose with complex Devanagari conjuncts which are uncommon in Hindi. (d) A Bodo poem with numbered stanzas and the ◆ section divider.

System	Output
Gold	A·DING-A·KONG/A·CHU
TESS	A”DING-A’KONG/A’CHU
GV	A’DING-A KONG/A CHU
GEMINI	A·DING-A·KONG/A·CHU
QWEN	A·DING-A·KONG/A·CHU

Table 3: System outputs for a Garo dictionary headword containing three middle dots (·). TESS and GV substitute apostrophes, spaces, or double-apostrophes; both VLMs correctly preserve the character.

āijo soloṁthāyni gonāṁthi as *āijo soloṁthāyni gonaṁsthī*, hallucinating a conjunct cluster (*sth*) where a simple aspirated stop (*th*) appears in the source. This pattern recurs throughout the Bodo data: QWEN confuses aspirated stops with conjunct clusters, drops vowel signs (e.g., omitting the *ā* matra), and substitutes visually similar characters—all consistent with Hindi-centric priors overriding the visual signal. We also observed Bangla and Gurmukhi characters appearing in otherwise Devanagari text, suggesting script confusion in the visual encoder. GEMINI handles Devanagari conjuncts and Bodo-specific characters more reliably, consistent with its lower CER on this dataset. GV also performs well on Devanagari, preserving conjuncts and vowel signs accurately, though it occasionally replaces the danda (the Devanagari full stop character which looks like a vertical line) with a pipe character.

Layout and reading order. Several Garo documents are multi-column newspaper pages (cf. Figure 1). Table 4 shows the first three lines of each system's output on this page. GEMINI and QWEN correctly identify the masthead as a header region and transcribe it before the column text, while TESS skips it entirely and jumps into a column midpage, and GV partially recovers the masthead but omits details. On AILLA, where some documents use interlinear glossing with multiple aligned tiers (source language, morpheme breakdown, grammatical gloss, free translation), GEMINI best preserves the multi-tier structure and the grouping of numbered examples with their glosses. TESS tends to read line numbers as a block first and then the text content separately, completely disconnecting examples from their annotations, while GV partially interleaves tiers from different examples.

VLM-specific failure modes. VLMs exhibit failure modes absent from traditional OCR systems. GEMINI occasionally injects markdown formatting (`**bold**`) when it encounters visually emphasized text such as dictionary headwords, inflating CER with characters that do not appear on the page. On the AILLA Kaqchikel dictionary, we observed over 40 spurious markdown markers in a single page. Both VLMs with minimal prompts sometimes produce extremely poor output (e.g., QWEN MIN on Bodo: 51.1% CER), likely reflecting cases where the model fails to recognize the task as transcription without more elaborate instructions. QWEN also occasionally inserts characters from the wrong script—we observed Turkish dotless-*i* and Cyrillic characters in Garo Latin text, likely an artifact of the multilingual training data. These failure modes are largely eliminated by the detailed prompts, underscoring the importance of prompt design.

System	First three lines
Gold	ACHIK SONGBAD / (Weekly) / Editor: N. M. Marak.
TESS	Vol. IV jy / / BRITISH ARO RUSSIA RIPENG
GV	ACHIK SONGBAD / Editor: N. M. Marak. / Vol. IV]
GEMINI	ACHIK SONGBAD / (Weekly) / Ondimani:
QWEN	ACHIK SONGBAD / (Weekly) / Editor: N. M. Marak.

Table 4: First three lines of output for the multi-column Garo newspaper page in Figure 1. TESS skips the masthead and begins mid-column; GV partially recovers it; both VLMs correctly identify and transcribe the header first.

6 Conclusion

Our results demonstrate that VLMs are effective for performing OCR for language archive materials in zero-shot settings, and identify good prompt design as an important criterion for success when using these models. A generic detailed prompt that specifies exact transcription, diacritic preservation, and layout handling captures most of the gain over a minimal prompt; language-specific information (e.g., the Garo middle dot or AILLA glottal stop conventions) provides further improvement on some datasets but not all.

The open-weight QWEN model nearly matches the frontier GEMINI on Latin-script data (Garo CER 5.0% vs. 4.6%; AILLA 19.9% vs. 19.6%), but lags on Devanagari (Bodo CER 11.2% vs. 2.2%). This gap likely reflects differences in Devanagari representation in training data, and we expect it to narrow as open-weight models continue to improve.

Nevertheless, the strong Latin-script results demonstrate that competitive zero-shot OCR is achievable on consumer hardware without reliance on commercial APIs.

From a practical standpoint, transcriptions with WER below approximately 10% may be usable for many purposes by language archive users with only light manual correction, while higher error rates (as seen on the AILLA data) could still reduce the effort required compared to transcribing from scratch.

Based on our findings, we offer the following practical guidance for researchers and archivists seeking to digitize endangered language materials:

1. **Always use a detailed prompt.** A generic prompt specifying exact transcription, diacritic preservation, and layout handling provides most of the benefit over a minimal prompt and requires no language-specific knowledge.
2. **Add language-specific information with care.** Providing the language name and orthographic details can further improve results (as seen on Garo and AILLA), but overly specific instructions may cause hypercorrection (as with Gemini over-generating the Garo middle dot). Test on a small sample before committing to a language-specific prompt.
3. **For Latin-script languages, small open-weight models are competitive.** QWEN running locally on a single consumer GPU achieved results within 1 percentage point of GEMINI on both Garo and AILLA. This avoids sending potentially sensitive archival materials to external APIs.
4. **For Devanagari and non-Latin scripts, commercial APIs seem to lead.** In our experiments, GEMINI substantially outperformed QWEN on Bodo, and GV also performed well. While we have not comprehensively surveyed all VLMs and OCR systems, we suppose that researchers working with non-Latin scripts ought to expect weaker open-weight model performance.

Our work has three limitations that we also identify as areas for future work. First, VLM output could be further enhanced by other components in a processing pipeline, such as post-correction or image segmentation. We note that VLM-based OCR and post-correction are complementary: a higher-quality first-pass transcription from a VLM should reduce the annotation burden required to train a post-correction model. Future work could also investigate the effect of document segmentation as a preprocessing step.

Second, the CoRSAL dataset covers only two languages; extending it to additional languages and scripts is a priority for future work. We are currently engaged in annotating more data for Bodo and Garo, to be released in future versions.

Third, we have limited our work to "well-behaved" textual genres. Language archives often also have more unusual textual material produced by linguists, such as handwritten sketches of vowel charts or fragmentary grammatical hypotheses. These may also be rewarding to digitize, though given that they are presumably much more difficult to process well than printed materials, we have left them out of scope for the present work.

In summary, we release a new OCR evaluation dataset for two endangered South Asian languages and show that vision language models, when appropriately prompted, provide a strong zero-shot baseline for digitizing endangered language archives. In future work, we plan to use the methods we have outlined here to improve CoRSAL's own OCR-derived transcriptions, and to further expand the CoRSAL-OCR dataset.

Limitations

Our CoRSAL dataset covers only two languages (Bodo and Garo) across two scripts (Devanagari and Latin). While the inclusion of the 8-language AILLA benchmark broadens coverage, our results may not generalize to all scripts, document types, or archival conditions. The dataset is also relatively small (207 pages), and languages have uneven representation: Garo has 154 pages while Bodo has 53.

Our evaluation relies entirely on automatic metrics (CER and WER). These do not capture all dimensions of transcription quality that matter for downstream use, such as preservation of document structure or handling of non-textual elements. We also do not evaluate the effect of document segmentation, which may improve results for multi-column layouts.

VLM performance is sensitive to prompt design, and we explore only three prompt conditions. Different prompt formulations or few-shot examples could yield different results. Additionally, VLM

outputs are not fully deterministic, and we do not report variance across multiple runs.

Ethical Considerations

The archival materials used in this work are drawn from publicly accessible digital archives (CoRSAL and AILLA) that are maintained by institutional repositories with established access and use policies. Our native speaker annotators were fairly compensated for their work.

We note that digitizing endangered language materials raises ethical considerations around data sovereignty and community consent. The materials we have drawn from CoRSAL were already publicly archived, and we have affirmed with our contacts who maintain CoRSAL that these languages' respective community members do not object to our use of these materials for the purposes described in this work.

Acknowledgements

We thank Mark Phillips and Shobhana Chelliah at CoRSAL for their help in getting access to and understanding the data used in this work. We also thank Prafulla Basumatary for his help in recruiting our native speaker annotators. We additionally thank our two native speaker annotators, Didwm Basumatary and Matsram Peter K. Sangma, for working with us to transcribe the images in this dataset. Finally, we thank our three anonymous reviewers for their helpful feedback, which we used as we produced the final form of this work.

References

Milind Agarwal and Antonios Anastasopoulos. 2024. A Concise Survey of OCR for Low-Resource Languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Milind Agarwal and Antonios Anastasopoulos. 2025. AILLA-OCR: A First Textual and Structural Post-OCR Dataset for 8 Indigenous Languages of Latin America. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 120–127, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Shobhana L. Chelliah and Mark Phillips. 2023. Computational resource for South Asian languages (CoRSAL). `https://digital.library.unt.edu/explore/collections/CORSAL/`. University of North Texas Digital Library.

Yan Hon Michael Chung and Donghyeok Choi. 2025. Finetuning Vision-Language Models as OCR Systems for Low-Resource Languages: A Case Study of Manchu. *arXiv preprint*. ArXiv:2507.06761 [cs].

Ijazul Haq, Yingjie Zhang, and Irfan Ali Khan. 2025. PsOCR: Benchmarking Large Multimodal Models for Optical Character Recognition in Low-resource Pashto Language. *arXiv preprint*. ArXiv:2505.10055 [cs].

Adithya Kolavi, Samarth P, and Vyoman Jain. 2025. Nayana OCR: A scalable framework for document OCR in low-resource languages. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 86–103, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically Aware Semi-Supervised Learning for OCR Post-Correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302. Place: Cambridge, MA Publisher: MIT Press.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-Centric Evaluation of OCR Systems for Kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, page 629–633, USA. IEEE Computer Society.

Muhammad Abdullah Sohail, Salaar Masood, and Hamza Iqbal. 2024. Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts. *arXiv preprint*. ArXiv:2412.16119 [cs] version: 1.

A Annotation

To assess transcription quality, 12 pages per language were independently double-annotated (annotators E and M for Bodo; Q and B for Garo).

	With norm.	Without
Bodo (E vs. M)	0.8%	1.0%
Garo (B vs. Q)	1.7%	1.8%

Table 5: Inter-annotator CER (%) on 12 double-annotated pages per language, with and without punctuation spacing normalization.

Table 5 reports inter-annotator agreement as micro-averaged CER, both with and without our punctuation spacing normalization (§C).
Agreement is high for both languages. The punctuation normalization has a larger effect on Bodo (reducing disagreement by 17%), consistent with the fact that annotators differed on spacing before the Devanagari danda. The effect on Garo is minimal, as expected.

B Per-Language AILLA Results

Performance varies substantially across languages, with best-system CER ranging from 3.8% (Mixe) to 55.4% (Cusco Quechua, quch). With detailed prompts, both VLMs match or outperform GV on most languages, with the largest gains on Mam (32.4% vs. 38.7%) and Kaqchikel (5.2% vs. 9.9%). Notably, GV achieves the best result on Mixe (3.8%) and TESS on South Bolivian Quechua (14.8%), indicating that traditional systems can still be competitive on individual languages even when VLMs lead in aggregate.

C Evaluation Details

Before computing CER and WER, both the system output and the gold-standard reference are passed through the following normalization pipeline, applied in order:

1. **Unicode normalization.** Both texts are converted to NFD (Canonical Decomposition) form. This ensures that characters with equivalent representations (e.g., a precomposed accented character vs. a base character followed by a combining accent) are compared consistently.

2. **Quote normalization.** Directionally-oriented quotation marks are replaced with their straight ASCII equivalents, as OCR systems vary in which form they produce.

3. **Punctuation spacing.** Whitespace immediately preceding punctuation marks (. , ; : ? ! and the Devanagari danda, U+0964) is removed. This normalization is motivated by observed disagreements between human annotators on whether to place a space before sentence-final punctuation, particularly the danda. Without this step, such formatting differences would inflate both CER and WER.

4. **Whitespace collapsing.** Newlines are replaced with spaces, and runs of multiple whitespace characters are collapsed to a single space. Leading and trailing whitespace is stripped.

CER is then computed as CER $= \text{Lev}(p, g)/|g|$, where $\text{Lev}(p, g)$ is the character-level Levenshtein distance between the preprocessed prediction p and gold g, and $|g|$ is the character length of g. WER is computed analogously at the word level: both texts are split on whitespace, and the word-level edit distance is normalized by the number of words in g.

We report **micro-averaged** metrics throughout. For a dataset of n documents, micro-averaged CER is $\sum_{i=1}^{n} \text{Lev}(p_i, g_i) \,/\, \sum_{i=1}^{n} |g_i|$, so that longer documents contribute proportionally more to the aggregate.

D Annotation Guidelines

Annotators were given the following instructions:

- Faithfully represent line breaks with newlines within paragraphs.
- Use a double newline to separate sections which are not part of the same typographical unit (e.g., page header, body text, page number).
- Include all text on the page, including page numbers, headers, and other marginal text.
- Use a standard reading order (top-down, left-to-right) to determine how to order different typographical units within the linear transcription.
- Write exactly the text that appears on the page, making no corrections during transcription.

E Decoding Parameters

GEMINI was accessed via the Google Gemini API using default parameters. QWEN was served locally using llama.cpp with the following parameters: 5-bit quantization (Q5_K_M), temperature 0.95, top-p 0.95, top-k 20, context size 8192 tokens

		Traditional		Gemini			Qwen		
Language	**N**	TESS	GV	MIN	GEN	LANG	MIN	GEN	LANG
Kaqchikel	40	9.1	9.9	25.1	**5.2**	5.5	22.9	6.2	6.1
Mam	47	37.6	38.7	53.4	32.6	**32.4**	42.3	33.1	34.2
Miskitu	50	30.1	32.6	37.4	28.4	28.4	33.6	**28.0**	28.1
Mixe	40	7.3	**3.8**	23.5	5.6	5.6	26.8	4.2	3.9
Quechua (quch)	17	57.1	66.3	56.8	55.9	55.9	72.7	**55.4**	55.5
Quechua (quh)	50	**14.8**	16.0	19.9	16.4	15.9	25.1	16.3	15.9
Tzeltal	13	86.8	30.2	32.4	**29.0**	29.1	35.7	29.8	30.1
Zoque	39	10.0	**8.5**	9.6	9.2	9.2	10.5	29.2	10.0
Macro avg.		31.6	25.8	32.1	22.7	**22.8**	33.7	25.3	23.0

Table 6: Per-language CER (%) on the AILLA benchmark, with macro-averaged means across languages. N = number of pages. Bold = best system per language (or per row for Avg.). The LANG prompt for AILLA describes the corpus as a whole (mentioning glottal stops, ejectives, and mixed Spanish/English content) rather than individual languages.

per slot, and a maximum of 4096 generation tokens per request.

F Prompts

We evaluate three prompt conditions for each VLM. The MIN (minimal) prompt and the first two paragraphs of the GEN (generic detailed) prompt are shared across all datasets. The LANG (language-specific) prompts extend GEN with a corpus-specific paragraph.

F.1 Minimal (MIN)

> Transcribe the text in this image.

F.2 Generic Detailed (GEN)

> Transcribe all text in this image exactly as written. Preserve the original spelling, punctuation, and diacritics. Do not correct, translate, or omit any text. If a character is ambiguous, transcribe your best interpretation. Output only the transcribed text with no commentary.
>
> If the document has multiple columns, transcribe in reading order: left to right, top to bottom within each column. Include all headers, page numbers, and titles. Skip photographs or illustrations but continue transcribing surrounding text.

F.3 Language-Specific (LANG)

Each language-specific prompt consists of the GEN prompt above followed by one of the following paragraphs.

Garo.

> This text is written in Garo, a Sino-Tibetan language spoken in North-East India. It uses a Latin-based script. The most important special character is a dot used as a letter, which appears frequently (in approximately 10% of all words). It should be transcribed as the Unicode middle dot (·). This dot may appear at the bottom, middle, or top of the line and may resemble a full stop, but full stops only appear at sentence ends. Some documents may also contain English text; transcribe it exactly as written.

Bodo.

> This text is written in Bodo, a Sino-Tibetan language spoken in Assam, India. It uses the Devanagari script with complex conjunct characters and agglutinative suffixes—every character and vowel sign (matra) must be accurate. Documents may contain English words or phrases mixed in; transcribe these exactly as written. Preserve any special symbols used as section dividers (such as ⧫[5] or *).

AILLA.

> This text is from a linguistic archive of Indigenous languages of Latin America (including Mayan languages, Quechua, Miskitu, and Zoque). Key features to preserve: glottal stops may be written as an apostrophe (‘), the numeral 7, or an accent mark—transcribe exactly as shown. Ejective consonants are marked with an apostrophe after the consonant (t’, k’, b’, q’)—the apostrophe is part of the letter, not punctuation. Documents often contain Spanish or English translations and linguistic annotations alongside the Indigenous text. Transcribe all of it. Preserve any metadata markers, line numbers, and annotation codes exactly as they appear.

---

[5] The actual symbol used in Bodo documents is U+25C8 (white diamond containing black small diamond), rendered here as ⧫ for compatibility with pdfLaTeX.