AmericasNLP 2026

Sixth Workshop on NLP for Indigenous Languages of the Americas

Proceedings of the Workshop

July 3, 2026

The AmericasNLP organizers gratefully acknowledge the support from the following sponsors.

Sponsor

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-415-6

Introduction

We would like to welcome you to AmericasNLP 2026, the Sixth Workshop on Natural Language Processing for Indigenous Languages of the Americas!

The main goals of the workshop are to:

- encourage research on NLP, computational linguistics, corpus linguistics, and speech around the globe to work on Indigenous American languages.
- promote research on both neural and non-neural machine learning approaches suitable for low-resource languages.
- connect researchers and professionals from underrepresented communities and native speakers of endangered languages with the machine learning and NLP communities.

In 2026, AmericasNLP will be held in San Diego, USA, on July 4th. Prior to the workshop, we introduced a new shared task: The AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages, the first shared task dedicated to generating captions for images depicting Indigenous cultures of the Americas, written in the Indigenous languages themselves. To support this, we introduced and publicly released a newly constructed dataset spanning five cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Orizaba Nahuatl, and Wixárika. Eight teams participated, submitting 27 systems in total.

We received a total of 39 submissions: 28 research papers, 3 non-archival works, and 8 shared task system description papers (across all shared tasks). 18 archival papers were accepted (acceptance rate: 64%) – in addition to the previously published and system description papers.

We would like to acknowledge all the time and effort put into the reviewing process, and thank the program committee members for helping us create a high-quality program in a short amount of time. AmericasNLP would not have been possible without the help of our sponsor Google. Finally, we would also like to thank all the authors who submitted their work to the workshop, the participants of the shared task, and everyone who will be at the workshop to exchange and discuss their ideas for improving natural language technologies for Indigenous languages of the Americas!

Manuel Mager, Abteen Ebrahimi, Minh Duc Bui, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Luis Chiruzzo, Rolando Coto-Solano, and Katharina von der Wense

AmericasNLP 2026 Organizing Committee

Program Committee

Program Committee

Heriberto Avelino, National Institute for Anthropology and History
Eduardo Blanco, University of Arizona
Ona De Gibert, University of Helsinki
Cristina España-Bonet, BSC/DFKI GmbH
Silvia Fernandez Sabido, CentroGeo
Andrew Fisher, University of New Brunswick
Luke Gessler, Indiana University Bloomington
Santiago Góngora, Universidad de la República
Lewis Howe, University of Georgia
Elwin Huaman, University of Cambridge
Simran Khanuja, Carnegie Mellon University
Éric Le Ferrand, Boston College
Zoey Liu, Department of Linguistics, University of Florida
Ali Marashian, University of Colorado Boulder
Daniela Moctezuma, Centrogeo
Remo Nitschke, University of Arizona
Angeles Belem Priego Sanchez, Universidad Autónoma Metropolitana
Nathaniel Robinson, Johns Hopkins University
Hossain Shaikh Saadi, Johannes Gutenberg University Mainz
Mario Sanz-Guerrero, Johannes Gutenberg University Mainz
Daan Van Esch, Google Research
Raul Vazquez, University of Helsinki

Keynote Talk

Jacqueline Brixey
University of Wisconsin–Madison

Bio: Jacqueline [Lina] Brixey is a postdoctoral researcher at the University of Wisconsin-Madison. She completed her PhD in Computer Science at the University of Southern California in December 2024. Her research focuses on Indigenous and endangered languages, dialogue systems, and bilingualism.

Table of Contents

Neural Text-to-Speech for Myaamia: Speech Synthesis for an Indigenous Algonquian Language

Anita Baral[1] **John Femiani**[1] **Hunter Lockwood**[2]
Daniela Inclezan[1] **Balaram Bhandari**[2]

[1]Department of Computer Science and Software Engineering, Miami University, USA
[2]Myaamia Center, Miami University, USA

barala@miamioh.edu, femianjc@miamioh.edu, lockwoht@miamioh.edu,
inclezd@miamioh.edu, bhandab@miamioh.edu

Abstract

We present the first neural text-to-speech (TTS) implementation for Myaamia (Miami-Illinois), an Indigenous Algonquian language of North America. Developed in collaboration with the Myaamia Center at Miami University, our approach upholds principles of data sovereignty. Using 14,358 utterances (10.4 hours total, 8.18 hours for training) from seven speakers, we train and evaluate FastSpeech, Glow-TTS, and VITS, assessing synthesis quality through objective (MCD, F0 RMSE, duration RMSE) and subjective (expert evaluation) metrics. VITS outperforms other models in spectral and prosodic accuracy, but challenges remain in phonetic precision and prosody modeling. Our results confirm the feasibility of neural TTS for Myaamia, with direct implications for language learning and revitalization. This work offers a replicable framework for other low-resource Indigenous languages while ensuring ethical, linguistic data governance.

1 Introduction

Since the 1990s, the global decline of Indigenous languages has driven revitalization efforts, and technology plays an increasingly important role in preserving linguistic and cultural heritage (Bird, 2020). The Myaamia (Miami-Illinois) language, traditionally spoken in the southern Great Lakes region of North America, became dormant after the last first-language speakers passed away in the mid-20th century following forced relocation. In recent decades, the Miami Tribe of Oklahoma has led systematic revitalization efforts, using linguistic expertise, education, and digital resources to reclaim the language (Baldwin et al., 2016).

Central to this work is the Myaamia Center at Miami University, which supports Myaamia language and cultural revitalization through research, education, and community partnerships. The Center developed and maintains the Indigenous Languages Digital Archive (ILDA), a web-based platform that brings together written and audio language materials to support archives-based language reclamation (Baldwin et al., 2016). Complementing ILDA, the Šaapohkaayoni community education portal provides Myaamia community members with access to self-directed learning modules regardless of geographic location. While these resources provide strong support for language learning, access to spoken language remains limited. An estimated 95% of the available audio recordings originate from just two individuals, which creates a bottleneck for learners seeking to develop listening and pronunciation skills. Neural TTS systems offer a promising way to address this gap by generating natural-sounding speech, reducing the burden on the few available speakers and expanding access to spoken language resources for learners (Brinklow, 2021). This study presents the first neural TTS system for Myaamia, developed in collaboration with the Myaamia Center at Miami University.

Our contributions to low-resource speech synthesis include:

1. We develop and evaluate the first neural TTS system for Myaamia, leveraging well-established neural TTS models (FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (Kim et al., 2021)) trained on 8 hours and 18 minutes of speech data.

2. We establish a performance benchmark for Myaamia TTS, assessing spectral (MCD), prosodic (F0 RMSE), and temporal (duration RMSE) accuracy alongside subjective evaluations by Myaamia linguists.

3. We uphold Indigenous Data Sovereignty principles, with all linguistic data remaining un-

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 1–10

der the ownership and governance of the Myaamia Center.

2 Related Work

The development of speech technology for endangered and Indigenous languages sits at the intersection of technological innovation and language preservation (Bird, 2020; Kraljevski et al., 2024; Pine et al., 2022).

2.1 Neural TTS Systems

Neural text-to-speech synthesis has advanced rapidly, from WaveNet's (van den Oord et al., 2016) neural waveform generation to Tacotron's (Wang et al., 2017) end-to-end approach. Subsequent architectures addressed efficiency and data constraints: FastSpeech (Ren et al., 2019) and FastSpeech 2 (Ren et al., 2021) introduced non-autoregressive generation, while Glow-TTS (Kim et al., 2020) applied flow-based modeling for efficient training. VITS (Kim et al., 2021) combined variational autoencoders with adversarial training in a fully end-to-end framework. Since then, further advances have emerged, including neural codec language models such as VALL-E (Wang et al., 2023), diffusion-based approaches like NaturalSpeech 2 (Shen et al., 2023), style-based models such as StyleTTS 2 (Li et al., 2023) and flow-matching methods like F5-TTS (Chen et al., 2025). However, these systems typically require substantially larger datasets and computational resources, and are not yet widely supported in open-source toolkits for low-resource language development. We implement and evaluate FastSpeech, Glow-TTS, and VITS for Myaamia, selecting these architectures for their proven effectiveness in low-resource settings and their availability in the Coqui TTS framework.

2.2 Low-Resource and Indigenous Language Speech Synthesis

Modern neural approaches have reduced the data requirements of TTS systems, making them increasingly viable for language preservation and education (Xu et al., 2020). However, challenges persist, including data scarcity, limited native speaker evaluation, and language-specific phonological complexities (Gumma et al., 2024; Hammerly et al., 2023). Earlier work includes rule-based synthesis for Navajo (Whitman et al., 1997), while more recent neural efforts include high-quality TTS for Kanyen'kéha (Mohawk), Plains Cree, SENĆOTEN (Pine et al., 2022, 2025), Võro (Rätsep and Fishel, 2023), Border Lakes Ojibwe (Hammerly et al., 2023), Mundari (Gumma et al., 2024), a multilingual system for Ojibwe, Mi'kmaq, and Maliseet (Wang et al., 2025), and Shipibo-Konibo (Menendez and Gomez, 2025). Our work builds on this foundation, extending neural TTS to Myaamia and contributing the first benchmark for this language.

3 Methodology

Figure 1 illustrates the Myaamia TTS pipeline, covering data preprocessing, model training and inference using FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS (Kim et al., 2021), and evaluation of synthesized speech.

3.1 Dataset: Myaamia-TTS

The dataset consists of 14,358 utterances (approximately 10.4 hours) of Myaamiaataweenki (Myaamia language) recordings created at the Myaamia Center since 2010. The recordings were collected in a controlled environment with minimal ambient noise and feature seven speakers, two of whom contributed approximately 95% of the recordings. Myaamiaataweenki employs a phonemic writing system derived from Americanist phonetic notation. The writing system has 20 basic symbols: 4 vowels, each of which can be short or long (a, aa, e, ee, i, ii, o, oo), and 12 consonants (p, t, k, c, s, š, h, m, n, l, w, y). Potential phonological challenges include preaspirated consonants (hp, ht, hk, hs, hš, hc), articulated with a brief puff of air before the consonant, and a set of vowel-devoicing rules. Still, the phoneme inventory overlaps heavily with English. Table 1 presents a representative subset of phonetic symbols along with example words where the symbol's sound is highlighted in blue.

The dataset consists of text sequences ranging from 3 to 163 characters. Because Myaamia employs a phonemic orthography with high grapheme-to-phoneme correspondence, we train our TTS systems at the character level, where each input character closely approximates a phoneme-level representation. This approach establishes a baseline for Myaamia TTS and allows us to empirically identify the specific cases where the character-phoneme mapping diverges, as discussed in Sections 4.3 and 5. Transcripts are drawn directly from the ILDA database, where they were authored and curated by Myaamia Center linguists using the same orthography taught to learners; no additional

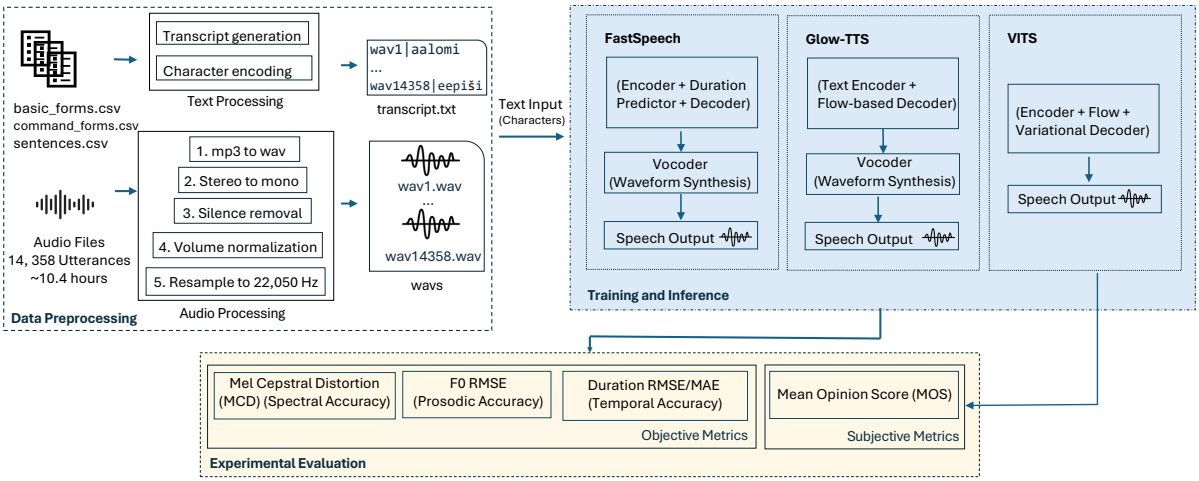

Figure 1: Overview of the text-to-speech (TTS) pipeline showing data preprocessing, training, and inference architectures (FastSpeech, Glow-TTS, and VITS), and evaluation metrics.

Table 1: Myaamia Phonetic Symbols and Examples. A representative subset is shown; the full inventory includes 20 symbols (see Section 3).

Symbol	Myaamia Example (Target Sound)	English Approximation (Similar Sound)
a	**a**ya	p**a**pa
ee	n**ee**we	b**ay**
p	aa**p**ooši	**p**ot
k	**k**iinte	**k**ey
hp	paa**hp**ilo	no English equivalent
ht	ee**ht**eeki	no English equivalent
nk	ii**nk**a	li**ng**er
nt	kii**nt**e	ti**nd**er

phonemic re-transcription was applied. Audio–transcript pairings and the exported character inventory were verified by Myaamia Center linguists before training, and no sub-utterance alignment was performed. Each model learns alignment internally from utterance-level pairs.

The majority of sequences fall within the shorter range, as shown in Table 2. Taken from dictionary recordings, the data falls into three major categories: basic expressions, command forms (imperatives), and example sentences. Below are examples from each category, all derived from the stem **ayaa-**, which means ‘go to a place’ in English.

- Basic form: **iiyaayani** (‘you go’)
- Command form: **ayaataawi** (‘let’s go!’)
- Example sentence: **nipwaantiikaaninkiši iiyaayaani** (‘I am going to school’)

Audio for the dataset was somewhat variable; files included three different sample rates: 44100 Hz (69.6%), 22050 Hz (29.8%), and 48000 Hz (0.6%). Recordings varied in duration from 0.52 to 20.06 seconds, with an average length of 2.61 seconds. Because two speakers contribute roughly 95% of the recordings, the synthesized voice, pronunciation, and prosody primarily reflect those two speakers rather than the broader community; we address this imbalance as a data limitation in Section 5.

Table 2: Distribution of Utterance Lengths in Characters

Character Count	Number of Utterances	Percentage	Total Characters
3-14	8,559	59.6%	96,117
15-25	5,131	35.7%	89,882
26-37	524	3.6%	15,651
38-83	140	1.0%	6,446
84-163	4	0.0%	511
Total	14,358	100.0%	208,607

3.2 Data Preparation and Standardization

The audio data consisted of stereo recordings in MP3 format. The corresponding transcripts were maintained in three database tables (*basic_forms*, *command_forms*, and *sentences*) within the Myaamia ILDA database and exported to CSV format. The preprocessing pipeline established a mapping between audio files and transcripts using unique identifiers from the database tables, generating a transcript.txt file containing paired audio filenames and transcription entries *(e.g., wav1|aahkohkimilo)*.

Stereo recordings were converted to mono format, and silence removal was applied using a threshold of -40 dB for segments exceeding 0.5 seconds. Volume normalization was performed to maintain consistent amplitude levels, and all audio was resampled to 22,050 Hz. Mono conversion was applied because the recordings are single-speaker and contain no meaningful stereo information. The 22,050 Hz target matches the Coqui TTS framework default and avoids upsampling the 29.8% of files already recorded at that rate, while retaining the full frequency range relevant to intelligible speech. Character encoding was applied to the transcript data to properly represent the special characters š and Š (IPA /ʃ/) in the Myaamia orthography. The standardized dataset was partitioned using an 80/20 split ratio, yielding 11,486 utterances for training and 2,872 for testing. The split was performed by random utterance-level sampling without explicit speaker or character stratification; given the size of the test set, the speaker distribution and character inventory in the training and test partitions are expected to closely mirror those of the full corpus. Total duration of the training data was 8 hours and 18 minutes, while the evaluation dataset was 2 hours and 4 minutes of audio. During training, the Coqui TTS framework automatically reserved a subset of the training partition for validation using its default *eval_split_size* of 0.01 (1%), yielding approximately 115 validation utterances per model. All models were trained for 1000 epochs, and convergence was monitored through training and validation loss curves.

3.3 Speech Synthesis For Myaamia

We explored three TTS architectures of increasing complexity: FastSpeech (Ren et al., 2019), Glow-TTS (Kim et al., 2020), and VITS (Kim et al., 2021). The text processing pipeline supported all Myaamia characters, including the special characters š and Š, along with punctuation and numbers. Our study began with FastSpeech's parallel sequence generation, followed by Glow-TTS's efficient training and robust voice conversion via normalizing flows, and concluded with VITS's end-to-end variational autoencoder (VAE) and adversarial training. VITS has shown promise for low-resource languages, with recent applications in African languages (Ogun et al., 2024), Mundari (Gumma et al., 2024), and Ojibwe (Hammerly et al., 2023). All models were trained on an NVIDIA A30 GPU (24GB) with a batch size of 32 for 1000 epochs. VITS used mixed precision (fp16) with an AdamW optimizer (lr = 0.001) and a text encoder featuring six layers, two attention heads, and 768 FFN channels. Its loss function prioritized mel loss (45.0) over KL, generator, discriminator, and duration losses. Glow-TTS used an RAdam optimizer (lr = 0.001) with a NoamLR scheduler (4,000 warmup steps) and a relative positional transformer encoder. FastSpeech applied the Adam optimizer (lr = 0.0001) with NoamLR and FFTransformers for both encoder and decoder. Hyperparameters follow the Coqui TTS default configurations for each architecture; no systematic hyperparameter search was performed.

The Real-Time Factor (RTF) measures processing time relative to audio duration, with values below 1.0 indicating real-time capability. To reflect realistic deployment conditions, RTF values were measured using CPU inference across all 2,872 test samples. FastSpeech achieved the best efficiency among the three models (RTF: 7.43 ± 3.00) due to its feed-forward architecture. Glow-TTS had moderate speed (RTF: 8.59 ± 3.29) as normalizing flows added computational overhead. VITS was the slowest (RTF: 12.71 ± 4.53) due to its complex VAE-GAN architecture, though this trade-off was justified by superior output quality (see Section 4). While none of the models achieve real-time CPU inference, all remain practical for offline generation in educational contexts. Training times followed a similar pattern: FastSpeech (27.69h), Glow-TTS (33.87h), and VITS (60.10h).

4 Experimental Evaluation

Our evaluation used 2,872 test samples (20% of the total dataset) to assess all three TTS architectures through objective and subjective metrics.

4.1 Objective Evaluation

We evaluated synthesis quality using three objective metrics. For all metrics, we extracted features from both reference and synthesized speech, applied dynamic time warping (DTW) for temporal alignment, and computed error measures between corresponding frames. Mel Cepstral Distortion (MCD) (Kubichek, 1993) assessed spectral quality by computing the difference between mel-frequency cepstral coefficients (MFCCs), excluding the 0th coefficient (Vasilijevic and Petrinović, 2011). Lower MCD values indicate greater spectral similarity. F0 RMSE (Tsanas et al., 2014) quantified prosodic accuracy by measuring pitch contour deviation between synthesized and reference speech (Luo et al., 2016). Duration RMSE (Henter et al., 2017) evaluated temporal accuracy by comparing segment lengths between synthesized and reference speech. Table 3 summarizes the results.

VITS achieved the lowest mean MCD (15.22) and F0 RMSE (17.50), indicating better spectral and pitch replication than other models. Glow-TTS showed intermediate performance (MCD: 17.91, F0 RMSE: 20.00), while FastSpeech had the largest deviations (MCD: 19.82, F0 RMSE: 30.51). All three models performed similarly in duration metrics, with RMSE values ranging from 0.29 to 0.31, suggesting comparable speech timing capabilities. However, MCD and F0 RMSE values were higher than those typically reported for well-resourced TTS systems (Kominek et al., 2008), likely reflecting the constraints of low-resource training. Despite VITS achieving the best results among the three models, further refinements are needed to enhance phonetic precision and natural prosody. Figure 2 presents a spectrogram comparison of the utterance *kiihkikaateeŝwilo (‘amputate my foot!’)* across all three models and the original recording, illustrating the spectral differences reflected in the objective metrics.

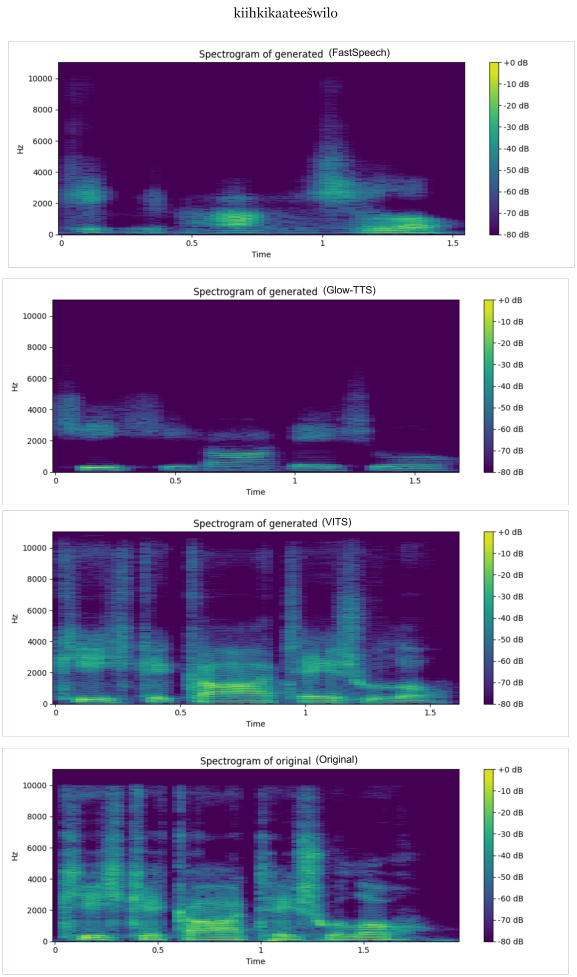

Figure 2: Spectrograms of the utterance *kiihkikaateeŝwilo* generated by FastSpeech, Glow-TTS, VITS, and the original reference audio.

4.2 Subjective Evaluation

To evaluate perceived quality and identify areas for improvement, we conducted a targeted subjective evaluation with four Myaamia experts using audio synthesized by the VITS model. Two of the experts were primary contributors to the recordings used to train the model, while the other two are linguists specializing in Myaamia. The Myaamia speaker community is extremely small, and only a limited number of additional speakers are available. Moreover, several evaluation criteria, including judgments on phonemic contrasts, prosodic alignment, and phonetic accuracy, require specialized linguistic expertise beyond conversational fluency, which further constrained participant selection. Future work will expand the evaluation to include additional community members, as discussed in Section 5. The study adhered to ethical research guidelines, as detailed in Section 6. Participants rated 20 audio samples: 10 evaluating general perceptual measures (naturalness, intelligibility, and overall quality) and 10 focusing on language-specific features (intonation, rhythm, and linguistic accuracy). Each set contained an equal number of synthesized and original recordings, with stratified randomization employed to minimize bias. Ratings were provided on a 5-point Likert scale, where 1 indicated “poor” quality (least natural, least intelligible, or least accurate), and 5 indicated “excellent” quality

Table 3: Objective Quality Metrics for FastSpeech, Glow-TTS, and VITS Text-to-Speech Models. The best score is represented in bold.

Metric	FastSpeech	Glow-TTS	VITS
MCD ↓	19.82 ± 2.33	17.91 ± 2.19	**15.22 ± 2.66**
F0 RMSE ↓	30.51 ± 29.93	20.00 ± 23.73	**17.50 ± 20.69**
Duration RMSE ↓	0.31 ± 0.37	0.30 ± 0.36	**0.29 ± 0.40**

(most natural, most intelligible, or most accurate).

Table 4 summarizes the subjective evaluation. Although naturalness and understandability scores are lower than those of the original recordings, they fall within the expected range given the limited data. Notably, the model achieves an intonation and rhythm score of 3.52 ± 0.81, indicating a strong capacity to capture essential prosodic features of Myaamia. However, to more closely approximate the fluidity and authenticity of native speech, further refinements in stress alignment and phonetic articulation are necessary.

As shown in Table 5, our system demonstrates competitive performance compared to other low-resource TTS models, despite being trained on a more limited dataset. Mundari TTS, trained on 24.76 hours of data, achieves a higher MOS for naturalness (3.69 ± 1.18) (Gumma et al., 2024), while our model scores 3.05 ± 0.89 with significantly less data. Similarly, Võro TTS uses 17 hours of training data, including 14 hours of Estonian, yet our model performs competitively despite relying solely on Myaamia data (Rätsep and Fishel, 2023).

4.3 Error Analysis

To identify specific phonological challenges in the synthesized speech, we analyzed qualitative feedback provided by expert evaluators on the VITS-generated samples. Each evaluator rated five synthesized utterances drawn from distinct subsets of the test data. The model exhibited difficulty with consonant distinctions specific to Myaamia. In the utterance *noonki šayiipaawe aalaankwiaani ('I'm tired this morning')*, one evaluator noted that the voiceless fricative *š* was realized as the affricate /tš/ (orthographic *c* in Myaamia), collapsing a phonemic contrast. In *maalami eelaamhsenki ('it is too windy')*, the same evaluator observed that the consonant cluster /mhs/ appeared to be reduced or realized closer to /nk/, suggesting the model may struggle to maintain multi-segment consonant sequences. In *weelaantaweeyani ('you are climbing')*, the lateral approximant /l/ was perceived as the labial-velar approximant /w/ by another evaluator.

Beyond consonant-specific issues, multiple evaluators noted reduced consonant clarity across their respective sample sets. Vowel length, which is phonemically contrastive in Myaamia (e.g., *a* vs. *aa*), was also affected. Two evaluators independently reported issues with vowel duration, including cases where vowels were perceived as longer than expected and instances of incorrect vowel length in some utterances (e.g., *iihia ('yes'), eetiliwatenki ('it is thin ice')*; *pinšiwa awiilawi meeneehwiki ('the cat food is gone')*). Given that the character-based input represents long vowels as doubled characters, this may reflect limitations in the duration model's handling of repeated graphemes. Evaluators also reported audible artifacts, including glitching, skipping, and unnatural segmentation at syllable boundaries. One evaluator identified specific positions within *aayaapweeyohsiaanki ('we take a walk')* where synthesis degraded, while another noted that the final syllable of *keekiipiinkweeholaci ('you blindfold him/her')* was choppy. These issues were particularly noticeable in longer, multi-word utterances. No comparable issues were reported in evaluations of the original recordings, suggesting these are model-specific limitations. These findings inform the directions for improvement discussed in Section 5.

5 Limitations and Future Work

While this work demonstrates the feasibility of neural TTS for Myaamia, several limitations remain.

Modeling Limitations. Our system uses character-based input rather than phoneme-based representations, which may have contributed to some of the phonetic issues identified in Section 4.3, particularly consonant conflation and vowel length errors. A hybrid character-phoneme approach, as explored for Mundari (Gumma et al., 2024), may improve pronunciation accuracy while maintaining flexibility. Additionally, the current system does not explicitly model prosodic features such as pitch, stress, or rhythm. Although the

Table 4: Subjective evaluation of synthesized Myaamia speech vs. original recordings, based on expert ratings. Mean Opinion Score (MOS) with standard deviations reported.

Metric	Generated Speech (MOS ↑ ± SD)	Original Speech (MOS ↑ ± SD)
Naturalness	3.05 ± 0.89	**4.79 ± 0.42**
Understandability	3.05 ± 1.10	4.55 ± 0.69
Overall Quality	3.20 ± 0.77	4.75 ± 0.44
Intonation/Rhythm	**3.52 ± 0.81**	4.52 ± 0.56
Linguistic Quality	3.40 ± 0.91	4.45 ± 0.56

Table 5: Naturalness comparison of TTS models across languages using Mean Opinion Score (MOS) and standard deviation (SD) on a 5-point Likert scale. Training data size (hours) included.

TTS Model (Language)	Training Data Size (Hours)	Naturalness (MOS ↑ ± SD)
Myaamia TTS (Ours)	8.18	3.05 ± 0.89
Võro TTS (Rätsep and Fishel, 2023)	17.00	3.62 ± 0.15
Mundari TTS (Gumma et al., 2024)	24.76	3.69 ± 1.18

VITS model achieved reasonable intonation scores (3.52 ± 0.81), evaluators noted audible artifacts in longer utterances (Section 4.3). Incorporating prosodic embedding layers and attention-based duration predictors may help address both issues (Henter et al., 2017).

Data and Deployment. The dataset imbalance noted in Section 3.1 limits generalization across speaker variation. Future work should prioritize expanding and diversifying the corpus through new recordings from additional speakers. Inference speed also remains a practical constraint for mobile or low-resource deployment; model compression techniques such as pruning or knowledge distillation could improve feasibility for integration into platforms like the Šaapohkaayoni Education Portal. Ongoing collaboration with the Myaamia Center remains essential to ensure that synthesized speech aligns with the linguistic expectations of speakers and learners (Baldwin et al., 2016; Brinklow, 2021).

Evaluation Limitations. The subjective evaluation was limited to four experts, and only the VITS model was assessed. While this reflects the small size of the Myaamia speaker community and the specialized linguistic expertise required for judgments on phonemic contrasts and prosodic alignment, it limits the generalizability of perceptual findings. Future evaluations should include additional community members at varying levels of Myaamia proficiency, to examine whether speakers at different proficiency levels perceive synthesized speech differently. Such evaluations should also assess all three models and report inter-rater agreement metrics to strengthen the evaluation methodology.

6 Ethical Considerations and Sovereignty

This research is grounded in the principles of Indigenous Data Sovereignty and aligns with the CARE Principles for Indigenous Data Governance (Carroll et al., 2020). The Myaamia Center maintains full ownership and governance over all linguistic and cultural data used in this study. The speech corpus was curated through the Indigenous Languages Digital Archive (ILDA), a platform purpose-built to support the reclamation goals of tribal communities, and all work was conducted under Miami University IRB protocol 05009e through a long-standing partnership with the Myaamia Center. Prior to the initiation of this research, the use of ILDA audio data for TTS development was approved by the director of the Myaamia Center, and consent was obtained from the primary speakers whose recordings comprise the dataset. All model training and evaluation were carried out on Miami University managed computing resources governed by the same institutional access controls as the ILDA database; no data was shared with third

parties outside this environment.

From its inception, this project has been developed in collaboration with the Myaamia Center, with center staff and affiliated experts providing guidance on data use, model development, and evaluation criteria. To ensure the resulting technology aligns with both cultural values and pedagogical needs, the subjective evaluation was carried out by linguistic experts and the primary speakers who contributed to the original recordings. Their direct participation means the synthesized speech is validated by the very individuals whose voices the models aim to represent. By centering this expertise at every stage, this work seeks to provide a sustainable, ethically governed tool in support of ongoing Myaamia language reclamation efforts.

7 Conclusion

This work presents the first neural text-to-speech system for the Myaamia language. We evaluated three architectures, FastSpeech, Glow-TTS, and VITS, trained on 8.18 hours of community-curated speech from the Indigenous Languages Digital Archive. VITS achieved the best performance across objective metrics (MCD: 15.22, F0 RMSE: 17.50), and subjective evaluation by Myaamia experts yielded encouraging results, particularly for intonation and rhythm (MOS: 3.52 ± 0.81). At the same time, error analysis revealed specific challenges in consonant distinctions, vowel length accuracy, and audible artifacts in longer utterances, indicating clear directions for future improvement.

By combining objective metrics with expert evaluation from original speakers and linguists, we established a performance baseline grounded in community standards. While further refinement is needed, particularly in phoneme-level modeling, prosodic control, and dataset diversification, this work provides a foundation for integration into community learning platforms such as the Šaapohkaayoni Education Portal. Our ongoing collaboration with the Myaamia Center ensures that development proceeds with appropriate oversight and accountability.

References

Daryl Baldwin, David J. Costa, and Douglas Troy. 2016. Myaamiaataweenki eekincikoonihkiinki eeyoonki aapisaataweenki: A miami language digital tool for language reclamation. *Language Documentation and Conservation*, 10:394–410.

Steven Bird. 2020. Decolonising speech and language technology. In *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pages 3504–3519. Association for Computational Linguistics (ACL).

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *WINHEC: International Journal of Indigenous Education Scholarship*, 16(1):239–266.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):1–12.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Varun Gumma, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri, and Kalika Bali. 2024. MunTTS: A text-to-speech system for Mundari. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 76–82, St. Julians, Malta. Association for Computational Linguistics.

Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous, and Chad Quinn. 2023. A text-to-speech synthesis system for border lakes ojibwe. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 60–65. Association for Computational Linguistics.

Gustav Eje Henter, Srikanth Ronanki, Oliver Watts, and Simon King. 2017. Non-parametric duration modelling for speech synthesis with a joint model of acoustics and duration. In *Proceedings of Interspeech*, pages 1213–1217, Stockholm, Sweden. ISCA.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *Advances in Neural Information Processing Systems*, volume 33.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

John Kominek, Tanja Schultz, and Alan W. Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proceedings of*

the First Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2008), pages 63–68, Hanoi, Vietnam.

Ivan Kraljevski, Frank Duckhorn, Daniel Sobe, Constanze Tschoepe, and Matthias Wolff. 2024. Preserving language heritage through speech technology: The case of upper sorbian. In *Proceedings of the 26th International Conference on Speech and Computer (SPECOM 2024)*, pages 3–12, Belgrade, Serbia. Springer Nature.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Preprint*, arXiv:2306.07691.

Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki. 2016. Emotional voice conversion using neural networks with different temporal scales of f0 based on wavelet transform. In *Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW9)*, pages 238–243, Sunnyvale, CA, USA. ISCA.

Daniel Menendez and Hector Gomez. 2025. Text-to-speech system for low-resource languages: A case study in Shipibo-konibo (a Panoan language from Peru). In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–7, Albuquerque, New Mexico. Association for Computational Linguistics.

Sewade Ogun, Abraham T. Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A. Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis. *Preprint*, arXiv:2406.11727.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for indigenous language education. *Computer Speech & Language*, 90:101723.

Aidan Pine, Dan Wells, Nathan Brinklow, Patrick Littell, and Korin Richmond. 2022. Requirements and motivations of low-resource speech synthesis for language revitalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 7346–7359. Association for Computational Linguistics.

Liisa Rätsep and Mark Fishel. 2023. Neural text-to-speech synthesis for võro. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 723–727, Tórshavn, Faroe Islands. University of Tartu Library.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations (ICLR)*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, volume 32.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *Preprint*, arXiv:2304.09116.

Athanasios Tsanas, Matías Zañartu, Max A. Little, Cynthia Fox, Lorraine O. Ramig, and Gari D. Clifford. 2014. Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering. *The Journal of the Acoustical Society of America*, 135(5):2885–2901.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *Preprint*, arXiv:1609.03499.

Antonio Vasilijevic and Davor Petrinović. 2011. Perceptual significance of cepstral distortion measures in digital speech processing. *Automatika*, 52:132–146.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Shenran Wang, Changbing Yang, Michael l Parkhill, Chad Quinn, Christopher Hammerly, and Jian Zhu. 2025. Developing multilingual speech synthesis system for Ojibwe, mi'kmaq, and maliseet. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 817–826, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *Preprint*, arXiv:1703.10135.

Robert Whitman, Richard Sproat, and Chilin Shih. 1997. A navajo language text-to-speech synthesizer. Technical Report 11222-930830-13TM, AT&T Bell Laboratories.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. *Preprint*, arXiv:2008.03687.

Evaluating Frontier LLM Translation Capability for Lakota

Lance Robertson
University of California, San Diego
lrobertson@ucsd.edu

Abstract

We evaluate seven large language models—four proprietary and three open-weight—on bidirectional Lakota–English translation using 200 sentence pairs from the New Lakota Dictionary. Each model is evaluated with and without extended reasoning, where the provider's API permits. The best model (Gemini 3.1 Pro) achieves a mean chrF++ of 59.4 on Lakota→English and 42.6 on English→Lakota; the strongest open-weight model trails the proprietary leaders, and no model produces reliable translation in either direction. Two independent LLM judges from different model families agree substantially (Cohen's $\kappa = 0.75$) that semantic equivalence ranges from 6% (GPT-5.2) to 60% (Gemini), diverging substantially from chrF++ scores. For the open-weight models, enabling reasoning changes refusal behavior far more than translation quality: it surfaces the limitation rather than overcoming it. Diacritic-normalization analysis shows models produce roughly correct base characters but place diacritical marks inconsistently. All results and evaluation code are publicly available at https://github.com/robotson/lakota-translation-benchmark.

1 Introduction

Lakota (*Lakȟótiyapi*) is a Siouan language spoken primarily on reservations in North and South Dakota. UNESCO classifies it as critically endangered (Moseley, 2010), with fewer than 2,000 first-language speakers, most over 65 years of age. Language revitalization efforts are ongoing across multiple tribal and academic institutions. The Lakota Language Consortium (LLC) publishes the New Lakota Dictionary (NLD) and associated pedagogical materials; the NLD serves as the source for this study's evaluation set.

Contemporary large language models (LLMs) support translation across dozens of languages, but the extent of training data for any given language is not disclosed. Users may reasonably expect similar quality across supported languages. For critically endangered languages like Lakota—where digitized text is scarce and no large-scale parallel corpus exists—this expectation is untestable without direct evaluation.

This paper reports the results of a direct evaluation. We tested seven LLMs—four proprietary and three open-weight—on 200 Lakota–English sentence pairs in both translation directions, under two experimental conditions designed to test the impact of extended thinking (chain-of-thought reasoning) on translation quality. We report chrF++ and BLEU scores, diacritic normalization analysis, model self-reported confidence, and refusal behavior. Related work on LLM translation of low-resource languages includes the MTOB benchmark (Tanzer et al., 2024), which evaluated frontier LLMs on Kalamang translation using in-context learning from a reference grammar, and the AmericasNLP shared tasks (Mager et al., 2021; Ebrahimi et al., 2023), which benchmark MT systems on indigenous American languages but do not include Lakota. The only prior Lakota-specific language model is LakotaBERT (Parankusham et al., 2025), a RoBERTa-based model trained for masked language modeling. To our knowledge, no prior study has evaluated zero-shot frontier LLM translation quality for Lakota.

Lakota presents specific challenges for LLM translation. It is polysynthetic, with extensive verbal morphology including person mark-

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 11–21

ing, aspect, mood, and evidentiality encoded through affixes and enclitics. The Standard Lakota Orthography (SLO) uses Latin script with diacritical marks: caron-marked consonants (č, š, ž, ǧ, ȟ), ogoneks on vowels for nasalization (ą, į, ų), acute accent for stress, and the glottal stop marker (’). The velar nasal is written with a separate character, ŋ. These diacritical marks distinguish phonemes—for example, *h* and *ȟ* represent different consonants—making them significant for both translation accuracy and character-level evaluation metrics.

2 Methodology

2.1 Evaluation Set

The evaluation set consists of 200 sentence pairs sourced from published Lakota learning materials for research purposes. Pairs were drawn from conversational examples in the New Lakota Dictionary (NLD) (Ullrich, 2008), published by the Lakota Language Consortium (LLC), and selected from entries appearing under dictionary headwords. The first author filtered pairs for naturalness and conversational relevance. Each pair contains one Lakota sentence and one English translation. Examples range from short questions (*Hé yačhíŋ he?* / “Do you want it?”) to multi-clause sentences with cultural context (*Wóčhičiyakiŋ kte ȟčiŋ.* / “I want to talk to you.”). The evaluation set is conversational register throughout.

The NLD uses the Standard Lakota Orthography (SLO), developed by the LLC. Multiple orthographic conventions have been in use across Lakota language communities, including the Buechel missionary system and the orthography used at Sinte Gleska University (White Hat, 1999). Standardization of Lakota orthography remains a contested issue within Lakota communities (Hauff, 2020). Our use of NLD materials was practical—it is the most accessible published collection of Lakota sentence pairs with English translations available to us—and should not be read as an endorsement of any particular orthographic standard or institutional approach to language documentation.

Each pair includes a conversational score (4–8) assigned during curation, where higher values indicate more everyday, conversational language and lower values indicate more formal or structurally complex sentences. The evaluation set also records the headword context from which each pair was drawn. The same 200 pairs were used for all models and both experimental conditions.

2.2 Models

Seven models were evaluated. Four are proprietary frontier models from three providers: Gemini 3.1 Pro Preview (Google), Claude Opus 4.6 (Anthropic), Claude Sonnet 4.6 (Anthropic), and GPT-5.2 (OpenAI). Three are open-weight frontier models, accessed through Together AI: DeepSeek-V4-Pro, Qwen3.6-Plus, and GLM-5.1. All models were accessed through litellm (v1.63), a unified API wrapper; the open-weight models were routed through an OpenAI-compatible endpoint to preserve the same structured-output contract. Selection criteria: frontier-tier models supporting structured JSON output with schema enforcement. Including open-weight models tests whether the closed–open capability gap observed on higher-resource tasks holds for Lakota.

2.3 Experimental Conditions

Two conditions were run on the same evaluation set:

Baseline: Temperature 0, no extended thinking. Maximum output tokens: 1,024. This condition uses near-deterministic decoding and represents baseline model capability.

Thinking: Extended thinking enabled at maximum provider settings. Maximum output tokens: 16,384 (to accommodate thinking token overhead). Claude models used `thinking.budget_tokens` = 8,192; GPT-5.2 used `reasoning_effort` = high; Gemini used `thinkingLevel` = high.

Gemini 3.1 Pro enables thinking by default and cannot disable it, so its baseline already includes default-level thinking; the baseline→thinking comparison isolates only the increase from default to high, not the enabling of thinking.

For Claude models, enabling thinking fixes temperature at 1 (it cannot be set independently while thinking is on); for

GPT-5.2, high-reasoning configurations ignore sampling parameters, so `temperature=1` likely had no effect. These proprietary baseline→thinking comparisons should therefore be read as "baseline vs. maximum supported reasoning settings" rather than clean temperature-controlled ablations. The three open-weight models (served via Together AI) instead toggle reasoning directly: thinking uses the provider's default reasoning and baseline disables it via the documented control—`reasoning.enabled` for DeepSeek, `chat_template_kwargs.enable_thinking` for Qwen and GLM—holding each model's output-format regime fixed, a cleaner within-model ablation.

2.4 Structured Output Schema

All models produced structured JSON output enforced via provider-native schema validation with three fields: `translation` (best translation or empty string), `confidence` (0.0–1.0), and `refusal_reason` (string or null).

The system prompt was identical across all models and conditions:

> You are a translation system. Translate the input text. Respond as JSON with three fields: "translation" (your best translation, or empty string if you cannot translate), "confidence" (a number from 0.0 to 1.0 representing your confidence in the translation quality), "refusal_reason" (null if you attempted a translation, or a brief explanation if you cannot translate).

User prompts followed the format "Translate from Lakota to English: {text}" and "Translate from English to Lakota: {text}" for the two directions.

2.5 Metrics

chrF++ (a character n-gram F-score extended with word n-grams) was the primary metric, computed via SacreBLEU (Popović, 2015, 2017) with default parameters: character n-gram order 6, word n-gram order 2, $\beta = 2$. chrF++ is preferred over BLEU for morphologically rich languages because it operates at the character level and does not penalize valid morphological variants as harshly as word-level metrics. **BLEU** was computed as a secondary metric for comparability.

Diacritic-normalized chrF++ was computed for English→Lakota translations by stripping Unicode combining marks (categories Mn and Mc) from both candidate and reference via NFD decomposition before scoring. NFD decomposition ensures canonical equivalence between precomposed characters (e.g., U+01CE, ǎ) and combining sequences (e.g., a + U+030C) before mark removal. This normalization removes carons, ogoneks, and stress marks in SLO; the character ŋ (eng) and the glottal stop marker (') are unaffected. The delta between raw and normalized chrF++ isolates orthographic variation attributable to these diacritical marks from broader lexical and morphological differences.

LLM semantic judgment was used to evaluate Lakota→English translations, where chrF++ is known to penalize valid paraphrases against single references. We initially ran BERTScore (Zhang et al., 2020) (roberta-large) on the proprietary models to capture semantic matches that chrF++ misses, but it was uninformative: F1 ranged from 0.906 to 0.959 regardless of actual translation quality. At the pair level, GPT-5.2's translation of "Do you have a car?" as "Do you have a horse?" received BERTScore 0.975. We did not extend the BERTScore pass to the open-weight models added later, having moved to LLM-based semantic judgment as our primary check. We instead used a separate LLM (Gemini 3 Flash Preview) to judge whether each hypothesis–reference pair conveys the same meaning, following the GEMBA framework (Kocmi and Federmann, 2023) but simplified to English-to-English semantic equivalence. Because both strings are English, the judge requires no knowledge of Lakota—it evaluates only whether the model's English output preserves the meaning of the English reference. Each pair was rated `equivalent`, `partially_equivalent`, or `not_equivalent`. All Lakota→English translation attempts across 14 model×condition combinations were judged. To check for family-level self-judging bias, every pair was also judged by Llama-3.3-70B-Instruct-Turbo, an open-weight model from a developer not among those evaluated; agreement is reported in §3.6.

Self-reported confidence (0.0–1.0) was extracted from the structured output.

Model	L→E chrF++ (σ)	E→L chrF++ (σ)
Gemini 3.1 Pro	63.1 (3.3)	33.6 (3.4)
Claude Opus 4.6	45.1 (2.6)	29.8 (4.8)
Claude Sonnet 4.6	42.7 (4.8)	20.8 (4.6)
GPT-5.2	25.5 (5.7)	15.9 (5.8)

Table 1: Pilot variance: mean chrF++ and cross-run standard deviation across 3 runs (20 pairs), four proprietary models (February 2026). The open-weight models, added in May 2026, were not separately re-piloted.

2.6 Variance

A pilot study was conducted before the full thinking evaluation to assess run-to-run variance under the thinking condition's decoding settings. Twenty sentence pairs were sampled and each of the four proprietary models was run 3 times on both directions (480 API calls). Cross-run standard deviation ranged from 2.6 (Opus L→E) to 5.8 (GPT-5.2 E→L). Model rankings were stable across all three runs (Table 1). Based on these results, the full evaluation was conducted as a single run per condition.

2.7 Data Contamination

A central concern with LLM-based evaluation on low-resource pairs is that the evaluation set may appear in pretraining data. The evaluation set here is drawn from the New Lakota Dictionary (Ullrich, 2008), whose contents are findable on the open web in OCR'd reproductions of varying quality; any web-scale training corpus could plausibly include some form of this material.

To bound the empirical risk we searched indexed open-web content for all 200 pairs. The Lakota source string appears verbatim for about 10% of pairs, but the Lakota sentence and its English reference *together* for only 0.5%—almost always in scanned copies of the dictionary itself. This distinction matters: controlled studies find that aligned source–target overlap can substantially inflate translation scores, while source- or target-only overlap produces smaller, less consistent inflation, at the 1B–8B scales tested (Kocyigit et al., 2025). The overlap we observe is almost all the one-sided kind.

The score distribution points the same way: a memorized test set would cluster near ceiling, whereas ours ranges widely (6–61% semantic equivalence across the seven models), consistent with genuine partial capability rather than wholesale recall—though, as §4 notes, the most common phrases are likely memorized.

2.8 Procedure

Each evaluation run called the model API once per (pair, direction) combination: 200 pairs × 2 directions = 400 calls per model, 1,600 calls per condition. A 1.5-second delay was inserted between calls. API timeout was set to 60 seconds for baseline and 180 seconds for thinking. Failed calls were retried up to 3 times with exponential backoff. Proprietary models were evaluated in February–March 2026 and open-weight models in May 2026; exact model identifiers (API versions current on those dates) are listed in the released code. Results were written to per-model JSONL files with per-pair metadata.

3 Results

3.1 Translation Quality

Table 2 reports the best-condition results for each model. For Gemini, baseline and thinking scores are nearly identical (§3.2); thinking is reported here for consistency since it includes confidence data. For all other models, thinking is reported as the higher-scoring condition.

Gemini leads in both directions. Opus and Sonnet score within 2 chrF++ points of each other on L→E (Sonnet slightly ahead), but Opus leads Sonnet by 7 points on E→L. GPT-5.2 trails the proprietary models in both directions. Among the open-weight models, DeepSeek-V4-Pro is strongest (38.0 / 29.2 chrF++), placing between GPT-5.2 and the Claude models, while GLM-5.1 and Qwen3.6-Plus score at or below GPT-5.2; no open-weight model approaches Gemini. Every model scores substantially higher on Lakota→English than English→Lakota. The gap ranges from 6.4 points (GPT-5.2) to 19.1 points (Sonnet). This asymmetry is consistent across all models: Lakota→English (producing English from Lakota input) scores higher than English→Lakota (producing Lakota from

Model	Dir	N	chrF++	σ (pairs)	Median	BLEU
Gemini 3.1 Pro	L→E	199	**59.4**	27.6	56.8	43.5
Claude Opus 4.6	L→E	200	45.9	25.0	43.5	27.2
Claude Sonnet 4.6	L→E	189	46.1	25.9	44.3	28.1
GPT-5.2	L→E	194	30.0	21.8	21.4	15.4
DeepSeek-V4-Pro	L→E	200	38.0	26.0	30.9	23.3
GLM-5.1	L→E	184	29.3	18.8	24.5	14.2
Qwen3.6-Plus	L→E	198	26.1	21.9	18.2	14.4
Gemini 3.1 Pro	E→L	199	**42.6**	22.1	38.9	24.7
Claude Opus 4.6	E→L	200	34.3	16.0	31.8	18.8
Claude Sonnet 4.6	E→L	196	27.0	13.2	24.2	15.7
GPT-5.2	E→L	184	23.6	12.5	20.5	14.3
DeepSeek-V4-Pro	E→L	195	29.2	14.7	26.9	17.0
GLM-5.1	E→L	193	14.6	7.2	12.8	9.3
Qwen3.6-Plus	E→L	164	18.4	8.7	17.2	13.3

Table 2: Translation quality by model and direction. Proprietary models are shown in their best condition; open-weight models (DeepSeek, GLM, Qwen) in their reasoning-enabled condition. N $<$ 200 reflects refusals, empty responses, or API errors excluded from scoring (§3.5). σ is across sentence pairs, not across runs.

Model	Dir	≥80	≥60	≥40	P10
Gemini	L→E	24%	47%	72%	21.6
Opus	L→E	12%	25%	56%	15.9
Sonnet	L→E	12%	25%	55%	16.0
GPT-5.2	L→E	5%	10%	23%	10.4
DeepSeek	L→E	8%	22%	36%	11.0
GLM	L→E	3%	7%	20%	11.8
Qwen	L→E	4%	10%	17%	8.2
Gemini	E→L	7%	20%	47%	17.2
Opus	E→L	1%	6%	29%	16.0
Sonnet	E→L	1%	1%	15%	12.9
GPT-5.2	E→L	1%	1%	8%	12.1
DeepSeek	E→L	1%	4%	21%	13.0
GLM	E→L	0%	0%	0%	7.4
Qwen	E→L	0%	0%	2%	8.9

Table 3: Score distribution by chrF++ threshold (thinking condition). Percentages over scored translations; P10 = 10th percentile.

Model	Dir	Sc. 4	Sc. 5	Sc. 6
Gemini	L→E	52.7	58.2	81.9
Opus	L→E	41.6	41.0	65.6
Sonnet	L→E	40.3	45.1	64.1
GPT-5.2	L→E	27.9	23.7	45.7
DeepSeek	L→E	33.0	33.9	59.6
GLM	L→E	27.3	26.0	41.0
Qwen	L→E	23.3	22.1	41.7
Gemini	E→L	37.1	42.7	61.3
Opus	E→L	31.2	32.7	45.6
Sonnet	E→L	24.7	26.2	35.2
GPT-5.2	E→L	22.2	21.5	30.4
DeepSeek	E→L	29.4	26.3	31.0
GLM	E→L	13.0	16.5	16.3
Qwen	E→L	17.3	17.3	21.9

Table 4: Mean chrF++ by conversational score (thinking condition). Score 4 = more formal/complex; score 6 = most conversational. Scores 7–8 excluded (n = 6 combined).

English input) for every model.

Standard deviations are large—22–28 points for L→E and 13–22 points for E→L—indicating that per-pair quality varies widely. Some pairs score 100.0 (exact character match to reference) while others score below 10. Table 3 reports the distribution of scores across chrF++ thresholds and tail behavior (10th percentile).

Table 4 stratifies chrF++ by the evaluation set's conversational score (§2.1). For Gemini L→E, the most conversational pairs (score 6) average 81.9 chrF++ while more complex pairs (score 4) average 52.7—a 29-point gap. The pattern holds across all models and both directions, suggesting that performance is concentrated on high-frequency phrasebook-like items, consistent with the finding that low-resource LLM translation quality tracks test-set coverage (Aycock et al., 2025).

Even the best model (Gemini L→E) produces translations scoring ≥80 chrF++—roughly usable quality—on only 24% of pairs. For E→L, only Gemini exceeds 5% at this threshold. If "reliable" is operationalized as ≥60 chrF++ on a majority of pairs, no model qualifies in either direction.

Model	Dir	Base	Think	Δ
Opus 4.6	L→E	45.0	45.9	+0.9
Opus 4.6	E→L	27.4	34.3	+6.9
Sonnet 4.6	L→E	45.4	46.1	+0.7
Sonnet 4.6	E→L	24.8	27.0	+2.2
GPT-5.2	L→E	26.4	30.0	+3.6
GPT-5.2	E→L	18.3	23.6	+5.3
Gemini	L→E	58.4	59.4	+1.0
Gemini	E→L	40.4	42.6	+2.2
DeepSeek	L→E	40.1	38.0	−2.1
DeepSeek	E→L	26.4	29.2	+2.8
GLM	L→E	27.8	29.3	+1.5
GLM	E→L	13.6	14.6	+1.0
Qwen	L→E	25.2	26.1	+0.9
Qwen	E→L	15.6	18.4	+2.8

Table 5: Baseline→thinking comparison

Model	Raw	Norm.	Δ
Gemini 3.1 Pro	42.6	48.7	+6.1
Claude Opus 4.6	34.3	40.8	+6.5
Claude Sonnet 4.6	27.0	32.3	+5.3
GPT-5.2	23.6	27.2	+3.6
DeepSeek-V4-Pro	29.2	33.6	+4.4
GLM-5.1	14.6	18.1	+3.5
Qwen3.6-Plus	18.4	21.7	+3.3

Table 6: Diacritic normalization, English→Lakota (thinking condition).

3.2 Effect of Extended Thinking

Table 5 compares baseline (near-deterministic decoding, no extended thinking) and thinking (maximum supported reasoning settings) for each model.

Among proprietary models, thinking improved chrF++ across all model-direction pairs, with E→L gains (2.2–6.9 points) consistently larger than L→E gains (0.7–3.6 points)—reasoning helps more on the harder generation direction. For Gemini, whose baseline already includes default-level thinking, the L→E delta is small (+1.0) while E→L still gains (+2.2). The open-weight models show only marginal chrF++ movement (-2.1 to $+2.8$ points); unlike the proprietary frontier, reasoning does little for their translation quality or semantic equivalence (§3.6) and instead changes their refusal behavior (§3.5).

3.3 Diacritic Normalization

Table 6 reports raw and diacritic-normalized chrF++ for English→Lakota (thinking condition).

Stripping diacritics adds 3.3–6.5 chrF++ points for E→L. The gap indicates that models produce roughly correct consonant and vowel sequences but place diacritical marks—stress, nasalization, and caron-marked consonants—inconsistently. Multiple orthographic conventions are in active use for Lakota (§2.1); models may have been trained on text in several of these systems, and the prompt did not specify an orthographic standard.

Model	L→E	E→L
Gemini 3.1 Pro	0.93 (σ = 0.06)	0.83 (σ = 0.11)
Claude Opus 4.6	0.71 (σ = 0.17)	0.46 (σ = 0.15)
Claude Sonnet 4.6	0.70 (σ = 0.15)	0.32 (σ = 0.11)
GPT-5.2	0.55 (σ = 0.12)	0.45 (σ = 0.11)
DeepSeek-V4-Pro	0.84 (σ = 0.10)	0.68 (σ = 0.13)
GLM-5.1	0.63 (σ = 0.18)	0.56 (σ = 0.22)
Qwen3.6-Plus	0.78 (σ = 0.16)	0.59 (σ = 0.20)

Table 7: Self-reported confidence (thinking condition).

3.4 Confidence

Table 7 reports mean self-reported confidence from the thinking condition.

Among the proprietary models, the confidence ranking largely matches the chrF++ ranking for L→E: Gemini > Opus ≈ Sonnet > GPT-5.2. The open-weight models are the most overconfident: DeepSeek-V4-Pro and Qwen3.6-Plus report L→E confidence of 0.84 and 0.78—above every proprietary model except Gemini—despite scoring well below the Claude models. All models report lower confidence for E→L than L→E.

Gemini reports the highest confidence in both directions (0.93 L→E, 0.83 E→L) and also achieves the highest chrF++ scores, so its confidence is directionally accurate. However, 0.83 confidence on E→L where chrF++ is 42.6 represents significant overconfidence in absolute terms.

GPT-5.2's E→L confidence (0.45) exceeds Sonnet's (0.32) despite scoring lower on chrF++ (23.6 vs 27.0), so E→L confidence does not track quality.

At the per-pair level, confidence is a poor guide to correctness, especially for the open-weight models without reasoning: Qwen3.6-Plus renders *Hamáčhola* ("I am naked") as "Spider" with 0.95 confidence, and the three open-weight models return three different

	L→E		E→L	
Model	**Ref**	**Emp**	**Ref**	**Emp**
GPT-5.2	6	0	16	0
Sonnet 4.6	0	11	2	2
Opus 4.6	0	0	0	0
Gemini	1	0	0	1
DeepSeek-V4-Pro	0	0	5	0
GLM-5.1	16	0	7	0
Qwen3.6-Plus	2	0	36	0

Table 8: Non-translation outcomes (thinking condition). Ref = refusal, Emp = empty response, Err = API error.

confident answers (0.8–0.95) for *Wakȟálya yo* ("Make coffee"). High confidence attached to output that does not preserve meaning is common rather than exceptional, so self-reported confidence cannot be used to filter unreliable translations.

3.5 Refusals and Failures

Table 8 reports non-translation outcomes from the thinking condition.

Among the proprietary models, GPT-5.2 refused the most translations (22 total, 16 on E→L, up from 4 at baseline); among the open-weight models, Qwen3.6-Plus refused far more (38 total, 36 on E→L). Refusal reasons typically cited inability to produce reliable Lakota text. Sonnet produced 11 empty responses on L→E (the model returned an empty translation string with no refusal reason) and refused 2 E→L translations. Opus produced complete translations on all 200 pairs in both directions. Gemini had one L→E refusal and one E→L empty response.

The open-weight models make the role of reasoning in refusal behavior especially clear. With reasoning enabled, Qwen3.6-Plus refuses 36 of 200 E→L pairs and GLM-5.1 refuses 16 of 200 L→E pairs; with reasoning disabled, these fall to 0 and 1 respectively. Disabling reasoning does not improve translation quality (§3.2, §3.6)—it removes the step at which the model recognizes it cannot translate. For these models, reasoning surfaces the limitation rather than overcoming it.

The structured schema lets each model decline with a `refusal_reason`, which could bias models toward refusal over a low-confidence attempt. The non-thinking results argue against this: disabling reasoning makes refusals nearly vanish with no gain in semantic equivalence, so refusals under reasoning track recognized difficulty rather than a prompt artifact.

Among the proprietary models, refusals are concentrated on E→L (16 of 22 GPT refusals, both Sonnet refusals)—the harder generation direction.

Reported scores exclude refusals, the standard conditional-on-attempting convention. Scoring refusals as chrF++ = 0 over all 200 pairs lowers the high-refusal cases (Qwen3.6-Plus E→L from 18.4 to 15.1, GLM-5.1 L→E from 29.3 to 27.0), with no change to the ranking. For the open-weight models it also tempers the apparent thinking gains: Qwen3.6-Plus's E→L thinking score no longer exceeds its baseline once reasoning-induced refusals are counted.

3.6 Semantic Equivalence

Table 9 reports LLM semantic judgments for Lakota→English translations across all seven models and both conditions, rating each hypothesis–reference pair `equivalent`, `partially_equivalent`, or `not_equivalent`. The two judges (Gemini 3 Flash and the independent Llama-3.3-70B) agree substantially (Cohen's $\kappa = 0.75$, 0.89 quadratic-weighted), with disagreement concentrated in the adjacent `partially_equivalent` tier rather than between the polar categories.

The judge reveals that chrF++ is a reasonable proxy for models with higher equivalence rates—Gemini's chrF++ of 58.4 corresponds to 60.4% semantic equivalence—but misleading for models with lower rates. GPT-5.2's chrF++ of 26.4 implies some partial overlap, but only 6% of its translations are semantically correct. The remaining 87% are fluent English that does not preserve the source meaning (Table 10).

Open-weight models do not close the gap. The strongest, DeepSeek-V4-Pro, reaches roughly 19–21% equivalence—between GPT-5.2 and the Claude models—while GLM-5.1 and Qwen3.6-Plus remain at GPT-5.2's baseline tier (7–9%). No open-weight model approaches Gemini.

For the stronger models, 16–24% of translations are *partially equivalent*—capturing the

Model	Cond.	Equiv	Partial	Not	κ
Proprietary					
Gemini	base	60.4	23.4	16.2	0.64
Gemini	think	61.3	23.1	15.6	0.65
Opus 4.6	base	31.5	18.0	50.5	0.72
Opus 4.6	think	37.0	21.0	42.0	0.63
Sonnet 4.6	base	29.0	16.0	55.0	0.69
Sonnet 4.6	think	30.7	23.8	45.5	0.67
GPT-5.2	base	6.0	7.0	87.0	0.63
GPT-5.2	think	13.4	9.8	76.8	0.75
Open-weight					
DeepSeek-V4-Pro	base	20.5	13.0	66.5	0.70
DeepSeek-V4-Pro	think	19.0	14.0	67.0	0.74
GLM-5.1	base	9.0	7.0	83.9	0.68
GLM-5.1	think	9.2	6.5	84.2	0.76
Qwen3.6-Plus	base	7.1	4.5	88.4	0.77
Qwen3.6-Plus	think	7.1	7.6	85.4	0.82

Table 9: LLM semantic judge results (Gemini 3 Flash Preview), L→E; Equiv/Partial/Not are percentages of judged pairs (equivalent / partially equivalent / not equivalent). κ = Cohen's inter-judge agreement with an independent Llama-3.3-70B judge on each row's pairs. `base/think` = reasoning disabled/enabled.

core meaning but missing elements or shifting emphasis—a band that binary scoring would obscure.

Extended thinking provides the largest benefit to the weakest *proprietary* model: GPT-5.2 improves from 6.0% to 13.4% equivalent (2.2×), while Gemini barely changes (60.4% to 61.3%), consistent with the chrF++ thinking deltas (§3.2). The open-weight models behave differently: reasoning leaves their semantic equivalence essentially unchanged (DeepSeek -1.5, Qwen 0.0, GLM $+0.2$ points) while increasing their refusal rates (§3.5).

3.7 Qualitative Error Analysis

Reading the lowest-scoring Lakota→English items reveals a consistent failure mode: compositional translation, rendering idioms and culturally specific vocabulary element by element rather than by conventional meaning. The word glossed "electricity" is rendered "sacred," and the idiom *Wakȟálya yo* ("make coffee") becomes "sanctify it" (Table 10). The starkest of these failures come from the open-weight models; on the same items the frontier models often recover the meaning—Gemini and Opus render *Wakȟálya yo* as "make coffee," and Sonnet translates the electricity sentence (chrF++ 58.2 vs. 18.6). The pattern is partly an artifact of evaluation design: dictionary example sentences over-represent idiomatic usage, and some literal renderings are defensible translations that the single idiomatic reference does not credit (*Wakȟálya yo* → "boil it"). Most failures, however, aren't defensible alternatives like this—the output simply doesn't match any plausible English translation of the source.

4 Discussion

The LLM judge results (§3.6) demonstrate that chrF++ alone cannot reliably characterize model capability on low-resource translation: it closely tracks semantic equivalence for the strongest models but is misleading for the weakest, where fluent English masks near-zero meaning preservation. The BERTScore check we set aside in §2.5 is nonetheless revealing. What was informative was the per-pair correlation between the two: chrF++ penalizes any surface deviation from the reference (a floor), while BERTScore rewards any fluent English (a ceiling), so the two diverge most for models producing fluent output unrelated to the source. For the proprietary models in the baseline condition, this correlation tracks the semantic-equivalence gradient (Gemini $r = 0.91$; Opus and Sonnet $r = 0.83$–0.86; GPT-5.2 $r = 0.79$).

For external calibration, strong systems reach chrF++ in the mid-50s on a high-resource pair such as Chinese→English (Jiao et al., 2023); Gemini's 59.4 on Lakota→English sits at or above that range, yet its 42.6 on English→Lakota and 60.4% semantic equivalence show the capability is far from what that single number suggests. For comparison within indigenous-language MT specifically, the AmericasNLP 2023 shared task reports ChrF of roughly 25–40 for the best fine-tuned systems translating into eleven indigenous languages (Ebrahimi et al., 2023); Gemini's zero-shot English→Lakota result is comparable to the strongest of these systems, though the metric (ChrF vs. chrF++) and system class (fine-tuned NMT vs. zero-shot LLM) differ.

For the proprietary models, extended thinking yields a consistent but small chrF++ gain (1–7 points) that does not change the prac-

Failure mode	Model	Reference	Model output	chrF++
Coined term read literally	DeepSeek	We live in a small trailer house with no electricity.	We are poor in a very small *sacred* tipi.	18.6
Idiom / phonological slide	DeepSeek	Make coffee (put the water to boil).	Sanctify it!	2.9
Defensible literal (dictionary-bound)	Opus	Make coffee (put the water to boil).	Boil it! / Heat it up!	6.5
Kinship confusion	GLM	Did my daughter call?	Did my younger brother come?	27.6
Wrong content word	DeepSeek	Do you have a car?	Do you have a lighter?	65.9
Fluent but unrelated	DeepSeek	I want to talk to you.	I will be very happy.	6.5

Table 10: Representative L→E failure modes across models. chrF++ shows that surface overlap and meaning preservation diverge: a one-word near-miss (65.9) outscores output unrelated to the source. The two *Wakȟálya yo* ("Make coffee") rows contrast a genuine error ("Sanctify it") with a defensible literal translation ("Boil it") that the idiomatic single reference penalizes.

tical assessment; it is largest on the harder English→Lakota direction and for the weakest model (GPT-5.2, $+7.4$ points of semantic equivalence). It is not what separates the leaders: raising Gemini's reasoning from default to high moves Lakota→English by only $+1.0$ (§3.2), suggesting its lead reflects base capability rather than reasoning budget. The open-weight models behave differently again—reasoning leaves their translation quality and semantic equivalence essentially flat while raising refusals (§3.5), declining items rather than improving them. Taken together, these patterns suggest that on a language this far outside the training distribution, the binding constraint is knowledge rather than inference, and reasoning's main contribution is calibration—recognizing what one does not know—rather than capability.

The wide per-pair variance (σ = 22–28 chrF++ for L→E) likely reflects the distribution of Lakota text in training data. Common phrases such as *Mázaská etáŋ luhá he?* ("Do you have any money?") score at or near 100.0 across nearly all models, suggesting these phrases appear verbatim in training data. Culturally specific constructions involving kinship terms, complex verb morphology, or ceremonial language score near zero. The conversational score stratification (Table 4) confirms this pattern quantitatively: for every model and direction, score-6 pairs substantially outperform score-4 pairs. This suggests that model performance is concentrated on high-frequency phrasebook-like items rather than reflecting general capability across Lakota sentence types.

5 Conclusion

As of spring 2026, no frontier LLM, proprietary or open-weight, can reliably translate Lakota. The best Lakota→English performance (chrF++ 59.4) corresponds to 60.4% semantic equivalence on conversational sentences, but the worst model (GPT-5.2) achieves only 6% equivalence despite producing fluent English throughout. Open-weight models do not close the gap: none approaches Gemini, and even the strongest (DeepSeek-V4-Pro) trails the proprietary leaders substantially. The best English→Lakota performance (chrF++ 42.6) is inadequate for unsupervised use. Extended thinking provides a modest improvement of 1–7 chrF++ points for the proprietary models but does not change the practical assessment; for the open-weight models it alters refusal behavior more than translation quality, surfacing the limitation rather than overcoming it. Diacritic normalization analysis shows models produce roughly correct base characters but place diacritical marks inconsistently, possibly reflecting orthographic heterogeneity in training data.

Limitations

Each pair has a single reference translation, and chrF++ against one reference underestimates quality for valid paraphrases; the LLM judge partially mitigates this but cannot fully

solve it. Lakota also marks the speaker's gender through distinct enclitics—the dataset's *Wakȟálya yo* ("make coffee") uses the men's imperative particle, where a woman would say *Wakȟálya ye*—so a single reference can penalize a valid rendering from the other speech register. The evaluation set consists of constructed conversational examples from the New Lakota Dictionary; performance on narrative or spontaneous text may differ.

Although our open-web analysis (§2.7) finds little verbatim overlap between the evaluation pairs and indexed text, contamination cannot be entirely excluded as a contributing factor. Constructing original Lakota–English pairs not derived from any published source would allow it to be ruled out decisively and is a priority for future work.

No fluent Lakota speaker participated in the evaluation. Quality is assessed via established automatic metrics (chrF++) and LLM-based semantic judges (§2.5); while both are standard practice in low-resource MT evaluation, neither substitutes for evaluation by fluent speakers, and qualitative analysis is correspondingly restricted to axes that do not require Lakota fluency. Collaboration with tribal-college language programs and fluent speakers is the most important methodological next step.

This work evaluates only general-purpose frontier and open-weight LLMs without a fine-tuned dedicated MT baseline such as a Lakota-adapted NLLB-200; comparing zero-shot LLM performance against a fine-tuned low-resource NMT system on the same evaluation set would situate these results within the existing low-resource MT literature.

The specific model versions evaluated reflect API availability in February–May 2026; provider model identities and behavior shift over time. The proprietary and open-weight models were evaluated in different months (§2.8); the cross-provider comparison therefore reflects the model snapshots available at each evaluation rather than a synchronized run, and the relative ordering of comparably scoring models should be read with that in mind.

Ethics Statement

The evaluation set was drawn from the New Lakota Dictionary (NLD), a commercially published reference work. The evaluation set is used for research evaluation and is not redistributed. We acknowledge that orthographic standardization is a contested issue within Lakota communities (Hauff, 2020), and that our use of NLD materials reflects availability rather than endorsement of any particular orthographic standard or institutional approach. Future work should evaluate against materials in other actively used orthographies, ideally in collaboration with language programs at tribal colleges and universities.

Acknowledgments

The author used Claude (Anthropic) for editorial feedback, coding assistance, and experimental design discussion.

Author Contributions

L.R. designed the study, curated the evaluation set, ran all evaluations, and made all methodological decisions.

Data Availability

Code and aggregate results are available at `https://github.com/robotson/lakota-translation-benchmark`. The evaluation set was sourced from the New Lakota Dictionary and is not redistributed; see `data/example_pairs.json` for the data schema. Aggregate results are provided in `results/comparison.csv` and `results/llm_judge_summary.csv`. All code requires only API keys and `pip install litellm sacrebleu python-dotenv bert-score`.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*. ArXiv:2409.19151.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, and 1 others. 2023. Findings of the Americas-NLP 2023 shared task on machine translation

into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219.

Tasha R. Hauff. 2020. Beyond numbers, colors, and animals: Strengthening Lakota/Dakota teaching on the Standing Rock Indian Reservation. *Journal of American Indian Education*, 59(1):5–25.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. ArXiv:2501.18771.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, and 1 others. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 202–217.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Kaushik Parankusham, Rodrigue Rizk, and K.C. Santosh. 2025. LakotaBERT: A transformer-based model for low resource Lakota language. *arXiv preprint arXiv:2503.18212*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Jan F. Ullrich. 2008. *New Lakota Dictionary*. Lakota Language Consortium, Bloomington, IN.

Albert Sr. White Hat. 1999. *Reading and Writing the Lakota Language*. University of Utah Press, Salt Lake City.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*.

Bridging Digital Tools for Linguistic Documentation and Revitalization

Christopher Haberland* **Carly Crowther*** **Jingnong Qu** **Anuk Centellas**
University of Washington
{haberc, carlyc88, jingnong, anukz}@uw.edu
*Joint first authors.

Abstract

Digital tools serving language revitalization tend to fall into two categories: 1) linguist-oriented documentation tools that prioritize annotation, morphological analysis, and archival preservation, and 2) community-facing applications that emphasize accessibility and language learning. Few systems integrate the former with the latter, and practical barriers — including the cost of computational expertise, single-user workflows, and limited data governance — further constrain their utility. These disconnects incur additional development and communication costs for revitalization teams consisting of linguists and community members. We introduce `langlit`, a collaborative web-based platform that attempts to tailor documentation workflows for the language revitalization context within a single system. The platform integrates a finite-state morphological analyzer with a three-tier human-in-the-loop annotation workflow, searchable corpus interfaces with multiple query modalities, interactive word construction guided by the morphological grammar, corpus-linked hypothesis tracking with provenance, and a grammar-derived editable dictionary. All components share a single underlying FST grammar, and the system supports configurable access controls, collaborative editing, and optional LLM integration with transparent data handling. Designed for redeployment across languages through a modular architecture, `langlit` is published as an open-source repository on GitHub. We situate our system within the existing landscape of revitalization tools through a comparative analysis and discuss how integrated, community-informed design can better serve the specific goals of language revitalization.

1 Introduction

Digital tools for language revitalization and language documentation serve different goals. Documentation tools prioritize annotation, archival standards, and linguistic analysis, while revitalization efforts focus on intergenerational transmission and everyday language use (Flavelle and Lachler, 2023; Gessler, 2022). Software designed principally for documentation may lack features desired by revitalization teams whose work involves not only collating linguistics knowledge, but also transmitting it in ways consistent with community values (Le Ferrand et al., 2022; Gessler, 2022). As a result, linguistic knowledge produced during documentation often remains inaccessible to teachers and learners (Neubig et al., 2020).

Practical barriers compound this disconnect. Documentation tools are largely unfamiliar outside academic linguistics (Skilton et al., 2025), and building accessible, community-facing applications requires developer time that most revitalization projects cannot sustain (Wagner, 2017). Few reusable templates exist that communities could adapt independently.

In response, we introduce `langlit`, an open-source web-based system designed to bridge documentation and revitalization workflows. The system makes annotations and corpus data accessible to teachers and learners through collaborative editing, corpus search, and configurable data governance. It is designed so that the products of linguistic documentation are immediately and transparently available to language community stakeholders as documentation work progresses.

Beyond its organizational and collaboration features, the system is designed to hasten linguistic knowledge discovery. A graphical interface allows for generation of morphological interpretations of a text, and optional tooling integrating large language models is provided to assist during the processes of annotation and linguistic discovery. The design emphasizes language generality and modularity so that it can be adapted for other efforts. We invite open contributions to our work, which is

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 22–32

published as an open repository on GitHub. [1]

Section 2 reviews the landscape of digital tools for language documentation and revitalization. We examine their features, limitations, and how they emphasize certain roles over others. Section 3 describes the system architecture and its core components. Section 5 discusses limitations and future work.

2 Background

Structured corpora, morphological parsing, and interlinear glossing are rarely incorporated into tools intended for teachers and learners, despite broad recognition that revitalization efforts benefit from example-rich, searchable materials. This section reviews the features and limitations of existing tools to motivate the design of systems that bridge documentation and community use.

2.1 Features of digital tooling for language revitalization

Digital tools expand both the reach and flexibility of language revitalization efforts. Online platforms, mobile applications, and multimedia resources enable geographically dispersed learners to engage with a language outside of traditional classroom settings, which is particularly valuable for diaspora communities lacking in-person access to fluent speakers (Chew et al., 2023; Mauger, 2025; Meighan, 2024). Against the backdrop of reduced intergenerational transmission, language technologies can support classroom teaching by providing accessible, searchable corpora and recordings of fluent speakers (Mainzinger, 2024). Multimedia materials further support culturally grounded learning by pairing linguistic data with narratives and oral histories (Meighan, 2021).

A central benefit of language technology is its ability to distribute authentic language materials across time and space, supporting learning even where fluent speakers are few or geographically dispersed (Richards et al., 2025; Tennell and Chew, 2024). These considerations motivate the design of systems that combine analytical rigor with accessibility, collaboration, and pedagogical support.

2.2 Linguist documentation tools

The most prominent documentation tools are technical environments developed by and for linguists. FieldWorks Language Explorer (FLEx) integrates a lexicon, text corpus, and morphological parser within a unified workflow, making it widely used in language documentation , and no other tool combines these three components in an integrated fashion (Skilton et al., 2025) (see Table 1). Another widely used tool, ELAN, supports multi-tier annotations of audio and video (Wittenburg et al., 2006), but its workflow assumes significant linguistic expertise and privileges researchers' needs over those of teachers or community members (O'Neil et al., 2024). ELAN's interlinearization features remain under development and do not yet provide the integrated lexicon–parser workflow valued in FLEx (Skilton et al., 2025).

Despite their important role, these tools have notable limitations. FLEx is difficult to install and operate without specialized training, is optimized for single-user workflows with limited version control, and lacks transparent version histories, complicating accountability for changes to language data (Skilton et al., 2025). Its management by SIL International and the requirement for data to be uploaded to an outside server raises data privacy and sovereignty concerns (Skilton et al., 2025). Both tools require powerful local computing and a desktop workflow.

The technological landscape has evolved significantly since these tools were first developed. Smartphones, web applications, and cloud-based collaboration are now widespread, and advances in natural language processing have expanded what is possible for both linguists and learners. However, few of these advances have been used to upgrade core documentation software.

Integrating NLP models into tools like FLEx and ELAN remains technically cumbersome, to the point of discouraging adoption even among experienced documentary linguists (Gessler, 2022). This suggests a need for systems that retain the analytical depth of existing documentation tools while supporting modern, collaborative workflows for the downstream users of language documentation outputs.

2.3 Community-facing tools

Community-facing tools such as mobile applications, web dictionaries, gamified learning tools, and multimedia storytelling platforms are intended for direct use by language communities (Ajani et al., 2024; Bettinson and Bird, 2021; Galla, 2016; Tennell and Chew, 2024). In communities with

[1] https://github.com/haberchr/langlit

few fluent speakers, digital platforms distribute recordings of Elders, extending the reach of limited linguistic resources (Meighan, 2021, 2024). For example, Littell et al. (2017) introduce a reusable framework for web dictionaries across languages and Kazantseva et al. (2018) describe a verb conjugation tool for Kanyen'keha designed to help learners navigate complex morphology. These tools are often effective for engagement and beginner learning but few consistently integrate morphological search functionality over annotated corpora, limiting utility for teachers and advanced learners who need contextualized examples of specific grammatical phenomena (Neubig et al., 2020; Taylor-Adams, 2019).

Beyond functional limitations, concerns about data sovereignty are central to the design of community-facing language technologies (Chew et al., 2023; Kukutai and Taylor, 2017; Schwab-Cartas, 2018; Tennell and Chew, 2024). These commitments are difficult to uphold when projects rely on external infrastructure such as third-party platforms or large language models. Few reusable, community-adaptable application frameworks exist, so each new project typically requires full custom development. This is a structural barrier because communities with the greatest need are often the least positioned to commission or maintain such tools (Chew, 2021; Wagner, 2017).

Despite their complementary strengths, documentation and community-facing tools have largely developed in parallel, and as Gessler (2022) notes, existing software has not kept pace with the collaborative, accessible workflows that revitalization teams need. Bridging this gap requires platforms with usable interfaces, collaborative capabilities, and community governance over language data.

2.4 Comparison of select work

Table 1 compares the documentation, pedagogical, and infrastructure features of a sample of work identified in our review of digital tools for language documentation and revitalization. Our review illustrates the general patchwork of features offered by tools purposed for revitalization that motivated the design of `langlit`.

Neither FLEx nor ELAN, the most widely used documentation applications, provide a pedagogical interface for language community members. FLEx offers concordance generation and regex filtering within its Texts & Words module, but this does not permit search functionality accessible to non-specialists (Skilton et al., 2025). SIL has recently released FieldWorks Lite (beta)[2], a cross-platform companion application that introduces real-time collaborative editing for lexicon data, though it does not support morphological parsing, text annotation, or the broader documentation workflows provided by FLEx. ELAN supports multi-tier regex search over annotated media, but a documented limitation in its Multiple Layer Search prevents reliable morphosyntactic queries across utterance boundaries, restricting its utility as a corpus search tool (Wilbur, 2019). Neither tool offers integrated word construction, collaborative hypothesis documentation, or configurable data governance.

Gessler (2022) presents Glam, a work-in-progress system that most directly shares `langlit`'s goal of bridging NLP models and documentary linguistics through shared software infrastructure. Glam targets documentary linguists and NLP integration, providing UIs for annotation and model interaction, but does not describe a pedagogical interface for teachers or learners. Plaid[3] constitutes a linguistic annotation backend that is designed to solve data management and collaboration for linguistic documentation applications, but is not deeply coupled with analytical or pedagogical tooling.

Among the community-facing tools in our comparison, Littell et al. (2017) and Debenport et al. (2023) both offer dictionary interfaces with lexical search, but neither supports morphological corpus search over annotated texts. Mukurtu, described by Debenport et al. (2023), is the most developed system for data governance, providing granular, protocol-based access control designed specifically for cultural sovereignty needs; however, Mukurtu supports metadata and media annotation within its archival framework rather than linguistic corpus annotation such as morphological glossing or interlinear glossed text. Richards et al. (2025) describe a mobile application with corpus search and multimodal features, but without collaborative editing, open-source availability, or morphological analysis. Kazantseva et al. (2018) stands out as the only system besides `langlit` that provides both a morphological analyzer and an

[2] https://software.sil.org/fieldworks/download/fieldworks-lite/

[3] https://larc-iu.github.io/plaid/manual.html

	Documentation and pedagogical capabilities									Infrastructure & governance				
Tool	Morph. Analyzer	Dict.	Corpus Search	Pedagog. UI	Corpus Annot.	Multi-modal	Phrase Builder	Hypoth. Doc.	LM Integ.	Collab. Editing	User Mgmt	Open Source	Data Gov.	Platform
FLEx	✓	✓	*limited	—	✓	—	—	—	—	—	*limited	✓	—	Desktop
ELAN	—	—	*limited	—	✓	✓	—	—	—	—	*single	✓	*?	Desktop
Littell et al. (2017)	—	✓	*lex	✓	—	—	—	—	—	—	—	✓	*?	Web/Mobile
Kazantseva et al. (2018)	✓	—	—	✓	—	—	✓	—	—	—	—	—	—	Web/Mobile
Gessler (2022)	—	—	—	—	✓	—	—	—	✓	✓	✓	✓	*✓[a]	Web
Debenport et al. (2023)	—	✓	*lex	✓	*limited	✓	—	—	—	✓	✓	✓	✓	Web
Pugh and Tyers (2023)	✓	—	—	—	—	—	—	—	—	—	—	✓	—	CLI
Mainzinger (2024)	—	✓	—	✓	—	*?	—	—	—	—	—	—	—	*mixed
Cox et al. (2025)	✓	✓	—	✓	—	✓	—	—	—	—	—	✓	*?	Web
Richards et al. (2025)	—	—	✓	✓	—	✓	—	—	—	—	—	—	*token	Mobile
Hammerly et al. (2026)	✓	✓	—	*ext	—	—	*ext	—	—	—	—	✓	—	CLI
`langlit`	✓	✓	✓	✓	✓	—	✓	✓	✓	✓	✓	✓	✓	Web

Table 1: Comparison of selected language revitalization tools. ✓ present; — absent; * partial or unclear (asterisk with label). *lex*: search limited to lexical lookup. *limited*: feature present but with significant constraints. *ext*: downstream application described but not integrated into a single UI. *single*: single-user only. *token*: token/password-based access. *?*: unclear from available documentation. *mixed*: no single integrated platform. LM Integ. refers to in-app integration with language models (e.g., LLMs, neural taggers) for annotation assistance or other tasks, not rule-based morphological parsing. *(a) Derivative work `Plaid` adeptly addresses infrastructure and governance issues.

integrated word/phrase construction interface for learners; however, it lacks corpus search, annotation, and collaborative infrastructure.

Several recent projects build directly on FST-based morphology. Pugh and Tyers (2023) describe a finite-state analyzer for Highland Puebla Nahuatl, but without an accompanying application layer, and Mainzinger (2024) presents a technology roadmap for Mvskoke drawing on multiple existing systems rather than a single integrated platform. Cox et al. (2025) describe a long-term community partnership to build an FST-backed intelligent dictionary for Tsuut'ina, with verb paradigms reviewed item-by-item by a community language committee. Hammerly et al. (2026)'s Ojibwe-Morph introduces an FST framework whose downstream applications include a verb conjugation tool for education, a spell-checker, and intelligent dictionary search. In each case, pedagogical and word-construction capabilities are built on FST output, but do not provide corpus annotation, corpus search, or collaborative editing capabilities.

Among the tools surveyed, no system combines a morphological analyzer, corpus annotation, corpus search, a pedagogical interface, word construction, and hypothesis documentation within a single collaborative, open-source, web-based platform with configurable data governance. `langlit` is designed to address this gap.

3 System Description

`langlit` is a Python application built with the `streamlit`[4] framework used for displaying data-centric applications as web interfaces. It is easily shareable online via Streamlit Community Cloud with minimal setup. `langlit` integrates an FST morphosyntactic grammar with corpus annotation, search, word construction, linguistic hypothesis tracking, and a dictionary. The user-defined FST grammar and phonological file backend (`.lexc`/`.xfscript`) are read by the Helsinki Finite State Transducer Python package `hfst` (Lindén et al., 2011) and and form the backbone of the morphological interpretation engine.

The system can be customized for a target language by creating a "language pack." This consists of a configuration module defining language-specific parameters that are declared in a single `config.py` file imported by all pages, along with other feature and language specific files to support the morphological analysis backend. This architecture supports cross-language applicability. Communities can enable or disable components as needed for the intended audience (linguists or community teachers and learners) or desired functionality.

In the following subsections, we describe the system's core components: human-in-the-loop annotation workflows (§3.1), multi-modal corpus

[4] `https://streamlit.io`

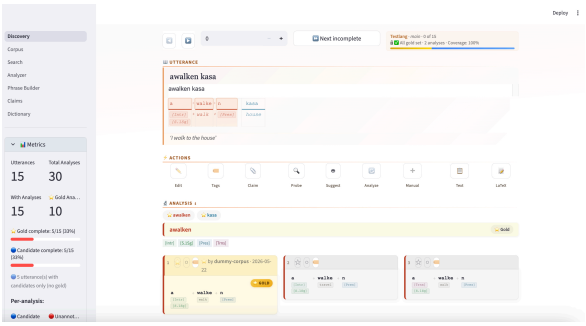

Figure 1: The `langlit` Discovery page.

search (§3.2), IGT export and pedagogical material generation (§3.3), corpus-linked hypothesis tracking (§3.4), a grammar-derived editable dictionary (§3.5), and collaborative editing with configurable access control (§3.6).

3.1 Human-in-the-Loop Annotation Workflows

The Discovery page (Figure 1) serves as the primary interface for corpus exploration and annotation management. It provides an overview of corpus-level metadata and annotation progress, including utterance analyses and gold annotation counts, allowing users to assess the state of the corpus at a glance.

For each utterance, Interlinear Glossed Text (IGT) is rendered inline as an HTML table aligned to the tokenized surface form, with spans lacking a gold annotation (*gaps*) highlighted in red to make annotation completeness visually salient. The table displays the normalized surface form, morpheme boundaries, and gloss. Utterances can be navigated sequentially or by jumping directly to incomplete entries, streamlining the annotation workflow. Critically, the underlying data model also supports multilingual comparison of utterances, addressing a documented limitation of FLEx (Skilton et al., 2025).

The sidebar computes and displays annotation progress metrics, including per-analysis counts of gold, candidate, and unannotated spans, and the rate of agreement between gold and candidate labels across all analyses with both labels set. Disagreements are flagged for reconciliation. A "next incomplete" navigation control moves the reviewer directly to the next utterance lacking a complete gold annotation, simplifying navigation through a large corpus.

The Discovery page is primarily designed for linguists, researchers, and annotators who need to systematically understand the morphosyntactic properties of a corpus and build a gold-annotated dataset that can be formatted as IGT. It also serves as a useful exploration tool for teachers and advanced learners seeking a deep understanding of the language's structure. Once labeled, the data becomes a valuable resource for teachers who may use the Search page to identify relevant linguistic examples for lesson creation.

3.1.1 FST-assisted analysis

An FST analyzer supports corpus review on the Discovery page. Finite-state transducers are well-suited for morphological analysis in revitalization contexts because they do not require large training corpora, which are often unavailable for low-resource languages (Kazantseva et al., 2018). Unlike neural approaches requiring expensive training runs and expansive amounts of digitized data resources, FSTs are rule-based, deterministic, and can be constructed directly from morphosyntactic knowledge. This allows for early-stage and continuous development as the documentation team learns more about the language, even in the context of resource scarcity. Many Indigenous languages have complex inflectional systems yielding extremely large numbers of surface forms (Mithun, 2001), making the concatenative approach (Hammerly et al., 2026) that `langlit` adopts particularly appropriate. FST-based parses discretize morphological transitions, keeping analysis transparent and correctable—especially important where analyses must remain interpretable by less technically inclined team members (Pugh and Tyers, 2023).

Beyond morphological analysis, the software used to create FST-based systems is highly extensible to a range of downstream modular applications, including spell checkers, grammar checkers (Pirinen et al., 2023), verb conjugators (Kazantseva et al., 2018), interactive transcription systems (Lane and Bird, 2022) and morphologically aware dictionary search (Hammerly et al., 2026; Cox et al., 2025). The app realizes several of these extensions directly: the Phrase Builder page uses the same `.lexc` grammar to guide interactive word construction (§3.3), and the Dictionary page derives its lexeme and morpheme inventory from the same source (§3.5).

3.1.2 Three-tier annotation workflow

Annotation enrichment is expensive, and revitalization programs rarely have the resources required to annotate a corpus from scratch (Nicolai et al., 2020; Neubig et al., 2020). The Discovery page implements a three-tier human-in-the-loop design. The first tier is human-labeled annotation, supporting tabula rasa span annotation for morphological or other properties according to user-defined conventions.

The second and third tiers are motivated by recent work that shows the potential for NLP tooling to improve the speed at which high quality annotations can be produced (Ginn et al., 2024; Liang et al., 2026). These tiers are optional to allow teams to annotate according to their preference and utility. The second tier allows morphological annotation using finite state transducers via `hfst` (Lindén et al., 2011), a package that references `.lexc` and `.xfscript` files to deterministically infer morphological annotations for spans of text.

The third tier, useful in parallel data contexts, prompts a server-based or locally hosted large language model (LLM) to distinguish from among morphological interpretations generated by the FST step to choose the most explanatory interpretation given its parallel translation of the referenced data in a high-resource language.

It is important to highlight that many groups working towards revitalization may prefer to not send any target language data to AI servers to maintain maximal data ownership and sovereignty; to this end, our example workflow does not provide any surface-form target language data to the AI server in this step - only the morphological glosses are passed. The external API call is opt-in, gated behind a UI and password toggle, and can be disabled entirely by communities that prefer fully manual review or have concerns about data exposure through third-party services, aligning with data privacy principles addressed in O'Neil et al. (2024). Indeed, data sovereignty concerns may also be allayed by selecting local LLM models for inference in this step, if desired.

A human reviewer may assign a *gold* label that may agree with or override *candidate* interpretations suggested by the FST and selected by then LLM. The UI ensures that annotations labeled by humans and machine systems are always distinguished and apparent. Declined analyses are tracked separately from unannotated ones.

3.2 Multi-Modal Corpus Search

Documentary corpora contain naturalistic, morphologically annotated examples drawn from actual usage, yet these remain practically out of reach for teachers without specialist training (Neubig et al., 2020; Taylor-Adams, 2019).

Low-resource language corpora "typically lack not only graphical search interfaces, but also the rich annotations (such as morphological and syntactic parses) that are conventionally required to support the function of a search interface" (Neubig et al., 2020). Compounding this, search interfaces that do exist typically require users to query in technical terms that are beyond the expertise of most teachers and learners (Taylor-Adams, 2019).

The Search page allows access to corpus entries containing a specified word or phrase, enabling users to efficiently locate and examine instances of target forms in context. Search results display the full utterance alongside its translation. Entries can be bookmarked and downloaded in CSV, PDF, or LaTeX format, making it straightforward to compile collections of examples for further analysis or instructional use.

Entries can be searched by translations, morphological glosses, tags, or a combination. Beyond the simple lexical lookup offered by other tools (see Table 1), the Search page supports regex and neural embedding queries over both monolingual and bilingual corpora, enabling comparative analysis and evidence gathering for linguistic research and teaching.

For linguists and researchers, the Search page provides a fast way to identify and examine instances of specific morphological phenomena across the corpus. It simplifies the process of finding examples with specific grammatical features or vocabulary, supporting the preparation of lesson materials such as handouts and study guides, which is useful for teachers. The Search page also offers an accessible entry point for learners to explore how particular words or concepts appear in the language.

All modes return paginated example cards displaying the surface form, gloss, and IGT of the gold analysis. Results of a search can be individually bookmarked within a session and staged for linking to a **hypothesis** being evaluated by a linguistic research team(§3.4), or immediately exported in a CSV for offline use. This design directly re-

sponds to the Teacher-in-the-Loop model of Neubig et al. (2020), which called for corpus retrieval that does not require users to express queries in technical terms and can function over partially annotated data.

3.3 IGT Export and Pedagogical Material Generation

Language documentation has historically prioritized researcher access over community-usable resource creation (Gessler, 2022). The app addresses this gap with two pages that convert annotation and grammar resources directly into exportable pedagogical materials.

3.3.1 Analyzer

The Analyzer page provides an interactive interface for morphological analysis of utterances. It accepts free-text input in the target language and runs it through the same preprocessing and FST pipeline. Users can examine the full set of interpretations for each span and designate one as the gold standard analysis, building an IGT entry for a word or phrase in real time.

Users can bookmark analyzed utterances, which are added to an export queue that can be downloaded, enabling ready-to-use IGT displays for external documents. The resulting three-line interlinear (surface / morphological / gloss) is exported as a standalone LaTeX document using a tabular layout compatible with standard linguistic IGT conventions, or as a PDF via a `pdflatex` subprocess call. An optional LLM call proposes a free translation from the gold gloss line.

The Analyzer is primarily aimed at linguists and researchers who need to inspect and validate FST output for specific forms. It is also a valuable tool for learners seeking a deeper understanding of the morphological structure of individual words, and for teachers looking to create IGT-based exercises or materials for translation and morphological analysis practice.

3.3.2 Phrase Builder

The Phrase Builder page provides a morphologically guided phrase construction interface that traverses the `.lexc` FST while exposing the FST's morphological generation functionality in a user-understandable way. Users begin by selecting a part of speech and searching for a specific root meaning, then are guided through answers to a series of grammatical questions that iteratively build a morphologically complex word from a root meaning. The interface reformulates each morphological choice as a grammatical question (e.g., a dropdown menu for person-number agreement), using question templates and tag-to-question mappings defined in a configuration file and comments in the `.lexc` file. `lexlit` currently supports adding phonological and infix rules via Python or `.xfscript`.

The Phrase Builder serves a range of users. For linguists and researchers, it provides an environment for exploring the productive morphology of the language. For learners, the guided step-by-step interface offers an accessible way to develop intuitions about the morphological structure of the language. For teachers, the page supports the construction of paradigms and morphologically varied word forms that can be exported and used as lesson materials.

3.4 Corpus-Linked Hypothesis Tracking

Neither the tools reviewed by Neubig et al. (2020) and O'Neil et al. (2024) nor any tool in Table 1 provide a mechanism to link specific corpus examples to in-progress linguistic hypotheses. The Claims page addresses this gap by implementing a structured system for documenting and tracking linguistic hypotheses or discoveries about the language. From both the Discovery and Search pages, any corpus utterance can be linked to a claim with an evidence relationship (*supports*, *contradicts*, or *neutral*), along with a mandatory note explaining the relevance of the example and a username and timestamp for provenance. A claim detail view displays the full evidence set and tracks supporting and contradicting evidence counts separately. Every update to a claim generates a timestamped revision record, providing an audit trail of how analysis evolved. Claims can be exported as structured LaTeX or PDF documents displaying the claim and associated evidence to support pedagogical communication.

Because evidence links are attributed to named users and claim status is visible to all collaborators, the system supports the transparent, reciprocal research process called for by O'Neil et al. (2024), which is particularly valuable for underdocumented languages where hypotheses may need to be revised as continued analysis of a corpus yields new insights.

The Claims page is a useful reference for teachers who want to incorporate verified linguistic

knowledge into lesson planning and for independent learners seeking a deeper understanding of the language's structure. Future work will 1) document agentic AI tools for automating linguistic discovery using a custom `langlit` harness, and 2) improved grammar reference distribution tools for members of the language community.

3.5 Grammar-Derived Editable Dictionary

A recurring problem in documentation is that dictionary materials exist as static files disconnected from the morphological resources used for analysis (Neubig et al., 2020; Cox et al., 2025). The Dictionary page addresses this by deriving its content directly from the same `.lexc` file used by the FST and Phrase Builder. Because the dictionary is generated from the same resource that powers analysis, there is no structural divergence between dictionary coverage and FST coverage: adding a morpheme to the analyzer makes it immediately visible in the dictionary. This tight coupling eliminates the need to keep a separate lexical database synchronized with a changing analyzer.

The Dictionary page displays each entry alongside a translation. Users can search for specific entries, filter by lexicon, and group results by lexicon or gloss, making it straightforward to navigate and organize the vocabulary of the language. Entries can be edited directly, allowing the lexical database to be updated as new information is discovered. Users may edit the categories from the `.lexc` file that are displayed via the language pack `.yaml` configuration file.

The Dictionary provides linguists and researchers an organized, searchable view of the lexicon with the ability to make edits. For teachers and learners, it functions as a practical vocabulary reference, enabling quick lookup of specific words and their translations for use in lessons or independent study.

3.6 Collaborative, Transparent Editing with Configurable Access Control

O'Neil et al. (2024) identify collaborative editing, user management, edit history, and data transparency as essential features of any cross-culturally applicable documentation tool, specifically noting that FLEx's single-user workflow and opaque version history as significant limitations (Skilton et al., 2025).

3.6.1 Web deployment

The application is deployable as a web application via Streamlit Community Cloud or other providers. For development purposes, computational linguists and developers may easily deploy the application locally or on custom servers.

3.6.2 Access control

User identity is established through OpenID Connect (OIDC) authentication on Streamlit Community Cloud and in local deployments after server configuration. Access is gated by Streamlit's authentication function (`st.login()`), which executes an OICD flow with Google as the identity provider, restricting app access to admin-defined accounts. Editing and LLM-usage is gated behind an admin-defined password, so non-privileged users have read-only access by default. The app currently supports SQLite as a backend, which is backed up to a Google Workspace account established by the admin. The local database allows for concurrent read access and serialized writes for users of the app. Future work could support cloud-native storage solutions.

3.6.3 LLM Integration and data transparency

External API calls are entirely opt-in: the LLM toggle's function is transparently presented to the user, and users may opt to avoid passing language data through third-party services by disabling any LLM inference without affecting other tool functionality.

4 Conclusion

Digital tools for language documentation and revitalization have historically directly served either linguists or community members, but rarely both. To address this gap, we present `langlit`, an open-source, web-based platform that combines a finite-state morphological analyzer, three-tier human-in-the-loop annotation workflow, multi-modal corpus search, interactive word construction, corpus-linked hypothesis tracking, and a grammar-derived dictionary in a single system. Because all components draw from one shared FST grammar, adding a morpheme to the grammar immediately propagates across the analysis, search, word construction, and dictionary pages. The platform is designed for redeployment across languages through a modular architecture, and its configurable access

controls reflect the data sovereignty priorities common in Indigenous language contexts. We hope that `langlit` can lower barriers for linguists and language community members to create, access, and use language documentation to benefit revitalization work.

5 Limitations

A major limitation of `langlit` is that its customization and integration in a language revitalization context may require involvement of a developer or computational linguist. We acknowledge that this assumption falls short of many real-world contexts and remains a problem to be addressed by future work.

We have not yet conducted user acceptance testing among various stakeholder roles identified in this work. Sustained use from users of varied perspectives is critical for corroborating many of the design assumptions introduced in this work.

We acknowledge that the software presented does not yet incorporate all desiderata for language documentation, analysis, and teaching. For example, our application does not currently support the inclusion of multimodal annotations. It is our hope that the application may serve as a starting point to allow others to suit the software to their needs.

References

Yusuf Ayodeji Ajani, Bolaji David Oladokun, Shuaib Agboola Olarongbe, Margaret Nkechi Amaechi, Nafisa Rabiu, and Musediq Tunji Bashorun. 2024. Revitalizing Indigenous Knowledge Systems via Digital Media Technologies for Sustainability of Indigenous Languages. *Preservation, Digital Technology & Culture*, 53(1):35–44.

Mat Bettinson and Steven Bird. 2021. Collaborative fieldwork with custom mobile apps. *Language Documentation & Conservation*, 15:411–432.

Kari A. B. Chew. 2021. #KeepOurLanguagesStrong: Indigenous Language Revitalization on Social Media during the Early COVID-19 Pandemic. *Language Documentation and Conservation*, 15:239–266.

Kari A.B. Chew, Sara Child, Jackie Dormer, Alexa Little, Olivia Sammons, and Heather Souter. 2023. Creating Online Indigenous Language Courses as Decolonizing Praxis. *The Canadian Modern Language Review*, 79(3):181–203.

Christopher Cox, Bruce Starlight, Janelle Crane-Starlight, Hanna Big Crow, and Antti Arppe. 2025. Creating an intelligent dictionary of tsuut'ina one verb at a time. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 110–119, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Erin Debenport, Mishuana Goeman, Maria Montenegro, and Michael Wynne. 2023. How a Dictionary Became an Archive: Community Language Reclamation Using the Mukurtu Content Management System. *Dictionaries: Journal of the Dictionary Society of North America*, 44(2):29–55.

Darren Flavelle and Jordan Lachler. 2023. Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Candace K. Galla. 2016. Indigenous language revitalization, promotion, and education: function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Luke Gessler. 2022. Closing the NLP Gap: Documentary Linguistics and NLP Need a Shared Software Infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Hammerly, Nora Livesay, Antti Arppe, Anna Stacey, and Miikka Silfverberg. 2026. OjibweMorph: An Approachable Finite-State Transducer for Ojibwe (and Beyond). *Language Resources and Evaluation*, 60:27.

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tahu Kukutai and John Taylor. 2017. *Indigenous Data Sovereignty: Toward an agenda*. ANU Press.

William Lane and Steven Bird. 2022. A finite state aproach to interactive transcription. In *Proceedings of the First Workshop on NLP applications to field linguistics*, pages 1–10, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Fashioning local designs from generic speech

technologies in an Australian aboriginal community. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4274–4285, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Siyu Liang, Talant Mawkanuli, and Gina-Anne Levow. 2026. Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 16–30, Rabat, Morocco. Association for Computational Linguistics.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2011. HFST—Framework for Compiling and Applying Morphologies. In *Systems and Frameworks for Computational Morphology*, pages 67–85, Berlin, Heidelberg. Springer.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150, Honolulu. Association for Computational Linguistics.

Julia Mainzinger. 2024. Technology and Language Revitalization: A Roadmap for the Mvskoke Language. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 7–12, St. Julians, Malta. Association for Computational Linguistics.

Stacey Mauger. 2025. The role of digital technology in Indigenous language revitalization: a systematic review of barriers, opportunities, and effective practices. Master's thesis, Ontario Tech University.

Paul J. Meighan. 2021. Decolonizing the digital landscape: the role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17.

Paul J. Meighan. 2024. Indigenous language revitalization using TEK-nology: how can traditional ecological knowledge (TEK) and technology support intergenerational language transmission? *Journal of Multilingual and Multicultural Development*, 45:3059–3077.

Marianne Mithun. 2001. *The languages of native North America*. Cambridge University Press.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, and 5 others. 2020. A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020. Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Alexandra O'Neil, Daniel Swanson, and Shobhana Chelliah. 2024. Computational Language Documentation: Designing a Modular Annotation and Data Management Tool for Cross-cultural Applicability. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 107–116, Bangkok, Thailand. Association for Computational Linguistics.

Flammie A Pirinen, Sjur N. Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Robert Pugh and Francis Tyers. 2023. A finite-state morphological analyser for Highland Puebla Nahuatl. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108, Toronto, Canada. Association for Computational Linguistics.

Mark Richards, Caroline Jones, Josephine Lardy, Anna Godden, Wanirr Godden, Sarah Bock, and Helena Lardy. 2025. Warrma Mangarrayi: Co-Designing an App for Learning Mangarrayi, an Indigenous Language of Northern Australia. *International Journal of Human–Computer Interaction*, 41(11):7172–7189.

Joshua Schwab-Cartas. 2018. Keeping Up with the Sun: Revitalizing Isthmus Zapotec and Ancestral Practices through Cellphilms. *The Canadian Modern Language Review*, 74(3):363–387.

Amalia Skilton, Sofia Gottlieb Pierson, Sunkulp Ananthanarayan, and Claire Bowern. 2025. Digital infrastructure and its impacts on language work: A case study of FieldWorks Language Explorer (FLEx). *Language*, 101(3):e136–e165.

Angela Taylor-Adams. 2019. Recording to revitalize: Language teachers and documentation design. *Language Documentation & Conservation*, 13:426–445.

Courtney Tennell and Kari AB Chew. 2024. Perspectives on relationality in online Indigenous language learning. *AlterNative: An International Journal of Indigenous Peoples*, 20(3):512–520.

Irina Wagner. 2017. New Technologies, Same Ideologies: Learning from Language Revitalization Online. *Language Documentation & Conservation*.

Joshua Wilbur. 2019. ELAN as a search engine for hierarchically structured, tagged corpora. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 90–103, Tartu, Estonia. Association for Computational Linguistics.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

A Systematic Comparison of Parameter-Efficient Fine-Tuning Techniques for Low-Resource Neural Machine Translation: Evidence from Indigenous Languages of the Americas

Drew Stackhouse, Justin DeBenedetto
Department of Computing Sciences, Villanova University
{dstackho,jdeben01}@villanova.edu

Abstract

We present the first systematic benchmark of parameter-efficient fine-tuning (PEFT) for low-resource neural machine translation (NMT) of indigenous languages of the Americas. We evaluate eight PEFT methods alongside full fine-tuning on NLLB-200-distilled-600M across 13 indigenous-to-Spanish language pairs spanning four resource tiers (357–125,008 training sentences). OFT (Orthogonal Finetuning) achieves the highest development-set chrF++ among PEFT methods (26.63) while training only 0.28% of parameters. LoRA (Low-Rank Adaptation) offers a strong efficiency–quality tradeoff (25.27 chrF++, 0.19%). On held-out test data, full fine-tuning ranks first (25.12) with OFT a close second (25.06; $p = 0.43$). VeRA (Vector-based Random Matrix Adaptation) and Prefix Tuning consistently underperform. These results demonstrate that PEFT is a viable alternative to full fine-tuning for indigenous-language NMT.

1 Introduction

Of the world's approximately 7,000 living languages, only a small proportion have sufficient digital resources to benefit from modern natural language processing (Joshi et al., 2020). The indigenous languages of the Americas are disproportionately affected. Hundreds of languages spanning dozens of language families are spoken across North, Central, and South America, yet most lack the parallel corpora, monolingual text, and standardized orthographies that contemporary NMT systems require (Mager et al., 2021). Machine translation has the potential to support language documentation, education, and access to information for these communities, but only if effective systems can be built from the limited data available.

Multilingual pretrained models have emerged as the dominant approach to low-resource NMT. Models such as NLLB-200 (NLLB Team et al., 2022) cover over 200 languages and can be fine-tuned on new language pairs with relatively small amounts of parallel data. However, full fine-tuning of these models by updating all 615 million parameters is computationally expensive, requires substantial GPU memory, and risks catastrophic forgetting of the pretrained representations that make transfer learning effective in the first place (Kirkpatrick et al., 2017). PEFT methods address these limitations by training only a small fraction of the model's parameters while keeping the pretrained weights frozen or nearly so (Houlsby et al., 2019).

The landscape of PEFT methods has expanded rapidly in recent years, with techniques ranging from low-rank weight decompositions (Hu et al., 2022) to orthogonal transformations (Qiu et al., 2023) to learned activation scaling (Liu et al., 2022). Each method makes different assumptions about how model weights should be adapted and each offers a different tradeoff between parameter efficiency and expressiveness. Despite this growing diversity, no systematic comparison of PEFT methods exists for low-resource indigenous-language NMT. Prior PEFT benchmarks have largely focused on high-resource or English-centric settings, leaving the question open of which methods are most effective when parallel data is measured in hundreds or thousands of sentences rather than millions.

This work addresses that gap. We make the following contributions:

1. The first comprehensive benchmark of eight PEFT methods for indigenous-language NMT, evaluated against a full fine-tuning baseline.

2. An analysis across 13 typologically diverse language pairs spanning four resource tiers (357–125,008 training sentences), revealing how data availability interacts with method effectiveness.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 33–45

3. A quantification of the parameter-efficiency vs. translation-quality tradeoff, identifying Pareto-optimal methods.

4. Practical guidelines for practitioners selecting PEFT methods for underserved languages based on resource tier and compute budget.

2 Background

2.1 Low-Resource Neural Machine Translation

Neural machine translation systems depend on large quantities of parallel text, and their performance degrades substantially when this data is scarce. For low-resource language pairs this dependence creates tension as the languages most in need of translation tools are precisely those for which training data is hardest to obtain. The problem is compounded by morphological complexity. Many of the indigenous languages of the Americas are agglutinative or polysynthetic. They form long, meaning-dense words through the concatenation of many morphemes, which means that word-level models encounter a larger effective vocabulary and more severe data sparsity than they would for analytic languages. Standard tokenization schemes developed for European languages perform poorly on these morphological profiles. Automatic metrics calibrated on word boundaries underestimate translation quality for systems that handle them well.

Responses to these challenges have taken several forms. Transfer learning from high-resource languages, particularly via multilingual pretrained models, has become the dominant paradigm, exploiting the observation that representations learned across many languages share useful structure (Joshi et al., 2020). The AmericasNLP shared tasks (2021–2025) have motivated community efforts specifically for indigenous language MT, producing parallel corpora, shared evaluation frameworks, and a growing set of competitive baselines. However, most published systems treat fine-tuning strategy as a secondary concern, focusing instead on data augmentation, back-translation, or ensemble methods. The question of which fine-tuning approach is most appropriate for this setting has received comparatively little systematic attention.

2.2 Multilingual Pretrained Models for MT

Phase 1 of this study compares three multilingual encoder-decoder candidates representing distinct pretraining philosophies: mBART-50 (Tang et al., 2021), a denoising-objective model covering 50 languages; ByT5-small (Xue et al., 2022), a byte-level model robust to orthographic variation at the cost of longer sequences; and NLLB-200 (NLLB Team et al., 2022), trained specifically for translation across 200 languages and the only candidate that natively includes several of our target languages (Quechua, Guarani, Aymara).

2.3 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning methods reduce the cost of adapting large pretrained models by training a small number of parameters while keeping the pretrained weights frozen or nearly so. For low-resource settings this both reduces memory and compute demands and limits the degree to which pretrained representations can be overwritten, partially mitigating catastrophic forgetting. We evaluate eight PEFT methods spanning four methodological families.

Low-Rank Reparameterization. LoRA (Hu et al., 2022) decomposes weight updates into a product of two low-rank matrices, initialized to zero so that training begins from the pretrained model's behavior; it has become the dominant PEFT paradigm and serves as our natural baseline. AdaLoRA (Zhang et al., 2023) adaptively reallocates the rank budget across weight matrices via SVD-based pruning. DoRA (Liu et al., 2024a) decomposes each weight into magnitude and direction, applying LoRA-style updates only to the directional component. VeRA (Kopiczko et al., 2024) shares a single pair of frozen random matrices across layers and learns only small per-layer scaling vectors, reducing trainable parameters dramatically.

Activation Scaling. IA^3 (Liu et al., 2022) learns three vectors per transformer layer that element-wise rescale keys, values, and feedforward activations, introducing fewer trainable parameters ($\sim$74K, 0.01% of base) than any other method in this benchmark.

Orthogonal Transformations. OFT (Qiu et al., 2023) constrains weight updates to be orthogonal transformations, preserving the pairwise angular relationships between neurons in the pretrained weight matrix. This constraint is motivated by the hypothesis that the relative geometry of learned representations matters more than their absolute positions, and preserving this geometry limits catastrophic forgetting. The orthogonal matrices are

parameterized via a block-diagonal Cayley transform which ensures orthogonality throughout training without requiring projection steps. BOFT (Liu et al., 2024b) factorizes this orthogonal transformation using butterfly matrices, a structured sparse decomposition that allows information to propagate across all dimensions of the weight matrix, thus achieving a better expressiveness–parameter tradeoff than OFT's block-diagonal structure.

Prompt-Based Methods. Prefix Tuning (Li and Liang, 2021) prepends learnable continuous "virtual tokens" to the key and value matrices at every layer, modifying behavior by steering attention patterns rather than altering any pretrained weights directly. A small MLP reparameterizes the prefix during training and is discarded at inference.

2.4 Evaluation Metrics for MT

We adopt chrF++ (Popović, 2017) as the primary evaluation metric and report BLEU (Papineni et al., 2002) as a secondary metric for comparability with prior work. chrF++'s character-level n-gram F-score (augmented with word unigrams and bigrams) is less sensitive to tokenization boundaries than BLEU and rewards partial morphological matches, a critical property for agglutinative and polysynthetic languages where a single word may encode information that would span an entire clause in an analytic language. For the indigenous languages in this study, many of which exhibit productive morphology and lack standardized orthographies, character-level evaluation captures meaningful overlap that word-level metrics miss entirely; BLEU additionally correlates poorly with human judgments under morphological richness (Bapna and Firat, 2019). All analytical conclusions are drawn from chrF++. Metric computations use SacreBLEU (Post, 2018).

3 Methodology

Our experimental design follows a three-phase structure.[1] Phase 1 selects a base model from three multilingual pretrained candidates. Phase 2 benchmarks eight PEFT methods plus full fine-tuning on development data. Phase 3 evaluates the trained models on held-out test sets to assess generalization. All experiments are repeated with three random seeds (0, 1, 2) to estimate variance.

[1]Code, configuration files, and experiment scripts are available at https://github.com/drewstackhouse/peft-nmt-americas.

Language	Family	Train	Dev	Tier
Quechua	Quechuan	125,008	996	High
Wayuu	Arawakan	59,715	6,635	High
Guarani	Tupian	26,032	995	High
Awajun	Jivaroan	21,964	1,018	Med.
Nahuatl	Uto-Aztecan	16,063	672	Med.
Raramuri	Uto-Aztecan	14,720	995	Med.
Shipibo-K.	Panoan	14,592	996	Med.
Wixarika	Uto-Aztecan	8,966	994	Low
Bribri	Chibchan	7,508	996	Low
Aymara	Aymaran	6,531	996	Low
Otomi	Oto-Manguean	4,889	599	Low
Ashaninka	Arawakan	3,883	883	V-Low
Chatino	Oto-Manguean	357	499	V-Low

Table 1: Summary of the 13 indigenous-language-to-Spanish translation pairs. Tier boundaries: very-low (<5K), low (5K–10K), medium (10K–25K), high (>25K training sentences).

3.1 Languages and Data

We use parallel corpora for 13 indigenous-language-to-Spanish translation pairs drawn from the AmericasNLP shared-task datasets (Mager et al., 2021; Ebrahimi et al., 2022; Chiruzzo et al., 2024). The languages span 10 language families and exhibit substantial typological diversity, including agglutinative (Quechua, Aymara, Nahuatl), polysynthetic (Ashaninka, Guarani), and tonal (Chatino, Bribri, Otomi) profiles. Training set sizes range from 357 sentences (Chatino) to 125,008 sentences (Quechua), a 350-fold difference that motivates grouping languages into four resource tiers for analysis: *very-low* (<5K: Chatino, Ashaninka), *low* (5K–10K: Otomi, Aymara, Bribri, Wixarika), *medium* (10K–25K: Shipibo-Konibo, Raramuri, Nahuatl, Awajun), and *high* (>25K: Guarani, Wayuu, Quechua). These tiers are defined relative to the data available in this study; even the high tier (up to 125K sentences) remains far below what is typically available for high-resource languages such as French or German. All data are formatted as tab-separated parallel sentences with NLLB-style language codes as column headers. Table 1 summarizes the languages and data sizes.

3.2 Phase 1: Base Model Selection

We compare three multilingual pretrained models as candidates for fine-tuning: NLLB-200-distilled-600M (NLLB Team et al., 2022), mBART-large-50-many-to-many (Tang et al., 2021), and ByT5-small (Xue et al., 2022). To control for the adaptation method, we apply LoRA with identical hyperparameters to each model and evaluate on a strati-

Method	Targets	Params	% Base
Full FT	all	615.1M	100.00
AdaLoRA	q, v	1,771K	0.29
OFT	q, k, v, o	1,714K	0.28
LoRA	q, v	1,181K	0.19
DoRA	q, v	665K	0.11
VeRA	q, v	616K	0.10
BOFT	q, k, v, o	553K	0.09
Prefix Tuning	–	492K	0.08
IA^3	k, v, wo	74K	0.01

Table 2: Trainable parameter counts and percentage of the 615M-parameter base model for each adaptation method. Methods are ordered by parameter count.

fied subset of four languages spanning the resource tiers: Chatino (very-low), Bribri (low), Nahuatl (medium), and Guarani (high). The model with the highest mean chrF++ across languages and seeds is selected for all subsequent experiments.

3.3 Model and Tokenizer Preparation

The selected base model is NLLB-200-distilled-600M, an encoder-decoder transformer with approximately 615 million parameters. Of the 13 source languages, three—Quechua, Guarani, and Aymara—are natively present in the NLLB-200 vocabulary. For the remaining 10 languages, we add a new language token to the tokenizer and initialize its embedding by copying from the Spanish (`spa_Latn`) token embedding. This initialization provides a reasonable starting point for the source language representation while requiring no additional training data. Embedding layers remain frozen during PEFT training.

3.4 PEFT Method Configurations

We evaluate eight PEFT methods spanning four methodological families, each configured following its original paper's recommendations. Full per-method hyperparameters are provided in Appendix A. Table 2 summarizes the trainable parameter counts.

3.5 Training Configuration

All experiments share a common training configuration. We train for a maximum of 15 epochs with an effective batch size of 64 (per-device batch size 16 with 4 gradient accumulation steps). The optimizer is AdamW with weight decay 0.01, warmup ratio 0.06, and label smoothing 0.1. The default learning rate is 1×10^{-3} for all PEFT methods except DoRA, which uses 5×10^{-4} for training stability; full fine-tuning uses 3×10^{-5}. Training uses mixed-precision bf16 arithmetic where supported.

We evaluate at the end of each epoch using greedy decoding (beam size 1) and save the model checkpoint that achieves the highest development-set chrF++. Early stopping with a patience of 5 epochs terminates training if no improvement is observed. The maximum sequence length is 128 tokens for both source and target.

3.6 Inference and Evaluation

At inference time, PEFT adapter weights are merged into the base model for all methods except Prefix Tuning, which retains its PEFT wrapper due to architectural incompatibility with weight merging. Translations are generated using greedy decoding (beam size 1) with a maximum output length of 128 tokens, consistent with the decoding strategy used during development-set evaluation.

We compute chrF++ (Popović, 2017) as the primary metric and BLEU (Papineni et al., 2002) as a secondary metric, both via the SacreBLEU implementation (Post, 2018). chrF++ uses word order 2 (i.e., word unigrams and bigrams as additional features beyond character n-grams).

3.7 Statistical Analysis

We assess statistical significance using paired bootstrap resampling (Koehn, 2004) with 10,000 resamples and a two-sided test. Each method comparison is based on 13 paired observations (one mean chrF++ score per language pair, averaged over three seeds). We adopt a significance threshold of $p < 0.05$.

4 Results

4.1 Phase 1: Base Model Selection

Table 3 compares the three candidate base models. NLLB-200-distilled-600M achieves the highest mean chrF++ (26.13) across the four evaluation languages, outperforming mBART-large-50 (24.32) by 1.81 points and ByT5-small (13.92) by 12.21 points. The gap between NLLB and mBART is moderate but consistent, while ByT5 lags substantially, likely due to its byte-level tokenization producing very long sequences that exceed the 128-token training limit. NLLB also trains fastest (0.47 hours/run vs. 0.59 for mBART and 0.70 for ByT5). Based on these results, we select NLLB-200-distilled-600M as the base model for all subsequent experiments.

Model	chrF++	BLEU	Hours
NLLB-200-600M	**26.13**	**7.78**	**0.47**
mBART-large-50	24.32	5.81	0.59
ByT5-small	13.92	1.21	0.70

Table 3: Phase 1 model selection. Mean chrF++ and BLEU across four languages (Chatino, Bribri, Nahuatl, Guarani) and three seeds, all using LoRA adaptation. Hours = mean training time per language-seed run.

Method	chrF++	BLEU	Hours
OFT	**26.63**	7.82	1.63
Full FT	26.13	**7.84**	0.96
LoRA	25.27	7.16	0.88
AdaLoRA	23.77	6.27	0.94
DoRA	23.50	6.10	1.29
BOFT	22.99	5.90	2.56
IA^3	21.03	5.00	0.86
Prefix Tuning	19.63	4.01	0.66
VeRA	18.69	4.14	0.97

Table 4: Phase 2 development-set results. Mean chrF++ and BLEU across 13 language pairs and 3 seeds. Hours = mean training time per language-seed run.

4.2 Phase 2: Development Set Performance

4.2.1 Overall Method Ranking

Table 4 presents the overall development-set results. OFT achieves the highest mean chrF++ (26.63), followed closely by full fine-tuning (26.13) and LoRA (25.27). The top three methods are separated by just 1.36 chrF++ points, while the gap between the best and worst methods (OFT vs. VeRA) spans 7.94 points. Notably, OFT surpasses full fine-tuning while training only 0.28% of the parameters, demonstrating that parameter efficiency need not come at the cost of translation quality.

The middle tier—AdaLoRA (23.77), DoRA (23.50), and BOFT (22.99)—achieves moderate performance, trailing LoRA by 2.3–3.6 points despite similar or fewer trainable parameters. The bottom tier—IA^3 (21.03), Prefix Tuning (19.63), and VeRA (18.69)—shows that the most parameter-efficient methods sacrifice substantial translation quality.

Full fine-tuning achieves the highest BLEU (7.84) despite ranking second in chrF++, reflecting chrF++'s greater sensitivity to character-level overlap in morphologically rich languages.

4.2.2 Per-Language Results

OFT wins the most language pairs on the development set, achieving the highest chrF++ for 9 of 13 languages. Full fine-tuning wins 3 languages (Aymara, Wayuu, Quechua—all in the high tier), and VeRA wins 1 (Guarani, by a negligible margin). No single method dominates across all languages, but the top-three methods (OFT, Full, LoRA) are remarkably consistent, each appearing in the top 3 for at least 10 of 13 language pairs. Lower-ranked methods show high variance—Prefix Tuning, for instance, achieves 23.75 chrF++ on Chatino (outperforming LoRA) but only 12.21 on Raramuri.

Method	chrF++	BLEU
Full FT	**25.12**	**7.18**
OFT	25.06	6.27
LoRA	22.70	5.32
BOFT	21.46	4.76
AdaLoRA	21.27	5.03
DoRA	20.94	4.76
IA^3	20.19	4.57
Prefix Tuning	19.57	3.80
VeRA	18.43	3.86

Table 5: Phase 3 test-set results. Mean chrF++ and BLEU across 13 language pairs and 3 seeds.

4.2.3 Performance by Resource Tier

Performance varies substantially across resource tiers. In the *very-low* tier (<5K sentences), OFT leads by a wide margin (23.98 chrF++), outperforming the second-best method (full fine-tuning, 20.69) by 3.29 points. This advantage narrows in the *low* tier (OFT 26.04 vs. full 25.57) and *medium* tier (OFT 24.94 vs. full 24.64). In the *high* tier (>25K sentences), full fine-tuning takes the lead (32.51 vs. OFT 31.45), suggesting that full parameter updates become advantageous when sufficient data are available.

LoRA consistently ranks third across all tiers, maintaining a 1–2 point gap behind the leader. The bottom-tier methods (Prefix Tuning, VeRA) show the steepest performance degradation as data decreases: VeRA drops from 27.94 chrF++ in the high tier to 11.07 in the very-low tier, a 16.87-point decline.

4.3 Phase 3: Test Set Performance

4.3.1 Overall Test Results

Table 5 presents the held-out test results. Full fine-tuning achieves the highest mean chrF++ (25.12), narrowly surpassing OFT (25.06). LoRA remains third (22.70), followed by BOFT (21.46) and AdaLoRA (21.27). The bottom three methods maintain their rankings: IA^3 (20.19), Prefix Tuning (19.57), and VeRA (18.43).

On the test set, full fine-tuning wins 8 of 13 language pairs, with OFT winning the remaining 5. OFT retains its advantage on very-low-tier languages (Chatino, Ashaninka) and several low-tier pairs, while full fine-tuning dominates the medium and high tiers.

4.3.2 Dev-to-Test Generalization

Method rankings are largely stable between development and test sets. Six of nine methods retain their dev-set rank on the test set; the largest rank change is BOFT, which rises from 6th to 4th. The top-two swap (OFT $\rightarrow$ 2nd, Full $\rightarrow$ 1st) reflects a difference of only 0.06 chrF++ on test, well within noise.

Dev-to-test chrF++ drops vary considerably across methods. LoRA (-2.57), DoRA (-2.56), and AdaLoRA (-2.51) show the largest degradation, suggesting some overfitting to development data. Full fine-tuning (-1.01) and OFT (-1.57) degrade moderately. Prefix Tuning (-0.06) and VeRA (-0.27) show minimal dev-to-test gaps, though this stability reflects consistently low performance rather than robust generalization.

4.4 Parameter Efficiency Analysis

Figure 1 shows that parameter count is a poor predictor of translation quality: OFT (1.71M parameters) outperforms AdaLoRA (1.77M) by 2.86 chrF++ despite nearly identical budgets, and IA^3 (74K) outperforms VeRA (616K) by 2.33 points with $8\times$ fewer parameters, confirming that adaptation mechanism matters more than parameter count. Evaluated against three objectives simultaneously—maximize chrF++, minimize trainable parameters, and minimize training time—seven of nine methods are Pareto-optimal, each taking a unique position on the tradeoff frontier: OFT at the quality front, IA^3 at the minimum-parameter front, Prefix Tuning at the minimum-time front, and LoRA, Full FT, DoRA, and BOFT at intermediate positions. Only AdaLoRA and VeRA are fully dominated: LoRA strictly beats AdaLoRA on all three dimensions, and IA^3 strictly beats VeRA; practitioners have no reason to prefer either method.

4.5 Statistical Significance

Paired bootstrap resampling (10,000 resamples, two-sided, $\alpha = 0.05$) was applied to both the development and test sets; full results are in Appendix B. On the development set, 32 of 36 pairwise comparisons are significant. The four non-significant pairs are: Full vs. OFT ($p = 0.097$), AdaLoRA vs. DoRA ($p = 0.052$), IA^3 vs. Prefix Tuning ($p = 0.135$), and Prefix Tuning vs. VeRA ($p = 0.229$).

On the test set, 29 of 36 comparisons are significant. Full vs. OFT becomes even less significant ($p = 0.425$), further supporting the central finding that there is no statistically significant difference between the two methods. Three additional pairs lose significance on test: AdaLoRA vs. BOFT ($p = 0.291$), IA^3 vs. Prefix Tuning ($p = 0.293$), and AdaLoRA vs. Prefix Tuning ($p = 0.111$), along with BOFT vs. Prefix Tuning ($p = 0.074$) and DoRA vs. Prefix Tuning ($p = 0.136$). These new non-significant pairs cluster around Prefix Tuning, reflecting that the bottom tier of methods are genuinely close in absolute performance and their small differences are overwhelmed by noise on held-out data.

Across both sets, the non-significance of Full vs. OFT is the most consequential result: there is no statistically significant difference in translation quality between OFT and full fine-tuning, while OFT trains fewer than 0.3% of the parameters. This is the central finding of the study.

4.6 Error Analysis

Aggregate metrics like chrF++ and BLEU summarize translation quality as a single number, but two methods with similar scores can exhibit qualitatively different failure modes. To characterize these differences, we compute four diagnostic metrics over the test-set translations: *repetition*, *source copy ratio*, *length ratio*, and *entity preservation F1*.

Repetition measures self-repetition as $1 - (\text{unique n-grams}/\text{total n-grams})$, macro-averaged over $n \in \{1, 2, 3\}$ (range $[0, 1]$; lower is better). **Source copy ratio** computes the multi-set intersection of hypothesis tokens with source tokens, divided by hypothesis length (range $[0, 1]$; lower is better), capturing the degree to which the model copies source text rather than translating. **Length ratio** divides hypothesis length by reference length (ideal $= 1.0$); values below 1 indicate under-generation and values above 1 indicate over-generation, aggregated via the median at the sentence level to resist outliers. **Entity F1** extracts numeric entities from hypothesis and reference via regex and computes F1 over their multi-set overlap (range $[0, 1]$; higher is better), restricted to the subset of sentences whose reference contains at least one entity.

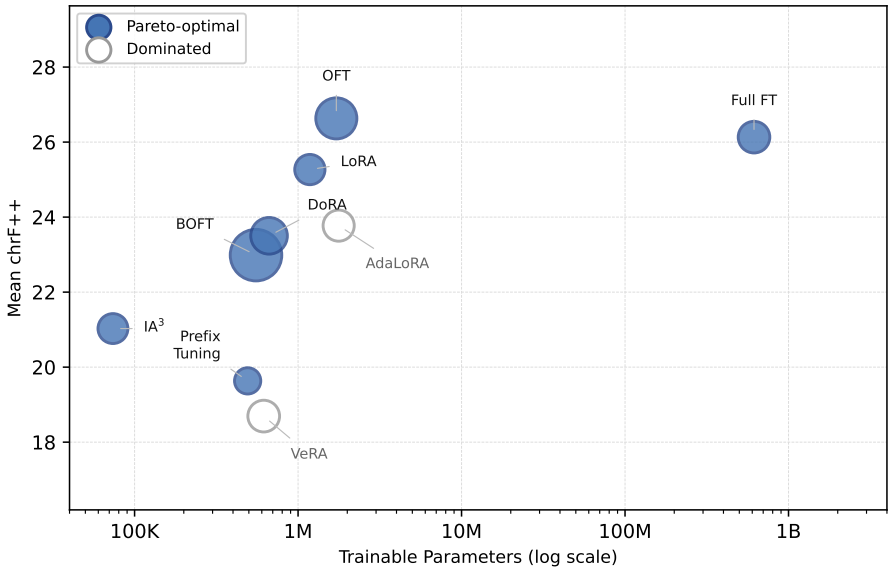

Figure 1: Efficiency–quality tradeoff across all nine methods (development set). Each bubble is one method; bubble area is proportional to mean training hours per language-pair run (BOFT largest at 2.6 h; Prefix Tuning smallest at 0.7 h). Filled circles are Pareto-optimal across three objectives simultaneously (maximize chrF^{++}, minimize trainable parameters, minimize training time); hollow circles (AdaLoRA, VeRA) are strictly dominated on all three. Full FT sits at 615M parameters (far right); all PEFT methods fall below 2M.

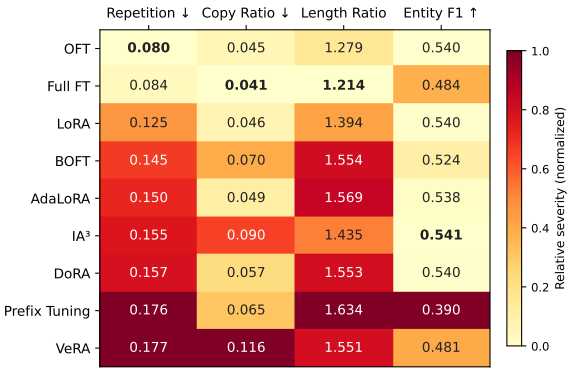

Figure 2: Heatmap of diagnostic metrics by method. Color intensity reflects relative severity (normalized per column); yellow indicates better performance and red indicates worse. Cell values are raw (un-normalized) scores.

Full diagnostic metrics are reported in Appendix C, and Figure 2 summarizes the failure profiles. OFT and full fine-tuning produce the least repetition (0.080 and 0.084 respectively) and lowest source copy ratios (0.045 and 0.041), confirming that their chrF++ advantage reflects genuinely cleaner translations rather than superficial n-gram matching. Full fine-tuning also achieves the length ratio closest to 1.0 (1.214), while all other methods over-generate by 28–63%.

However, aggregate chrF++ rankings do not predict all dimensions of quality. Full fine-tuning achieves the best chrF++ but the *third-worst* entity F1 (0.484), behind only Prefix Tuning (0.390) and VeRA (0.481), indicating that it drops or distorts numeric content more often than most PEFT methods. Conversely, IA^3—which ranks 7th in chrF++—achieves the *best* entity F1 (0.541), suggesting that its minimal activation-scaling intervention preserves factual content more faithfully than methods that modify weight matrices directly. VeRA exhibits the worst repetition (0.177) and source copy ratio (0.116), consistent with its low chrF++ and suggesting that its shared-random-matrix parameterization struggles to learn a genuine translation function under low-resource conditions.

Failure modes intensify as training data decreases. In the very-low resource tier (<5K sentences), repetition scores roughly triple compared to the high tier for most methods: AdaLoRA rises from 0.105 to 0.427, and BOFT from 0.123 to 0.404. Length ratios become extreme, with AdaLoRA (2.99), BOFT (2.85), and DoRA (2.83) producing hypotheses nearly three times the reference length—a signature of repetitive over-generation. OFT degrades most gracefully: its very-low repetition (0.216) and length ratio (1.59) remain the best in the tier, which explains the 3-point chrF++ margin it holds over all other methods in that setting. The full breakdown by resource tier is provided in Appendix D.

5 Conclusion

5.1 Summary of Findings

This study compared eight PEFT methods and full fine-tuning for neural machine translation of 13 indigenous languages of the Americas into Spanish. Four findings stand out.

First, OFT and full fine-tuning do not differ significantly in translation quality ($p = 0.097$), with OFT training only 0.28% of the base model's parameters. On the development set, OFT leads (26.63 chrF++); on the held-out test set, full fine-tuning leads marginally (25.12 vs. 25.06). This demonstrates that orthogonal transformations preserve pretrained knowledge effectively under data scarcity.

Second, PEFT method effectiveness interacts strongly with data availability. OFT excels in the very-low tier (<5K sentences), leading full fine-tuning by 3.29 chrF++ points on dev and 3.00 points on test. Full fine-tuning takes the lead in the high tier (>25K sentences), where sufficient data mitigate overfitting. LoRA offers consistent third-place performance across all tiers.

Third, parameter count alone does not predict translation quality. OFT (1.71M parameters) outperforms AdaLoRA (1.77M) by 2.86 chrF++ points despite similar parameter budgets, and IA^3 (74K parameters) outperforms VeRA (616K) by 2.33 points despite having 8× fewer parameters. The adaptation mechanism matters more than the number of trainable parameters.

Fourth, error analysis reveals that aggregate scores mask method-specific failure profiles. Full fine-tuning achieves the best chrF++ but the third-worst entity preservation F1 (0.484), while IA^3 shows the opposite pattern (7th in chrF++, best entity F1 at 0.541). This suggests that method selection should consider which dimensions of translation quality matter most for a given application, not just a single aggregate score.

5.2 Practical Recommendations

For practitioners working with low-resource indigenous languages, we offer the following guidance:

- **Base model:** NLLB-200-distilled-600M provides the strongest starting point among the models tested, particularly for languages not in its pretraining vocabulary.
- **Default PEFT method:** OFT is recommended as the default choice, achieving the best or near-best quality across resource levels with a 359× parameter reduction.
- **Speed-constrained settings:** LoRA offers a strong alternative with training time half that of OFT and consistently strong performance.
- **Very-low-tier languages:** OFT is particularly advantageous when training data is extremely scarce (<5K sentences), outperforming all alternatives by a substantial margin.
- **Methods to avoid:** VeRA and Prefix Tuning consistently underperform, ranking 8th–9th across conditions. BOFT requires the most training time while achieving below-median quality.

5.3 Future Work

Future work should explore bidirectional translation (Spanish to indigenous languages), larger base models (NLLB-200-1.3B, 3.3B), per-method hyperparameter optimization (e.g., rank sweeps for LoRA, target module ablations for OFT), combinations of PEFT methods, data augmentation via back-translation for very-low tiers, and human evaluation with indigenous language community members.

Limitations

Target module configurations differ across PEFT methods, following each method's original paper recommendations rather than a single standardized configuration. This means methods differ in both adaptation mechanism and scope of modified parameters, which is realistic for practitioners but prevents a pure apples-to-apples comparison of mechanisms.

Learning rates differ across conditions: 1×10^{-3} for most PEFT methods, 5×10^{-4} for DoRA, and 3×10^{-5} for full fine-tuning. These rates reflect necessary tuning for training stability but introduce an additional variable.

No per-method hyperparameter search was conducted (e.g., rank sweeps for LoRA, target module ablations). All methods use a single recommended configuration; results therefore reflect default performance rather than each method's ceiling.

Statistical power is limited by $n = 13$ language pairs for the paired bootstrap test. While 32 of 36 pairwise comparisons reach significance, only large systematic differences are detectable at this sample size.

ByT5-small (∼300M parameters) is substantially smaller than NLLB (∼615M) and mBART (∼611M), making the Phase 1 comparison not purely architecture-controlled.

Evaluation relies on automatic metrics only (chrF++, BLEU). Human evaluation of translation quality, adequacy, and fluency would provide complementary insight, particularly for morphologically complex languages where automatic metrics may not capture meaning preservation.

All translations are unidirectional (indigenous language to Spanish). The reverse direction poses different challenges (generation in morphologically rich languages) and may yield different method rankings.

Greedy decoding (beam size 1) was used throughout for consistency. Beam search could alter relative method rankings, particularly for methods that produce higher-entropy output distributions.

Only one base model scale was tested (600M distilled). Results may not transfer to larger NLLB variants (1.3B, 3.3B), where the relative advantage of PEFT over full fine-tuning could differ.

Ethical Statement

All parallel corpora used in this work were drawn from the AmericasNLP shared-task datasets, which were compiled in collaboration with indigenous language communities and released for research purposes. We have not collected, redistributed, or modified any community data, and we follow the terms under which these datasets were made available.

The goal of this research is to lower the computational barrier to building NMT systems for underserved languages rather than to produce deployment-ready translation tools. The chrF++ scores reported here, while meaningful for benchmarking, reflect the inherent difficulty of low-resource MT: translations generated by these systems may be fluent but inaccurate. Any practical use of models trained on these data should involve review by speakers of the target language community.

We recognize that the development of NLP tools for endangered and minority languages carries both promise and risk. At best, such tools support language documentation, education, and access to information. At worst, they can be used to displace human translators, misrepresent community voices, or create a false impression of language vitality. We encourage future work in this space to be conducted in partnership with the communities whose languages are involved.

Acknowledgments

This work used the Augie High-Performance Computing cluster, funded by award NSF 2018933, at Villanova University.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1538–1548. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second AmericasNLP competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6282–6293. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 388–395. Association for Computational Linguistics.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. VeRA: Vector-based random matrix adaptation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, volume 1, pages 4582–4597. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and David Ha. 2024b. Parameter-efficient orthogonal finetuning via butterfly factorization. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2021)*, pages 202–217. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, and Al Youngblood. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT 2017)*, pages 612–618. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 186–191. Association for Computational Linguistics.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. 2023. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics (TACL)*, 10:291–306.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

A PEFT Method Configurations

Each method is configured following the recommendations in its original paper rather than imposing a single standardized configuration across all methods. We consider this the most informative comparison for practitioners as it reflects the performance each method achieves under its intended operating conditions.

LoRA: rank 8, $\alpha = 16$, dropout 0.1, targeting query and value projections (q, v).

AdaLoRA: initial rank 12, target rank 8, $\alpha = 16$, dropout 0.1, targeting q and v, with orthogonality regularization weight 0.5.

DoRA: rank 4, $\alpha = 16$, dropout 0.1, targeting q and v, with weight decomposition into magnitude and direction components enabled.

VeRA: rank 256, targeting q and v, with shared frozen random matrices and learned per-layer scaling vectors initialized to 0.1.

IA^3: targeting key, value, and feedforward output projections (k, v, wo), with scaling vectors initialized to 1.0 (identity).

OFT: rank 32, targeting all attention projections (q, k, v, o), with block-diagonal Cayley parameterization initialized to the identity transformation.

BOFT: block size 4, butterfly factor 1, targeting q, k, v, and o projections.

Prefix Tuning: 20 virtual tokens with MLP reparameterization during training.

Full fine-tuning serves as the baseline, updating all 615 million parameters.

A noteworthy design choice is that low-rank methods (LoRA, AdaLoRA, DoRA, VeRA) target only query and value projections, while orthogonal methods (OFT, BOFT) target all four attention projections. This follows the respective authors' recommendations and reflects a real-world practitioner scenario, although it means the methods differ in mechanism and scope.

B Pairwise Bootstrap Significance Tests

Tables 6 and 7 report all 36 pairwise bootstrap significance tests on the development and test sets respectively. Each test uses 10,000 paired resamples over 13 language pairs, two-sided, at $\alpha = 0.05$. ΔchrF++ = chrF++(Method A) $-$ chrF++(Method B); Method A is always the higher-scoring method so $\Delta > 0$ throughout. Rows are sorted by ΔchrF++ descending within each significance tier; a horizontal rule separates significant ($p < 0.05$) from non-significant pairs.

Method A	Method B	ΔchrF++	p	Sig.
OFT	VeRA	7.94	<0.001	✓
Full FT	VeRA	7.44	<0.001	✓
OFT	Prefix Tuning	7.00	<0.001	✓
LoRA	VeRA	6.57	<0.001	✓
Full FT	Prefix Tuning	6.50	<0.001	✓
LoRA	Prefix Tuning	5.63	<0.001	✓
OFT	IA^3	5.60	<0.001	✓
Full FT	IA^3	5.10	<0.001	✓
AdaLoRA	VeRA	5.08	<0.001	✓
DoRA	VeRA	4.81	<0.001	✓
BOFT	VeRA	4.29	<0.001	✓
LoRA	IA^3	4.24	<0.001	✓
OFT	BOFT	3.65	<0.001	✓
Full FT	BOFT	3.15	<0.001	✓
OFT	DoRA	3.13	<0.001	✓
OFT	AdaLoRA	2.86	<0.001	✓
Full FT	DoRA	2.63	<0.001	✓
AdaLoRA	IA^3	2.74	<0.001	✓
Full FT	AdaLoRA	2.36	<0.001	✓
LoRA	BOFT	2.28	<0.001	✓
DoRA	IA^3	2.47	<0.001	✓
IA^3	VeRA	2.33	<0.001	✓
BOFT	IA^3	1.96	<0.001	✓
LoRA	DoRA	1.77	<0.001	✓
LoRA	AdaLoRA	1.49	<0.001	✓
OFT	LoRA	1.36	<0.001	✓
Full FT	LoRA	0.86	<0.001	✓
AdaLoRA	Prefix Tuning	4.14	0.002	✓
DoRA	Prefix Tuning	3.87	0.003	✓
BOFT	Prefix Tuning	3.35	0.003	✓
AdaLoRA	BOFT	0.79	0.016	✓
DoRA	BOFT	0.52	0.017	✓
AdaLoRA	DoRA	0.27	0.052	
OFT	Full FT	0.50	0.097	
IA^3	Prefix Tuning	1.39	0.135	
Prefix Tuning	VeRA	0.94	0.229	

Table 6: All 36 pairwise bootstrap significance tests, development set (Phase 2). Sorted by ΔchrF++ descending within each tier.

C Aggregate Error Analysis

Table 8 reports four diagnostic metrics averaged across all 13 language pairs and 3 seeds on the test set. Metric definitions are given in Section 4.6.

D Error Analysis by Resource Tier

Table 9 provides the full breakdown of diagnostic metrics by method and resource tier, aggregated across languages within each tier and across 3 seeds.

Method A	**Method B**	**ΔchrF++**	p	**Sig.**
Full FT	VeRA	6.69	<0.001	✓
OFT	VeRA	6.64	<0.001	✓
Full FT	Prefix Tuning	5.55	<0.001	✓
OFT	Prefix Tuning	5.49	<0.001	✓
Full FT	IA^3	4.93	<0.001	✓
OFT	IA^3	4.87	<0.001	✓
LoRA	VeRA	4.27	<0.001	✓
Full FT	DoRA	4.17	<0.001	✓
OFT	DoRA	4.12	<0.001	✓
Full FT	AdaLoRA	3.85	<0.001	✓
OFT	AdaLoRA	3.80	<0.001	✓
Full FT	BOFT	3.66	<0.001	✓
OFT	BOFT	3.60	<0.001	✓
LoRA	Prefix Tuning	3.12	<0.001	✓
BOFT	VeRA	3.03	<0.001	✓
AdaLoRA	VeRA	2.84	<0.001	✓
DoRA	VeRA	2.52	<0.001	✓
LoRA	IA^3	2.50	<0.001	✓
Full FT	LoRA	2.42	<0.001	✓
OFT	LoRA	2.37	<0.001	✓
IA^3	VeRA	1.77	<0.001	✓
LoRA	DoRA	1.75	<0.001	✓
LoRA	AdaLoRA	1.43	<0.001	✓
BOFT	IA^3	1.27	<0.001	✓
AdaLoRA	IA^3	1.07	<0.001	✓
LoRA	BOFT	1.24	0.004	✓
DoRA	IA^3	0.75	0.005	✓
BOFT	DoRA	0.52	0.026	✓
AdaLoRA	DoRA	0.32	0.043	✓
BOFT	Prefix Tuning	1.89	0.074	
AdaLoRA	Prefix Tuning	1.69	0.111	
DoRA	Prefix Tuning	1.37	0.136	
Prefix Tuning	VeRA	1.15	0.144	
BOFT	AdaLoRA	0.19	0.291	
IA^3	Prefix Tuning	0.62	0.293	
Full FT	OFT	0.06	0.425	

Table 7: All 36 pairwise bootstrap significance tests, test set (Phase 3). Sorted by ΔchrF++ descending within each tier. Note that Full FT and OFT are the final non-significant pair ($p = 0.425$), with Full FT leading OFT by only 0.06 chrF++ on held-out data.

Method	Repetition	Copy Ratio	Length Ratio	Entity F1
OFT	**0.080**	0.045	1.279	0.540
Full FT	0.084	**0.041**	**1.214**	0.484
LoRA	0.125	0.046	1.394	0.540
BOFT	0.145	0.070	1.554	0.524
AdaLoRA	0.150	0.049	1.569	0.538
IA^3	0.155	0.090	1.435	**0.541**
DoRA	0.157	0.057	1.553	0.540
Prefix Tuning	0.176	0.065	1.634	0.390
VeRA	0.177	0.116	1.551	0.481

Table 8: Error analysis: diagnostic metrics averaged across 13 language pairs and 3 seeds on the test set. Bold indicates best value per column (lowest for Repetition and Copy Ratio, closest to 1.0 for Length Ratio, highest for Entity F1).

Method	Tier	Repetition	Copy Ratio	Length Ratio	Entity F1
Full FT	high	0.049	0.035	1.043	0.510
OFT	high	0.077	0.035	1.122	0.552
LoRA	high	0.101	0.040	1.178	0.584
AdaLoRA	high	0.105	0.038	1.161	0.590
BOFT	high	0.123	0.040	1.192	0.520
VeRA	high	0.129	0.033	1.109	0.535
DoRA	high	0.131	0.041	1.205	0.596
Prefix Tuning	high	0.143	0.037	1.225	0.421
IA³	high	0.146	0.037	1.180	0.569
OFT	low	0.044	0.055	1.240	0.615
Full FT	low	0.060	0.037	1.228	0.542
BOFT	low	0.062	0.083	1.279	0.603
LoRA	low	0.072	0.046	1.320	0.587
AdaLoRA	low	0.076	0.063	1.287	0.607
DoRA	low	0.077	0.077	1.288	0.596
IA³	low	0.083	0.105	1.282	0.606
Prefix Tuning	low	0.091	0.094	1.465	0.475
VeRA	low	0.127	0.150	1.428	0.531
Full FT	medium	0.043	0.051	1.139	0.430
OFT	medium	0.051	0.047	1.281	0.484
LoRA	medium	0.102	0.054	1.396	0.493
BOFT	medium	0.114	0.092	1.451	0.494
AdaLoRA	medium	0.119	0.049	1.448	0.484
IA³	medium	0.122	0.132	1.347	0.499
DoRA	medium	0.131	0.058	1.441	0.494
VeRA	medium	0.195	0.155	1.552	0.422
Prefix Tuning	medium	0.222	0.069	1.951	0.283
OFT	very-low	0.216	0.036	1.589	0.487
Full FT	very-low	0.268	0.035	1.595	0.433
Prefix Tuning	very-low	0.305	0.038	1.950	0.388
VeRA	very-low	0.310	0.094	2.456	0.418
LoRA	very-low	0.315	0.040	1.862	0.473
IA³	very-low	0.377	0.056	2.301	0.450
BOFT	very-low	0.404	0.047	2.852	0.435
DoRA	very-low	0.406	0.043	2.827	0.436
AdaLoRA	very-low	0.427	0.041	2.990	0.429

Table 9: Diagnostic metrics by method and resource tier (test set). Within each tier, methods are sorted by repetition (ascending).

Linguistic Feature Tagging for Automatic Classification of 27 Closely-Related Quechua Varieties

Claire Benét Post and Alexis Palmer
University of Colorado, Boulder
Department of Linguistics
{benet.post,alexis.palmer} @ colorado.edu

Abstract

This paper presents a multi-dialect text classifier for Quechua that augments neural models with rule-based, linguistic information to address challenges in low-resource, morphologically complex settings. The approach is built on a carefully curated dataset spanning multiple genres, including annotated parallel bible corpora, and encodes manually annotated lexical variation and polypersonal verbal agreement as explicit features within a transformer-based classifier. Results show that neural models substantially outperform statistical baselines, enabling highly accurate multi-class classification across 27 Quechua dialects. The impact of linguistic augmentation is context-dependent: gains are minimal in high-resource settings but more pronounced in low-resource and cross-domain conditions. Overall, this work aims to contribute to the development of dialect-sensitive NLP methods for Quechua and other low-resource, morphologically rich languages.

1 Introduction

Quechua is a family of languages spoken across the Andean region, comprised of approximately 40 dialects and 9 million speakers (Adelaar, 2020; Adelaar and Muysken, 2004; Grimes, 1985; Hornberger and King, 1998). These dialects vary substantially in orthography, morphology, and syntax (Hornberger and Limerick, 2019; Limerick, 2018), creating challenges for natural language processing (NLP), particularly given the limited availability of dialect-specific resources.

Most existing NLP approaches treat Quechua as a monolithic language, collapsing dialectal distinctions into a single standardized form. As illustrated in Table 1,[1] machine translation (MT) systems like Google Translate produce hybridized morphosyntactic outputs that fail to reflect any individual variety. This dialectal homogenization not only reduces linguistic fidelity but also reinforces systemic biases in language technologies, further marginalizing underrepresented speaker communities (Blasi et al., 2022; Liu et al., 2022; Ziems et al., 2022). Similar issues arise in other indigenous language families of the Americas, such as Nahuatl and Maya, which exhibit rich dialectical variation but remain computationally underrepresented and homogenized (García et al., 2021; Riemland, 2023).

ISO	Family	"He sent me."
inb	CU	Pai-mi kacha-mu-wa-rka.
		3sg-DIR send-MOV-3sg→1sg-PST
quw	CU	Pi-mi ñuca-ra cacha-mu-ca.
		3sg-DIR 1sg-DO send-MOV-3sg→1sg.PST
qub	Q1	Pay-mi noga-ta-ga cacha-masha.
		3sg-DIR 1sg-DO-EMP send-3sg→1sg.PST
qvn	Q1	Pay-mi cachra-ra-yä-man noga-ta-ga.
		3sg-DIR send-3sg→1sg-PST-from 1sg-DO-EMP
quh	Q2	Pay-taj kacha-mu-wa-rqa.
		3sg-CON send-MOV-3sg→1sg-PST
quz	Q2	Pay-taq-mi kacha-mu-wan-pas.
		3sg-CON-DIR send-MOV-3sg→1sg-ADD
Google Translate		Pay-mi kacha-mu-wa-rqa.
		3sg-DIR send-MOV-3sg→1sg-PST

Table 1: Automatic translation of "He sent me" vs. text from manually glossed Bible corpora. Note that the MT output does not align with any particular variety, nor with any subfamily.

Even when NLP tools for Quechua exist, they are typically limited in scope, often focusing on Southern Quechua (Zevallos et al., 2022; Rios et al., 2008; Rios Gonzales and Castro Mamani, 2014), only a handful of dialects (Medina, 2013; Melgarejo et al., 2022; Vergara, 2022), or lacking dialectal specificity altogether (Chen et al., 2024; Monson et al., 2006).[2] This concentration of resources risks encoding Southern Quechua dialect norms as defaults, obscuring variation across the language

[1] Gloss explanations are in Table 8 in Appendix A.

[2] Table 9 in Appendix A provides a summary of currently available NLP tools.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 46–63

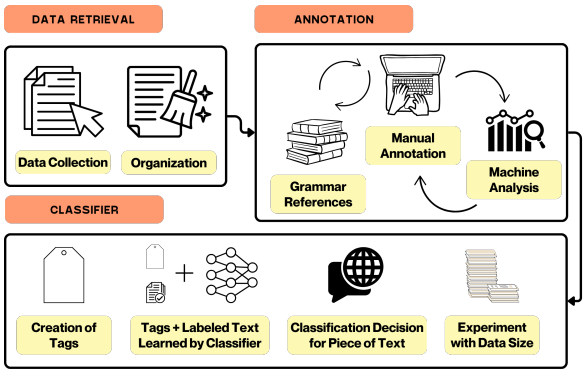

Figure 1: Dialect classification workflow.

family and limiting broader applicability. As well, the issue posed by lack of dialect-specific tools is amplified by Quechua's morphological richness, where grammatical meaning is encoded at the level of affixes. While neural models have advanced NLP significantly, purely data-driven approaches are often insufficient for capturing dialectal variation especially in morphologically rich contexts.

To address these problems, this paper adopts a rule-augmented approach that integrates neural models with linguistically informed representations on the task of dialect classification. Such approaches have shown promise in low-resource contexts by improving both performance and interpretability (Li et al., 2020; Škrlj et al., 2021; Sheth et al., 2023). Here, linguistic knowledge is incorporated not as an end in itself, but as a means of improving dialect classification. Dialect classification itself is important for being able to separate data for other downstream NLP tasks.

Thus, this paper investigates the following question: **Can linguistically informed representations improve neural dialect classification in low-resource settings?** To answer, this paper delivers the following key contributions:

1. A manually verified dataset of Quechua texts sorted by ISO dialect across multiple genres, including annotated parallel bible corpora.
2. A highly accurate multi-dialect classification framework covering 27 Quechua varieties across multiple language families.[3]
3. A systematic evaluation of linguistically informed, rule-augmented neural models, showing that their benefits are context-dependent: gains are minimal in high-resource settings but more pronounced in low-resource and cross-domain conditions.[4]
4. An exploratory analysis of segmentation strategies, showing no consistent improvements across conditions and suggesting a secondary role relative to data and feature design.

2 Related work

Early work on Quechua dialect classification is limited, with Medina (2013) providing a foundational approach using traditional machine learning methods (e.g., Naive Bayes, JRip) to distinguish between Cuzco and non-Cuzco varieties. While this work highlights challenges such as data scarcity and substantial dialectal variation, it is restricted to binary classification.

More recent advances in Quechua NLP have demonstrated the effectiveness of transformer-based models. QuBERT (Zevallos et al., 2022), a RoBERTa-based model trained on Southern Quechua, achieves strong performance on downstream tasks such as part-of-speech (POS) tagging and named entity recognition (NER). However, such models are typically trained on a limited subset of dialects and do not explicitly account for dialectal variation, limiting their applicability in multi-dialect settings.

Tokenization plays a central role in NLP for morphologically rich, low-resource languages, where effective segmentation can reduce sparsity and improve generalization. Approaches range from data-driven methods such as Byte Pair Encoding (BPE) (Shibata et al., 1999) and unigram language models (Kudo and Richardson, 2018) to linguistically informed methods like Prefix-Root-Postfix Encoding (PRPE) (Zuters et al., 2018; Chen and Fazio, 2021), which explicitly model morphological structure. Prior work shows that morphology-aware segmentation can improve downstream performance, including within the QuBERT framework (Zevallos et al., 2022).

PRPE builds on earlier work by Zuters et al. (2018) and Chen and Fazio (2021), which also explore hybrid approaches combining PRPE with BPE and unigram models for morphological neural machine translation (NMT). Additional work such as Ortega et al. (2020) proposes BPE-guided segmentation that aligns linguistic boundaries with sub-word merges, though this approach relies on

[3] All resources and code for this project may be found at the GitHub Repository: https://github.com/clairepost/Quechua_Classifier.

[4] Workflow and architecture shown in Figure 1.

ISO	Variety	No Bible	Only Bible	All Data
inb	Colombian Inga	707	151,198	151,905
qub	Huallaga Huánuco	3,350	119,909	123,259
quf	Lambayeque		160,059	160,059
quh	Southern Bolivian	214,240	135,012	349,252
quk	Chachapoyas	7,194		7,194
qul	San Martín		139,975	139,975
qup	North Bolivian		177,265	177,265
quw	Southern Pastaza		116,704	116,704
qux	Tena Lowland	203,888		203,888
quy	Yauyos	564,034	564,820	1,128,854
quz	Ayacucho	2,301,431	574,539	2,875,970
qvc	Cusco		160,816	160,816
qve	Cajamarca		167,740	167,740
qvi	Eastern Apurímac		145,704	145,704
qvh	Imbabura Highland		115,875	115,875
qvm	M-Y-L		131,047	131,047
qvn	North Junín		138,226	138,226
qvo	Napo Lowland		115,908	115,908
qvs	Huaylla Wanca		153,378	153,378
qvw	Northern Pastaza		112,921	112,921
qvz	Huaylas Ancash		157,628	157,628
qwh	Panao Huánuco	31,054	244,059	275,113
qxl	Salasaca Highland		127,034	127,034
qxh	Panao Huánuco		119,326	119,326
qxn	Northern Ancash		507,098	507,098
qxo	Southern Ancash	9,982	136,530	146,512
qxr	Cañar Highland		506,958	506,958
Total words		3,335,880	5,179,729	8,515,609
Percent		39.2%	60.8%	100%

Table 2: Word counts across datasets by ISO code and Quechua variety.

a limited set of suffixes.[5] Other preprocessing approaches include normalization pipelines for Quechua II (Rios Gonzales and Castro Mamani, 2014), though these are not always easily reproducible or generalizable across dialects.

3 Data

We collect a corpus of more than 8.5M words, across 27 varieties of Quechua as seen in Table 2.

3.1 Quechua Corpus Collection

The corpus was constructed through a combination of archival data collection, web scraping, and automated filtering. Initial data were gathered from publicly available linguistic resources, including AILLA, Runasimi, Ethnologue, Glottolog, and OLAC as well as tools such as Corpus Crawler.[6]

The data collected was cataloged with information on its source, type, and ISO code. This initial set included domains such as spoken word transcriptions (over 200k words), legal texts, and educational materials.[7] Collecting data from linguistic repositories has the advantage of providing dialectal metadata, including standard ISO code information, source location, and resource-specific dialect names.[8] These metadata were used to assign gold labels for the classifier: texts were labeled according to existing classifications in the source materials, explicit dialect names in resource titles or descriptions, or through regional information associated with the data. All dialect assignments were manually checked before inclusion in the final corpus. The final classifier evaluated in this paper was therefore trained and evaluated against manually verified dialect labels, rather than labels generated automatically by the model.

All collected data were processed to retain only Quechua content, remove non-Quechua material (e.g., Spanish and English), and ensure reliable dialect labeling. To improve data quality and obtain coherent document-level texts, a second round of collection was also conducted using a custom scraper targeting full-text Quechua bible sources across 25 dialects.

An initial version of the classifier was then used to label previously unclassified materials, enabling iterative expansion of the dataset. This process yielded a larger and more balanced corpus suitable for multi-dialect classification. The refined dataset includes texts from 27 Quechua varieties, spanning both religious and non-religious genres, with statistics shown in Table 2.

3.2 Parallel Data as Annotation Support

Additional Spanish and English bible data were collected to support cross-lingual analysis and annotation tasks. Specifically, we structured bible data into a three-way parallel format. All texts were converted into CSV files with metadata fields including `iso`, `resource`, `book`, `verse`, and `text`. Corresponding Spanish and English texts were processed in parallel, with additional linguistic annotation applied using off-the-shelf spaCy models (`es_core_news_md` and `en_core_web_md`). Alignment was performed at the verse level, allowing each Quechua segment to be paired with its Spanish and English equivalents.

To support targeted linguistic analysis, we reduce the dataset to a subset of bible texts (Matthew, Mark, and John). Selection was guided by the presence of morphosyntactic features relevant to polypersonal agreement. Specifically, a Spanish morphological parser (`es_core_news_md`) was used to identify dative and accusative clitic constructions, which often correspond to object marking in Quechua. This filtering strategy enables efficient identification of relevant constructions (Sec. 4) while maintaining cross-lingual alignment.

[5] See Table 10 in Appendix A for links to all resources.
[6] See Table 10 in Appendix A for links to all resources.
[7] See Figure 2 in Appendix A.
[8] For further information on ISO codes and family information see Table 11 in Appendix B.

While Spanish clitics provide a useful proxy for object marking, some ambiguities remain. For example, the clitic *nos* may correspond to either inclusive or exclusive first-person plural in Quechua, requiring disambiguation based on context and reference to dialect-specific grammars. Third-person clitics (e.g., *le*, *lo*, *la*) were not explicitly targeted due to their limited role in object agreement, though some instances were retained when co-occurring with other relevant features.

For each classification experiment, data were split into training and evaluation sets using an 85/15 partition at the level of text chunks. Documents were segmented into overlapping chunks of 250 words (with 50-word overlap), and these chunks served as the unit of classification. Splits were performed randomly without stratification, resulting in class distributions that reflect the natural imbalance of the dataset. The full-data setting includes 27 dialects, while the no-bible condition contains only 10 dialects.

4 Annotation

This section outlines the linguistic annotation and feature development used to support dialect classification. We consider: (i) lexical variation across dialects, (ii) polypersonal verbal agreement, and (iii) construction of a morphological inventory used to derive segmentation heuristics.[9] Together, these annotations inform both feature design and preprocessing for the classifier.

Some forms are used across multiple dialects. When counting, we consider both the number of **unique** surface forms, regardless of dialect, and the **total** number of terms identified for all dialects, which includes duplicated surface forms. The total number of collected items for each of these annotation sets is:

- **Lexical terms:** 86 unique, 330 total
- **Polypersonal morphemes:** 110 unique, 157 total
- **Additional morphemes:** 448 total

4.1 Lexical Variation

Our approach to analyzing lexical variation across Quechua dialects follows Medina (2013), who identify differences in core vocabulary items across varieties. We extend her analysis to additional dialects using both corpus data and dialect-specific grammatical resources. Lexical items under comparison include translations for terms (referred to here by their English words) such as *father*, *land*, *sun*, and *moon*. We exclude items exhibiting high uniformity across dialects (e.g., *mother* = *mama*, *water* = *yacu*, *woman* = *warmi*), due to their limited diagnostic value.

For each dialect, we identify the relevant lexical forms through corpus search, cross-referenced with dictionaries and grammars. When multiple lexical variants are attested for a single concept-variety pair, we record all variants (e.g., *tayta/taita* for *father*). These variants are compiled into a comparative lexical chart (Figure 3 in Appendix B), enabling systematic comparison across varieties.

We supplement the collected lexical data through extensive consultation of dialect-specific resources: a grammar of Huallaga Huánuco Quechua (Weber, 1989), a dictionary of Margos-Yarowilca-Lauricocha Quechua (Bean, 1986), a grammar of North Junín Quechua (Adelaar, 1977), a grammar of Huaylla Wanca Quechua (Cerrón-Palomino, 1976), a dictionary of Huaylas Ancash Quechua (Carranza Romero, 2003), a dictionary of Panao Huánuco Quechua (Smith, 1994), a grammatical sketch of Northern Conchucos Ancash Quechua (Wroughton, 1988), a grammar sketch of Southern Ancash Quechua (Hintz, 2017), and a dictionary of Ayacucho Quechua (Parker, 1969). In addition to core vocabulary, pronominal roots encoding person and number distinctions (`1SG`, `2SG`, `3SG`, `1PL.INCL`, `1PL.EXCL`, `2PL`, `3PL`) were also incorporated to ensure broader lexical coverage across dialects.

4.2 Polypersonal Verbal Suffixes

One of the most salient morphosyntactic features across Quechua dialects is *polypersonal verb agreement*, in which a single verb encodes both subject and object through suffixation. This system, widely documented in typological and descriptive work (Adelaar and Muysken, 2004; Lakämper and Wunderlich, 1998; Camacho Rios, 2020; Rataj, 2015), reflects the agglutinative structure of Quechua and provides a rich source of dialectal variation.

To capture this variation, we manually annotate polypersonal constructions across 25 Quechua varieties using the collected parallel bible corpora.[10] For each sentence-level instance, annotations include subject and object person/number (e.g., `1SG`

[9] More information on ISO codes and dialects in Table 11 in Appendix B.

[10] See Figure 4 in Appendix B.

Family	Added Morphs
Quechua I	+144
Colombia-Ecuador	+92
Cajamarca-Lambayeque	+70
San Martín-Amazonas	+62
Quechua II	+80

Table 3: Morphemes added to PRPE per Quechua family

`Subj → 2PL Obj`), verb stem, and relevant suffixes, along with aligned Spanish and English translations. Given the complexity of Quechua verbal morphology, annotation focuses primarily on present indicative forms, which most consistently encode subject–object agreement. Ambiguous cases (particularly those involving tense, mood, or aspect distinctions) are excluded from final analyses when reliable interpretation is not possible. Reference grammars and linguistic descriptions were consulted extensively to support annotation decisions,[11] particularly for identifying object-marking suffixes and "transitions" between subject and object forms. When necessary, additional resources such as SAILS[12] were used to confirm morphological patterns. As an example, we use this resource to determine that in qvn, the affix for 1sgQB is $-wa$.

Cross-dialect comparison reveals some typological trends. Conservative dialects, particularly those spoken in highland regions of Peru and Bolivia, tend to preserve rich agreement paradigms, including distinct markers for first- and second-person objects. In contrast, dialects in Ecuador and Colombia often display morphosyntactic simplification, with reduced or absent object agreement marking. These patterns provide informative features for dialect classification.

4.3 Morphological Inventory and Heuristics

To support linguistically informed segmentation, we construct an expanded inventory of Quechua morphemes through manual analysis of grammatical resources. While some grammars provide explicit morpheme lists, many require close examination of descriptive text to identify suffixes and inflectional patterns, as inconsistent formatting and multi-language scripts often render OCR ineffective. This process involved extracting both common and dialect-specific morphemes, including case markers, tense and aspect suffixes, evidentials, and derivational affixes. The resulting inventory extends prior PRPE resources (Chen and Fazio, 2021) and forms the basis for dialect-aware segmentation heuristics used in subsection 5.1.

We derive heuristics from this inventory by treating morphs marked with a hyphen (e.g., `-nchi`) as suffixes and those without as roots. Given that Quechua lacks productive prefixation, suffix-based segmentation is particularly central. Morphemes are first grouped by dialect and then consolidated into family-level inventories corresponding to major Quechua families: Colombia-Ecuador Quechua (cu: `qxl, inb, qup, quw, qvi, qvo, qvz, qxr`), Cajamarca-Lambayeque Quechua (ca: `qvc, quf`), Quechua I (q1: `qub, qux, qvh, qvm, qvn, qvw, qwh, qxh, qxn, qxo`), Quechua II (q2: `quy, quh, qul, quz, qve`), and San Martín-Amazonas Quechua (sm: `qvs, quk`).

Morphemes with multiple meanings (e.g., `-pi`) are counted only once per family. Compared to the 64 morphemes used in Chen and Fazio (2021), this work substantially expands coverage across all families, adding a total of 448 morphemes (Table 3).

5 Methodology

We investigate a range of modeling strategies, varying both data settings and, crucially, degree to which linguistic information is incorporated into the model. This section describes modeling options, and we present results in Section 6.

5.1 Segmentation Strategies

To evaluate the impact of input representation on dialect classification, we explore multiple segmentation strategies, including BPE, unigram language models (Kudo and Richardson, 2018), PRPE (Zuters et al., 2018; Chen and Fazio, 2021), and a hybrid PRPE+BPE approach (Chen and Fazio, 2021; Ortega et al., 2020).

We extend PRPE with the dialect-informed morphological inventory described in subsection 4.3, enabling segmentation to better capture variation across Quechua families. These heuristics guide suffix identification and improve alignment between sub-word units and linguistically meaningful morphemes.

In addition, we train our segmentation at the level of major Quechua language families, allowing representations to capture shared morphological patterns within families while preserving cross-dialect distinctions. This setup enables controlled comparison of how segmentation and linguistic granularity affect classification performance.

[11] See Table 12 in Appendix B.

[12] https://sails.clld.org/languages/qvc

5.2 Model Architectures

QuBERTa Classifier. Our primary neural classifier is built on QuBERTa, a RoBERTa-based model adapted for Quechua (Zevallos et al., 2022). Input texts are preprocessed by removing numeric tokens and segmenting documents into overlapping chunks (250 tokens with 50-token overlap) to satisfy the model's 512-token input constraint. Chunks are tokenized and padded using the `Llamacha/QuBERTa` tokenizer.

The model is fine-tuned for dialect classification using standard optimization procedures, with hyperparameters (batch size, learning rate, number of epochs) tuned on validation data. We evaluate performance with standard classification metrics.

Rule-Augmented QuBERTa. Our rule-augmented variant extends the base QuBERTa model by incorporating linguistically motivated features derived from lexical and polypersonal annotation. We encode these features as tags and append them to each text chunk during preprocessing, allowing the model to access explicit morphosyntactic and lexical cues alongside learned representations.

Lexical tags are generated by matching dialect-specific vocabulary items and appending corresponding markers (e.g., `<TAG_tayta>`). Polypersonal tags are derived through suffix matching over a curated list of verbal affixes, using a longest-match strategy to prioritize more specific morphological forms. Identified polypersonal verbal suffixes are encoded as tags (e.g., `<VERB_wanku>`) and added as additional tokens to the input.

We use the same training procedure for both the base model (standard QuBERTa classifier) and the rule-enhanced version, allowing direct comparison of performance with and without linguistic feature injection. Separate configurations evaluate the contribution of lexical tags, polypersonal tags, and their combination.

Naive Bayes Baseline. We implement a Multinomial Naive Bayes classifier as a statistical baseline, following Medina (2013). The model operates over TF-IDF representations, which encode term frequency and corpus-level importance to capture distributional patterns in word usage.

To ensure comparability with the neural models, the same cleaned and chunked inputs are used. TF-IDF features are extracted from each text segment and used to train a multi-class classifier over dialect labels. To mitigate class imbalance across dialects, a limit is placed on the number of chunks per document, preventing over-representation of high-resource sources. This constraint after experimentation proved to improve stability and provides a more balanced comparison with neural approaches.

Model	Rules *(# Dialects)*	All *(27)*	NO BIBLE *(10)*	ONLY BIBLE *(25)*
Bayes	No rules	**.9385**	**.7551**	**.9713**
Bayes	Lexical+Verb	.9064	.7506	.9406
Neural	No rules	.9907	.9662	.9957
Neural	Lexical	.9916	.9664	.9949
Neural	Verb	**.9936**	**.9670**	**.9969**
Neural	Lexical+Verb	.9928	.9643	.9961

Table 4: Weighted F1 scores across models, rules, and data settings with standard QuBERTa tokenization.

5.3 Experimental Settings

To evaluate model performance under varying resource conditions, we compare several data settings. In the high-resource ALL setting, all available corpora (see Table 2) from 27 dialects are included, comprising religious texts, other written materials, and previously unclassified data labeled through the pipeline described in Section 5.2. This setting is relatively balanced across varieties and domains. In the limited-data setting (NO BIBLE, or NB), bible data are removed to simulate a low-resource and domain-mismatched environment. This results in a reduced set of 10 dialects with highly imbalanced distributions, where resource availability ranges from large corpora (over 2 million words) to very limited data (fewer than 10,000 words). A third configuration utilizing only bible data (ONLY BIBLE, or B) is confined to a single domain, with data for 25 dialects. This corpus is close to fully-balanced; for some dialects, we have the complete bible and for others only the New Testament.

Model performance is measured using standard classification metrics, including accuracy, precision, recall, and F1-score, along with confusion matrices to analyze dialect-specific errors.[13]

6 Results

6.1 Overall Performance

Overall results across all experimental conditions are summarized in Table 4. The QuBERTa-based models consistently outperform the statistical baseline (modeled after Medina (2013)) across all set-

[13] See Appendix C.

tings, maintaining F1 scores above .96 even in the most constrained condition, compared to approximately .75 for Naive Bayes.

In the ALL setting, the base QuBERTa classifier achieves near-ceiling performance without linguistic augmentation (.9907 F1). Rule-based features yield only marginal gains, with the best configuration (verb rules) reaching .9936 F1.[14] Statistical comparison confirms that this improvement is small and not significant (ΔF1 = +0.0015, $p = 0.133$).[15] Similar patterns hold in the ONLY BIBLE setting, where performance remains uniformly high due to domain homogeneity.

In the limited-data setting (NO BIBLE), the neural baseline drops to .9662 F1. Rule augmentation produces more variable effects, with the best-performing model (verb rules) reaching .9670 F1. While average per-dialect improvements are larger (ΔF1 = +0.0649), high variance results in non-significant tests ($p = 0.398$). Notably, the weighted change is slightly negative, suggesting gains are concentrated in lower-resource dialects (e.g., qvo, qwh) rather than uniformly distributed.

quz	quh	English
<hallp'a>	<pacha>	earth
<inti>	<indi>	sun
<quilla>	<killa>	moon
<runa>	<ullku>	person

Table 5: Dialectal variation encoded through lexical document-level tagging.

Examination of misclassification patterns shows a concentration of errors among closely related Southern Quechua varieties.[16] The baseline no-rules model frequently confuses Cuzco Quechua (quz) with Southern Bolivian Quechua (quh), while the rules-augmented model reduces these errors, possibly by leveraging fine-grained lexical distinctions between texts, such as those seen in Table 5.[17]

Averaging across languages, classification performance is consistently high. This setting, though, assumes availability of significant amounts of data across varieties. To get a more nuanced picture of performance, we next investigate performance in more realistic settings.

[14] Further details on morpheme exclusion in the verb rules are provided in Appendix C.

[15] See Table 14 in Appendix C.

[16] See Table 11 in Appendix A.

[17] For a confusion matrix comparison see Table 15 and Table 16 in Appendix C.

Data settings				*Rule settings*			
Train	**#**	**Eval**	**#**	**No**	**Lex**	**Verb**	**Lex+V**
NB	10	ALL	27	.4032	.3805	**.4039**	.3936
NB	10	ALL-F	27	.4999	.4429	**.5344**	.4754
NB	10	ALL ∩ NB	10	.7646	.7663	**.7745**	.7701
NB	10	ALL ∩ NB-F	10	.8328	.8360	**.8424**	.8303
B	25	NB	10	.7105	**.8094**	.8054	.8038
B	25	NB ∩ B	8	**.8288**	.8283	.8195	.8227
B ∩ NB	8	NB	10	**.7594**	.7587	.7556	.7526
B ∩ NB	8	NB ∩ B	8	**.8182**	.8015	.8171	.8127

Table 6: Cross-domain weighted F1 results across training and evaluation conditions. # is number of dialects per setting. (-F) indicates family level evaluation. NB/B= NO BIBLE/BIBLE ONLY data. $X \cap Y$ represents the intersection of X and Y datasets.

6.2 Cross domain experiments

To assess generalization, we conduct cross-domain experiments with training and evaluation data from different distributions (e.g., NB vs. B). These settings better reflect real-world conditions where domain mismatch is common, especially in low resource settings for Indigenous languages. Cross-domain performance (Tab. 6) is substantially lower than in-domain results (Tab. 4), reflecting differences in lexical choice, genre, and dialect coverage.

We first examine a low-resource model trained on NB (10 dialects) and evaluated on ALL (27 dialects). Because of the mismatch between labels, we introduce a relaxed evaluation condition (-F), counting family-level matches as correct (groupings in App. B.1). This partially compensates for missing dialect coverage, though some families remain absent in training data (e.g., San Marín).

Under this setting, performance largely degrades, but differences emerge across rule configurations. In the relaxed condition, the no-rules model achieves .4999 F1, while the verb-augmented model improves to .5344, representing the largest gain observed in this configuration. This suggests that verb-based features provide useful structure when generalizing beyond the training distribution. Across dialects, the mean improvement is positive (ΔF1 = .0511, $d = 0.21$), though not statistically significant (Wilcoxon $p = 0.50$).[18]

Restricting evaluation to dialects present in the training set (10 dialects) improves performance across all configurations (F1 $> .76$), indicating that much of the degradation is driven by label space mismatch. In this controlled setting, verb rules still provide a small but consistent improvement, and provide more evidence that linguistic features are

[18] See Table 14 in Appendix C.

particularly valuable when models must generalize beyond their training label space, rather than simply interpolate within it when moving from a smaller model setting.

The final set of experiments trains on the larger but domain-restricted BIBLE ONLY corpus and evaluates on heterogeneous NO BIBLE data. Performance again drops substantially, indicating that bible text does not provide a fully representative training distribution. This observation aligns with prior work: although bible corpora are widely used due to accessibility and multilingual coverage (Christodouloupoulos and Steedman, 2015), they reflect formal or archaic registers, translation-driven inconsistencies, and a narrow range of genres (Levshina, 2022; Hutchinson, 2024).

Despite this domain mismatch, rule augmentation improves performance for the no-rules model, which achieves .7105 F1, while lexical tagging performs best (.8094), outperforming verb (.8054) and combined rules (.8038). This finding stands in contrast to prior experiments, in which verb-based rules yielded the highest effectiveness, thereby suggesting that the utility of specific linguistic features depends on the direction of the domain shift.

One explanation is that BIBLE ONLY training already exposes the model to relatively consistent verbal morphology, reducing the added value of verb-based rules at test time. Lexical tagging instead helps compensate for lexical and genre differences between bible and non-bible data. This is supported by per-dialect results, where gains are concentrated in higher-support dialects such as *quz* (F1 .7525 no rules to .8894 verb rules), while many low-resource dialects show minimal change.[19]

Looking at the bottom of Table 4, when evaluation is restricted to overlapping dialect subsets, though, differences between rule configurations are negligible (e.g., .7594 vs. .7587). Thus the effectiveness of linguistic augmentation seems to depend not only on domain shift, but also on label space alignment and data coverage.

6.3 Effect of Segmentation

Segmentation strategy does not produce consistent improvements across conditions. While morphology-aware approaches (e.g., PRPE+BPE) show gains in some low-resource settings, these effects are not stable (see Table 7). In high-resource settings, differences between segmentation methods are minimal and models seem to have sufficient data to learn effective representations regardless of segmentation strategy. In lower-resource conditions, results are more variable, with no single method consistently outperforming others.[20]

Model	Segment	Low Data		All Data	
		F1w	F1m	F1w	F1m
ca	bpe	**.945**	.533	**.972**	**.935**
ca	prpe	.933	.468	.964	.922
ca	prpe+bpe	.931	.454	.971	.930
ca	unigram	.941	**.534**	.967	.929
q1	bpe	.937	.485	**.978**	.942
q1	prpe	.937	.458	.977	.941
q1	prpe+bpe	**.949**	.534	.977	**.943**
q1	unigram	.948	**.560**	**.978**	.941
q2	bpe	.966	.644	.986	**.967**
q2	prpe	**.968**	.649	**.988**	.952
q2	prpe+bpe	**.968**	**.663**	.987	.951
q2	unigram	.967	.626	**.988**	.963
sm	bpe	.923	.384	**.960**	**.920**
sm	prpe	.923	.429	.945	.903
sm	prpe+bpe	**.929**	**.438**	.939	.895
sm	unigram	.921	.395	.955	.913

Table 7: Segmentation results grouped by dialect family under low-resource and full-data settings. F1w = weighted F1 (accounts for class imbalance), F1m = macro F1 (equal weighting across dialects).

7 Conclusion & Future Work

This paper presents a multi-dialect classification framework for Quechua that augments neural models with linguistically informed features. Across all settings, neural models substantially outperform statistical baselines, enabling accurate multi-class classification across 27 Quechua dialects.

The impact of linguistic augmentation, however, is nuanced. In high-resource and homogeneous (bible) settings, performance is already near ceiling, and rule-based features provide only marginal gains. In contrast, in low-resource and cross-domain conditions, linguistic features become more valuable, though their effects are uneven and depend on both data availability and evaluation setup. In particular, verb-based features are most beneficial when generalizing from limited training data to broader label spaces, while lexical tagging was found to be more effective under domain shift, especially when transferring from bible-only to heterogeneous corpora. The utility of linguistic features then may be context-dependent rather than uniformly additive.

More broadly, this work aims to contribute to on-

[19] For a confusion matrix comparison see Table 18 and Table 17 in Appendix C.

[20] A segmentation results breakdown is in Appendix C.

going efforts to integrate linguistic knowledge into modern machine learning pipelines. By advancing dialect-sensitive NLP tools, it hopes to support more accurate and inclusive language technologies, helping to address the marginalization of Quechua speakers in digital spaces.

Future work will extend this approach in several directions. First, we plan to develop family-specific RoBERTa models that better capture variation beyond Southern Quechua. Second, we aim to expand and refine morphological inventories used in segmentation and tagging. Additional directions include evaluating segmentation in downstream morphological tasks and revisiting alternative sub-word strategies such as BPE-guided methods (Ortega et al., 2020). Third, this approach would greatly benefit from expansion beyond bible corpora, and work is underway to annotate more varied domains across additional dialects. Finally, for applications to other language families, our results suggest that lexical tagging offers the most favorable trade-off between annotation effort and performance gains, particularly in cross-domain and low-resource settings.

Limitations

A limitation of this work is the reliance on domain-specific data, particularly bible texts, which constitute a large portion of the available corpora. While we explicitly evaluate a NO BIBLE setting to simulate low-resource conditions, the properties of religious text may not fully reflect naturalistic language use and might have some effect on classification of non-bible texts. Current work is aimed at expanding the corpus to more diverse domains for a wider range of dialects.

As well, although this work expands classification to 27 varieties of Quechua, there are still more within the language family. Varieties were included if *any* data could be found and preprocessed into text files. Even still, this leaves room for improvement by finding additional textual resources.

Ethics

This work involves the use of textual data from various publicly available sources for different Quechua dialects. We recognize the importance of indigenous data sovereignty and the need to handle language data in a way that respects the communities from which it originates. Wherever possible, this work relies on publicly available sources and keeps record of metadata and authorship in order to further support language documentation efforts.

At the same time, automated dialect classification systems carry potential risks as misclassifications may obscure dialectical distinctions, reinforce inaccurate generalizations, or privilege better-represented varieties over lower-resource ones. These risks are particularly relevant in cross-domain settings, where model performance is less stable and uneven across dialects. As such, the models presented here should not be used as authoritative tools for linguistic identification, but rather as assistive technologies whose outputs require careful interpretation and are up for welcome debate by community members.

Acknowledgments

Most sincere thanks to the two anonymous reviewers for thorough, insightful, and helpful reviews. This work would not have been possible without the help and advice of our colleagues. First, thanks to Wilma Doris Loayza for her Quechua teaching materials that inspired the study into polypersonal verbal agreement patterns for this project. Next, the modeling portion of this paper was assisted by Saksham Khatwani, who helped give feedback on the classifier and advice on implementing the different segmentation methods. Additional thanks are in order to Andrew Cowell, Hannah Haynie, and Jim Martin for their feedback at various points throughout this project. Parts of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another."

References

Willem Adelaar. 1977. *Tarma Quechua: Grammar, texts and dictionary*. Ph.D. thesis, Universiteit van Amsterdam, Lisse.

Willem F. H. Adelaar and Pieter C. Muysken. 2004. *The Inca Sphere*, page 165–410. Cambridge Language Surveys. Cambridge University Press.

Willem FH Adelaar. 2020. Morphology in Quechuan Languages. In *Oxford Research Encyclopedia of Linguistics*.

Mark Bean. 1986. *Pequeño diccionario de palabras útiles: Quechua-castellano, castellano-quechua*, primera edición edition. Dirección Departmental de Educación - Huánuco e Instituto Lingüístico de Verano, Perú. Published. Available online: https://www.sil.org/resources/archives/27945.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Gladys Camacho Rios. 2020. *Verb morphology in South Bolivian Quechua: A case study of the Uma Piwra rural variety*. Ph.D. thesis.

Lawrence Kidd Carpenter. 1982. *Ecuadorian Quichua: Descriptive sketch and variation*. University of Florida.

Francisco Carranza Romero. 2003. *Diccionario quechua ancashino-castellano*. Iberoamericana.

Rodolfo Cerrón-Palomino. 1976. *Gramática quechua: Junín/Huanca*. Ministerio de Educación/IEP, Lima.

Junhao Chen, Peng Shu, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Zhengliang Liu, Lewis C Howe, and Tianming Liu. 2024. QueEn: A Large Language Model for Quechua-English Translation. *arXiv preprint arXiv:2412.05184*.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Luis Cordero Crespo. 1955. *Diccionario de Quichua*. Biblioteca Hernán Malo González.

S Khalil Bello García, E Sánchez Lucero, E Bonilla Huerta, J Crispín Hernández Hernández, J Federico Ramírez Cruz, and B Estela Pedroza Méndez. 2021. Implementation of neural machine translation for Nahuatl as a web platform: a focus on text translation. *Programming and Computer Software*, 47:778–792.

Joseph E Grimes. 1985. The interpretation of relationships among quechua dialects. *Oceanic Linguistics Special Publication*, pages 271–284.

Manuel Guzmán. 1920. *Gramatica de la lengua Quichua (dialecto del Ecuador)*. Prensa Católica, Quito.

Marleen Haboud de Ortega, Luis Montaluisa Chasiquiza, Fabián Muenala Pineda, and Froilan Viteri Gualinga. 1982. *Shimiyuc-Panca, Caimi Ñucanchic*. Pontificia Universidad Católica and Ministerio de Educación y Cultura, Quito. Yanapaccuna: Filemón Aguinda Díaz, Mariano Cerda Chimbo, Enrique Contreras Ponce, Manuel Díaz Cajas, Agustín Jérez Jérez, Luis Macas Ambuludi, César Shiguango Grefa, Consuelo Yánez Cossío. Shuyuccuna: José Aviléz López, José Higuera Rosero. Quillcac: Humberto Cachihuango, Manuel Serrano, Gladys Muenala Vega.

Daniel J. Hintz. 2017. *El Aspecto Verbal en Quechua Campos Semánticos Entretejidos y el Surgimiento de Sistemas Gramaticales*, volume 58 of *Serie Lingüística Peruana*. Instituto Lingüístico de Verano, Lima.

Nancy H Hornberger and Kendall A King. 1998. Authenticity and unification in Quechua language planning. *Language Culture and Curriculum*, 11(3):390–410.

Nancy H Hornberger and Nicholas Limerick. 2019. Teachers, textbooks, and orthographic choices in Quechua: Bilingual intercultural education in Peru and Ecuador. In *Perspectives on Indigenous writing and literacies*, pages 141–164. Brill.

Ben Hutchinson. 2024. Modeling the sacred: Considerations when using religious texts in natural language processing. *arXiv preprint arXiv:2404.14740*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Renate Lakämper and Dieter Wunderlich. 1998. Person marking in Quechua—A constraint-based minimalist analysis. *Lingua*, 105(3-4):113–148.

Natalia Levshina. 2022. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

Tao Li, Parth Anand Jawale, Martha Palmer, and Vivek Srikumar. 2020. Structured tuning for semantic role labeling. *arXiv preprint arXiv:2005.00496*.

Nicholas Limerick. 2018. Kichwa or Quichua? Competing alphabets, political histories, and complicated reading in Indigenous languages. *Comparative Education Review*, 62(1):103–124.

Zoey Liu, Crystal Richardson, Richard Hatcher Jr, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. *arXiv preprint arXiv:2204.05541*.

Rosemary Jiménez Medina. 2013. *Clasificación Por Dialecto De Documentos Escritos En Quechua*. Ph.D. thesis, Universidad Nacional De San Antonio Abad.

Nelsi Melgarejo, Rodolfo Zevallos, Héctor Gómez, and John E Ortega. 2022. WordNet-QU: Development of a lexical database for Quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell, and Alon Lavie. 2006. Building NLP systems for two resource-scarce indigenous languages:

Mapudungun and Quechua. *Strategies for developing machine translation for minority languages*, page 15.

Carolyn Orr. 1973. *Dialectos quichuas del Ecuador con respecto a lectores principiantes*. ILV, Quito.

Carolyn Orr. 1992. *Runa shimi (Gramática quichua de Tena)*. ILV, Quito.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Gary John Parker. 1969. *Ayacucho Quechua grammar and dictionary*, volume 82 of *Janua linguarum : Series practica*. Mouton, The Hague. Bibliography: p. [226].

Félix Quesada. 1976. *Gramática quechua: Cajamarca-Cañaris*. Ministerio de Educación. Instituto de Estudios Peruanos.

J. F. M. Raez. 1917. *Gramáticas en el quichua-huanca y en el de Ayacucho*. Sanmarti y Ca, Lima.

Vlastimil Rataj. 2015. Marcación de la segunda persona objeto en las transiciones del quechua cusqueño. *IBERO-AMERICANA PRAGENSIA*, 43(1):27–46.

Matt Riemland. 2023. Theorizing sustainable, low-resource MT in development settings: Pivot-based MT between Guatemala's indigenous Mayan languages. *Translation Spaces*, 12(2):231–254.

Annette Rios. 2010. Applying finite-state techniques to a native american language: Quechua. *Institut für Computerlinguistik, Universität Zürich*.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A Quechua-Spanish parallel treebank. *LOT occasional series, Netherlands Graduate School of Linguistics*, 12:53–64.

Annette Rios Gonzales and Richard Alexander Castro Mamani. 2014. Morphological disambiguation and text normalization for Southern Quechua varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Amit Sheth, Kaushik Roy, and Manas Gaur. 2023. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intelligent Systems*, 38(3):56–62.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Blaž Škrlj, Matej Martinc, Nada Lavrač, and Senja Pollak. 2021. autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. *Machine Learning*, 110(5):989–1028.

Terrence Smith. 1994. *Alli rimay ashina (Un pequeño diccionario de palabras utiles en el quechua de Panao)*. Dirección Regional de Educación-Huánuco and Instituto Lingüístico de Verano, Huánuco.

Nelsi Belly Melgarejo Vergara. 2022. Desarrollo de recursos léxicos multi-dialécticos para el quechua. Master's thesis, Pontificia Universidad Catolica del Peru (Peru).

David Weber. 1989. *A grammar of Huallaga (Huánuco) Quechua*, volume 112. Univ of California Press.

James F. Wroughton. 1988. Major Clause Constituents of Conchucos (Ancash) Quechua. Master's thesis, University of Texas at Arlington, Arlington.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Yoshikawa, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. *arXiv preprint arXiv:2204.03031*.

Jānis Zuters, Gus Strazds, and Kārlis Immers. 2018. Semi-automatic quasi-morphological word segmentation for neural machine translation. In *International Baltic conference on databases and information systems*, pages 289–301. Springer.

A Appendix: Background

Gloss	Meaning
1sg	first person singular
3sg	third person singular
DIR	direct evidential
DO	direct object marker
MOV	directional motion away
EMP	emphatic marker
CON	contrastive marker
PST	past tense
ADD	additive marker ('also')
3sg→1sg	third person subject acting on first person object

Table 8: Gloss abbreviations explanations for the morphological parses in Table 1.

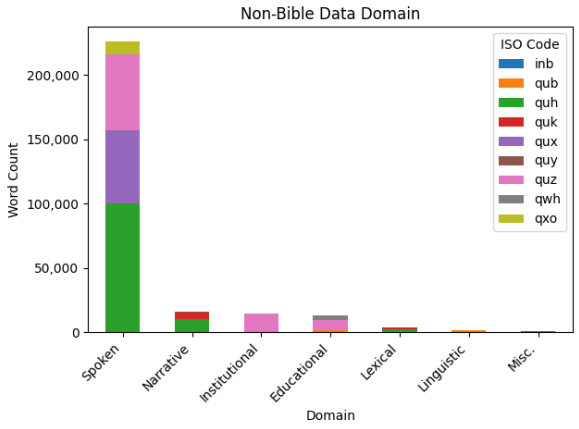

Figure 2: Word counts broken up by domain type and ISO code for non-bible data corpora.

B Appendix: Data

B.1 Dialect grouping

The dialect groupings used in this study follow the family hierarchy and clade structure outlined by Glottolog.[21] This classification allows for hierarchical grouping of varieties at the family, subfamily, and dialect levels. For instance, dialects from the Quechua I group (e.g., Cajamarca Quechua) are modeled separately from those in Quechua II-A or II-B (e.g., Ayacucho or Ancash Quechua). Dialect identity and family membership were assigned using this system because they are widely used by other resources as classification for data. A table of family composition, summarizing clade-level distinctions and vocabulary comparisons, is shown in Table 11.

[21] https://glottolog.org/resource/languoid/id/quec1387

Dialects	NLP Tools
Cuzco Quechua	Text classifier (Medina, 2013), AntiMorfo: finite state transducer morphological analyzer (Rios, 2010)
Southern Quechua	QuBERT: monolingual BERT model (Zevallos et al., 2022), SQUOIA: finite state transducer morphological analyzer & spell checker (Rios Gonzales and Castro Mamani, 2014), Quechua-Spanish Treebank (Rios et al., 2008)
Southern, Central, Northern, and Amazonian Quechua	WordNET-QU: lexical database (Melgarejo et al., 2022), POS tagging models (Vergara, 2022)
Unspecified	QueEn: LLM machine translation (Chen et al., 2024), Quechua-Spanish machine translation & morphological analyzer (Monson et al., 2006)

Table 9: Quechua Dialects and Corresponding NLP Tools

B.2 Linguistic Annotation

Drawing from the Medina (2013) analysis of lexical differences across Quechua dialects, a thorough examination of vocabulary discrepancies was conducted, as shown in Figure 3. This included dialects not originally covered by Medina, utilizing both corpus data and dialect-specific grammar resources. If there were multiple words that were used for a lexical term, we indicated that with a / in the chart.

As well, Figure 4 gives a summary of all the polypersonal verbal suffixes found during the verbal annotation phase. Grammars used for help in finding specific polypersonal verbal agreement information is seen in Table 12.

Resource	Link
Ortega et al. (2020)	https://github.com/johneortega/mt_quechua_spanish/tree/master/bpe_guided
AILLA	https://ailla.utexas.org
Runasimi	https://runasimi.de
Ethnologue	https://www.ethnologue.com
Glottolog	https://glottolog.org
OLAC	http://www.language-archives.org
Bible data	https://live.bible.is
QuBERTa tokenizer	https://huggingface.co/Llamacha/QuBERTa

Table 10: Resources and links

				Lexical Terms					
ISO	Dialect Name	Family	Variety	father	land	water	sun	moon	man
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	tayta	allpu	yacu	inti	quilla	runa
qux	Yauyos Quechua	Quechua 1	Yauyosic	tayta	pacha	yaku	rupay/inti	killa	runa
qvh	Huamalíes-Dos de Mayo Huánaco Quechua	Quechua 1	Central Quechua 1	tayta	pacha	yaku	inti	killa	runa
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	inti	quilla	runa
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	tayta	pacha/alpa	yacu	inti	quilla	runa
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1	taytä	pacha/allpa	yacu	inti	quilla	nuna
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	yaya/papä	pacha/allpa	yacu	inti	quilla	nuna
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	tayta	pacha	yacu	inti	quilla	runa
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	rupay	quilla	runa/ollgo
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	tayta	patsa	yacu	rupay	quilla	runa
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua	tayta	pacha	yaku	inti	killa	runa
quh	Southern Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	tata	pacha	yaku	inti	kuilla	runa
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	tata	pacha	unu	inti	killa	runa
quz	Cusco Quechua	Quechua 2	Cuscan Quechua	tayta	hallp'a	unu	inti	killa	qhari
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua	tayta/papa	pacha/allpa	unu	inti	killa	qari
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua	tayta	allpa/pacha	yaku	rupay	killa	runa
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua	tayta	pacha	yaku	rupay	killa	runa
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	yaya/tayta	pacha	yaku	indi	killa	kari
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	taita	patsa	iaku	indi	killa	runa/kari/jinti
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa	yaku	inti	killa	kari
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa	yacu	inti	quilla	runa
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	jahua/papá	pacha	yacu	inti	quilla	runa
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	pacha/allpa	yacu	inti	quilla	runa
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	yaya	allpa/pacha	yacu	indi	quilla	runa
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	taita	pacha	yacu	inti	quilla	runa
qvs	San Martín Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua	tata	pacha	yaku	inti	killa	runa
quk	Chachapoyas Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua	tata	allpa	yaku	inti	killa	runa/ullku

Figure 3: Common words for each dialect organized by Quechua families and varietal sub-families.

						Verbs															
						1sg object				1pl.incl object	1pl.excl object				2sg object				2pl object		
ISO Code	Dialect Name	Family	Clade 1	Clade 2	Clade 3	2sg1sg	3sg1sg	2pl1sg	3pl1sg	3sg1pl.incl	2sg1pl.excl	3sg1pl.excl	2pl1pl.excl	3pl1pl.excl	1sg2sg	3sg2sg	1pl.excl2sg	3pl2sg	1sg2pl	3sg2pl	3sg2pl
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Huallaga Huánuco Quechua	-manqui	-masha	-manqui				-manchi	-chimasha			-shunqui / -shcanqui		-shunqui	-mushcă		-shunqui
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union	-manqui		-manqui								-shunqui		-shunqui			
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union	-manqui	-masha				-mashcanqui					-shcanqui					
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos	-manqui	-mushga		-manga			-mantsic						-yăshunqui			-yăshunqui
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos	-manqui	-mushga		-manga			-mantsic						-yăshunqui			-yăshunki
qvh	Huamalíes-Dos de Mayo Huánaco Quechua	Quechua 1	Central Quechua 1	Huaylay		-munki / -manki	-man				-ykanki									-yăshayki	-yăshunki
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay		-huanqui	-huarca														
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	Yaru Quechua		-manqui	-masha				-mashcanqui					-shcanqui					
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1	Jauja-Huanca		-manqui	-man									-shunqui		-cushunqui	-chwanpis	-cushunqui	-cushunqui
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua			-wanki	-wan	-wankichik	-wanku		-wankiku	-wanku		-wanku	-ki	-sunki		-sunki	-ykichik	-sunkichik	
quh	Southern Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua	South Bolivian-Argentinian Quechua		-wanki	-wan		-wanku	-wanchej					-yki				-ykichej		
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua			-wanki	-wan	-wankichis			-wankichis	-wanku		-wanku	-yki	-sunki			-ykichis		
quz	Cusco Quechua	Quechua 2	Cuscan Quechua			-wanki	-wan	-wankichis	-wanku	-wankichis	-wankiku	-wanku	-wankiku	-wanku	-yki	-sunki	-ykiku	-sunkiku	-ykichis	-sunkichis	-sunkichis
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua			-wanki	-wan		-wanku		-wankikú				-yki				-ykichis		
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua	Cajamarca Quechua		-wangi		-washpapis	-wananqa		-wangilla				-yki	-shushqa			-ykillapas	-shushpa	
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua	Lambayeque Quechua		-manki	-masha								-yki				-shaykillapa		
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	Tungurahua Highland Quichua		NONE				-gangui	NONE								-kiga		
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Cañar-Azuay-South Chimborazo Highland Quichua		NONE	-manga	-huanguichic / -huarcani	NONE		-huangui				NONE	-huarca	NONE			-huanguichic	
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Imbabura Highland Quichua	-huangi			-jun	-chingui									-jpipash		
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-wankungi / -wangi	-wanka				-rkangi	-murka			-iki				-ikichita		
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-wanki		-wankichi			-shkanki								-ptiypas		
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huangi	-huaca / -shcami												-jpis	-shcami	-canchi
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huanqui	-huarca														
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia -Oriente Quechua	Colombia-Oriente Quechua	-huangi	-hua		-huanaun										-cplga		
qvs	San Martín Quechua	San Martin-Amazonas Quechua	San Martín Quechua			-wanki	-washka				-shkanki				-ptinika						

Figure 4: Dialects are organized by their typological classification and with each clade in the family tree. Under the "verbs" row are columns of subject-object polypersonal verbal suffixation patterns.

ISO	Dialect	Family	Sub-family	Clade 1	Clade 2	Clade 3	Clade 4
qvc	Cajamarca Quechua	Cajamarca-Lambayeque Quechua	Cajamarca Quechua				
quf	Lambayeque Quechua	Cajamarca-Lambayeque Quechua	Lambayeque Quechua				
qxl	Salasaca Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua A	Tungurahua Highland Quichua			
qxr	Cañar Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Cañar-Azuay-South Chimborazo Highland Quichua			
qvo	Napo Lowland Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Napo Low-land Quechua
qup	Southern Pastaza Quechua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
quw	Tena Lowland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
qvz	Northern Pastaza Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Colombia-Oriente Quechua	Oriente Quechua	Pastaza Quechua
qvi	Imbabura Highland Quichua	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Colombia-Oriente Quechua	Imbabura Highland Quichua		
inb	Colombian Inga	Colombia-Ecuador Quechua	Ecuadorian Quechua B	Imbabura-Columbia-Oriente Quechua	Colombia-Oriente Quechua		
qub	Huallaga Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Huallaga Huánuco Quechua		
qvm	Margos-Yarowilca-Lauricocha Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union		
qxh	Panao Huánuco Quechua	Quechua 1	Central Quechua 1	AP-AM-AH	Panao-Union		
qxn	Northern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos		
qxo	Southern Conchucos Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay	Conchucos		
qvh	Huamalíes-Dos de Mayo Huánuco Quechua	Quechua 1	Central Quechua 1	Huaylay			
qwh	Huaylas Ancash Quechua	Quechua 1	Central Quechua 1	Huaylay			
qvw	Huaylla Wanca Quechua	Quechua 1	Central Quechua 1				
qvn	North Junín Quechua	Quechua 1	Central Quechua 1	Yaru Quechua			
quy	Ayacucho Quechua	Quechua 2	Ayacuchan Quechua				
qul	North Bolivian Quechua	Quechua 2	Bolivian-Argentinian Quechua				
qve	Eastern Apurímac Quechua	Quechua 2	Cuscan Quechua	Cusco Quechua	Eastern Apurímac Quechua		
quz	Cusco Quechua	Quechua 2	Cuscan Quechua	Cusco Quechua			
quk	Chachapoyas Quechua	San Martin-Amazonas Quechua	Chachapoyas Quechua				
qvs	San Martín Quechua	San Martin-Amazonas Quechua	San Martín Quechua				

Table 11: ISO codes and their corresponding family information, including dialect classification, family, sub-family, and clade structure.

Language	ISO	Source
Huallaga Huánuco Quechua	qub	(Weber, 1989)
Cajamarca Quechua	qvc	(Quesada, 1976)
Colombia–Ecuador Quechua	qvi, qxr, qvo	(Carpenter, 1982)
Southern Conchucos Ancash Quechua	qxo	(Hintz, 2017)
Tena Lowland Quichua	quw	(Haboud de Ortega et al., 1982), (Orr, 1992), (Orr, 1973)
Northern Conchucos Ancash Quechua	qxn	(Wroughton, 1988)
Huaylla Wanca Quechua	qvw	(Raez, 1917), (Cordero Crespo, 1955)
Cañar–Azuay–South Chimborazo Highland Quichua	qxr	(Cerrón-Palomino, 1976), (Guzmán, 1920)

Table 12: Grammars with dialect-specific information on polypersonal verbal agreement.

C Appendix: Results

C.1 Neural Model Variants

In addition to the configurations reported in the main results, we conducted a series of exploratory experiments to better understand the contribution of linguistic features and input formatting in the neural model under the full data setting (see Table 13).

One consistent finding across these experiments is that the ordering of linguistic tags relative to the raw text has a sizeable effect on performance. When both polypersonal and lexical rules were included, an initial configuration appended tags after the text (text + tags), yielding slightly lower performance. Reversing this order (tags + text) resulted in consistent improvements across runs, with the strongest configuration achieving the highest overall performance of over 99% accuracy. Presenting linguistically enriched features at the beginning of the input sequence seems to guide the model more effectively by foregrounding relevant structural cues prior to processing the surface form.

A further refinement involved examining the contribution of individual affixes within the polypersonal rule set. In particular, the suffixes *man* and *ki* were hypothesized to introduce noise. The suffix *man* can encode non-polypersonal meanings such as directional or locative functions across dialects, while *ki* is both orthographically and phonologically short and appears in a wide range of nominal forms, potentially reducing its discriminative value. Empirical results, such as those in Table 13, supported this hypothesis: excluding these affixes led to a consistent improvement in performance. Consequently, the final neural configuration omits *man* and *ki* from the polypersonal feature set.

C.2 Naive Bayes Chunking Strategy

For the Naive Bayes classifier, we conducted additional experiments to evaluate the effect of document chunking on model performance (see Table 13). Specifically, we compared two configurations: one in which all available chunks from each document were used, and another in which the number of chunks per document was capped at 50.

The results show that limiting the number of chunks per document leads to a substantial improvement in performance. Without any restriction, documents with larger amounts of text contribute disproportionately more training instances, potentially biasing the classifier toward dialects with greater data volume. By capping the number of chunks per document, the dataset becomes more balanced across dialects, reducing this source of skew.

C.3 Detailed Segmentation Results

A detailed breakdown of segmentation performance by dialect family and data condition is shown in Table 7.

In high-resource settings (all data), differences between segmentation methods are relatively small. As shown in Table 7, performance remains consistently high across all segmentation strategies, with only marginal variation in both weighted and macro F1. In these conditions, standard frequency-based methods such as BPE and Unigram achieve the highest or near-highest scores across most families, which may indicate that sufficient data allows models to learn effective sub-word representations without explicit morphological guidance.

In contrast, segmentation choice has a slightly more pronounced effect in the low-resource setting. Across families, morphology-aware and hybrid approaches consistently outperform BPE when looking at macro F1. In particular, PRPE+BPE and Unigram yield the strongest and most stable performance. For example, in the Quechua I and Quechua II families, PRPE+BPE achieves the highest or tied-highest macro F1, while Unigram performs best for Quechua I under low-resource conditions. These results indicate that morphology-aware or proba-

Model	Data	Rules	Notes	Accuracy	Precision	Recall	F1
bayes	all	none	unlimited chunk size	.9145	.9062	.9145	.8971
bayes	all	none	limited chunks = 50	.9461	.9539	.9461	.9385
neural	all	verb+lexical	tags + text; no “man” or “ki”	.9953	.9953	.9953	.9951
neural	all	verb+lexical	text + tags	.9906	.9909	.9906	.9906
neural	all	verb	tags + text; all affixes	.9933	.9933	.9933	.9932
neural	all	verb	tags + text; no “man” or “ki”	.9940	.9940	.9940	.9940

Table 13: Additional experimental results on impact of chunk size constraints, rule-based linguistic features, and input ordering of tags versus text.

Setting	n	ΔF1	p (Wilcoxon)	d	F1 (rules)
All Data	27	+0.0015	0.133	0.324	0.957
Bible Only	25	+0.0008	0.285	0.148	0.996
No Bibles	10	+0.0649	0.398	0.342	0.570
NB x All (f)	27	+0.0241	0.502	0.234	0.534
B-only x NB	25	+0.0061	0.465	0.215	0.809

Table 14: Effect of linguistic rule augmentation across different data conditions. Settings using verb rules includes: all data, bible only, no bibles, and no bibles train by all data evaluation relaxed to family accuracy. Setting under lexical rules include: bible-only train by no bible evaluation. n denotes the number of dialects evaluated. ΔF1 indicates the average change in F1 score relative to the no-rules baseline. p (Wilcoxon) reports the significance of paired differences across dialects. d (Cohen’s d) measures effect size. F1 (rules) shows the macro F1 score of the rule-augmented model.

Dialect	Classifier Choice											Total
	qub	quf	quh	quk	qul	qup	quw	qux	quy	quz	other	
qub	**91**	0	0	0	0	0	0	0	0	1	1	93
quf	0	**121**	0	0	0	0	0	0	0	0	0	121
quh	0	0	**254**	0	0	0	0	1	0	9	0	264
quk	0	2	0	**0**	0	0	0	1	0	0	2	5
qul	0	0	0	0	**106**	0	0	0	0	0	0	106
qup	0	0	0	0	0	**133**	0	0	0	0	0	133
quw	0	0	0	0	0	0	**89**	0	0	0	0	89
qux	0	0	6	0	0	0	0	**144**	0	3	0	153
quy	0	0	0	0	0	0	0	0	**823**	1	0	824
quz	0	0	21	0	0	0	0	2	0	**2081**	1	2105
other	0	0	1	0	0	0	0	0	0	0	0	
Total	91	123	282	0	106	133	89	148	823	2095		

Table 15: Truncated confusion matrix for dialect classification on all data without rules.

bilistic segmentation may better support generalization when training data is limited.

Performance differences also vary by dialect family (see Figure 5 and Figure 6). The Quechua II model achieves the highest scores across nearly all conditions, reflecting its larger and more balanced training data. In contrast, the San Martín model shows consistently lower performance across segmentation strategies, suggesting greater sensitivity to data sparsity. However, even in these lower-resource families, PRPE+BPE improves macro F1 relative to BPE, which suggests that linguistically informed segmentation provides some gains.

Dialect	Classifier Choice											Total
	qub	quf	quh	quk	qul	qup	quw	qux	quy	quz	other	
qub	**91**	0	0	0	0	0	0	0	0	1	1	93
quf	0	**121**	0	0	0	0	0	0	0	0	0	121
quh	0	0	**247**	0	0	0	0	0	0	17	0	264
quk	0	1	0	**0**	0	0	0	1	0	0	3	5
qul	0	0	0	0	**106**	0	0	0	0	0	0	106
qup	0	0	0	0	0	**133**	0	0	0	0	0	133
quw	0	0	0	0	0	0	**89**	0	0	0	0	89
qux	0	0	4	0	0	0	0	**147**	0	2	0	153
quy	0	0	0	0	0	0	0	0	**823**	1	0	824
quz	0	0	1	0	0	0	0	0	0	**2104**	0	2105
other	0	0	1	0	0	0	0	1	0	0	–	
Total	91	122	253	0	106	133	89	149	823	2125		

Table 16: Truncated confusion matrix for dialect classification using verb rules on all data.

Dialect	Classifier Choice										Total
	quf	quh	qul	quy	quz	qve	qvo	qvw	qwh	other	
quf	**0**	0	0	0	0	0	0	0	0	0	0
quh	9	**8**	16	47	19	3	1	12	11	35	161
qul	0	0	**0**	0	0	0	0	0	0	0	0
quy	0	0	0	**394**	0	0	0	1	1	0	396
quz	165	3	24	93	**1020**	198	14	92	24	38	1671
qve	0	0	0	0	0	**0**	0	0	0	0	0
qvo	0	0	0	0	0	0	**0**	0	0	14	14
qvw	0	0	0	0	0	0	0	**0**	0	0	0
qwh	0	0	0	2	0	0	0	1	**4**	7	14
other	0	0	0	0	1	0	0	0	0	–	
Total	174	11	40	536	1040	201	15	106	40		

Table 17: Truncated confusion matrix for dialect classification on no rules model trained on only-Bible data tested on no-Bible data.

Dialect	Classifier Choice										Total
	quf	quh	qul	quy	quz	qve	qvo	qvw	qwh	other	
quf	**0**	0	0	0	0	0	0	0	0	0	0
quh	8	**7**	17	50	25	1	3	3	21	26	161
qul	0	0	**0**	0	0	0	0	0	0	0	0
quy	0	0	0	**394**	0	0	0	0	1	1	396
quz	35	8	29	97	**1359**	49	17	18	33	26	1671
qve	0	0	0	0	0	**0**	0	0	0	0	0
qvo	0	0	0	0	0	0	**0**	0	0	14	14
qvw	0	0	0	0	0	0	0	**0**	0	0	0
qwh	0	0	0	3	0	0	0	0	**7**	5	15
other	0	0	0	0	1	0	0	0	0	–	
Total	43	15	46	544	1385	50	20	21	62		

Table 18: Truncated confusion matrix for dialect classification on lexical model trained on only-Bible data tested on no-Bible data.

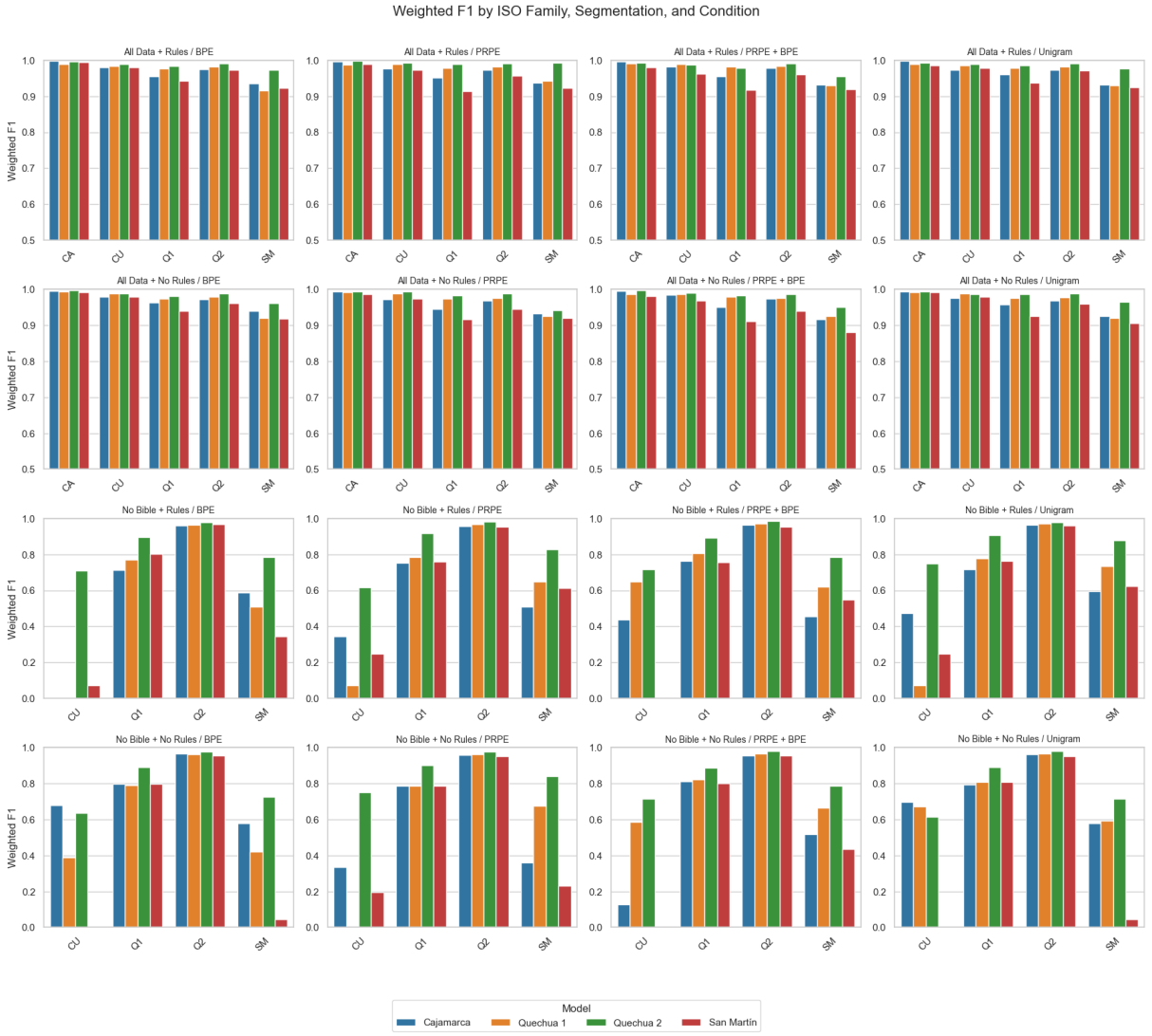

Figure 5: Weighted F1 scores by language family, segmentation method, and experimental condition. Each subplot is a combination of data and rules; bars show performance trained on different Quechua families.

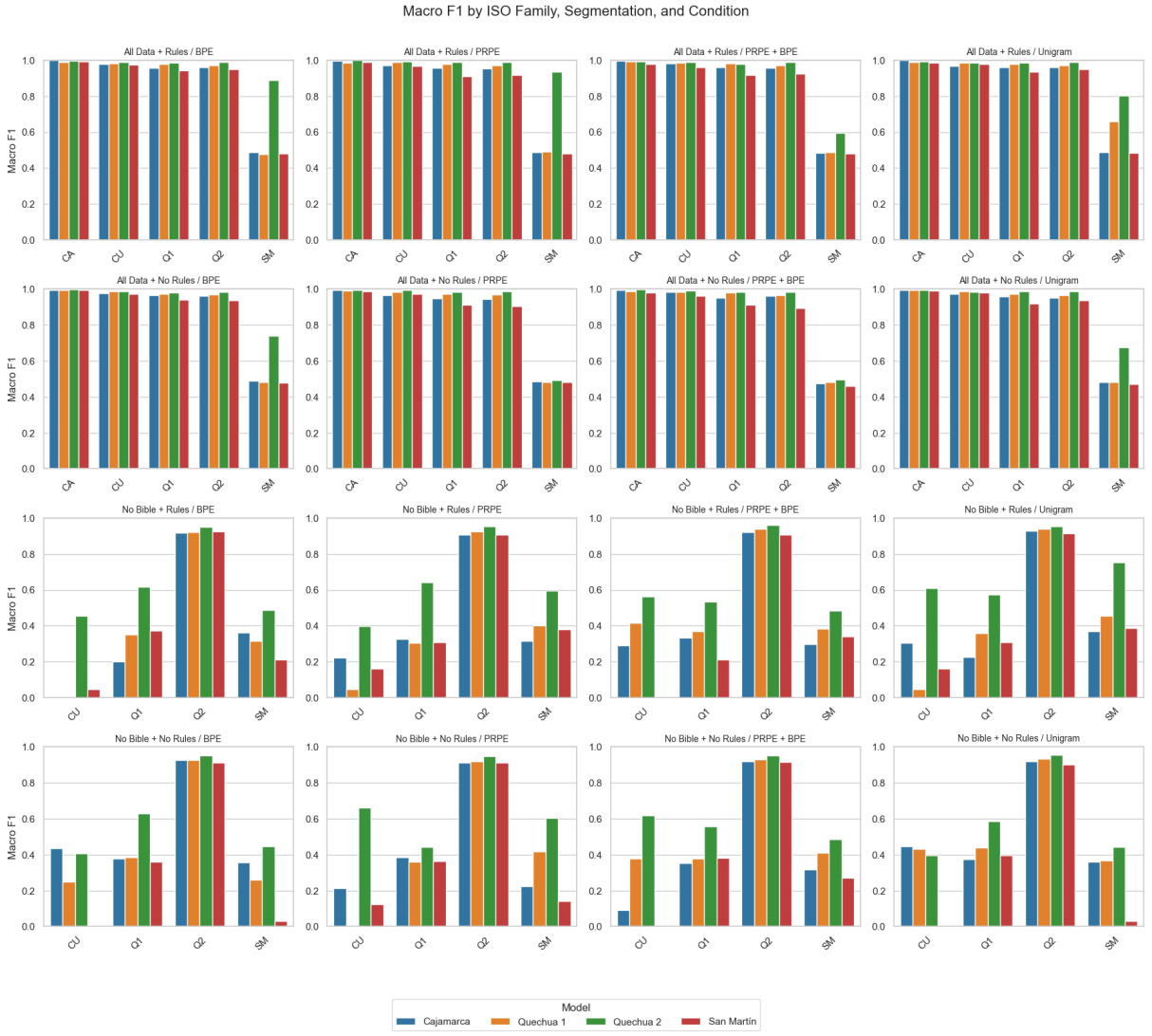

Figure 6: Macro F1 scores by ISO language family, segmentation method, and experimental condition. Each subplot is a combination of data and rules; bars show performance trained on different Quechua families.

What Resources Matter for Interlinear Glossing? Using LLMs and RAG for the Low-Resource Mapudungun Language

Anaís Almendra[1], Arianna Bisazza[2], Claudio Gutierrez[3], Felipe Hasler[1]
[1] Department of Linguistics, Universidad de Chile, Santiago, Chile
[2] CLCG, University of Groningen, Groningen, Netherlands
[3]Department of Computer Science, Universidad de Chile, Santiago, Chile
Correspondence: anaisalmendra21.aa@gmail.com

Abstract

Interlinear glossing is essential for the study and revitalization of endangered languages. However, it remains a time-consuming process that requires extensive linguistic expertise. Recent advances in Large Language Models (LLMs) offer a potential solution. In this research, we study the case of Mapudungun, an endangered language spoken in Chile and Argentina, to generate automatic interlinear glosses using the Gemini 2.5 Pro model. Our study investigates which information configuration through Retrieval-Augmented Generation (RAG) yields the best results. We compare the integration of a formal grammar, a dictionary, a small annotated corpus, and a combination of all these resources. Our evaluation shows that while dictionary integration causes a significant degradation in performance, grounding the model with a structured corpus maximizes accuracy relative to the resources employed. Notably, we find that a remarkably small dataset of 589 meaning units provides enough normative guidance to significantly improve the morphological tagging task. This work highlights the viability of utilizing minimally annotated corpora to assist in the documentation of morphologically complex languages.

1 Introduction

Currently, numerous languages worldwide are in a vulnerable situation (SIL International, 2026). Among them is Mapudungun (Gundermann et al., 2011; Zúñiga, 2019), an indigenous language of Chile and Argentina with approximately 380,000 speakers in the former country (Instituto Nacional de Estadísticas, 2024). This language has a morphological profile classified as agglutinative and polysynthetic (Golluscio and Hasler, 2017; Zúñiga, 2017) with a templatic structure for morpheme incorporation (Fortescue et al., 2017).

In the field of linguistic documentation, there are data annotation techniques through which text, audio, or video are structured and aligned with metadata (Woodbury, 2011). Among these procedures, interlinear glossing stands out as a technique commonly used in projects of this nature (Ginn and Palmer, 2023). This annotation format operates through the segmentation of morphemes and the assignment of grammatical tags (Elsner and Liu, 2025), which allows capturing the morphosyntactic features of the language under study and performing precise analyses of them (Ginn et al., 2024b). In this context, the Leipzig Glossing Rules provide a series of syntactic and semantic rules, in addition to a standardized set of tags, to systematize the interlinear glossing process (Comrie et al., 2015).

In the case of endangered languages, interlinear glossing has established itself as a documentary tool capable of facilitating their revitalization (Elsner and Liu, 2025). Despite the benefits that the generation of interlinear glossing brings to endangered languages, this task constitutes a highly complex endeavor, heavily dependent on experts, and requiring large amounts of time for its generation (Ginn et al., 2024b). To illustrate this task, the following example corresponds to a Mapudungun text organized according to the Leipzig Glossing system:

(1) kuydaufisayawürkey
kuyda -ufida -ya -w -ürke
take.care -sheep -FUT -REFL -EVID
-y
-IND.[3]

'It is say he was tending his sheep.'

In this example, Mapudungun phenomena can be observed, such as object incorporation into the verbal root and the suffixation of elements such as evidentiality, tense, and person.

Despite the utility of interlineal glossing, Mapudungun lacks a corpus with these characteristics.

In response to this limitation, and in relation to

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 64–73

previous works (Ginn and Palmer, 2023; Ginn et al., 2024a,b; Elsner and Liu, 2025), we propose the use of Large Language Models (LLMs) for automatic interlinear glossing. Since Mapudungun, like most languages in the world, lacks a sufficient corpus to train large language models (Zhang et al., 2024), our study leverages the linguistic capabilities of these models to generate glosses for Mapudungun without the need for further training.

We investigate what is the most efficient approach for generating automatic interlinear glossing for Mapudungun. Specifically, we evaluate the Gemini 2.5 Pro model and explore the use of the Retrieval-Augmented Generation (RAG) technique to provide the LLM with three different sources of language information: (i) grammar, (ii) dictionary, and (iii) a Mapudungun corpus newly annotated according to the Leipzig Glossing system.

Our work identifies key issues regarding the generation of automatic glossing for Mapudungun. Firstly, the results indicate that none of the additional information sources provide statistically significant improvements in the text and segmentation tiers of the Leipzig glosses. Secondly, the tagging task is identified as the most complex. In this tier, the integration of additional information from the annotated corpus and the use of all materials combined do provide a statistically significant difference. Thirdly, it is more effective to use solely standard information, such as the annotated corpus, rather than the combination of multiple resources (dictionary, grammar, and corpus together), since the former produces better results and requires fewer resources. Consequently, these findings contribute to research seeking to perform automatic interlinear glossing for languages with scarce annotated corpora, such as Mapudungun, establishing the most efficient alternative with the lowest environmental impact.

2 Background and Related Work

2.1 Mapudungun Language

Mapudungun is a language of isolated genealogy (Golluscio and Hasler, 2017; Zúñiga, 2017) spoken in Chile and Argentina, which is currently in a vulnerable situation (Gundermann et al., 2011; Zúñiga, 2017). In Chilean territory, it has approximately 380,000 speakers (Instituto Nacional de Estadísticas, 2024). From the point of view of its morphological typology, it is classified as an agglutinative and polysynthetic language (Zúñiga, 2017). In turn, its affix integration structure responds to a templatic model (Fortescue et al., 2017; Zúñiga, 2017), given that its morphological incorporation is rigid regarding the positions in which morphemes can be added in relation to the verbal root. This morphology of the language has been classified as complex due to the diversity of linguistic phenomena it exhibits (Zúñiga, 2017). In this regard, processes such as noun incorporation, the adjunction of multiple morphemes onto a single root, and the large number of elements that possess the capacity to operate as roots within a morphological construction stand out (Zúñiga, 2017).

Regarding the available linguistic resources for Mapudungun, although grammars, dictionaries, and various digital tools exist, there is a lack of a specialized corpus in interlinear glossing. In this context, the automation of annotated corpus generation emerges as a strategy with the potential to accelerate the creation of resources for the study and revitalization of the language.

2.2 Automatic Interlinear Glossing

Given the complexities associated with the creation of interlinear glossing, recent research has explored the application of LLMs to accelerate the processes of generating corpora with interlinear glossing. Along these lines, Ginn and Palmer (2023) approached the task through the fine-tuning of a pretrained model, reporting an increase of two percentage points in performance compared to the unrefined architecture. This study highlights the inherent complexity that the automation of this labor represents. Subsequently, Ginn et al. (2024b) performed a model adjustment for gloss generation in multiple languages, employing a database of approximately 450,000 examples distributed across 1,800 languages. The results show an improvement of about 7 percentage points against state-of-the-art models. Despite this, the authors note that performance varies largely with the amount of available data.

Instead of model refinement, Ginn et al. (2024a) evaluated the use of few-shot prompting, exploiting the in-context learning abilities of LLMs. In this design, the system receives as input a transcription line and its respective translation, from which it must generate the corresponding gloss. The findings indicate that providing interlinear glossing examples in the prompt substantially impacts the model's performance. Likewise, it is concluded that, although interlinear gloss generation remains

highly complex, the strategic selection of pertinent examples can yield significant improvements.

Finally, Elsner and Liu (2025) also applied the prompting technique for automatic interlinear glossing, adopting the Leipzig Glossing Rules as a normative standard. They worked with different languages and used a prompting system with examples for each one that contained sentences, glosses, and translations. The study shows promising results and highlights the potential of LLMs as support tools for linguists.

2.3 In-context Learning for Low-Resource Languages

In the context of LLM use and low-resource languages, recent research has opted to utilize the linguistic capabilities of the models instead of further training them (Court and Elsner, 2024; Zhang et al., 2024; Spencer and Kongborrirak, 2025; Zhu et al., 2025). In the aforementioned studies, the models are operated through prompting with different types of linguistic information, depending on the expected task and the language being worked with.

Specifically, Court and Elsner (2024) work with the translation task from a Quechua variant to Spanish. They utilize the RAG and prompting technique to work with three types of materials, both separately and jointly: translations of morphemes and words, grammatical descriptions, and usage examples from parallel corpora. Spencer and Kongborrirak (2025) experiment with different RAG and prompting configurations with the goal of assisting the creation of a grammar for a low-resource language.

Zhu et al. (2025) use a puzzle-based methodology with features of varying complexity to identify whether LLMs can capture linguistic features of unseen languages. To deliver the language data, they work with step-by-step prompting. Finally, Zhang et al. (2024) test the translation task between English and languages unseen in the training corpora. For this, they utilize three materials through prompting: dictionary, grammar, and morphologically analyzed text.

3 Methodology

The present research works with the RAG technique on the Gemini 2.5 Pro model and uses as materials *A Grammar of Mapuche* (Smeets, 2008), *Diccionario mapudungun-español español-mapudungun* (Augusta, 2017), and adapts part of the AVENUE corpus (Levin et al., 2002), enriching it through the use of Leipzig Glosses.

Our interaction system, as shown in the Prompt Used Appendix A.1, instructs the model to process an input text and return its analysis in a single message response according to the Leipzig Glossing Rules. In this way, we encourage the model to treat the four Leipzig tiers as a single unit, thereby aiming to minimize potential hallucinations or the omission of units during the segmentation and tagging processes.

3.1 Evaluation

The results of the experiment were evaluated according to three Leipzig Glossing tiers: **text**, **segmentation**, and **tagging**. The translation tier is excluded from the analysis given that the task of evaluating how well a translation is performed requires theoretical and practical approaches beyond the scope of this study, which we leave to future work. We assess text, segmentation, and tagging in a binary manner based on the manually annotated evaluation corpus. That is, only an exact match of the expected meaning is considered correct, while any deviation is counted as an error.

McNemar's test was used to compare the models, and analyses of Accuracy and p-values were conducted. Each model was compared against the evaluation corpus; subsequently, the baseline was compared independently with corpus, dictionary, grammar and all the materials models. Because McNemar's test strictly requires paired data of equal length, we developed a tier alignment system to handle the LLMs' tendency to omit information or hallucinate content. In cases where the model generated an unnecessary morpheme or omitted a required one, our system aligned the output by inserting empty spaces, which were scored as incorrect predictions. This alignment method ensured that the evaluation sets maintained the same length.

In Table 1, we present representative examples for each evaluation tier using the dictionary and all materials settings. This exact cell-by-cell alignment and evaluation protocol was systematically applied to all experimental configurations.

3.2 Materials Used

We experiment with delivering three different linguistic resources to the model via RAG, as well as their combination:

Tier / Setting	Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6
1. Text Tier						
Evaluation	nierpuafuy	–	–	–	–	–
Baseline	nierpuafuy (✓)	–	–	–	–	–
Dictionary	nierpuafuy (✓)	–	–	–	–	
All	nierpuafuy (✓)	–	–	–	–	–
2. Segmentation Tier						
Evaluation	nie	r	pu	a	fu	y
Baseline	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓)
Dictionary	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓))
All	nie (✓)	Ø (×)	rpu (×)	a (✓)	fu (✓)	y (✓)
3. Tagging Tier						
Evaluation	have	ITR	TRANS	FUT	FRUS	IND.[3]
Baseline	have (✓)	Ø (×)	TRNSL (×)	FUT (✓)	IRR (×)	IND.[3] (✓)
Dictionary	have (✓)	Ø (×)	TRNSL (×)	FUT (✓)	IRR (×)	IND.[3] (✓)
All	have (✓)	Ø (×)	TRANS (✓)	FUT (✓)	IRR (✓)	IND.3 (×)

Table 1: A representative example *nierpuafuy* 'They would gradually have' from the evaluation corpus, illustrating the cell-by-cell alignment and binary evaluation across all tiers. The table contrasts the baseline, dictionary, and all materials configurations against the evaluation corpus. This alignment method provides equal sequence lengths across outputs, yielding the exact paired binary metrics required to compute McNemar's test.

- Grammar: *A Grammar of Mapuche* (Smeets, 2008) has been used by already existing morphological analyzers for Mapudungun (Almendra, 2025) and presents a chapter entirely dedicated to the morphology of the language and its structuring.

- Dictionary: *Diccionario mapudungun-español español-mapudungun* (Augusta, 2017) has been classified as one of the most comprehensive lexical works of the language (Augusta, 2017) and has been used for the development of a morphological analyzer for Mapudungun (Almendra, 2025).

- Corpus: the AVENUE project includes an open-access corpus of transcribed spoken Mapudungun (Levin et al., 2002). For this research, a part of it was selected and enriched with annotation according to Leipzig Glosses (cf. section 3.3)

Each of these materials provides different kinds of information regarding Mapudungun morphology. The grammar presents examples alongside explanations of the rules that Mapudungun follows regarding its morpheme structuring. The dictionary constitutes a repository of equivalencies between Mapudungun and Spanish. This resource presents the lemmas containing the language's morphemes together with their equivalencies in the majority language. Finally, the corpus is a set of demonstrations regarding how to gloss Mapudungun according to the Leipzig system. This information provides annotation examples in specific contexts of use; however, it does not present explicit explanations regarding the functioning of Mapudungun morphology.

3.3 Corpus

The corpus was obtained from a total of four audios that were aligned in ELAN (Max Planck Institute for Psycholinguistics, n.d.) with their transcriptions and translations for their subsequent analysis in FLEx (SIL International, 2026). It consists of a series of texts in Mapudungun glossed according to the Leipzig rules. Each sentence or word in Mapudungun appears in the corpus with four tiers of information: text, segmentation, tagging, and translation (see example gloss in 1). As Mapudungun has multiple orthographic systems (Llanquimán et al., 2025), we standardized the orthography using the tools and guidelines of *KMT - Kümewirin Mapudüngun Trapümwe* (Chandía, n.d.).

Then, we applied the following annotation principles:

1. The corpus was cleaned by eliminating interjections, incomplete or erroneous speech onsets, and proper nouns.

2. The inclusion of lexical borrowings was permitted, provided they did not imply code-

switching to Spanish within the same sentence, thus maintaining the focus on the structure of Mapudungun.

3. For the generation of Leipzig glosses, truncated or abbreviated forms in speech were represented by their full form in the gloss. For example, a form such as *feli* was segmented as *feley* for its subsequent tagging.

4. The glossing process allowed the correction and validation of translations and the assignment of tags to morphemes. Samples in which there was doubt along any of these two aspects were excluded.

The annotation process was carried out by two of the authors in two stages. First, a linguist specialized in ELAN and FLEx with intermediate knowledge of Mapudungun grammar performed the audio annotation and generated the preliminary analyses. Subsequently, a linguist specializing in the language's grammar reviewed and finalized the data in FLEx.

Given the polysynthetic and agglutinative nature of Mapudungun, the corpus split into the RAG retrieval subset and the evaluation subset was not measured by individual words. Instead, we used the concept of unit of meaning as the base metric. We define a unit of meaning as any element that possesses an independent linguistic tag. For example, the lexeme *ñuke* 'mother' counts as one unit. In contrast, a complex verbal form such as *tunualengün* is segmented into *tu-nu-a-*l=*engün*, receiving the tags [grab-NEG-FUT-NMLZ-3.PL], and is therefore counted as five meaning units. Once the entire corpus was processed, we counted the total number of meaning units and divided the RAG and evaluation data based on this metric.

Since the corpus was derived from audios, to avoid biases in the use of RAG and the evaluation, we aimed to allocate half of the content of each audio to each subcorpus, thus ensuring equitable representativeness. In its entirety, the total corpus corresponds to 40 speech turns and 1,175 units of meaning. The RAG subset contains 589 units and 22 speech turns, while the evaluation subset consists of 586 units and 18 speech turns.

In comparison to other studies, where hundreds of thousands of examples are utilized (Ginn et al., 2024b), the size of the corpus in this research is minute. Given the cost of manual annotation, we are interesting in assessing whether LLMs could accelerate the annotation process of further data.

3.4 Infrastructure Selection

The model used was Gemini 2.5 Pro on Google Cloud (Vertex AI), specifically within the Model Garden infrastructure employing the RAG Engine technique. This was selected for two essential reasons: (i) data management, and (ii) accessibility to the use of RAG. On the one hand, the Vertex AI platform provides guarantees regarding data use, indicating that these will not be used for training its models and that the generated results do not belong to the company either (Google LLC, 2025). On the other hand, Vertex AI offers the possibility of applying RAG without the need to develop code. This aspect is crucial to our approach, since it facilitates access for individuals who are not specialized in this area, such as linguists, and allows working with specialized models and data like grammars, dictionaries, or other materials. The model temperature is set to 0 in all experiments.

4 Results

We present the results using accuracy and statistical significance. Since each tier of the Leipzig gloss was evaluated independently, we report the performance separately for each one. In all experiments, 'Baseline' refers to the results obtained by querying Gemini 2.5 Pro with a fixed prompt containing 3 glossing examples (see full prompt in A.1). The 'All' setting refers to the combination of the three materials (grammar, dictionary, corpus) provided altogether to the LLM via RAG.

4.1 Text

In the text tier, only the corpus configuration outperforms the baseline, but by less than one percentage point. While the combined materials configurations match the baseline performance, the dictionary and grammar configurations underperform.

As shown in Table 2, the observed differences were only significant in the case of the dictionary, which worsened the performance ($p < 0.05$). None of the other settings showed a significant difference in the text tier ($p > 0.05$). These results indicate that the baseline already achieves high performance in the text processing task, which limits the observable margin of improvement from material additions via RAG.

Setting	Text	Segmentation	Tagging
Baseline	98.26%	78.62%	41.08%
Corpus	**98.96%**	74.92%	**56.41%***
Grammar	96.18%	76.88%	39.46%
Dictionary	90.62%*	70.95%*	34.91%*
All	98.26%	**80.60%**	**55.69%***

Table 2: Evaluation results (Accuracy) across the three tasks: text, morphological and tagging. The table compares zero-shot execution against various RAG augmented contexts. Bold values indicate results that outperform the baseline, and * denotes statistical significance.

4.2 Segmentation

We first observe that the segmentation results are overall lower than those achieved in the text task. This indicates a greater complexity for the morphological segmentation task.

In the segmentation tier, only the use of all materials outperforms the baseline. The corpus, dictionary, and grammar configurations individually underperform.

As shown in Table 2, once again the observed differences were only significant in the case of the dictionary, which worsened the performance ($p < 0.05$). None of the other settings showed a significant difference in the segmentation tier ($p > 0.05$). Notably, even the configuration that outperformed the baseline (All) did not yield a statistically significant improvement.

4.3 Tagging

The tagging task appears as the most complex among the three evaluated tiers with accuracies below 60% in all configurations.

In this tier, the corpus configuration and the combined materials configuration outperform the baseline, with the former achieving better performance. On the other hand, the individual use of the dictionary and grammar underperforms.

As shown in Table 2, the observed pattern was maintained, and the use of the dictionary significantly worsened the performance ($p < 0.05$). Unlike the previous tiers, the all materials and corpus settings significantly improved the performance ($p < 0.05$). Only the grammar configuration did not show a significant difference, neither improving nor worsening the performance ($p > 0.05$).

5 Discussion

Our results show that the inclusion of the grammar does not yield statistically significant gains in the system's performance compared to the baseline. This can be due to the fact that the model already possesses in its weights knowledge of Mapudungun that overlaps with the one provided by the grammar. Alternatively, this could mean that the model is not capable of exploiting the information provided by grammar and applying it to novel examples.

Dictionary integration produces a significant degradation of results across all analyzed cases. Even in the text tier, which should not present complexities for the model given that it constitutes a replication of the text provided by the user, the use of the dictionary worsened performance.

In contrast, the configurations based on the use of the corpus and all materials consistently record the highest levels of accuracy.

Regarding the statistical significance of these findings, we found that no configuration made a significant improvement in the text processing and segmentation tasks. In these areas, the performance of the model without additional information overlaps with the all materials and corpus configurations. Despite this, in the tagging task, we identify statistically significant improvements when employing the corpus and all materials configurations. Among them, the former not only exhibits superior performance but also achieves this through the use of a smaller volume of data compared to the use of all materials. These results show that the greater the complexity of the task, the greater the relevance of using RAG; in this sense, the use of additional information only makes a real contribution to improving the tagging tier.

In the tagging tier, the all and corpus configurations exhibit the best performance, with the latter outperforming the rest. Whereas the dictionary and grammar fail to reach the baseline, the aforementioned configurations exceed it with statistical significance. We attribute the performance gains of

the all configuration to the inclusion of the corpus, given that it yields the best overall results when used in isolation.

The performance gains achieved through the corpus integration can be observed in the mitigation of recurring baseline errors. Examples of this involve the morphemes *-fu* and *-nie*. In the baseline output, *-fu* was tagged as IRR instead of the correct target tag, FRUS. The all and corpus configurations resolved this by correctly tagging the *-fu* morpheme as FRUS. Likewise, the baseline confused the morpheme *-nie* with the verbal root *nie* 'have'. Once again, the all and corpus configurations correctly assigned the PROG.PS tag to this morpheme. Despite the gains achieved, we note common errors in the tagging tier, primarily driven by homophonous morphemes. In Mapudungun, the suffix *-tu* can take different meanings depending on the context, functioning as a verbalizer, a repetitive marker, or a transitivizer. In cases where *-tu* appeared, the model exhibited a frequent error by assigning the verbalizer label even when incorrect, probably due to an overrepresentation of this specific function in the data. To prevent this, we believe that refining the selection of examples to ensure a balanced representation of these syntactic phenomena could improve performance in future work.

Our findings align with the high complexity of automatic interlinear glossing highlighted in previous research (Ginn and Palmer, 2023), especially within the tagging tier. Consistent with recent literature (Ginn et al., 2024a; Elsner and Liu, 2025), we observed that utilizing prompting and the examples provided by our developed corpus offers an effective alternative to model fine-tuning. Furthermore, our RAG methodology improved the baseline results for Mapudungun, aligning with its successful use in tasks with other languages (Spencer and Kongborrirak, 2025). Nevertheless, these outcomes emphasize that the nature of the augmented material is critical, as external information does not inherently guarantee better results, as also observed in other studies (Court and Elsner, 2024). For instance, while some studies found the use of grammar beneficial (Zhang et al., 2024), the grammar configuration tested in this study failed to produce statistically significant gains. Additionally, this contrasts with previous work (Court and Elsner, 2024), which observed performance drops when using a corpus in translation tasks. These discrepancies suggest that the effectiveness of specific RAG resources is task-dependent.

In view of the above, we propose that to optimize the automatic tagging of Mapudungun using LLMs, the most effective strategy consists of employing an annotated corpus that acts as a normative guide for the model. This finding is relevant, as it contradicts a possible hypothesis that the use of diverse linguistic resources favors better generalization or morphological analysis. In this sense, we believe that combining a corpus structured under the Leipzig Glossing Rules with a RAG approach is a methodological novelty that allows us to improve the performance in automatic interlinear glossing generation. For low-resource languages with complex morphology like Mapudungun, this approach prioritizes information density over large data volumes, providing multiple specialized tiers from just a single word or sentence. Despite these insights, the findings of our study should be interpreted within the scope of a single LLM configuration. While this framework can inform future work across other architectures, it is important to acknowledge this limitation.

Finally, the size of the corpus employed in this research provides relevant evidence regarding the processing of Mapudungun morphology by Gemini 2.5 Pro. As mentioned, the quantity of examples provided by the corpus in this study is considerably smaller than that utilized in other works. In this context, our results indicate that the use of 589 meaning units, equivalent to 22 speech turns, is sufficient to improve automatic interlinear glossing for Mapudungun. Consequently, these findings establish a minimum threshold of data necessary to replicate this technique in other low resource languages like Mapudungun.

6 Conclusions

In the context of the automatic generation of interlinear glossing for Mapudungun using RAG, the results indicate that the best cost-result strategy consists of utilizing a previously glossed corpus that operates as a normative standard, rather than integrating multiple documentary sources. Our findings show that integrating LLMs and external data is a viable approach for glossing low-resource languages such as Mapudungun.

Likewise, we observe that the usefulness of providing extra linguistic materials via RAG depends on the complexity of the task addressed. While in the text and segmentation tiers, external information sources do not yield substantial improvements,

their impact is statistically significant in the tagging tier, which constitutes the highest level of difficulty within the glossing process.

Consequently, these findings provide an empirical framework to guide the work and methodological decisions of those dedicated to glossing data-scarce languages, like Mapudungun. The evidence obtained delivers concrete results to guide the course of action in those projects and research endeavors seeking to accelerate the processes of interlinear glossing generation for this type of languages.

7 Ethical considerations

For the development of this research, an annotated corpus was constructed from the online resource AVENUE Project (Levin et al., 2002). The use of this resource, enriched with morphological annotation, is proposed as a methodological strategy for the reuse of preexisting materials. This approach is substantiated as an alternative to over-intervention in speaking communities, relying exclusively on already consolidated data.

In turn, the integration of the resources used seeks to establish a pathway that facilitates the work of glossers of this language. The purpose of this experimental design does not aim to replace the work of human annotators, but rather to optimize and accelerate their processes, thus allowing the generation of a larger volume of useful corpus for the study of Mapudungun morphology. Under this logic, we propose using a single annotated corpus rather than combining multiple materials, as this approach requires significantly fewer computational resources. This reduction of the processing footprint is fundamental from an ethical perspective, considering that marginalized communities are precisely those who suffer the most from the environmental costs involved in training LLMs (Bender et al., 2021).

Finally, the application of the RAG architecture and the identification of the most efficient technique represent a contribution to the democratization of access to LLM-based technologies. Given its low technical barrier, this strategy enables non-engineering profiles, such as linguists, documentarians, and community members, to leverage these technologies for their own languages. Ultimately, we hope that the results and techniques implemented here will contribute to the efforts of documentation, study, and revitalization of languages globally.

Acknowledgments

Claudio Gutierrez thanks funding from IMFD, ANID - Millennium Science Initiative Program - Code ICN17_002.

Arianna Bisazza is funded by the Talent Programme of the Dutch Research Council (NWO) under project VI.Vidi.221C.009.

Felipe Hasler and Anaís Almendra were supported by ANID FONDECYT Project No. 1251110 *Estructura argumental y cambio de valencia en lenguas de los andes del sur*.

Anaís Almendra thanks to National Center for Artificial Intelligence CENIA FB210017, Basal ANID for their support during the development of this research.

References

Anaís Almendra. 2025. ¿cómo estudiamos y aportamos a las lenguas indígenas desde la computación? In Nicolás Albornoz, Valentina Espinoza, Camila Cortez, and Joaquín Vásquez, editors, *Cartografías lingüísticas: Un abordaje desde y hacia la interdisciplinariedad*, pages 169–182. Ediciones Colegas.

Félix José de Augusta. 2017. *Diccionario mapudungún-español, español-mapudungún*. Universidad Católica de Temuco and Centro de Investigaciones Diego Barros Arana. Compilado y editado por Belén Villena Araya.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623.

Andrés Chandía. n.d. KMT – Kümewirin Mapudüngun Trapümwe: Unificador ortográfico de mapudüngun. Sin fecha. Accedido: 08-Abril-2026.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Micha Elsner and David Liu. 2025. Prompt and circumstance: A word-by-word LLM prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the 22nd SIGMORPHON workshop*

on Computational Morphology, Phonology, and Phonetics, pages 1–14, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Michael Fortescue, Marianne Mithun, and Nicholas Evans. 2017. Introduction. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pages 1–23. Oxford University Press.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Michael Ginn and Alexis Palmer. 2023. Robust generalization strategies for morpheme glossing in an endangered language documentation context. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, page 89–98. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024b. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Lucía Golluscio and Felipe Hasler. 2017. Jerarquías referenciales y alineamiento inverso en Mapudungun. *RASAL Lingüística*, pages 69–93.

Google LLC. 2025. Términos de servicio específicos de Google Cloud (es-419). Sin fecha. Accedido: 08-Abril-2026.

Hans Gundermann, Jaqueline Canihuan, Alejandro Clavería, and César Faúndez. 2011. El mapuzugun, una lengua en retroceso. *Atenea (Concepción)*, (503):111–131.

Instituto Nacional de Estadísticas. 2024. Resultados – censo 2024. Accedido: 15 de abril de 2026.

Lori Levin, Rodolfo Vega, Jaime G. Carbonell, Ralf D. Brown, and Carolina Huenchullan. 2002. Data collection and language technologies for Mapudungun. In *Proceedings of the LREC-2002 Workshop*, Las Palmas, Spain.

Eduardo Llanquimán, Cristian Lagos, and Elizabeth Torrico-Ávila. 2025. Los grafemarios de la lengua mapuche como herramienta de revitalización lingüística: una revisión bibliográfica. *Revista de Lenguas y Literatura Indoamericanas*, 26(01–02):28–57.

Max Planck Institute for Psycholinguistics. n.d. ELAN. Accessed: 08-Abril-2026.

SIL International. 2026. Ethnologue: Languages of the world. Accessed: April 13, 2026.

SIL International. 2026. FieldWorks. Accessed: 08-Abril-2026.

Ineke Smeets. 2008. *A Grammar of Mapuche*, volume 41 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin and New York.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Hongpu Zhu, Yuqi Liang, Wenjing Xu, and Hongzhi Xu. 2025. Evaluating large language models for in-context learning of linguistic patterns in unseen low resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 414–426, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fernando Zúñiga. 2017. Mapudungun. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, pages 1–24. Oxford University Press.

Fernando Zúñiga. 2019. Grammatical relations in Mapudungun. In A. Witzlack-Makarevich and B. Bickel, editors, *Argument selectors: A new perspective on grammatical relations*, pages 39–67. John Benjamins Publishing Company.

A Appendix

A.1 Prompt used

"Eres un lingüista experto en el análisis morfológico del mapudungun, una lengua genealógicamente aislada hablada en Chile y Argentina. Debes aplicar el formato de las Glosas de Leipzig para entregar tus análisis morfológicos del mapudungun al usuario. Los inputs del usuario pueden ser de los siguientes formatos:

Ejemplo 1:
Input del usuario: ilotukelan
Tu respuesta al usuario:
ilotukelan
ilo-tu-ke-la-n

carne-VRBZ-HAB-NEG-IND.1SG
“Yo no como carne”
Ejemplo 2:
Input del usuario: chaw
Tu respuesta al usuario:
chaw
chaw
padre
“padre”
Ejemplo 3:
Input del usuario: tañi chaw müley tañi rukamew
Tu respuesta al usuario:
tañi chaw müley tüfi rukamew
ta= ñi chaw müle-y tüfey ruka-mew
DET= 1SG.POSS padre estar-IND.[3] ese casa-PPOS
“Mi padre está en esa casa”

Si el input es un elemento que no presenta morfemas (por ejemplo, un sustantivo sin afijos), no realices segmentación morfológica y repite la forma tal como aparece, tal como se presenta en el ejemplo 2.

Para la entrega de tus respuestas piensa paso a paso para realizar la segmentación y análisis del input entregado por el usuario, pero solo entrega tu respuesta, es decir, el análisis. Ciñe tus respuestas a los ejemplos entregados. Las equivalencias entre el mapudungun y las etiquetas y traducción debe ser en español, no en inglés u otra lengua.

Cuando el usuario te entregue un texto en mapudungun responde solo con la estructura analizada según los lineamientos de las Glosas de Leipzig. No añadas información adicional, solo entrega tu análisis.”

Deer, Deities, and Dancing: Culturally Biased LLM Hallucination in Low-Resource Wixárika Translation

Henry Gagnier
Pittsford Sutherland High School
Pittsford, New York, USA
henrygagnier9@gmail.com

Ashwin Kirubakaran
Edison Academy Magnet School
Edison, New Jersey, USA
ashwinkiru10@gmail.com

Abstract

Large language models (LLMs) struggle with low-resource polysynthetic languages, yet the nature of their failures remains underexplored. We evaluate GPT-4o-mini, Gemma 3 27B, Llama 3.3 70B, and NLLB-200 on Spanish↔Wixárika translation using zero-shot and 5-shot prompting. All systems are unusable, scoring below 3 BLEU and 21 chrF. Qualitative analysis reveals that LLMs largely ignore source content and instead generate fluent hallucinations. Spanish outputs frequently include indigenous cultural stereotypes such as deer, deities, rain dance, and shamans, regardless of the input, while Wixárika outputs are repetitive across different inputs and morphologically implausible. Few-shot prompting yields model-dependent improvements, with Gemma and Llama improving substantially at higher shot counts while GPT-4o-mini remains flat. These results demonstrate that current LLMs are unable to represent polysynthetic morphology and instead default to exoticizing Indigenous culture and identity. We call for the development of inclusive morphological-aware modeling strategies and increased resource creation to ensure that Indigenous languages of the Americas are represented safely and accurately.

1 Introduction

Large language models (LLMs) have expanded rapidly, producing high-quality machine translation (MT) systems for many high-resource languages (Zhu et al., 2024). LLMs trained on high-resource data underperform on low-resource languages, and endangered low-resource indigenous languages are excluded from this progress (Pucinskaite and Mitkov, 2025; Sindhujan et al., 2025). Improving LLM support and accuracy of indigenous languages is vital for access to language technologies and for indigenous cultural and linguistic preservation (Coleman et al., 2024).

Wixárika or Huichol is a Uto-Aztecan language spoken primarily in the Sierra Madre Occidental mountains of western Mexico, across the states of Jalisco, Nayarit, Durango, and Zacatecas. Wixárika has approximately 50,000 speakers (Leza and López, 2006) and is currently classified as vulnerable by UNESCO (Moseley, 2010). The language presents many challenges for NLP as it is a polysynthetic language where single words can contain the information of multiple clauses in a fusional language like Spanish (Mager et al., 2019). This complexity, combined with the severe lack of resources for Wixárika, makes it an important test case for current language technologies for indigenous languages (Mager et al., 2018b).

Despite growing interest in NLP for indigenous languages of the Americas, Wixárika has received little progress and attention in MT outside of AmericasNLP shared tasks (De Gibert et al., 2025; Ebrahimi et al., 2024, 2023). Fine-tuned models have achieved chrF++ scores of up to 28 (Ebrahimi et al., 2024) in recent tasks. LLMs have been shown to struggle with polysynthetic and morphologically complex indigenous languages of the Americas, including Southern Quechua (Court and Elsner, 2024; Stap and Araabi, 2023). Prior work has not documented and analyzed cultural biases in LLM-based indigenous language MT.

While this work has not been performed, it is extremely necessary. This work makes three main contributions: (1) an evaluation of LLMs on Wixárika MT, (2) a comparison of dedicated multilingual MT with LLMs on Wixárika, and (3) a qualitative analysis of model outputs, failures, and biases. We aim to reduce biases and improve the inclusion of Wixárika and polysynthetic languages in NLP.

2 Data

We use the Wixárika–Spanish parallel corpus[1] (Mager et al., 2018a), which consists of human

[1] https://github.com/pywirrarika/wixarikacorpora

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 74–81

Statistic	Wixárika	Spanish
Sentence pairs	8,966	
Total tokens	48,896	68,569
Vocabulary size	17,386	11,678
Avg. sentence length	5.5	7.6
Type-Token Ratio	0.356	0.170

Table 1: Statistics of the Wixárika–Spanish parallel corpus

translations of many Grimm and Andersen stories from Spanish to Wixárika. We first present corpus statistics (Table 1).

Wixárika has a significantly higher Type-Token Ratio (TTR) of 0.356 compared to 0.170 in Spanish. This is a consequence of polysynthesis, as inflectional and derivational morphology create a large number of distinct word forms. This has direct implications for LLM performance, as models are less likely to have encountered Wixárika word forms during pretraining due to their polysynthetic nature, which is compounded by the limited language resources for Wixárika.

We randomly sample 100 sentence pairs as a test set, and we reserve 50 additional pairs as a pool from which 5 examples are drawn for the few-shot prompting. We use the random seed `6` for all sampling.

3 Experimental Setup

3.1 Models

We evaluate three large language models (LLMs) which are GPT-4o-mini (`openai/gpt-4o-mini`) (OpenAI et al., 2024), Gemma 3 27B (`google/gemma-3-27b-it`) (Team et al., 2024), and Llama 3.3 70B Instruct (`meta-llama/llama-3.3-70b-instruct`) (Grattafiori et al., 2024). We selected these models to represent a diverse set of large language models that are publicly accessible, computationally efficient, and in two cases, open-weight. We additionally evaluate NLLB-200 (600M distilled) (Team et al., 2022), a dedicated multilingual MT system on zero-shot Wixárika (`hch_Latn`) translation, as a non-LLM baseline.

3.2 Prompting

We evaluate in two conditions for the LLMs. First, we evaluate in a zero-shot setting where the prompt identifies and briefly describes Wixárika and instructs the model to translate the text. Second, we evaluate in 5-shot, 10-shot, and 25-shot settings with the same system prompt with five, ten, or twenty-five parallel examples drawn from the few-shot pool. We evaluate NLLB-200 in zero-shot only. We release the prompts used in Appendix A.

Model	Dir	0-BLEU	0-chrF	5-BLEU	5-chrF
NLLB-200 (600M)	ES→HCH	0.64	10.19	–	–
	HCH→ES	1.16	13.44	–	–
GPT-4o-mini	ES→HCH	0.47	10.64	1.27	14.38
	HCH→ES	0.72	14.52	2.18	16.25
Gemma-3-27B	ES→HCH	0.11	11.99	**2.45**	13.80
	HCH→ES	1.58	16.75	**2.29**	17.29
Llama-3.3-70B	ES→HCH	0.41	**15.66**	1.94	**16.03**
	HCH→ES	0.57	17.01	1.28	**17.69**

Table 2: BLEU and chrF for Spanish↔Wixárika translation

3.3 Metrics

We report BLEU (Papineni et al., 2002) and chrF (Popović, 2015), which were both computed using the `sacrebleu` Python package (Post, 2018). We set `effective_order` to `True` in BLEU, and all other settings for BLEU and chrF remained as their default. chrF is considered more informative in the case of Wixárika as it is polysynthetic (Zheng et al., 2021). We additionally report BERTScore (Zhang et al., 2020) F1 using `bert-base-multilingual-cased` on Wixárika→Spanish outputs to quantify the semantic distance of hallucinated outputs from reference translations.

4 Results

4.1 Overall Results

We first look at the overall results for all models in the zero-shot and five-shot settings (Table 2). All systems score below 3 for BLEU and below 21 for chrF in both directions, confirming that LLMs cannot usefully translate between Spanish and Wixárika. While Gemma-3-27B achieves the highest BLEU scores on both directions of 2.45 for Spanish to Wixárika and 2.29 for Wixárika to Spanish, Llama 3.3 70B achieves the highest chrF scores for Wixárika to Spanish and Spanish to Wixárika. We observe a gap in chrF between the Spanish to Wixárika and the Wixárika to Spanish results in all models in the zero-shot setting, with all models having the highest chrF in Wixárika to Spanish translation. Gemma and NLLB exhibit the largest gaps, with Gemma showing a gap of 4.76 points.

We next look into the differences between zero-shot and few-shot performance in all LLMs. Five-

Model	Shot	P	R	F1
NLLB-200 (600M)	zero	0.608	0.652	0.629
GPT-4o-mini	zero	0.708	0.690	0.699
	five	0.718	0.710	0.714
Gemma-3-27B	zero	0.688	0.685	0.687
	five	0.710	0.703	0.706
Llama-3.3-70B	zero	0.695	0.691	0.693
	five	0.722	0.720	0.720

Table 3: BERTScore for Wixárika→Spanish outputs

Model	Dir	5-shot	10-shot	25-shot
GPT-4o-mini	ES→HCH	14.38	14.03	15.55
	HCH→ES	16.25	15.25	16.38
Gemma-3-27B	ES→HCH	13.80	15.51	19.08
	HCH→ES	17.29	17.79	**18.52**
Llama-3.3-70B	ES→HCH	16.03	18.83	**20.86**
	HCH→ES	17.69	18.21	18.12

Table 4: chrF at 5, 10, and 25 shots in all three LLMs

shot prompting yields increases in BLEU and chrF across LLMs. In Gemma 3 27B, BLEU increased by 2.34 points, and chrF increased by 2.80 points in Spanish to Wixárika translation, while in Wixárika to Spanish translation, BLEU increased by 0.71 and chrF increased by 0.54. Overall, few-shot prompting increased MT performance in LLMs for Wixárika consistently, but performance was still not usable.

Viewing BERTScore outputs for Spanish, we can see that LLMs score 0.69-0.72 F1 on Spanish references, with NLLB scores only 0.63, indicating that LLM outputs are fluent Spanish but distant from the source, while NLLB outputs are not fluent or accurate.

Looking at results for 5-shot, 10-shot, and 25-shot MT (Table 4) we are able to see that a larger number of shots yields improvements depending on the model. In Spanish-to-Wixárika translation, Gemma and Llama improve substantially with more examples. Using 25 examples, Gemma reaches 19.08, or a 7.09 point improvement over zero-shot translation, and Llama reaches 20.86 chrF, or 5.20 points over zero-shot. GPT-4o-mini remains fairly flat across all numbers of shots. For the Wixárika to Spanish translation, improvements are smaller. Despite these gains with a larger number of shots, all systems are below usability.

4.2 Error Analysis

We now zoom in on example translations and provide a qualitative analysis of these results to better understand the nature of system failures (Tables 5-6). In the Wixárika-to-Spanish direction (Table 5), LLMs' output is almost completely detached from the source. Instead of translating the sentences, the models generated unrelated Spanish sentences based on harmful cultural stereotypes. NLLB-200 frequently outputs partially or fully untranslated strings, reflecting a different failure mode compared to the LLMs. We further investigate LLM results. In Llama 3.3 70B Instruct 19% of zero-shot Wixárika-to-Spanish translations mention deer in the output (63% of which mention the deer eating maize), 33% of outputs mention gods, spirits, or shaman, 32% of outputs mention singing or dancing, and 14% of outputs mention the sun. In GPT-4o-mini results, 29% of outputs mention the sun shining. In Gemma 3 27B, 40% of outputs mention deer, and 12% of results mention peyote. Each of the three LLMs focuses on separate stereotypes to different extents. Some outputs, such as "El chamán va a hacer una ceremonia para que llueva en el río," (The shaman is going to perform a ceremony to make it rain in the river), are harmful stereotypes towards indigenous peoples. This particular output reduces complex traditions to a rain dance, and it also uses exotic-sounding words such as raining in the river, despite the ground truth being unrelated to this.

In the Spanish-to-Wixárika direction (Table 6), LLM errors were more pronounced. Errors often include repeated words. In GPT-4o-mini, many outputs included "naka" and "jíkuri," not corresponding to the Grimm story content whatsoever. Gemma produced similarly short, uniform outputs, and Llama, which achieved the highest zero-shot chrF, produced more diverse outputs, but these outputs failed to reflect the source text. NLLB produced primarily English and Spanish outputs. Across both directions and all systems, there is no evidence of correct morphological structure in Wixárika outputs consistent with the practical impossibility of learning its complex morphology with limited pretraining data.

5 Discussion

We evaluated NLLB-200, GPT-4o-mini, Gemma 3 27B, and Llama 3.3 70B Instruct for Wixárika↔Spanish MT without fine-tuning. We found that no evaluated system can usefully translate between Spanish and Wixárika. All systems scored below 3 BLEU and below 21 chrF. LLMs

#	Ground truth (es)	gemma-3-27b-it	gpt-4o-mini	llama-3.3-70b-instruct	nllb-200-600M
1	mientras que la refinada doncella se caerá y se...	Nosotros estamos viendo el agua.	el sol brilla hoy,	El venado está comiendo maíz.	Mexi tsimu'ane +imari kanaka wem+k+ meta kanita...
2	Al lado opuesto del lago se presentaba	Nosotros vemos el venado.	El sol brilla.	El venado se comió la mitad de la milpa de mi hermano.	+ta+ye hapa pata an'ukai
3	Pero Hansel la confortaba diciéndole:	Está lloviendo mucho.	te veo, estoy feliz de verte.	Cuando yo estaba en el camino, vi a mi abuela que...	Por el mango peik+n+t+wakai reik+h+awet+:
4	-"Vamos a la casa del señor Korbes."- replicó el gallo.	Nosotros estamos yendo a la iglesia ahora.	el barril está lleno de agua.	El venado está comiendo maíz.	"tepa+k+ kie uki korbes."-petay+ ukiri.
5	el gato se puso sus botas galantemente,	¡Qué bien cantas, perezosa, así me gusta!	El sol está saliendo.	El venado está bailando.	Mitsu paukex+a yu xapatuxitsie temaik+t+,

Table 5: Example Wixárika to Spanish translations (zero-shot) and ground truth Spanish sentences

#	Ground truth (hch)	gemma-3-27b-it	gpt-4o-mini	llama-3.3-70b-instruct	nllb-200-600M
1	mexi tsimu'ane +imari kanaka wem+k+ meta kanita...	múixie útsuri xíki marie.	Naka, ka nika jíkuri uñi,	Tukari neixa piku teku.	while the refined soella will fall and break. "-
2	an+ta+ye hapa pata'ukai	múe'u úni tsitsímá.	naka xukari, nika xukari:	ne tsikíri teitei niwaxi ne tsikíri:	On the opposite side of the lake was present
3	per+ta hansel peik+n+t+wakai reik+ h+awet+:	múiní nié úruami, ma tsámie.	Naka jíkuri wíka jiyá:	Tee kaa xapari puyeeka tsikári ne ra'utu.	Hansel however comforted her by saying:
4	"tepa+k+ kie uki korbes."-petay+ ukiri.	Muéka, kieri miea ukuri tsaméka.	naka jíkuri, ta jíkuri wíri,	Ne tsikári te+ku ne wixarika tewei ya+ne:	"Let's go to the house of Monsieur Korbes". replied...
5	mitsu paukex+a yu xapatuxitsie temaik+t+,	múixie úhie mie útsïri tsïriákei.	naka jíkuri neiká.	teku teweku paa tsikiri,	El gato puso sus botas con galancia,

Table 6: Example Spanish to Wixárika translations (zero-shot) and ground truth Wixárika sentences

produced hallucinated outputs disconnected from source content, producing culturally biased outputs consistently focused on nature and religion. NLLB-200 generated untranslated outputs. These findings extend the observation that even modern, large-scale systems are unable to support low-resource polysynthetic languages when language data are scarce and morphologically complex (Yahan and Islam, 2025).

Our results align with previous work demonstrating that the LLM translation performance degrades severely for low-resource languages (Lin et al., 2025; Ghazvininejad et al., 2023). LLMs are vulnerable when generating low-resource languages as they lack the representations needed to generate morphologically complex text (Anh et al., 2024). Wixárika's high TTR demonstrates this, displaying that polysynthesis produces a large increase in word forms, decreasing the likelihood that a form has been observed during pretraining.

The LLM outputs in the Wixárika-to-Spanish direction did not produce nonsensical strings, but plausible but unrelated Spanish sentences. Frequently, the same models output near-identical sentences despite diverse input sentences. This aligns with Southern Quechua, where LLMs have been observed to produce stereotypical outputs, and this aligns with findings that LLMs amplify societal biases related to indigeneity (Court and Elsner, 2024; Delgado and Toxtli, 2023). These findings raise a concern that hallucinated translations may misrepresent indigenous languages and their speakers in potentially harmful ways, with 33% of outputs mentioning gods, spirits, or shamans, 32% of outputs mentioning singing or dancing, and 14% of outputs mentioning the sun in Llama 3.3 70B Instruct. Previous work has shown that multilingual translation models are prone to hallucinations (Guerreiro et al.,

2023), which our results confirm. We also found that few-shot prompting improved results variably in each model. This may reflect that models are unable to extract useful morphological patterns from only five examples of a polysynthetic language (Anh et al., 2024), consistent with findings that few-shot learning is difficult for morphologically complex languages (Ismayilzada et al., 2025). The hallucinated LLM outputs are not generally exotic. They are specifically related to Wixárikan culture, suggesting that models draw on cultural associations from their training data rather than the domain of the text. Deer, corn, and peyote are three important and sacred species in Wixárikan culture. This means that mentions of deer and peyote are not random, but LLM biases that insert cultural stereotypes and elements into MT outputs despite no mention of this information in the source text.

Future work should investigate retrieval-augmented generation (RAG) approaches that inject morphological information, dictionary entries, or grammar descriptions into prompts, which have shown benefits in similar low-resource languages (Chang et al., 2025; Coleman et al., 2024). Although fine-tuned systems for Wixárika have achieved much higher usable scores in shared tasks (Ebrahimi et al., 2024; De Gibert et al., 2025), they have not been systematically evaluated for the types of stereotypical hallucinations documented in this study. Future work should extend fine-tuning approaches to explicitly mitigate cultural bias and exoticism. Synthetic data approaches should be worked on for Wixárika, given its low-resource status (de Gibert et al., 2025). Finally, human evaluation is needed as automatic metrics are poorly suited for polysynthetic languages (Kumar et al., 2026).

Our findings provide a baseline for zero-shot and few-shot LLM-based Spanish↔Wixárika MT, confirm that Wixárika remains out of reach of LLMs, and find that current LLM-based Wixárika translation commonly produces culturally insensitive and harmful hallucinations. This reflects an absence of fundamental data, tools, and computational approaches for Wixárika NLP needed to support speakers. Enabling Wixárika support in MT and LLMs requires significant technical advances in low-resource NLP, but the increase in resources, tools, and evaluation standards that Wixárika currently lacks.

6 Conclusion

This study evaluates GPT-4o-mini, Gemma 3 27B, Llama 3.3 70B Instruct, and NLLB-200 on Spanish↔Wixárika machine translation in zero-shot, 5-shot, 10-shot, and 25-shot settings. We qualitatively analyze model outputs and failures in LLMs.

We find that all systems are unsuccessful in both directions, scoring below 3 BLEU and below 21 chrF, with qualitative analysis revealing severe hallucinations, detachment from the source text, and an inability to generate morphologically plausible Wixárika forms. Spanish outputs commonly reflect indigenous stereotypes, relating to nature, dancing, and gods, despite the diverse ground truth text. Wixárika outputs do not reflect Wixárika and instead commonly repeat the same words across distinct translation tasks. Few-shot prompting yields improvements, although models are still unusable.

These results display that LLMs are unable to support Wixárika, as a low-resource polysynthetic language, and instead output unrelated cultural stereotypes. We hope this work will serve as a baseline for future research and a call for resource-building efforts and capable models for Wixárika and other indigenous polysynthetic languages.

Limitations

Several limitations should be considered in this study. First, our test set consists of 100 sentence pairs drawn from a single domain, which may not be representative of other domains of Wixárika or Spanish. Second, we evaluate only three LLMs and one dedicated MT system. Other models, such as larger proprietary systems or models fine-tuned on Uto-Aztecan languages, may perform differently. Third, we evaluate LLMs using only simple zero-shot, 5-shot, 10-shot, and 25-shot prompting. More sophisticated strategies, such as retrieval-augmented generation or chain-of-thought prompting, are not explored here. Finally, our evaluation relies exclusively on automatic metrics (BLEU and chrF), which are known to be poorly calibrated for polysynthetic languages (Zheng et al., 2021) and cannot assess meaning preservation, fluency, or grammatical adequacy in the way that human evaluation can.

Ethics

Ethical considerations are vital in the development of language technologies for indigenous languages.

This work uses a publicly available Wixárika-Spanish corpus (Mager et al., 2018a). We seek to contribute to the accessibility and visibility of Wixárika and other indigenous languages of the Americas within computational linguistics and emphasize the historical and cultural significance of these languages. We acknowledge that Wixárika is a vulnerable language, and work must support rather than extract from revitalization efforts. Technology developed without community consultation risks causing harm, including by producing inaccurate and stereotypical information. We document this extreme failure in the current model rather than propose these models for use. We encourage future work to involve native speakers and community leaders in resource construction, error analysis, and evaluation standards to ensure advances serve their preservation goals.

References

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Chen-Chi Chang, Chong-Fu Li, Chu-Hsuan Lee, and Hung-Shin Lee. 2025. Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances. *Preprint*, arXiv:2505.10829.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Cecilia Delgado and Carlos Toxtli. 2023. Evaluating machine perception of indigeneity: An analysis of chatgpt's perceptions of indigenous roles in diverse scenarios.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2026. Evaluating extremely low-resource machine translation: A comparative study of chrf++ and bleu metrics. *Preprint*, arXiv:2602.17425.

José Luis Iturrioz Leza and Paula Gómez López. 2006. *Gramática wixarika*, volume 1. Lincom Europa.

Kaiying Kevin Lin, Hsi-Yu Chen, and Haopeng Zhang. 2025. FormosanBench: Benchmarking low-resource Austronesian languages in the era of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16527–16539, Suzhou, China. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2005–2011, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018b. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Julita JP Pucinskaite and Ruslan Mitkov. 2025. Evaluating the LLM and NMT models in translating low-resourced languages. In *Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models*, pages 123–133, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Mahshar Yahan and Dr. Mohammad Islam. 2025. Leveraging large language models for Spanish-indigenous language machine translation at AmericasNLP 2025. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 126–133, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A LLM Prompts

We use the following four prompts for all LLM evaluations:

Zero-Shot Spanish to Wixárika

> **System:** You are a translator. Wixárika (also called Huichol) is an indigenous Uto-Aztecan language of Mexico with polysynthetic morphology. Translate the following Spanish sentence into Wixárika. Output ONLY the translation, no explanation.
>
> **User:** {Spanish sentence}

Zero-Shot Wixárika Spanish

> **System:** You are a translator. Wixárika (also called Huichol) is an indigenous Uto-Aztecan language of Mexico with polysynthetic morphology. Translate the following Wixárika sentence into Spanish. Output ONLY the translation, no explanation.
>
> **User:** {Wixárika sentence}

Few-Shot Spanish to Wixárika

> **System:** You are a translator. Wixárika (also called Huichol) is an indigenous Uto-Aztecan language of Mexico with polysynthetic morphology. Below are some example translations.
>
> **Examples:**
> Spanish: $\{s_1\}$ Wixárika: $\{t_1\}$
> Spanish: $\{s_2\}$ Wixárika: $\{t_2\}$
> ⋮
> Spanish: $\{s_n\}$ Wixárika: $\{t_n\}$
>
> Now translate the following Spanish sentence into Wixárika. Output ONLY the translation, no explanation.
>
> **User:** {Spanish sentence}

Few-Shot Wixárika to Spanish

> **System:** You are a translator. Wixárika (also called Huichol) is an indigenous Uto-Aztecan language of Mexico with polysynthetic morphology. Below are some example translations.
>
> **Examples:**
> Wixárika: $\{t_1\}$ Spanish: $\{s_1\}$
> Wixárika: $\{t_2\}$ Spanish: $\{s_2\}$
> ⋮
> Wixárika: $\{t_n\}$ Spanish: $\{s_n\}$
>
> Now translate the following Wixárika sentence into Spanish. Output ONLY the translation, no explanation.
>
> **User:** {Wixárika sentence}

IndigiEval: Evaluating LLMs in North American Indigenous Languages

Julia Mainzinger
Charles Darwin University
Darwin, Australia

Jacqueline Brixey
University of Wisconsin - Madison
Madison, WI, USA

Abstract

This paper presents IndigiEval, a framework for evaluating the language and cultural proficiency of several commercially available large language models (LLMs) across five North American Indigenous languages (Mvskoke, Choctaw, Cherokee, Cheyenne, and Hawaiian). This framework is a qualitative evaluation method intended for communities with small speaker populations to be able to critically evaluate LLM performance with minimal data and human effort. IndigiEval includes tasks such as answering cultural questions, translation, text generation, and speech recognition. The results of our experiments indicate that no currently available LLM performs well across all evaluation categories, and that LLMs frequently hallucinate orthographies, grammatical structures, cultural knowledge, and vocabulary for all languages and cultures considered. Our proposed evaluation framework is not intended as a comprehensive score, but rather a qualitative and flexible framework to inform language communities about a given LLM's potential as a resource, since each language has unique environments, strengths, and availability of resources.

1 Introduction

With the increasing capabilities of large language models (LLMs) in some low-resource languages, there has been growing interest in how AI, and specifically LLMs might support language revitalization efforts (Moshagen et al., 2024). In the North American context, limited access to first-language speakers leads people to turn to technology to fill gaps (Meighan, 2024).

Although LLMs demonstrate impressive performance in English and other majority languages, the same is not true for minority, Indigenous, and endangered languages (Choudhury, 2023; Stap and Araabi, 2023). Even in English, LLMs are known to make things up, providing both truth and fallacies without qualification (Hicks et al., 2024). These hallucinations and errors are magnified in languages that are underrepresented in the training data (Song et al., 2026a).

Concerns about accuracy are not unique to endangered languages; similar risks also exist in other high-stakes domains where accuracy is essential (Elliott et al., 2025). This includes domains such as law and medicine (Cheong et al., 2024; Quttainah et al., 2024; Singhal et al., 2025), where incorrect information can have profound real-world consequences. Accuracy in teaching language to second-language learners is likewise essential, as learners have no way to judge when mistakes occur, making them vulnerable to learning errors permanently.

Recent scholarship calls for a re-centering of AI research on Indigenous practices, including knowledge-making and storytelling (Lewis et al., 2024), as well as inter-generational language transmission and language learning applications (Hinton, 2013; Pitawanakwat, 2018; Neubig et al., 2020). There is also concern that people will turn to AI over local knowledge holders – the "Ask your Auntie, Not AI" campaign[1] has been created to challenge the idea that AI can replace elders.

Despite these concerns, limited work has been conducted on how to systematically evaluate LLM performance in truly low-resource Indigenous language contexts. Existing benchmarking approaches typically rely on large-scale datasets, such as thousands of evaluation questions or extensive human feedback (Vayani et al., 2025; Myung et al., 2024; Zhang et al., 2023a). For small speaker communities, generating such datasets can be prohibitively labor- and resource-intensive (Wiechetek et al., 2024; Pitawanakwat,

[1] https://www.honorearth.org/

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 82–94

2018). Moreover, given the limited demonstrated benefits of LLMs in low-resource settings, these forms of evaluation may be perceived as extractive, particularly when they require substantial contributions from fluent speakers without clear reciprocity.

For these reasons, we see value in demonstrating a small-scale evaluation that serve as a model for how communities might assess the limitations of LLMs and make informed decisions about whether and how these technologies should be used. In this work, we propose IndigiEval, a framework for exploring language and cultural capabilities and limitations of LLMs in North American Indigenous contexts. Toward this end, we design our framework according to the following principles:

- **Small-scale:** Fits within resource constraints without requiring large-scale datasets
- **Non-extractive:** Minimizes human annotation demands in order to reduce burden on small speaker communities
- **Accessible:** Evaluations can be carried out by community members or those familiar with the language situation without necessarily requiring fluent speakers

As illustrated in Figure 1, we evaluate LLMs across four categories: cultural knowledge, translation, text generation, and speech recognition. We choose tasks that leverage the types of materials commonly available in Indigenous language documentation and revitalization work. However, we also recognize that each community has its own strengths and priorities; as such, this framework is not one-size-fits-all, but can be adapted by other communities to assess whether LLM-based tools align with local needs (Zhang et al., 2022; Liu et al., 2022).

The authors of this work are citizens of the Muscogee Nation (first author) and the Choctaw Nation of Oklahoma (second author). Julia Mainzinger collaborates with the College of the Muscogee Nation on language revitalization projects. Jacqueline Brixey is a learner of the Choctaw language and has built language technology that supports revitalization efforts.

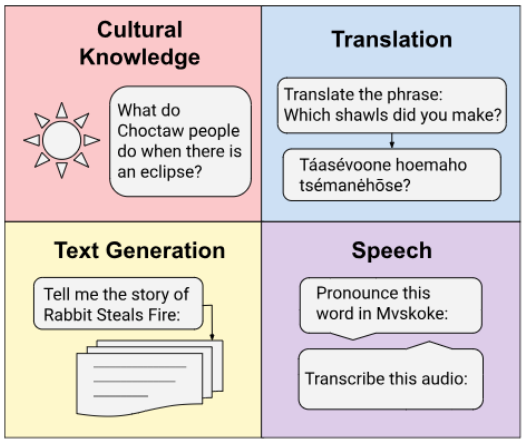

Figure 1: Evaluation categories of IndigiEval

2 Related Work

Several prior benchmarks assess LLM performance for lower-resource languages. Many of these benchmarks use questions from standardized exams, such as grade-school exams (Zhang et al., 2023a) or national language institutes (Song et al., 2026b). Results vary by language and task: one study has found that LLMs could not pass any exam beyond a primary school level in Indonesian (Koto et al., 2023), while another has shown how models can competently handle understanding but not generation tasks at the high school level in Latvian (Darģis et al., 2024).

Benchmark tasks to assess language proficiency often include vocabulary tests (Song et al., 2026b), short text generation (Chaka, 2024), and translation both to and from a given language (Zhang et al., 2023b). Task formats from previous work often favor multiple-choice formats (Darģis et al., 2024; Myung et al., 2024) and open-ended questions (Chaka, 2024), with only a few benchmarks utilizing true/false question formats (Vayani et al., 2025).

The benchmarks All Languages Matter and BLEnD probe LLMs for cultural competency and language proficiency in a diverse array of languages (Myung et al., 2024; Vayani et al., 2025). Other studies target specific contexts, such as FilBench, which evaluates LLM cultural knowledge and language proficiency in Filipino languages (Miranda et al., 2025), and ProverbEval, which tests LLMs for knowledge of proverbs in five African languages (Azime et al., 2025).

Many Indigenous American languages lack standardized tests or large collections of question/answer pairs that could be used in this man-

Language	ISO-3	Speakers	WKP	CC
Mvskoke	mus	300	0	0
Choctaw	cho	1,000	0	0
Cheyenne	chy	344	721	0
Cherokee	chr	2,000	1,003	1,025
Hawaiian	haw	7,000	2,968	6,343

Table 1: Languages included in this study, with speaker count, Wikipedia (WKP) article count, and Common Crawl (CC) count.

ner, and many of our communities lack the manpower and resources to carry out large-scale human evaluation. Our work therefore differs from previous work in that we present a small-scale qualitative evaluation framework. This framework provides representative tasks that could be feasible with small amounts of data, serving as an example of how LLM evaluation might be performed without burdening small speaker populations.

3 Languages

As we are most familiar with Choctaw (cho) and Mvskoke (mus), we evaluate LLMs across all categories for these two languages. We include three additional languages—Cheyenne (chy), Cherokee (chr), and Hawaiian (haw)—in the language proficiency section, as we can reference expertise in language documentation resources to assess responses. We selected the languages because they have similar speaker counts and representation in public datasets (Table 1). All five languages considered in our experiments are indigenous to the present-day United States, and all are endangered.

Choctaw. The Choctaw language is spoken by the Choctaws, who are the third most populous US tribal group, with approximately 223,000 people identifying as Choctaw[2]. However, the language has endangered status (Simons and Fennig, 2018), and it is estimated that there are fewer than 1,000 fluent speakers in the Choctaw Nation of Oklahoma (Rogers, 2021).

The Choctaw language is part of the Muskogean language family (Haas, 1979), and has subject-object-verb word order. The language is relatively well-documented, with numerous grammars (Byington, 1870), dictionaries (The Choctaw Nation of Oklahoma Dictionary Committee, 2016; of Choctaw Indians, 2026), and printed learning materials. We refer to these materials for our experiments.

Mvskoke. The Mvskoke (also known as Muscogee or Creek) language is spoken by members of the Muscogee and Seminole tribes. It is estimated that fewer than 300 first-language speakers remain, and nearly all are over the age of 60[3].

The language is synthetic and agglutinative, with a traditional orthography of 20 Latin letters. The orthography is relatively transparent and allows for spelling variations (Martin, 2011). We reference the dictionary by Martin and Mauldin (2000) and a collection of texts by Haas et al. (2015).

Cheyenne. Cheyenne is an Algonquian language, spoken by the Cheyenne people, a Great Plains Tribe (Leiker and Powers, 2011). There are two federally recognized tribes, now located in Montana and Oklahoma[4]. The language is highly agglutinative and polysynthetic (Badhorse, 2023). There are approximately 300 fluent speakers today (Littlebear, 2024). We draw vocabulary from Fisher et al. (2023) and sentences from Leman (1980).

Cherokee. The Cherokee people originate from the Southeastern United States, from North Carolina to Georgia (Justice, 2025). They were removed largely to Oklahoma, with one federally recognized tribe remaining in North Carolina [5]. The language is Iroquoian (Julian, 2010), with a unique 85-character syllabary[6]. There are an estimated 2,000 first-language Cherokee speakers (Zhang et al., 2022). We reference vocabulary from a collection of Cherokee dictionaries[7], and collect sentences from Howard and Eby (1990).

Hawaiian. Hawaiian is spoken by people indigenous to the Hawaiian Islands. It is in the Eastern Polynesian branch of the Austronesian language family (Parker Jones, 2018). Although there was no large-scale removal from ancestral

[2] https://archive.ncai.org/tribal-vawa/sdvcj-today/the-choctaw-nation-in-oklahoma

[3] This estimate is from personal communication with a member of Ekvn-Yefolecv, a community of Mvskoke people.

[4] http://www.cheyennenation.com and https://www.bia.gov/regional-offices/southern-plains/concho-agency

[5] https://www.doi.gov/tribes/cherokee

[6] https://georgiahistory.com/wp-content/uploads/2023/11/NIE-2017-web.pdf

[7] https://www.cherokeedictionary.net/about

lands as with many US continental tribes, linguistic loss occurred due to 'English-only' language ideology, social injustice, and population decline from disease (Brenzinger and Heinrich, 2013). Revitalization of the language began in the 1970s, and through school and immersion programs, there are now around 7,000 fluent speakers, including new first-language speakers (Brenzinger and Heinrich, 2013). We reference vocabulary from Pukui and Elbert (2003) and Leo and Kuamo o (2003), and sentences from Bardwell et al. (2020).

4 Evaluation categories

The tasks included in this framework are not meant to be comprehensive or exclusive. Rather, they are intended to demonstrate gaps in language and cultural competency of the LLMs. Failures observed at a small scale can indicate limitations in the ability to support community language needs. We present one task that evaluates cultural knowledge, and three tasks that evaluate language proficiency.

4.1 Cultural Knowledge

Studies have found that LLMs often exhibit Western biases (Myung et al., 2024). These types of biases can harmfully reproduce stereotypes towards Indigenous people with cultures distinct from Western ones (Hanson, 2025), or propagate misinformation about Indigenous communities.

To evaluate LLMs' cultural knowledge, we developed 10 multiple-choice questions in English about Mvskoke and Choctaw culture. These questions assess knowledge and values, including material knowledge, historical figures, events, and traditional food. Each multiple-choice question has four choices presented with only one correct answer. We determined the answer to these questions based on our own knowledge or by referencing published sources. We did not pose any questions that would be unacceptable to share with those outside of our communities. As both of our communities are fairly open to outsiders and many of the cultural details asked about are documented online, we hypothesized that the LLMs would be able to correctly answer many of these questions.

4.2 Language Proficiency

The language proficiency category comprises machine translation, vocabulary questions, text generation, and speech tasks.

4.2.1 Machine Translation

Previous research has shown that LLM-based machine translation for low-resource languages does not outperform traditional neural machine translation (Ebrahimi et al., 2022; Stap and Araabi, 2023; Robinson et al., 2023). Because almost all commercial LLMs do not publicly share their datasets, it is unclear exactly how much language-specific data is contained in a given model. However, performance is often directly correlated with the number of Wikipedia pages in a given language (Robinson et al., 2023). Additionally, for languages known to have low representation in the LLM's training data, the output is almost entirely nonsensical (Zhang et al., 2022). Table 1 provides the resource counts in the evaluation languages. While Mvskoke and Choctaw do not have any pages entirely in these languages on Wikipedia, we have found that many LLMs nevertheless attempt to translate to and from them.

In this task, we selected 10 sentences for translation in each of the five languages. The ten sentences display a range of syntactical structures and vocabulary. We utilize chrF++ as our scoring system.

It should be noted that we are not suggesting that LLMs should be used for Indigenous language translation over traditional machine translation methods. Here, we use translation as a task because it has a well-established scoring system, and we can utilize language documentation for high-quality references. This provides a more objective measure of language proficiency. We also anticipate that many language learners might ask for translation assistance from an LLM.

4.2.2 Vocabulary

In this task, we prompt the LLMs to translate single English terms into each language. While this is a sub-task of MT, prompting for isolated words enables us to assess specific vocabulary, such as modern vocabulary that may not be formalized in digitized corpora. Additionally, it helps us to identify hallucinated words more easily.

We translated 50 English words into each language by referencing the relevant dictionaries. We also consider several additional words that may not have a formal translation, such as "microwave" and "yogurt". Some of these words have locally-defined terms but are not present in

dictionaries. We do not include these as part of the scoring, but rather to observe the behavior when a model is asked for a word that may not exist in the language.

We prompt each model for a translation of the given English word. During scoring, we also reference these same dictionaries, allowing for synonyms or closely related terms to be marked correct.

4.2.3 Text Generation

Prompting for open-ended text generation allows us to consider the capabilities of the language model of the LLM by evaluating the sensibility, errors, and hallucination of the generated text (Chaka, 2024). In this task, we prompt the LLMs to tell a traditional story in Mvskoke and Choctaw and then compare their responses with published versions. The published versions are part of language documentation that has been written or reviewed by fluent speakers. We examine the overall fluency and contrast the outputs with the language documentation. We only prompt for one story for each model per language, as the stories are several paragraphs long, and the long-form generation gives a thorough sense of the models' command of the language.

4.2.4 Speech

Speech technologies can be meaningful tools for low-resource languages. Automatic speech recognition (ASR) is a valuable technology, as it can be used for automatic captioning, voice typing, and improving transcription efficiency (Ćavar et al., 2016). Additionally, in the context of North American Indigenous languages, orthographies are often varied, and standardization may not be consistently adopted. As a result, text-to-speech (TTS) can be helpful both for language learners and for overcoming challenges in writing systems. To this end, we include **ASR** and **TTS** tasks in our evaluation framework.

5 Setup

5.1 Prompting

To test our framework, we developed a Python script for each evaluation category that makes API calls to each of the LLMs. Each API call includes an instruction and a prompt. The instruction gives the LLM context for the prompt, such as "You are a helpful assistant knowledgeable in the [target] language." This is included along with the prompt for each task. The prompts are zero-shot with no web search functionality.

5.2 Models

For the text-based tasks, we tested large commercially available LLMs with the highest performance at the time of testing: OpenAI's GPT gpt-5.2-2025-12-11 (Singh et al., 2025), DeepSeek deepseek-reasoner (Guo et al., 2025), Anthropic's Claude claude-opus-4-6 (Anthropic, 2026), and Google's Gemini gemini-3.1-pro-preview (Team et al., 2025). For speech tasks, we tested gemini-3.1-pro-preview and gpt-4o-transcribe[8]. We leave it to future work to test other LLMs, including open source models.

6 Results

6.1 Cultural Knowledge

All of the LLMs tested are largely able to answer basic questions about popular sports, foods, and traditional clothing in the Mvskoke and Choctaw cultures. The results are in Table 2. GPT-5.2 and DeepSeek make more mistakes overall than Claude and Gemini, missing questions such as "How did clans get their names in Mvskoke culture?" and "What is the name of a traditional Choctaw dance?"

Interestingly, Claude and Gemini can correctly answer detailed and specific questions about stories documented in the books Haas et al. (2015) and Gouge et al. (2004), showing that these resources, which are two of the most extensive collections of written Mvskoke language documentation, are likely contained in the training data for these two models.

The most-missed question, which all four models missed, is "In the Dawes Commission short film by Bob Hicks, what does the grandmother tell the little girl she used to use for dancing?" This film is publicly available on YouTube, and the majority of it is in Mvskoke with English subtitles. The second most common mistake was a tie for a question about a Mvskoke historical figure, and a question about when the Okla Chahta Clan of California holds its annual gathering. These mistakes demonstrate that despite the vast training data, these LLMs are not infallible sources of cultural information.

[8]https://developers.openai.com/api/docs/models/gpt-4o-transcribe

	GPT	DS	Claude	Gemini
mus	4	5	9	8
cho	9	8	9	9
Total	13	13	18	17
	65%	65%	90%	85%

Table 2: Number and percent of correct answers on 20 multiple-choice cultural knowledge questions.

6.2 Machine Translation

We performed a zero-shot evaluation on 10 sentences per language. All of the examples come directly from language documentation or language learning textbooks and were thus produced and/or verified by fluent speakers. Because of this, the MT dataset is not multiparallel (as the vocab quiz is), but rather high-quality samples of each language. An example output of the task is given in Table 3. Outputs are evaluated by the chrF++ score with β=2 (Popović, 2015).

Overall, Gemini outperforms every other model. Our findings agree with Zhu et al. (2024) and Enis and Hopkins (2024) that the LLMs generally perform better from the language to English. The results for this task, illustrated in Figure 2, can be summed up by a few broader points.

The LLMs have data in every language. Even though Mvskoke and Choctaw are not listed in Common Crawl or other public datasets, it is clear that all the LLMs have at least minimal training data for these languages, as they produced text in the correct orthographies. In the lower-resourced languages, the responses range from complete gibberish (GPT-5.2) to somewhat understandable sentences (Gemini).

Language underrepresented online. Hawaiian's representation in publicly accessible datasets such as Common Crawl and Wikipedia (see Table 1) shows that more representation in the training data via larger representation online increases accuracy, as every model performs best for Hawaiian in the language -> X direction. The lesser-represented languages generally perform more poorly, with some variability in performance perhaps due to linguistic complexity or small sample size.

Nonsensical output. Even though chrF++ scores for our evaluation are on par with what would be expected from a baseline neural model (De Gibert et al., 2025), one issue is that LLMs often generate nonsensical output when producing translations from English into the given language. Table 3 shows an example of hallucinated words. However, when translating in the direction from the Indigenous language to English, there are no nonsensical words.

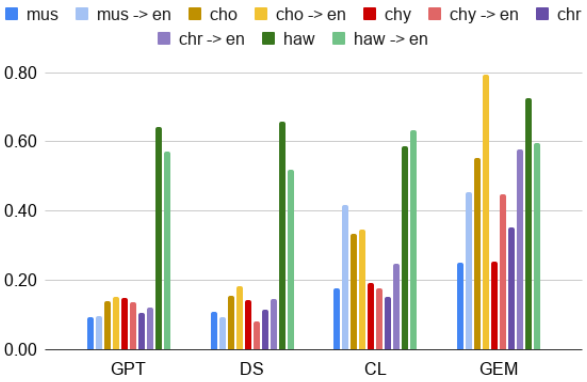

Figure 2: MT Results: chrF++ scores for translations in two directions, En -> X (darker color) and X -> En (lighter color), for four models (GPT 5.2, DeepSeek Reasoner, Claude, and Gemini 3.1 Pro).

6.3 Vocabulary Quiz

We prompt the models to translate 50 words from English into the target language. We hand-graded the responses by looking up each response in the corresponding dictionary. Synonyms or commonly accepted alternate spellings are counted as correct answers. Because we are only familiar with Mvskoke and Choctaw, we are unable to provide value judgments on alternative answers in languages not represented in the dictionaries. Because of this, the results for Mvskoke and Choctaw may be higher than those of the other three languages. Nonetheless, some interesting trends can be observed. Figure 3 shows the results of prompting the models for individual words. Overall, Gemini Pro 3.1 outperforms other models in every language, getting a perfect score for Hawaiian.

Nonsense and unrelated words. Many responses, especially in the lower-performing models, are simply jumbles of letters in the given orthography. Other responses are words that do exist in the language, but are unrelated to the term. For example, Claude provides Cherokee "ᎤᎮᎵᏏᏱ" (pretty) for "to know".

Inventing modern words. We also separately tested for modern words that are less formalized in the languages, such as "computer", "microwave", and "yogurt". These are not counted as part of the vocab quiz, but are intended to ex-

Source:	Yv likat mēkken omat hēr̄ēs.
Ref Translation:	This one sitting here would be a good king.
Mus->En:	It would be good if this one sitting here were king.
En->Mus:	Heyv likat meco here tvres.
Back translation:	This one sitting here [unknown] good will be.

Table 3: Example output from Gemini for Mvskoke translation. The word "meco" is nonsensical, and the sentence is grammatically incomplete.

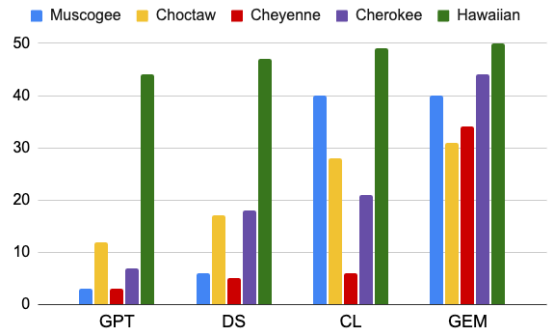

Figure 3: Vocabulary Quiz: Number of correct answers out of 50 for five models (GPT 5.2, DeepSeek Reasoner, Claude, and Gemini 3.1 Pro).

amine model behavior. Some of these terms have been normalized by community usage. A few of the words have published translations, such as Cherokee's "ᎠᎦᏙᎯᏍᏗ ᎠᏍᎪᎯᏰᎯᏍᏗᎩ" for "computer". No model was able to produce this term. For words that do not exist in a language, we found that the LLM will still try to come up with something. Sometimes the inventions, such as "esvhonkvtkv" (*thing you count with*) as Mvskoke for "computer", are not ideal but still reasonable. Others are completely nonsensical letter sequences.

Answer refusal. A notable behavior is Claude's refusal to provide responses for words with very low certainty. For example, when prompted to provide the word for yogurt, Claude responded "I'm not aware of a specific, established Muscogee (Creek) word for 'yogurt.' This is a modern food item that may not have a standardized translation in the Muscogee language. You may want to consult the Muscogee (Creek) Nation's language department for an official term." In terms of preventing hallucinations and errors, this might be preferred over providing low-quality translations. The only answer refusals by Claude were contained to the "modern" words, and did not affect the vocabulary quiz results.

6.4 Text Generation

Text generation performance is similar for Mvskoke and Choctaw languages for each LLM. GPT-5.2 and DeepSeek make use of some related terms, but the grammar and word choices overall are largely not understandable. Claude and Gemini can use both related terms and the correct grammar form (distant past) to conjugate verbs. Furthermore, some general aspects of the stories align with traditional stories, such as similar characters and storylines. Even so, the outputs contain many hallucinated words and grammatical mistakes, making the text difficult to read and, at times, meaningless. See Table 4 for an example of text generation in Mvskoke.

6.5 Speech

DeepSeek and Claude do not natively support TTS or ASR. Thus, we conduct the speech category evaluation only for GPT-4o and Gemini 3.1 Pro.

ASR. We test two short audio clips for each language Mvskoke and Choctaw. The samples are of clear, read speech recorded during language documentation (Gouge et al., 2004; Haas et al., 2015). The word error rates and character error rates are given in Table 5. Further details can be found in Appendix A. The WERs for both models are not better than previously developed ASR models in either language (Brixey and Traum, 2022; Mainzinger and Levow, 2024). One notable finding is that the models do not demonstrate competence in using the languages' orthographies for this task, unlike results found in other evaluation categories.

TTS. We prompt each model with a short phrase in Mvskoke and Choctaw to be said aloud. We provide the target language in both the instruction and the prompt. Based on listening to the audio, we determine that GPT and Gemini are not able to produce audio that sounds like these languages. GPT does not remotely follow the ex-

Output	Back-translation
Hofonof, ponvttv sulkē omof, kaspes. Totkv sekon. Ponvttv vtekat kvapvtes. Uewv tvpvlv, este hopvye totkv ocvtes. Mv este totkv vcayecvtes. Cufe makvtes, "Totkv horkoparēs. Vnheckv herēs, ce!"	A long time ago, when animals many, [unknown, possibly an attempt at "cold"]. There is no fire. All the animals [unknown]. Across [misspelled] the water, person far away had fire. That person stored the fire. Rabbit said, "I will steal the fire. My appearance is good!"

Table 4: Sample of Gemini response to the prompt, "Tell me the story about Rabbit Steals Fire. Tell it in Muscogee, in the style of a Muscogee traditional story." In the traditional story, the rabbit travels across a great body of water to steal fire. There are several instances of hallucinated words (red), grammatical issues, and awkward wording. The word order is very English-like - for example, putting "the rabbit said" before the quote, when a Mvskoke person would always put it after.

	GPT		Gemini	
	WER	CER	WER	CER
mus	0.92	0.47	**0.91**	**0.34**
cho	**0.82**	**0.26**	0.98	0.35

Table 5: Word Error Rates (WER) and Character Error Rates (CER) for GPT 4o Transcribe and Gemini 3.1 Pro for the ASR task.

pected phonemes for the orthographies – most egregiously, it pronounces the Mvskoke 'v' as /v/ instead of the vowel /ə/. Gemini produces this vowel correctly but makes other phonetic mistakes that render the audio unintelligible.

7 Conclusions and Future Directions

In this work, we demonstrated tests that an Indigenous language community could perform to evaluate a given LLM's acceptable language proficiency in a specific language without extensive labor or large-scale datasets. While some of our samples for certain tests were limited, the tests included in our framework provide a holistic understanding of a given LLM's proficiency that could be generalizable to other communities. Studies that demand significant effort from communities with small speaker populations without offering reciprocal benefits can be perceived as extractive. Our framework, which can be conducted by a second-language speaker, is thus less extractive, as comparisons can be made with language documentation, and consultation with elders could be limited to a subset of output for review.

We present four categories of evaluation: cultural knowledge, language proficiency, text generation, and speech. Overall, the results reveal substantial limitations across all models for the five Indigenous languages considered. Cultural knowledge, especially, is incomplete and inconsistent across the four LLMs evaluated, which, for minority communities, is essential for preventing harmful stereotypes and misinformation. In terms of language proficiency, performance is consistently higher in Hawaiian than in the other languages. GPT and DeepSeek show very low proficiency in all languages except Hawaiian. Claude and Gemini show slightly better performance, but their outputs still contain frequent errors and hallucinations, especially in the four lower-resourced languages. Text generation for Mvskoke and Choctaw shows that LLMs can use correct orthography and some limited grammatical constructs. However, the outputs still exhibit frequent hallucinations and grammatical errors. We found that ASR underperforms when compared to traditional neural models. Finally, text-to-speech capabilities are not able to produce intelligible audio in the given languages.

These evaluation tasks are not meant to be comprehensive or exclusive. Rather, they are offered as a starting point for communities to adapt according to their own priorities and goals. Communities with more existing linguistic resources may be able to curate larger datasets for evaluation. More-resourced language groups may find that LLMs are performant enough to use as part of an end-to-end system. For example, performance in Hawaiian might be good enough for use with a RAG system, as demonstrated by the Kumu Connect implementation (Baker-Ramos et al., 2025).

Future work could include fine-tuning models to better estimate the extent of performance improvement achievable with limited data, although our findings suggest that existing models may already incorporate much of the available

printed data for the five languages we reviewed. More granular methods, such as quantifying the number of hallucinations and mistakes (for example, Chaka (2024)), could offer additional insight. At the same time, these directions raise ethical and practical considerations. Curating data for LLMs requires significant effort, often from fluent speakers whose time may be better spent on community-based activities such as teaching (Wiechetek et al., 2024). Moreover, hallucinations are an inherent limitation of current LLM architectures and are unlikely to be fully eliminated, even in high-resource languages (Hicks et al., 2024).

These findings suggest that while LLMs may offer limited utility in certain contexts, they may not be appropriate for endangered-language situations where second-language learners are unable to identify errors in AI-generated outputs. Careful, community-driven evaluation remains essential in determining whether and how these technologies should be used.

8 Ethical Considerations

By prompting the LLMs with data from language documentation, there is the risk that these models may store and use that data without permission. However, it is likely the LLMs already scrape most of the language documentation from online sources, without regard for copyright, intellectual property rights, or tribal data sovereignty.

At present, none of the AI companies that develop these LLMs offer methods for speakers of any language to correct language proficiency issues or cultural misunderstandings. Additionally, to our knowledge, our tribes were not contacted to be included in any current LLM versions, nor were any culturally appropriate, modern, and representative linguistic data requested from our communities by LLM companies for inclusion. Unregulated use of such data by AI companies risks undermining community language authorities, and may harm the vitality of Indigenous languages.

Finally, while we have designed this framework to be as least extractive as possible, there may be necessary involvement from elders or fluent speakers, which may distract from other important language revitalization activities.

Limitations

The findings presented here are not equally applicable across all Indigenous languages. While North American Indigenous languages may share some broad characteristics, each language has distinct linguistic features, and communities differ in terms of available resources and speaker populations. As such, the relevance of our evaluation framework may vary significantly across contexts.

We recognize that English is implicated in all evaluation tasks, and as a result, this study is not a within-culture monolingual evaluation. We also acknowledge that "culture" is not equivalent to language, and we do not intend to define any singular or essentialized notion of culture. Rather, we recognize that there is diverse cultural, social, and linguistic variation within language communities.

Our evaluation is also limited by our own experiences in our respective language communities. Many additional questions could be asked during evaluation, including different tasks, use cases, and linguistic considerations. The selected tasks represent only a subset of possible evaluations and should not be interpreted as exhaustive. Likewise, the datasets used in this study are small; as a result, our findings should be interpreted as indicative rather than definitive and do not provide statistically precise estimates of model performance.

Finally, our analysis is shaped by our own knowledge and positionality as researchers and community members. We offer insight into the Mvskoke and Choctaw communities, but we are not elders and do not claim authoritative perspectives on every aspect. Our inclusion of three additional Indigenous languages is intended for comparative purposes only, and we do not claim authority on those languages. There may be linguistic, cultural, or contextual nuances that are not fully captured in our evaluation.

Acknowledgments

Thank you to our elders and knowledge holders. We acknowledge those who have gone before us, who have carried and passed down our languages, and those who continue to sustain them. Thank you to Steven Bird for advising, and to Tad Hosford and Ian Iglesias for reviewing Mvskoke output. Finally, we thank the anonymous reviewers

for their helpful feedback.

References

Anthropic. 2026. Claude opus 4.6.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266. Association for Computational Linguistics.

Rosalia Badhorse. 2023. *Tsetsėhestâhese and So'Taeo'o (Cheyenne) Language: Grammar Sketch for Learners*. Ph.D. thesis, University of Arizona.

Rachel Baker-Ramos, Will Gelder, Leah Cho, Jahnavi Kolakaluri, and Josiah Hester. 2025. Kumu connect: Design thinking for place-based generative educational technology in hawaiian immersion schools. In *Proceedings of the 2025 Conference on Research on Equitable and Sustained Participation in Engineering, Computing, and Technology*, pages 186–195. ACM.

Anita Bardwell, Joe Bardwell, Lopaka Weltman, and Hoaloha Westcott. 2020. *He Papa Kuhikuhi Pilina'ōlelo: Reference Grammar of the Hawaiian Language*. University of Hawai'i.

Matthias Brenzinger and Patrick Heinrich. 2013. The return of hawaiian: language networks of the revival movement. *Current Issues in Language Planning*, 14(2):300–316.

Jacqueline Brixey and David Traum. 2022. Towards an automatic speech recognizer for the choctaw language. In *Proc. S4SG 2022*, pages 6–9.

Cyrus Byington. 1870. Grammar of the Choctaw language. *Proceedings of the American Philosophical Society*, 11:317–367. Edited by Daniel Garrison Brinton. Also published as a monograph by McCalla and Stavely, Philadelphia, 1870.

Chaka Chaka. 2024. Currently available GenAI-powered large language models and low-resource languages: Any offerings? wait until you see. *International Journal of Learning, Teaching and Educational Research*, 23:148–173.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2454–2469, New York, NY, USA. Association for Computing Machinery.

Monojit Choudhury. 2023. Generative AI has a language problem. *Nature Human Behaviour*, 7:1802–1803.

Roberts Darģis, Guntis Barzdins, Inguna Skadiņa, Normunds Gruzitis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Marc T J Elliott, Deepak P, and Muiris Maccarthaigh. 2025. Evolving generative ai: entangling the accountability relationship. *Digital Government: Research and Practice*, 6(1):1–13.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude.

Louise Fisher, Lenora Holliman, Wayne Leman, Leroy Pine Sr., and Marie Sanchez. 2023. *Cheyenne Dictionary*. Chief Dull Knife College.

Earnest Gouge, Edited, Translated by Jack B. Martin, and Juanita McGirt. 2004. *Totkv Mocvse / New Fire: Creek Folktales*. Norman: University of Oklahoma Press.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and

175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Mary R. Haas. 1979. Southeastern languages. In Lyle Campbell and Marianne Mithun, editors, *The Languages of Native America: Historical and Comparative Assessment*, pages 299–326. University of Texas Press.

Mary R. Haas, James H. Hill, Jack B. Martin, Margaret McKane Mauldin, and Juanita McGirt. 2015. *Creek (Muskogee) Texts*. University of California Publications.

Zachary Arao Hanson. 2025. Indigenous (mis)representation in emerging LLM research methodologies. *UC Riverside Undergraduate Research Journal*, 19.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*, 26(2):1–10.

Leanne Hinton. 2013. *Bringing our languages home: Language revitalization for families*. Heyday Books.

Gregg Howard and Rick Eby. 1990. *Introduction to Cherokee*. Various Indian Peoples Publishing Co.

Charles Julian. 2010. *A History of the Iroquoian Languages*. Ph.D. thesis, University of Manitoba.

Daniel Heath Justice. 2025. *Our Fire Survives the Storm: A Cherokee Literary History*. University of Minnesota Press.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374. Association for Computational Linguistics.

James N Leiker and Ramon Powers. 2011. *The northern Cheyenne exodus in history and memory*. University of Oklahoma Press.

Wayne Leman. 1980. *A reference grammar of the Cheyenne language*. Chief Dull Knife College.

Aha Pūnana Leo and Hale Kuamo o. 2003. *Māmaka Kaiao*. University of Hawai i Press.

Jason Edward Lewis, Hundefinedmi Whaanga, and Ceyda Yolgörmez. 2024. Abundant intelligences: placing ai within indigenous knowledge frameworks. *AI & Society*, 40(4):2141–2157.

Richard Littlebear. 2024. Neneehove'tanonėstse tsehe'enėstsestse; tsehe'enėstsetse neneehove'tanone: We are our languages; our languages are us. *Tribal College*, 35:1–3.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Jack B Martin. 2011. *A grammar of creek (Muskogee)*. UNP - Nebraska.

Jack B Martin and Margaret Mckane Mauldin. 2000. *A Dictionary of Creek/Muskogee*. University of Nebraska Press.

Paul J Meighan. 2024. Indigenous language revitalization using *TEK-nology* : how can traditional ecological knowledge (TEK) and technology support intergenerational language transmission? *Journal of Multilingual and Multicultural Development*, 45:3059–3077.

Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. FilBench: Can LLMs understand and generate Filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah

Moeller, Jeffrey Micher, and 5 others. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

Mississippi Band of Choctaw Indians. 2026. Choctaw language dictionary.

Ōiwi Parker Jones. 2018. Hawaiian. *Journal of the International Phonetic Association*, 48(1):103–115.

Brock Pitawanakwat. 2018. Strategies and methods for anishinaabemowin revitalization. *The Canadian Modern Language Review*, 74(3):460–482.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Pukui and Elbert. 2003. *Hawaiian Dictionary*. University of Hawai i Press.

Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, and Shlomo Mark. 2024. Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: Narrative review and qualitative study. *JMIR AI*, 3:e51834.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Mike Rogers. 2021. Choctaw Nation members talk about impact of losing native speakers to COVID-19. *News 12*.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-first edition. SIL International, Dallas, Texas.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo' Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2026a. Are small language models the silver bullet to low-resource languages machine translation? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 1–26. Association for Computational Linguistics.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo' Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2026b. Are small language models the silver bullet to low-resource languages machine translation? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 1–26. Association for Computational Linguistics.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

The Choctaw Nation of Oklahoma Dictionary Committee. 2016. *Chahta Anumpa Tosholi Himona: New Choctaw Dictionary*, 1st edition. Choctaw Print Services.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating LMMs on culturally diverse 100 languages. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575. IEEE.

Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup, and

Sjur Nørstebø Moshagen. 2024. The ethical question – use of indigenous corpora for large language models. In *Proceedings of the Language Resources and Evaluation Conference*, pages 15922–15931. European Language Resources Association (ELRA) and ICCL.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language. In *ACL 2022*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781. Association for Computational Linguistics.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, ASR. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4004–4011. European Language Resources Association (ELRA).

A Appendix

This section provides additional detail about the ASR task.

mus	
reference	Este hvmket likvyvntvs, makesasvtēs, mahokvnts.
prediction	Este hvmket legayantis mvgisazetis mahagets.
cho	
reference	Haklo hattak mʊt kanima il ia chi ka pim anoli
prediction	Haklo, hattakbat kanimma il ia chinka pim anoli.

Table 6: ASR example output of short samples given by GPT 4o transcribe. Punctuation is stripped before error rate calculation. Output contains characters that are outside of the orthographies.

The ASR experiment comprised of two audio samples per language, one short and one longer sample. Each sample is from high-quality recordings from language documentation. Table 7 gives the length of each audio clip and the corresponding WER and CER of the outputs. Table 6 shows an example output from a short clip.

		GPT 4o transcribe		Gemini 3.1 pro	
lang	length	WER	CER	WER	CER
mus	12	0.83	0.41	1.00	0.45
mus	37	1.00	0.52	0.83	0.23
cho	5	0.70	0.15	1.00	0.26
cho	38	0.94	0.37	0.96	0.45

Table 7: Audio clip lengths (in seconds) and ASR error rates per model.

A data-centric approach to performance improvement in under-resourced ASR: The case of Dënë Sųłné

Olga Kriukova[1], Olga Lovick[1], Antti Arppe[2]

[1]University of Saskatchewan, [2]University of Alberta

Correspondence: olga.kriukova@usask.ca

Abstract

This paper presents a study focused on advancing Automatic Speech Recognition (ASR) for the under-resourced language Dënë Sųłné through data-centric approaches. We explore multiple strategies to enhance the quality of training data—both audio recordings and transcriptions—to address the challenges posed by mixed-quality datasets. Our experiments investigate which data preparation techniques most effectively improve ASR performance in this context. Our findings show that reducing spelling variants of the same lexeme in the corpus significantly improves model generalization, resulting in a substantial increase in recognition accuracy. Additionally, we demonstrate that increasing manually reviewed transcriptions consistently improves word and character error rates, while audio enhancement slightly reduces performance, highlighting the complex trade-offs in low-resource ASR development.

1 Introduction

Neural models have significantly expanded the tools and resources available for processing major languages. However, recent advances in machine learning have enabled languages with scarce linguistic data to also benefit from these models. Performance of neural models heavily depends on the quality of the training data (Sambasivan et al., 2021), yet under-resourced languages face a compounding problem: data is not only scarce but also particularly hard to obtain in high-quality form (Kjartansson et al., 2018; Liang and Levow, 2025).

Researchers working with major languages often take data quality for granted, relying on well-curated, validated benchmark datasets (cf. Joshi et al., 2020) such as—in the case of speech recognition technologies—Mozilla Common Voice (Ardila et al., 2020) or VoxPopuli (Wang et al., 2021). These datasets, while not of uniform quality, benefit from systematic curation processes and validation that under-resourced languages typically lack. This abundance of relatively high-quality data has naturally led the field toward model-centric approaches in machine learning, where researchers focus primarily on architectural innovations and hyperparameter optimization (Bhatt et al., 2024; Jakubik et al., 2024) or adding more data. However, this model-centric paradigm is ineffective for under-resourced languages, where factors such as non-standardized orthography, inconsistent transcriptions, and variable audio quality compound data scarcity (Lin et al., 2024; Liu et al., 2022; Wisniewski et al., 2020; Zhong et al., 2024). When training data contains numerous incorrect labels, even optimal model configurations will yield poor results. This is why, data-centric approaches—which prioritize improving the quality and accuracy of training data over model tuning—are essential for effective machine learning (Whang et al., 2023).

While the field increasingly recognizes this challenge (Luger et al., 2025), data-centric approaches for under-resourced languages remain largely underexplored. Through our work on fine-tuning an Automatic Speech Recognition (ASR) model for Dënë Sųłné, an under-resourced, endangered language spoken in Canada, we demonstrate how systematic dataset improvement can yield significant gains. We present experiments across data refinement stages, identify key preprocessing steps, and highlight the need for the computational linguistics community to engage more deeply with data-centric methodologies for advancing under-resourced language technologies.

2 Backround

2.1 Dënë Sųłné

Dënë Sųłné (ISO 639-3: chp), belongs to the Dene (Athabaskan) language family and spoken across Alberta, Saskatchewan, Manitoba, and the Northwest Territories by approximately 10,000 people

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 95–106

(Statistics Canada, 2021). The data for this study were collected from speakers of the neighbouring communities of Clearwater River and La Loche in Saskatchewan.

Like other Dene languages, Dënë Sųłné is polysynthetic and predominantly prefixing, with a large consonant inventory and phonemic contrasts of tone and nasality (Cook, 2004). These typological properties, combined with the absence of a fully standardized orthography, create specific obstacles for corpus-based work that are detailed in Sections 2.2 and 2.3.

2.2 Language data

The dataset for the present study was compiled from three sources. The first and most significant portion (85%) comes from the *Talking Dene* project (2020-2024; PI Olga Lovick). The second portion (9%) was collected by Kriukova for this study. The third portion (4.4%) consists of verb paradigm recordings collected by Nial Willems (2025) from one speaker. In all three cases, participating speakers explicitly consented to the use of their language data for training the ASR model. In total, the whole corpus for this study contained language data from 22 speakers. The total length of the audio data for the final iteration of the model was 12 hours 35 minutes, and the number of utterances was 18,779; however, these numbers changed over the course of the study, as explained in detail in Section 4.

The corpus contains predominantly Dënë Sųłné utterances, along with some English and Dënë Sųłné-English mixed utterances. To preserve the model's ability to recognize both languages, we included all utterance types in the training and testing datasets. This approach helps to prevent catastrophic forgetting of English that can occur when fine-tuning on a single language (Simmons, 2025).

The audio recordings in the dataset also vary in quality. The majority of them have good or satisfactory quality, which is sufficient for successful ASR training. Nevertheless, some recordings feature interviews with multiple speakers and thus contain occasional overlapping speech by two or more speakers. Additionally, some audio clips have significant background noise that, in rare cases, makes a speech signal inaudible.

2.3 Orthographic challenges in Dënë Sųłné

The orthographic instability of Dënë Sųłné has direct, measurable consequences for ASR training data. Our training data reflects the natural variation that arises when a language lacks a common written standard. As a result, the same word frequently appears in multiple written forms, with transcribers independently applying their own perceptual spelling strategies, especially regarding nasality and tone (Kriukova et al., 2026). Generational variation compounds this: younger speakers often produce innovative verb forms—deleting prefixes, altering morpheme order—which transcribers may render phonetically or "correct" to conservative forms, generating additional surface variants for the same pronunciation (Lovick et al., 2023, 2024).

The practical effect is a training vocabulary inflated by spelling variants without semantic distinction. Prior to any standardization, our corpus contained 17,355 unique token types; targeted standardization of frequent forms reduced this to 14,892 (see Section 4.1.1). This reduction directly decreased label noise and improved model generalization, motivating the data-centric approach described in the remainder of this paper.

3 Literature Review

There is a growing body of research on ASR for under-resourced and endangered languages. Several studies have demonstrated that pre-trained multilingual models generally achieve higher accuracy in low-resource settings than monolingual models (Jimerson et al., 2023; Sadeque, 2022; Yadav and Sitaram, 2022). However, Jimerson et al. (2023) found that no single ASR architecture consistently outperforms others across under-resourced languages. In their comparison of Whisper, Wav2Vec2, Kaldi DNN, and ESPnet2 across 11 under-resourced languages with datasets ranging from 26 minutes to 19 hours, they found no correlations between Word Error Rates (WER) and morphological properties, dataset size, or recording quality. For Whisper, the resulting WER was ranging from 5% to almost 75%. Notably, Whisper demonstrated superior performance for languages with large phone sets (>37 phones).

Another critical consideration in low-resource ASR is data partitioning, which presents unique challenges when datasets contain few speakers or recordings from a single speaker. Liu et al. (2023) demonstrated that random splits provide better training and test sets than the common "hold speaker out" method when data is scarce. Nonetheless, they recommend testing multiple partitioning methods given the individual characteristics of

small datasets.

Furthermore, Wisniewski et al. (2020) emphasize that training data preprocessing for under-resourced language corpora presents different challenges than preprocessing for well-resourced languages. Unlike standard NLP preprocessing pipelines, which can rely on established tools and conventions, preprocessing endangered language data is highly non-trivial and often requires developing language-specific solutions (as Section 4.1.1 illustrates). These tasks are not only time-consuming but also demand substantial familiarity with both the target language and the specific corpus characteristics. As the authors note, the severity of these challenges can sometimes even discourage computational linguists from working with the data.

Beyond these methodological considerations, certain linguistic factors complicate ASR development for under-resourced languages. For instance, code-switching, common in minority-language communities (cf. Moore, 2018), poses significant challenges for ASR systems (Coto-Solano et al., 2022; Guillaume et al., 2022; Simmons, 2025). Similarly, languages with complex orthographies or large phoneme inventories pose particularly demanding challenges for ASR (Adams et al., 2019; Gauthier et al., 2016; Prud'hommeaux et al., 2021). As Liang and Levow (2025) demonstrate, even state-of-the-art multilingual models struggle to accurately capture nuanced phonological contrasts such as tone distinctions, nasality, consonant length, and vowel length—features that are phonemic in many under-resourced languages (cf. Jimerson et al., 2023). Furthermore, Liang and Levow (2025) and Ćavar et al. (2016) also emphasize that fieldwork speech data—characterized by spontaneous speech and varied recording conditions—poses challenges for ASR models trained on standardized corpora.

The corpus we used in this study exemplifies many of the abovementioned challenges. Dënë Sųłiné lacks a fully standardized orthography and employs written representation of nasality and tone. Additionally, the corpus contains code-switching, and most audio recordings were captured in fieldwork settings with variable quality. The following section details our approach to navigating these obstacles.

4 Methodology

4.1 Dataset improvement

The results of the baseline model iteration were occasionally plausible but mainly unsatisfactory. It became evident that without altering the training data, we would be unlikely to improve the transcription results. Therefore, we carefully examined our training corpus and identified the main issues to be addressed to achieve higher-quality transcriptions. First, spellings of individual words varied widely across transcribers, making the training data very noisy. Second, our initial dataset contained too many sentences that were not reviewed by a linguist, resulting in inconsistent word boundary distributions and missing sentence-final enclitics in some transcripts. Most critically, innovative verb forms were spelled inconsistently and often required a manual review. Third, some audio clips had excessive background noise or significant speech overlap, which could also interfere with a model's pattern generalization.

In the following subsections, we describe how we addressed these data inconsistencies, while Section 5 demonstrates the effect of these actions on our ASR model performance.

4.1.1 Standardization of orthography

As discussed in Section 2.3, orthographic variation in Dënë Sųłiné transcriptions stems from both phonetic transcription practices and inconsistent spelling conventions. High variation makes training data noisy, hindering the model's ability to generalize speech-to-text patterns. This issue is not exclusive to word-level representations: subword tokenizers, which segment words into recurring character sequences, are equally sensitive to orthographic inconsistency. When the same word form appears under multiple spellings, it may be segmented into different subword units across instances, further compounding the noise in the training data. To address this, we reduced non-phonemic variation across multiple standardization stages and model iterations, focusing on the most frequent word types (for full details on the decision-making process and variant identification, see Kriukova et al. (2026)).

In the first stage, we targeted verbal enclitics and enclitic combinations, writing them separately from verbs and each other to reduce verbal vocabulary size (e.g., *łí nį á*—habitual+past+assertive—instead of *łínįá*),

along with frequent postpositions and adverbs. In the second and third stages, we orthographically standardized pronouns, possessive prefixes, numerals, remaining adverbs and postpositions, and frequent nouns and verbs—for example, replacing reduced numeral spellings with full forms (e.g., *tae* → *taghë*) and standardizing frequent verb forms such as *nëzǫ* 'it is fun', *bënasnį* 'I remember', *nësthën* 'I think'. In the fourth stage, we standardized word tokens belonging to frequent verb paradigms such as 'to be', 'to say', and 'to talk'. In total, these changes reduced the ASR training vocabulary (word-level) from 17,355 to 14,892 unique types.

It is worth noting that orthographic inconsistency affects not only the number of word-level representations but also the number of subword units. The model used in this study (see Section 4.4) employs a subword tokenizer, which segments words into recurring character sequences rather than treating each word as an atomic unit. When the same word form appears under multiple spellings, it may be segmented into different subword units across instances, producing inconsistent token representations that hinder model learning. Orthographic standardization is therefore important for both word-level and subword-level tokenization. For more information, see Kriukova et al. (2026).

4.1.2 Review of data

To address other inconsistencies in our corpus, such as reduced verb forms, differences in word boundaries, and missing sentence-final enclitics, we used two strategies. The main goal here was to increase the number of fully reviewed transcriptions in our corpus to improve the overall data quality. One strategy, performed by Lovick, involved manually reviewing and standardizing all utterances. This strategy was the most time-intensive, requiring 30–60 minutes per minute of multi-speaker recording and up to 20 minutes per single-speaker recording (see also Amith et al. (2021) for another example of manual data correction). The second strategy, performed by Kriukova, involved partial review of the utterances: fixing typing mistakes in both Dënë Sųłiné and English, correcting word boundaries, and adding missing enclitics. This approach took significantly less time to complete.

As a result, during this study, we completed a partial review, targeting particular tokens and types, of all utterances in the corpus before the last model iteration was trained. While time constraints prevented a complete manual review, over 400 utterances from different speakers were reviewed. Additionally, some previously reviewed transcriptions were added to the training corpus after the first model iteration was trained.

4.1.3 Review of audio clips

Our last step in improving the dataset was the review of audio clips. Our first dataset contained 14,974 audio clips with their corresponding transcriptions. The majority were of satisfactory quality; however, some clips exhibited excessive background noise or speech overlap, rendering individual speakers nearly unintelligible. Since speech overlap is a well-known cause of deterioration in transcription quality (Alharbi et al., 2021; Meng et al., 2024), we excluded or shortened audio clips that contained cross-talk. Furthermore, we removed clips from recordings with poor audio quality to assess the impact on model performance. Since full manual review was not feasible given the corpus size, these recordings were identified through listening to samples, inspecting spectrograms, and number of speakers on each recording, targeting those characterized by excessive background noise and speech overlap.

Nevertheless, during audio cleaning, we removed only speech overlaps and the noisiest segments, adjusting transcriptions when necessary, and limited removal to clips with substantial quality issues (n=129), since Whisper models benefit from varied audio quality (Radford et al., 2023). These removed clips predominantly featured impulsive noise, such as doors opening or closing, objects falling, phones ringing, or children shouting during play.

4.2 Training dataset

As a result of the dataset improvement efforts described in Section 4.1, our training dataset was continuously enhanced and updated. Therefore, each model iteration was trained on a slightly modified dataset. All training utterances could be divided into four categories: 1) reviewed utterances – utterances that were manually reviewed and standardized by Lovick; 2) partly reviewed utterances – utterances that underwent only a partial review by Kriukova; 3) partly standardized utterances–utterances that contain tokens that were standardized through targeted, corpus-wide standardization process, but were not comprehensively reviewed; and 4) unreviewed utterances – utterances

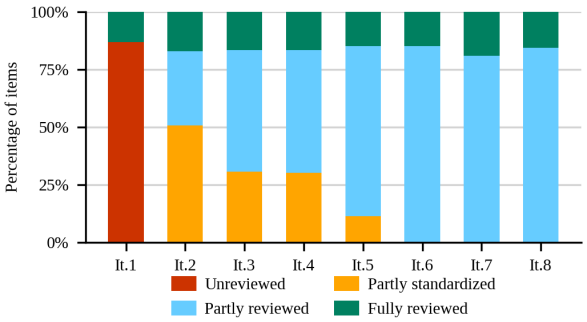

Figure 1: Status of utterances across iterations.

that were neither reviewed nor standardized. The changes in the distribution of utterances' statuses from iteration to iteration are illustrated in Figure 1. The log of the changes made to the dataset is outlined in Table 1.

Iter.	Utts.	Improvements
1	14,974	None
2	16,442	The number of fully reviewed utterances was increased; the first round of standardization was completed.
3	16,936	The number of reviewed and partly reviewed utterances was increased; the second round of standardization was completed.
4	16,807	Bad-quality audio files were removed from the corpus. Some audio files were edited to remove speech overlap.
5	18,615	Newly partly reviewed utterances were added to the corpus.
6	18,615	The partial review was completed for all the utterances. The third round of standardization was completed.
7	19,583	Reviewed recordings of paradigms were added to the corpus.
8	18,779	The fourth round of standardization was completed. Recordings of paradigms were concatenated into longer files.

Table 1: A summary of the ASR model's iterations.

4.3 Testing dataset

To track how changes to the dataset affect model performance, we prepared a dedicated test dataset of 100 utterances (3min 44sec). We decided to keep the testing dataset small, so we could thoroughly review all the test transcriptions and audio files within a reasonable time. Following Liu et al.'s (2023) recommendation to avoid the "hold-speaker out" partitioning method (see Section 3) in low-resource settings, we selected test utterances randomly from different speakers. The number of utterances required for each gender and age group was precalculated to ensure a representative test set. This decision is rooted in the fact that Dënë Sųłınné exhibits significant age-based variation (Jung et al., 2025), and it is important to assess how well the model generalizes across age groups. The division by gender was done to account for the acoustical differences between male and female voices.

Using a Python script, utterances were randomly sampled to preserve both gender and age balance. Each utterance filename encodes a unique three-letter speaker identifier, which the script used to group utterances by speaker and retrieve their corresponding gender and age decade. From each gender–decade group, the maximally even number of utterances was drawn across the 22 speakers, ensuring every speaker is represented in the test set. The resulting test set comprises 50 utterances from female speakers and 50 from male speakers. Speaker age decades were calculated from each speaker's age as of 2025.

After the testing utterances were selected, each recording–utterance pair was reviewed to ensure that the sound quality was good enough for the testing set and that there was no significant voice overlap. The recordings discarded for failing to meet the aforementioned criteria were replaced with randomly selected utterances from the same speaker. All transcriptions were then reviewed by Lovick.

The utterances in the final testing dataset were then tagged for the presence of codeswitching (n=35) and proper nouns (or named entities; n=16) and marked if they were English sentences (n=4). This coding system was applied to determine whether the presence of codeswitching or proper nouns affects Word and Character Error Rates (WER and CER). Additionally, to measure the effect of the audio quality on the transcription results, each recording was tagged as "Good" or "Satisfactory". The tag "Satisfactory" was given in the cases when: 1) the audio from a speaker is faint (usually due to distance from the microphone); 2) low to moderate volume static noise is present in the background (e.g., running water, TV in another room, refrigerator, etc.), 3) low to moderate volume non-speech vocalizations by another participant are present. All other recordings were tagged as "Good" (n=64). To estimate the effect of each tag, we used linear mixed-effects models with the lme4 package (Bates et al., 2015) in R, using lmerTest (Kuznetsova et al., 2017) to

obtain p-values via Satterthwaite's method. WER and CER were modeled as dependent variables. Fixed effects included speaker gender (M/F), audio quality (Quality: Good/Satisfactory), presence of code-switching (CDSW: TRUE/FALSE), presence of proper nouns (NE: TRUE/FALSE), fully English sentences (Full_ENG: TRUE/FALSE), speaker age (calculated from year of birth), and iteration (1–9) (added in R). Speaker (n = 20) and audio clip (n = 100) were included as random effects to account for the repeated-measures structure of the data.

4.4 ASR architecture

Since this study's focus was data-centric, the model architecture for all the iterations described below was identical. All model versions were fine-tuned from the Whisper-medium multilingual ASR model (Radford et al., 2023) with a learning rate of 5e-5, using the model's default subword tokenizer. The fine-tuning was performed on a high-performance computing cluster at the University of Saskatchewan. The GPU used for the training was Nvidia A100 (80G). Training time averaged 24 hours.

5 Results

5.1 Iteration 1

The initial model iteration served as a baseline for measuring improvements in transcription quality in subsequent iterations. This model was trained on 14,974 utterances, of which only 1,974 were fully reviewed, while the rest were normalized but not reviewed or standardized.

The baseline model achieved a WER of 81.2% and a CER of 46.3%. These error rates, notably higher than for major languages (cf. Prud'hommeaux et al., 2021), stem primarily from substantial spelling variation in the training dataset and inconsistent word boundaries between verbs and enclitics. Statistical analysis revealed no significant correlation between the presence of code-switching or proper nouns, or audio quality and the observed error rates. However, fully English utterances demonstrated significantly lower WER (p = .002).

5.2 Iteration 2

The second iteration was trained after our dataset underwent the first stage of orthographic standardization (see Section 4.1.1), which included standardizing the most frequent tokens in the corpus: enclitics, pronouns, some postpositions, and some adverbs. In total, the dataset for this iteration contained 16,442 utterances. For this iteration, we increased the number of reviewed sentences in the dataset from 1,974 to 2,805 by reviewing additional raw utterances and adding one already-reviewed transcript, for which we obtained re-consent after our experiments began. Moreover, 5,321 sentences were partly reviewed. As a result of our corpus manipulations, WER and CER for this iteration improved to 68.6% and 38.5%, respectively. Fully English sentences continued to show significantly lower WER in this iteration (p = .010).

5.3 Iteration 3

Due to time constraints with full reviewing, we prioritized increasing the number of partly reviewed utterances in the third iteration. Additionally, we expanded the dataset from 16,442 to 16,936 utterances by adding new, already partly reviewed utterances and initiated a second round of standardization (see Section 4.1.1). Model retraining revealed that WER improved to 65% in this iteration, while CER remained almost unchanged at 38.6%. Compared to Iteration 2, WER dropped by almost 4 percentage points, indicating that more words were now transcribed correctly. These improvements can be partly attributed to the slightly increased dataset but mainly to the higher number of partly reviewed utterances, which provided a better ratio of standardized tokens in the corpus. Fully English sentences again showed significantly better WER (p = .011).

5.4 Iteration 4

For the fourth iteration, we assessed how removing poor-quality recordings and cleaning of recordings with episodic noise would affect our model performance. We removed 129 audio files and their corresponding transcriptions from the training dataset due to poor quality (see Section 4.1.3). Additionally, 2,177 audio clips were shortened to remove loud noises overlapping with the speakers. As a result, the WER and CER scores for this iteration slightly worsened to 68% and 40.4%, respectively. This decrease in transcription quality can be attributed to a reduced training dataset by 129 (0.7%) deleted utterances. Furthermore, many recordings were shortened, reducing the overall length of the training dataset. A mixed-effects model for this iteration also showed that good audio quality became a significant factor for transcription accuracy,

having lower WER and CER (p = .006, p = .008). The effect of fully English sentences on WER was no longer significant in this iteration.

5.5 Iteration 5

By Iteration 5, an additional set of transcribed files became available for training, following a participant's consent to include their recordings. To compensate for data removed during the audio cleaning process in Iteration 4, we partly reviewed and added 1,810 new transcriptions, thereby increasing the overall proportion of partly reviewed data in the dataset. The testing results for this iteration showed 65.6% WER and 38.7% CER, improving in comparison to both Iterations 3 and 4. These improvements can be attributed entirely to the expanded training dataset and the increased proportion of the partly reviewed data. The significant effect of good audio quality on lower WER and CER remained in this iteration (p = .024 and .045, respectively).

5.6 Iteration 6

For the sixth iteration, we continued experimenting with transcription quality. During this stage, we fully completed the partial review of the utterances, so that the training dataset for this iteration consisted entirely of reviewed and partly reviewed utterances. Additionally, we completed the third round of standardization (see Section 4.1.1), focusing on the remaining frequent adverbs, nouns, and verb forms. Both WER and CER for this iteration decreased slightly to 65.3% and 37.9%, further demonstrating that standardization and review of transcription consistently reduce the model's error rates.

5.7 Iteration 7

For the seventh iteration of the model, we focused on improving the transcription of verbs. To achieve this, we added paradigm recordings (made by Willems (2025)) because their transcriptions were fully reviewed and covered all inflectional forms for multiple verbs. The WER and CER for this iteration increased to 67.25% and 40.1%, respectively. This deterioration in quality was somewhat unexpected. In investigating this increase, we discovered that short, single-word recordings can negatively affect transcription outcomes for longer multi-word recordings (Trabelsi et al., 2024). Given that the Dënë Sųłıné recordings we worked with were predominantly in this format (one verb or verb phrase per recording), the observed deterioration may be attributable to the short duration of the newly added files (see Section 6 for discussion).

5.8 Iteration 8

Although the addition of paradigms in Iteration 7 led to a deterioration in model performance, we did not remove them from the dataset for the fourth and final round of standardization (see Section 4.1.1). Instead, following standardization, we conducted an experiment using two datasets: one with paradigm recordings in their original format (one recording per verb form) and one with verb paradigms concatenated into longer files (4–10 seconds), with silences between individual words reduced to a minimum. The model trained on the dataset with concatenated paradigms yielded lower error rates (WER 63.7%, CER 36.8%) than the model trained on individual paradigm recordings (WER 64.6%, CER 39.8%). Therefore, we decided to use concatenated paradigms, and chose the better-performing model as our result for Iteration 8.

5.9 Observations across iterations

Across all iterations, speaker gender, code-switching, and the presence of proper nouns did not show significant effects on transcription error rates. Fully English utterances in the testing dataset had significantly lower WER in the first three iterations (p = .002, p = .010, p = .011); however, this effect disappeared by Iteration 4. Audio quality significantly influenced transcription error rates only in Iterations 4 and 5 (see Section 5.4 and 5.5).

While individual speakers showed some variation in WER (ranging from -6.9% to +6.0% relative to average), these differences were not statistically significant. Notably, WER and CER were increasing with speaker age in Iterations 5 (p = .045) and 8 (p = .022). However, significant variation existed among individual audio files (SD = 0.25, 95% CI [0.21, 0.29]), indicating that some recordings were inherently more difficult to transcribe regardless of speaker characteristics, audio quality, or other factors.

6 Discussion

In this study, we applied multiple strategies to improve the training dataset, which mostly led to reductions in model error rates (see Figure 1). Below, we first discuss strategies that improved performance, then those that proved ineffective, and

finally examine additional factors that may influence transcription accuracy.

The method that yielded the best results for our model is orthographic standardization (Iter. 2, 3, 6, 8). It reduced inconsistencies in written representations of words, decreased the vocabulary size and allowed the model to generalize more efficiently. From Iterations 1 to 3 only, spelling standardization contributed to decreases of 16.2 and 7.7 percentage points in overall WER and CER, respectively.

Increasing the number of reviewed and partly reviewed utterances (Iter. 2, 3, 5, 6, 7)—either by reviewed existing utterances or adding new verified data—also boosted ASR performance. These dataset reviews helped reduce typing mistakes (in both Dënë Sųłiné and English), made word boundaries more consistent, and slightly decreased spelling variation. For Iteration 5, data reviewing allowed for 2.4 and 1.7 percentage points decreases in WER and CER, respectively. Overall, our systematic data reviewing approach directly enabled performance gains across model iterations.

Our concatenation experiment (Iter. 8) also proved effective, demonstrating that short-format recordings can be retained in the dataset without increasing error rates. Combining paradigm recordings into longer audio files (4–10 seconds) mitigated the negative effects of single-word recordings. This finding may be particularly relevant for under-resourced languages where large portions of existing audio data were recorded in short formats (e.g., for talking dictionaries), though it should be noted that such transcripts are not suitable for language model training.

In contrast, manual audio cleaning and the removal of poor-quality recordings (Iter. 4) proved inefficient, leading to higher WER and CER metrics. Moreover, this process caused our model to perform significantly better on high-quality test recordings for two iterations, thereby reducing its robustness to the sound quality of the fieldwork recordings. These results can be attributed to the Whisper-medium model's documented robustness to varying audio quality (Gong et al., 2023), which appears to enable our fine-tuned model to handle background noise effectively even in low-resource settings. However, Trabelsi et al. (2024) note that for Whisper-base and Whisper-small, transcription quality may benefit from enhanced audio quality in the dataset, at least for English and French. Therefore, we suggest that the decision to enhance audio quality should be guided by the characteristics of the dataset, the size of the Whisper model, and the intended use cases for the fine-tuned model. Given that our model will be used to transcribe fieldwork recordings, which are typically noisier (Liang and Levow, 2025), and may also be deployed in classroom settings, robustness to background noise is essential.

Similarly, retraining the model on the dataset containing paradigm recordings in their original format (Iter. 7) increased error rates, likely due to the single-word-per-recording structure. Although ASR models can be fine-tuned on various utterance types, (Trabelsi et al., 2024) demonstrated that including single-word recordings can negatively affect the recognition of multi-word utterances. The inability to use such recordings—particularly those with high audio quality and verified transcriptions—is undesirable in low-resource settings, which motivated our concatenation approach described above.

Beyond these dataset optimization strategies, we examined whether specific linguistic and speaker characteristics affected transcription accuracy. Based on our earlier experiments with Whisper-small conducted before this study, we anticipated that proper nouns and code-switching would pose challenges for the model. Specifically, Whisper-small had difficulties recognizing Saskatchewan- and Canada-related proper nouns (e.g., Turnor Lake, Ile-a-la-Crosse, Manitoba), and speech with code-switching to English often resulted in fully English transcriptions. However, contrary to our expectations, neither factor significantly affected transcription accuracy in any Whisper-medium iteration. For instance, out of 76 English words in the test set 66 were transcribed as English words (though not always correctly), 8 were transcribed as Dënë Sųłiné words (e.g., *nowadays* → *now dë́* '(lit.) now if/when'), and one was omitted entirely. In the reverse direction, only two Dënë Sųłiné words were transcribed as English (e.g., *bazë́* 'regarding' → *pause*. This improvement in code-switching detection, along with better transcription of proper nouns, may be attributable to the more optimal model size and the larger training set used in this study.

Fully English sentences had significantly better transcriptions during the first three iterations, but later this effect had disappeared. It is most likely related to the fact that, after some standardization and data reviewing, the model improved at transcribing Dënë Sųłiné to the point where fully English

sentences were no longer advantaged. Speaker gender did not affect transcription quality in any iteration, despite the training dataset's gender imbalance (68% female speakers). However, speaker age showed a sporadic effect on WER, reaching significance only in Iterations 5 and 8. The changes made to the dataset before these iterations were unlikely to have triggered this effect. Hence, given the inconsistent pattern across iterations and the absence of a theoretical explanation for such a result, this may reflect a Type I error due to multiple comparisons rather than an actual effect. Further testing on a larger dataset is required to determine whether speaker age really affects WER.

7 Conclusions

Overall, this study contributes to the growing discourse on data-centric approaches for low-resource datasets and aims to inspire further exploration of computational methodologies for enhancing noisy, resource-constrained data. While the importance of data quality is widely acknowledged, our work provides quantitative evidence of exactly how much data-centric approaches can improve model training outcomes. Our findings demonstrate that effective ASR dataset preparation for an under-resourced language should prioritize quality of transcriptions and spelling standardization (if applicable) over audio enhancement. Crucially, out results suggest that the researchers working with imperfect audio recordings may use them for fine-tuning the Whisper-medium model without dramatically compromising performance. These insights require validation across other multilingual pre-trained ASR architectures, including Wav2Vec2 (Baevski et al., 2020), to establish their broader applicability.

As Jimerson et al. (2018) note, under-resourced communities need to make informed decisions about how to invest their resources when developing ASR tools for their languages. This also applies to time investment during dataset preparation: most language communities have unlimited time and financial resources for data curation, making it essential to identify which preprocessing steps lead to the best results. Through this study, we hope to have shown which aspects of dataset preparation should be prioritized, particularly for languages with mixed-quality datasets. Moreover, our work has revealed specific features of Whisper's behaviour—such as the minimal optimal recording length in training datasets—that are relevant to decisions about dataset structure. We hope that these insights will help language communities, linguists, and developers prioritize their dataset development efforts more efficiently.

In our work on the *Talking Dene* corpus, we plan to continue standardizing orthography to further improve transcription quality. We also plan to investigate the model's feasibility for retranscribing yet unstandardized transcriptions to reduce the corpus-reviewing workload and accelerate processing of the *Talking Dene* corpus. Finally, we plan to test the model with Dënë Sųłıné speakers and assess its effectiveness for real-life transcription scenarios.

Limitations

This study has several limitations that should be acknowledged. First, our orthographic standardization efforts require continuation to yield more statistically significant and comprehensive results. The current findings, while promising, represent only an initial exploration. Second, our analysis was conducted using only one ASR architecture: the Whisper-medium acoustic model. To establish the generalizability of our findings, validation across other state-of-the-art ASR architectures, such as Wav2Vec2, is necessary. The performance patterns observed here may be model-specific and require broader empirical validation. Finally, this study did not provide solutions to the out-of-vocabulary (OOV) problem arising from a combination of the morphological richness of Dënë Sųłıné verbs and orthographic inconsistencies in their spelling, as addressing this challenge was beyond the scope of the current work. However, we recognize this as a critical limitation and plan to investigate potential solutions in future research.

Ethical considerations

This study is approved by the University of Saskatchewan Board of Ethics (Beh-REB-4918). All speakers participating in this research gave their explicit consent for the use of their audio recordings for the Dënë Sųłıné Automatic Speech Recognition model training. The model cannot be made publicly available until the Clearwater River Dene Nation and La Loche communities decide how they want to distribute it.

Acknowledgments

We are grateful to the Clearwater River Dene Nation and La Loche (SK, Canada) Dene communities for the opportunity to work with their language. We especially want to thank the research assistants from the Clearwater River for their help in the data collection and transcription for this study: Trina Lemaigre and Chastity Sylvestre. Moreover, we want to thank all participants, whose recordings were used for the training of the Automatic Speech Recognition model (some referred to by pseudonym): Rebecca Dene, Teresa Dene, Mitchell Guetre, Gerald E. Haineault, Brenda Herman, Rhonda Herman, Sharon Kennedy, Allison Lemaigre, Andrea Lemaigre, Antoinelle Lemaigre, Edainya Lemaigre, Jeannie Lemaigre, Jennifer Lemaigre, Johny Lemaigre, Mikki Lemaigre, Miranda Lemaigre, Randall Lemaigre, Taitlyn Lemaigre, Taylon Lemaigre, Tina Lemaigre, Trina Lemaigre, Tyanne Lemaigre, Doreen Moise, Ernie Piche, Heather Piche, Ursula Piche, and Jeff Toulejour. We also want to thank Nial Willems for his help with verb-paradigm reviewing and for providing his reviewed transcriptions and recordings for this study. This study was funded by the SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages". The data collection for the *Talking Dene* study was funded by the SSHRC Insight Grant 435-2020-1197.

References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively Multilingual Adversarial Speech Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Altruki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9.

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-end automatic speech recognition: Its impact on the workflow in documenting Yoloxóchitl Mixtec. In *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222. European Language Resources Association.

Alexei Baevski, Henri Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on Neural Information Processing Systems*, pages 12449–12460, Vancouver, Canada.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Nikita Bhatt, Nirav Bhatt, Vishal Sorathiya, Samah Alshathri, and Walid El-Shafai. 2024. A data-centric approach to improve performance of deep learning models. *Scientific Reports*, 14.

Eung-Do Cook. 2004. *A grammar of Dëne Sųłiné (Chipewyan)*. Number 17 in Algonquian and Iroquoian Linguistics. University of Manitoba, Winnipeg.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3863–3867, Portorož, Slovenia. European Language Resources Association (ELRA).

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers. In *Proceedings of Interspeech 2023*, pages 2798–2802.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. 2024. Data-centric artificial intelligence. *Business & Information Systems Engineering*, 66:507–515.

Robbie Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1008–1016.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for Documenting Acutely Under-Resourced Indigenous Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dagmar Jung, Olga Lovick, Alison Lemaigre, Jakaterina Mazara, and Olga Kriukova. 2025. Mixed constructions across ages: Comparing two Dene corpora [Conference presentation]. Presented at the Conference of the Society for the Study of the Indigenous Languages of the Americas (SSILA), January 2025.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proceedings of the 6th Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55.

Olga Kriukova, Gabrielle Fontaine, Alison Lemaigre, Dagmar Jung, Antti Arppe, and Olga Lovick. 2026. Using automatic speech recognition to assist with standardization of Dënë Sųłıné transcripts. *(Submitted)*.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82:1–26.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. *arXiv preprint*. ArXiv:2506.17459 [cs].

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves NLP performance for Nigerian pidgin. In *LREC-COLING 2024*, pages 11510–11522.

Zoey Liu, Crystal Richardson (Karuk), Richard Hatcher Jr, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:3933–3944.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131.

Olga Lovick, Dagmar Jung, Olga Kriukova, Allison Lemaigre, and Barb Hannah. 2023. Variation and change across generations in current Dene: Reduction in verbs [Conference presentation]. Presented at the Annual Conference of the Canadian Linguistic Association, Toronto, Canada, May 2023.

Olga Lovick, Dagmar Jung, Olga Kriukova, Allison Lemaigre, and Barb Hannah. 2024. Variation and change in Dene verbs [Conference presentation]. Presented at the Conference of the Society for the Study of the Indigenous Languages of the Americas (SSILA), New York, USA, January 2024.

Sarah Luger, Rafael Mosquera-Gómez, Alex Miłowski, Thom Vaughan, Sara Hincapie-Monslave, Pedro Ortiz Suarez, and Kurt Bollacker. 2025. Building data infrastructure for low-resource languages. In *Proceedings of the 8th Workshop on Technologies for Machine Translation of Low-Resource Languages*, pages 154–160.

Lingwei Meng, Jiawen Kang, Yuejiao Wang, Zengrui Jin, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Empowering Whisper as a joint multi-talker and target-talker speech recognition system. In *Proc. Interspeech 2024*, pages 4653–4657.

Patrick James Moore. 2018. Re-valuing code-switching: Lessons from Kaska narrative performances. In Julia Christensen, Christopher Cox, and Lisa Szabo-Jones, editors, *Activating the heart: storytelling, knowledge sharing, and relationship*, pages 53–88. Wilfrid Laurier University Press.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Zarif al Sadeque. 2022. Automatic speech recognition for documenting endangered First Nations languages. Master's thesis, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not

the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Mark Simmons. 2025. Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper. In *Proceedings of the 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 155–161, Honolulu, HI, USA.

Statistics Canada. 2021. Mother tongue by geography, 2021 Census.

Asma Trabelsi, Laurent Werey, Sebastien Warichet, and Emanuel Helbert. 2024. Is noise reduction improving open-source ASR transcription engines quality? In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*, volume 3, pages 1221–1228.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 993–1003. Association for Computational Linguistics.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, 32:791–813.

Nial Austen Willems. 2025. The ts'ë- passive in Dëne Sųtłıné. Master's thesis, University of Saskatchewan, Saskatoon, Canada.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 306–315, Marseille, France. European Language Resources association.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 5071–5079, Marseille, France.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research. *arXiv preprint*. ArXiv:2412.04497 [cs] version: 2.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced Aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 4004–4011, Slovenia.

Towards a Community-accessible Cahuilla corpus: Developing HTR for J.P. Harrington's handwritten fieldnotes on Mountain Cahuilla

Ray Huaute *
University of California, Los Angeles
ray.huaute@ucla.edu

Jacqueline Brixey *
University of Wisconsin-Madison
brixey@wisc.edu

Abstract

This paper describes ongoing work to develop a corpus of Cahuilla language from the John Peabody Harrington collection, which contains linguistic and ethnographic fieldnotes documenting Indigenous languages of California and other regions across the Americas. Handwritten notes present numerous processing challenges, including scratch-outs, multilingual entries in Spanish and other Indigenous languages, unique abbreviations, and varying script orientations. We compare the efficacy of deep learning text recognition models to convert images of the notes into a machine-readable format, with a focus on respecting tribal data sovereignty in our methods. We find that Pylaia is the most accurate model for our data. Finally, we present the preliminary findings and indicate future directions for developing a Cahuilla corpus.

1 Introduction

The John Peabody Harrington (J.P., for short) collection represents a monumental record of over 100 Indigenous languages (Harrington, 1907). While the collection has been digitized by the Smithsonian Institution[1], most of the handwritten notes have not been converted into a machine-readable format. Of interest to this paper is the estimated 6,000 pages of notes on the Cahuilla language, an endangered language of Southern California (Simons and Fennig, 2018).

The goal of this work is to create and share a searchable corpus of the Harrington notes on Cahuilla with the Cahuilla community. A searchable corpus will support linguistic and downstream Natural Language Processing (NLP) research, as well as support community goals in the urgent work of language revitalization. A secondary objective is to create a broadly applicable and replicable model for transcribing the remainder of the J.P. Harrington fieldnotes. This initiative has the potential to substantially assist the more than 100 Indigenous communities with whom Harrington collaborated in gaining access to a wealth of knowledge about their languages. This paper presents preliminary work to identify the most accurate and efficient Handwritten Text Recognition (HTR) approach for converting the Cahuilla notes in the collection into a machine-readable format. We also describe a data sovereignty framework for working with the Cahuilla community that can be applied more generally to similar projects with other Indigenous language communities.

2 Overview of Cahuilla Language and Tribal Communities

Cahuilla (chl - ISO 639-3, henceforth referred to by the endonym: ˈiviʌuʔat) is a Native American language of Southern California with few first-language speakers remaining. Today, ˈiviʌuʔat is being spoken and reclaimed across many Cahuilla reservations and communities[2]. ˈiviʌuʔat, along with Luiseño, Acjachemem (Juaneño), and Cupeño, comprise the Cupan sub-group of languages that are part of the larger Takic branch of Uto-Aztecan languages (Hill and Hill, 2019). There are three dialects of ˈiviʌuʔat: Mountain, Desert, and Pass or Wanakik. ˈiviʌuʔat is an agglutinative, head-final language with SOV word order (Seiler, 1977).

The orthography in the J.P. Harrington collection is the Americanist Phonetic Notation (APN)[3].

*Equal contribution

[1] https://sova.siedu/record/naa.1976-95/contents

[2] There are currently nine federally recognized tribes that identify themselves as Cahuilla: Agua Caliente Band of Cahuilla Indians, Augustine Band of Cahuilla Indians, Cabazon Band of Mission Indians, Cahuilla Band of Indians, Los Coyotes Band of Cahuilla and Cupeño Indians, Morongo Band of Mission Indians, Ramona Band of Cahuilla Indians, Santa Rosa Band of Cahuilla Indians, and Torres Martinez Desert Cahuilla Indians.

[3] Also known as the North American Phonetic Alphabet

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 107–114

For this project, these characters will be converted into a code suitable for integration into text files within our database, and later transliterated into the community-preferred orthography (see (Huaute, 2023) for orthographies).

3 Overview of J.P. Harrington Collection

Recognized as one of the most prolific documentarians of California Indian languages, J.P. Harrington compiled over one million pages of cultural and linguistic fieldnotes covering more than 135 languages in California and the Far West from 1915 to 1954 (Mills and Ann, 19876). However, the communities whose ancestors collaborated with him have yet to achieve comprehensive access to, or benefit from, this valuable knowledge, a gap this project seeks to address.

Numerous scholars and workshops over the years aimed to make the data easily searchable[4] (Golla, 1991). A large-scale effort to manually transcribe, annotate, and format the collection into a database was undertaken at the University of California, Davis (Macri, 2010), resulting in the transcription of 67 reels of data, or roughly 235,000 sentences, representing 16 languages[5]. The main work concluded in 2013, and while the resulting database is not publicly available, the transcribed files are available upon request[6]. To the best of our knowledge, no prior work has attempted to convert any portion of the JP Harrington images into a machine-readable format using HTR techniques.

For this project, we aim to convert approximately 6,000 pages on the Mountain Cahuilla dialect. The bulk of the entries from this series were provided by Adan (Adam) Castillo, a Mountain Cahuilla speaker from the Soboba Indian Reservation.

3.1 Challenges of the collection

A significant challenge is that the notes are handwritten, with characters of varying shapes and sizes. An example page is shown in Figure 2. Harrington also wrote sporadically in cursive and in varying orientations, both of which automatic recognition approaches often struggle with (Khan et al., 2023; Pavlenko and Blackledge, 2004).

An additional challenge is that the collection is multilingual, with English, Spanish (Anderton, 1991), and multiple Indigenous languages present. Finally, Harrington's prolific use of abbreviations (Woodward and Macri, 2005) and inconsistencies in orthographic representation raise interpretability issues.

4 Review of Relevant Literature

4.1 Data sovereignty

In response to Indigenous communities' concerns about the development of large data centers and the expansion of AI technologies (Cox, 2025), we adopt data sovereignty as a foundational principle guiding this project. Central to the Indigenous Data Sovereignty movement is the idea of self-determination for Indigenous peoples and their sovereign right to own and control their data, which, in this work, includes linguistic data (Holton et al., 2022). The development of international standards for data sovereignty governance, such as the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) (Carroll et al., 2023), are also informative for our data sovereignty policy. Finally, a recent publication (Holton et al., 2022) notes potential licensing issues associated with "Terms of use" clauses when using third-party apps, such as Google Drive and iCloud. This was a consideration as we reviewed large platforms for our project, such as Transkribus (Kahle et al., 2017), eScriptorium (Kiessling et al., 2019), and OCR4ALL (Reul et al., 2019). However, many platforms we reviewed were not explicit and transparent in how shared data is stored and protected, leading us to not pursue these platforms as a viable approach.

4.2 Text recognition

Deep learning models, such as convolutional neural networks (CNNs) (Alam et al., 2025), recurrent neural networks (RNNs) (Keshri et al., 2018), and, more recently, the two combined as convolutional recurrent neural networks (CRNNs) (Shi et al., 2016), are effective at HTR tasks (Idris and Taha, 2022; AlKendi et al., 2024; Balci et al., 2017; Dash et al., 2024). Important aspects of the Harrington collection are multilingualism, abbreviations, and cursive, which deep learning models have demonstrated the ability to recognize (Alam et al., 2025; Romein et al., 2025; Al-Saffar et al., 2021).

Some popular online platforms that utilize HTR have been developed specifically for converting

(NAPA)

[4] http://www.rock-art.com/jph/nl04.htm

[5] Obtained via personal communication

[6] https://nas.ucdavis.edu/jp-harrington-database-project

archival documents to machine-readable text (e.g., Transkribus, OCR4all) but may not be universal, as we find that they have mostly been trained on large Indo-European languages. Automatic text recognition has meaningfully aided in the documentation of other Indigenous languages with digitized text resources (Carrera et al., 2024; Agarwal and Anastasopoulos, 2025, 2024).

5 Methods

Due to the collection's volume and the time and effort required for manual transcription, we propose developing an HTR approach to efficiently convert the scanned images into a machine-readable format. In this current work, we implemented three deep learning models that align with our data sovereignty policy. We compare the models to determine the most accurate approach, potentially reducing the need for additional post-processing steps.

5.1 Data Sovereignty Policy

To ensure maximal collective benefit of our project, our primary goal is to generate a machine-readable and searchable database of the J.P. Harrington ˈiviʌuʔat fieldnotes that can be provided to Cahuilla community members in a format they can easily access and understand.

Given the potential issues indicated in Section 4.1 and (Holton et al., 2022), we utilized only locally hosted models for this stage of the project. At later stages, we will work closely with the Cahuilla communities to develop a plan for data storage, curation, and access protocols. This approach aligns with the second CARE principle, authority to control (Carroll et al., 2023).

5.2 Preprocessing Images

We used a sample of 66 pages for our experiments. All downloaded PDF files[7] were converted into JPEG files of each page. We implemented a Python interface to create sliced images by selecting just a word, line, or single letter from a given JPEG. This approach allowed us to resolve issues in orientation changes and to omit scratch-outs. However, this resulted in slices of different dimensions.

The Pylaia and Jax libraries automatically correct image size differences. For our baseline CRNN model, we added a padding strategy to standardize the sizes. Figure 1 illustrates padding for 3 different slices.

[7] https://sova.si.edu/record/naa.1976-95/search?q=cahuilla&t=W&o=doc_position

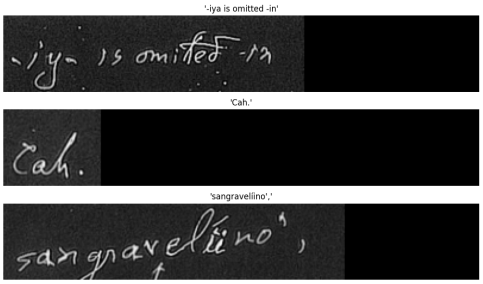

Figure 1: Example of padding to standardize image "slices" sizes. Slices can be lines, words, or singular characters.

5.3 Training and testing data

Next, we created ground-truth labels for each slice. Decomposed characters, such as s̥, were substituted for a composed character for ease and accuracy.

We then reviewed the dataset for word and character frequencies. There are 390 unique and 845 total words in the training set. "Cah", shorthand for "Cahuilla", appears the most (46 times). The majority of words are in English, 46 in Cahuilla (5% of total), and 21 in Spanish (roughly 2%).

The prediction character set contains at least 86 glyphs, including upper and lowercase English, Spanish, Cahuilla, and some IPA characters. The most commonly occurring characters are: space (369), e (310), a (284), o (262), and t (230). Thirteen characters appear only once. Not present in the data are: U, V, X, Z, z, 6, and Cahuilla ʌ.

The 66-page sample resulted in 494 slices. We allocated roughly 5% of the slices as a limited evaluation set. Given the small set of training examples at this preliminary stage of the research, we proposed comparing the models' performance on an unbalanced (raw) data set to that on a balanced data set. The merits of using the real-world representation (i.e., unbalanced) of characters are that the models will focus on learning the most highly occurring characters. In contrast, oversampling, or creating a balanced data set, is thought to encourage models to learn all characters (Kaur et al., 2019). As it is crucial to recognize Cahuilla, which occurs less frequently, we proposed testing the models on both balanced and unbalanced datasets. We created a balanced dataset using a Python script, ensuring that each character appears at least 60 times.

5.4 Models

All experiments were completed on a MacBook Pro M2. Training and testing sets were manually configured so that model results would be compara-

ble. At this early research stage, we did not explore additional fine-tuning of the models.

1. Baseline CRNN We implemented a CRNN model in Python using the TensorFlow (Abadi et al., 2015) and Keras (Chollet et al., 2015) libraries. The model is a three-block CNN with ReLU activations. The recurrent component is a single bidirectional LSTM.

2. Pylaia: Pylaia (Tarride et al., 2024) is a popular Python text recognition library that uses the deep learning framework PyTorch (Paszke, 2019). Pylaia powers historical text documentation platforms like Transkribus (Park, 2025).

Our Pylaia model comprises a four-block CNN, each block incorporating LeakyReLU activations and batch normalization. The recurrent component consists of three bidirectional LSTM layers.

3. Jax: Jax (Bradbury et al., 2018) is a Python library recently created by Google that is optimized for memory usage for large machine learning model development (Sapunov, 2024). Limited previous research using Jax has shown promising performance on computer vision tasks, such as medical imaging classification (Bećirović et al., 2025) and recognizing handwritten structured medical notes (Kale et al., 2025). Given that it can be run locally, we proposed to include it in our comparison experiments.

Our model consists of a 3 CNN layers followed by a bidirectional GRU and a linear classification layer trained with CTC loss. The model is optimized with Adam using a warmup schedule and gradient clipping.

6 Results

We found that the baseline CRNN approach was the most time-efficient of the models; the Pylaia and Jax models took 6-10 hours to train and test on the balanced dataset. Pylaia also frequently ran into issues with exceeding memory allocation. It may be a consideration for language communities with fewer technological resources to know the time and memory requirements for each model. We also reviewed the models' performance for overall accuracy and by language.

1. Overall performance: The results are given in Table 1. Pylaia was the best model in terms of both overall word error rate (WER) and character error rate (CER), and then the baseline CRNN model and Jax on the balanced data. Performance across the models mostly declined on the unbalanced data. This indicates that the balanced data had a positive effect on the models.

2. Performance by language: Next, we reviewed each model's performance by language (also in Table 1). Again, there was an improvement by all the models when using the balanced data. No model had zero errors on the unbalanced data. It is assumed that a language model is produced as a result of the RNN layer(s) in a CRNN (Dash et al., 2024). Both the Pylaia and baseline CRNN use LSTMs as the RNN layers; it is notable how much better the Pylaia model recognizes letters overall and performs on the two languages that are underrepresented in the data than the baseline CRNN.

7 Conclusions and Future Work

This initial work towards creating a corpus of the Cahuilla language compared the performance of deep learning models in converting handwritten text into a machine-readable format. We found that the Pylaia model achieved lower CER than a baseline CRNN and Jax models when trained and tested on the same data.

In future work, we will prepare additional training data covering the missing letters indicated in Section 5.3. Our results indicate that we should use the balanced data for the rest of the HTR work. We will explore computing power resources (such as servers or a more powerful computer) that align with our data sovereignty policy in the next step using Pylaia, as we anticipate that the computer used for the experiments in this paper will be insufficient for a larger training set. We will also consider post-processing correction approaches.

Our data sovereignty policy will continue to be refined and determined in future steps. *Collaborative consultation* with Cahuilla communities for guidance, reflection, and sharing throughout the project will ensure that researchers behave ethically and follow cultural protocols throughout the project and post-project, ensuring responsibility to the community (Leonard and Haynes, 2010). Our data management plan will include provisions for using cloud and server storage solutions that respect data sovereignty. Finally, we acknowledge the importance of maintaining flexibility and will revise the project's data sovereignty policy in close consultation with the Cahuilla community.

Limitations

Our data sovereignty policy helped to determine the models selected. LLMs were necessarily excluded because they did not align with the policy. We note that the computer used for the experiments was a limitation, as it was not the most powerful or the most recent MacBook. The performance of the models, especially with regard to running times, would differ on a different (more powerful) computer. However, we also recognize that technological limitations may be a factor for other Indigenous communities considering this type of work, who may have limited financial capacity and access to powerful computing resources.

As there are few HTR publications using Jax to compare our results to, we are limited in our ability to hypothesize about the model's performance; it may be attributable to some aspect of our limited sample data.

Acknowledgments

The authors wish to acknowledge the significant contributions of renowned linguist and ethnologist JP Harrington and his diligent and knowledgeable Cahuilla speakers and collaborators whose efforts provided the data utilized in this paper. We also thank the anonymous reviewers for their helpful feedback.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, and 21 others. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

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.

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.

Ahmed Al-Saffar, Suryanti Awang, Wafaa Al-Saiagh, Ahmed Salih Al-Khaleefa, and Saad Adnan Abed. 2021. A sequential handwriting recognition model based on a dynamically configurable crnn. *Sensors*, 21(21):7306.

Mahanur Alam, Md Johirul Islam Tutul, Md Anwar Hussen Wadud, Md Jakir Hossen, and MF Mridha. 2025. Bilingual Bangla ocr for rural empowerment: Detecting handwritten queries and agricultural assistance. *IEEE Open Journal of the Computer Society*.

Wissam AlKendi, Franck Gechter, Laurent Heyberger, and Christophe Guyeux. 2024. Advancements and challenges in handwritten text recognition: A comprehensive survey. *Journal of Imaging*, 10(1):18.

Alice J Anderton. 1991. Kitanemuk: Reconstruction of a dead phonology using John P. Harrington's Transcriptions. *Anthropological Linguistics*, pages 437–447.

Batuhan Balci, Dan Saadati, and Dan Shiferaw. 2017. Handwritten text recognition using deep learning. *CS231n: convolutional neural networks for visual recognition, Stanford University, Course Project Report, Spring*, pages 752–759.

Merjem Bećirović, Amina Kurtović, Nordin Smajlović, Medina Kapo, and Amila Akagić. 2025. Performance comparison of medical image classification systems using tensorflow keras, pytorch, and jax. *arXiv preprint arXiv:2507.14587*.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Shadya Sanchez Carrera, Roberto Zariquiey, and Arturo Oncevay. 2024. Unlocking knowledge with ocr-driven document digitization for Peruvian indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 103–111.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, and 1 others. 2023. The care principles for indigenous data governance. *Open Scholarship Press Curated Volumes: Policy*.

François Chollet and 1 others. 2015. Keras. `https://keras.io`.

Evelyn Cox. 2025. AI in a tribal context: A brief review of the literature.

Saswata Kumar Dash, Sompalli Pranay, Pervela Hemanth, and Aravindkumar Sekar. 2024. Multi-lingual

handwritten recognition using convolutional recurrent neural networks. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–7. IEEE.

Victor Golla. 1991. Introduction: John P. Harrington and his legacy. *Anthropological Linguistics*, pages 337–349.

John Peabody Harrington. 1907. Volume three: A guide to the field notes: Native american history, language, and culture of Southern California/basin. *The Papers of John Peabody Harrington in the Smithsonian Institution*, 1957.

Jane H Hill and Kenneth C Hill. 2019. Comparative takic grammar.

Gary Holton, Wesley Y Leonard, and Peter L Pulsifer. 2022. Indigenous peoples, ethics, and linguistic data. *The open handbook of linguistic data management*, pages 49–60.

Incamu Ray Huaute. 2023. *Topics in the phonology and morphology of Torres Martinez Desert Cahuilla*. Ph.d. dissertation, University of California, San Diego.

Ahmed A. Idris and Dujan B. Taha. 2022. Handwritten text recognition using crnn. In *2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pages 329–334.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th iapr international conference on document analysis and recognition (icdar)*, volume 4, pages 19–24. IEEE.

Apoorwa Kale, Yash Khandelwal, Vibhor Pandhare, Atreyee Ghosh, Nidhi Pathak, Bhure Singh Saitya, and Bhupesh Kumar Lad. 2025. A scalable, low-cost framework for multilingual intelligent document processing for continuity of care. In *IET Conference Proceedings CP942*, volume 2025, pages 161–166. IET.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM computing surveys (CSUR)*, 52(4):1–36.

Pooja Keshri, Prabhat Kumar, and Rajib Ghosh. 2018. Rnn based online handwritten word recognition in Devanagari script. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 517–522. IEEE.

Sulaiman Khan, Shah Nazir, and Habib Ullah Khan. 2023. Analysis of cursive text recognition systems: A systematic literature review. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–30.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: an open source platform for historical document analysis. In *2019 international conference on document analysis and recognition workshops (icdarw)*, volume 2, pages 19–19. IEEE.

Wesley Y Leonard and Erin Haynes. 2010. Making "collaboration" collaborative: An examination of perspectives that frame linguistic field research.

Martha J Macri. 2010. Working with language communities in unarchiving: Making the JP harrington notes accessible. In *Language Documentation: Practice and values*, pages 213–220. John Benjamins Publishing Company.

Elaine Mills and J Brickfield Ann. 19876. The papers of john peabody harrington in the smithsonian institution, 1907-57. *Millwood NY: Kraus International*.

Fiona Park. 2025. What are super models and how do they work?

A. et al. Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Aneta Pavlenko and Adrian Blackledge. 2004. Introduction: New theoretical approaches to the study of negotiation of identities in multilingual contexts. In Aneta Pavlenko and Adrian Blackledge, editors, *Negotiation of Identities in Multilingual Contexts*, pages 1–33. Multilingual Matters LTD.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. Ocr4all—an open-source tool providing a (semi-) automatic ocr workflow for historical printings. *Applied Sciences*, 9(22):4853.

Christel A Romein, Achim Rabus, Gundram Leifert, and Phillip Benjamin Ströbel. 2025. Assessing advanced handwritten text recognition engines for digitizing historical documents: Romein et al. *International journal of digital humanities*, 7(1):115–134.

Grigory Sapunov. 2024. *Deep learning with JAX*. Simon and Schuster.

Hansjakob Seiler. 1977. *Cahuilla grammar*. Banning: Malki Museum Press.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Gary F. Simons and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-first edition. SIL International, Dallas, Texas.

Solène Tarride, Yoann Schneider, Marie Generali, Melodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving automatic text recognition with language models in the pylaia open-source library. In *Submitted at ICDAR*.

Lisa L Woodward and Martha J Macri. 2005. JP harrington database project: an archival resource for anthropologists, archaeologists, and Native communities. *Journal of California and Great Basin Anthropology*, 25(2):235–240.

A Appendix

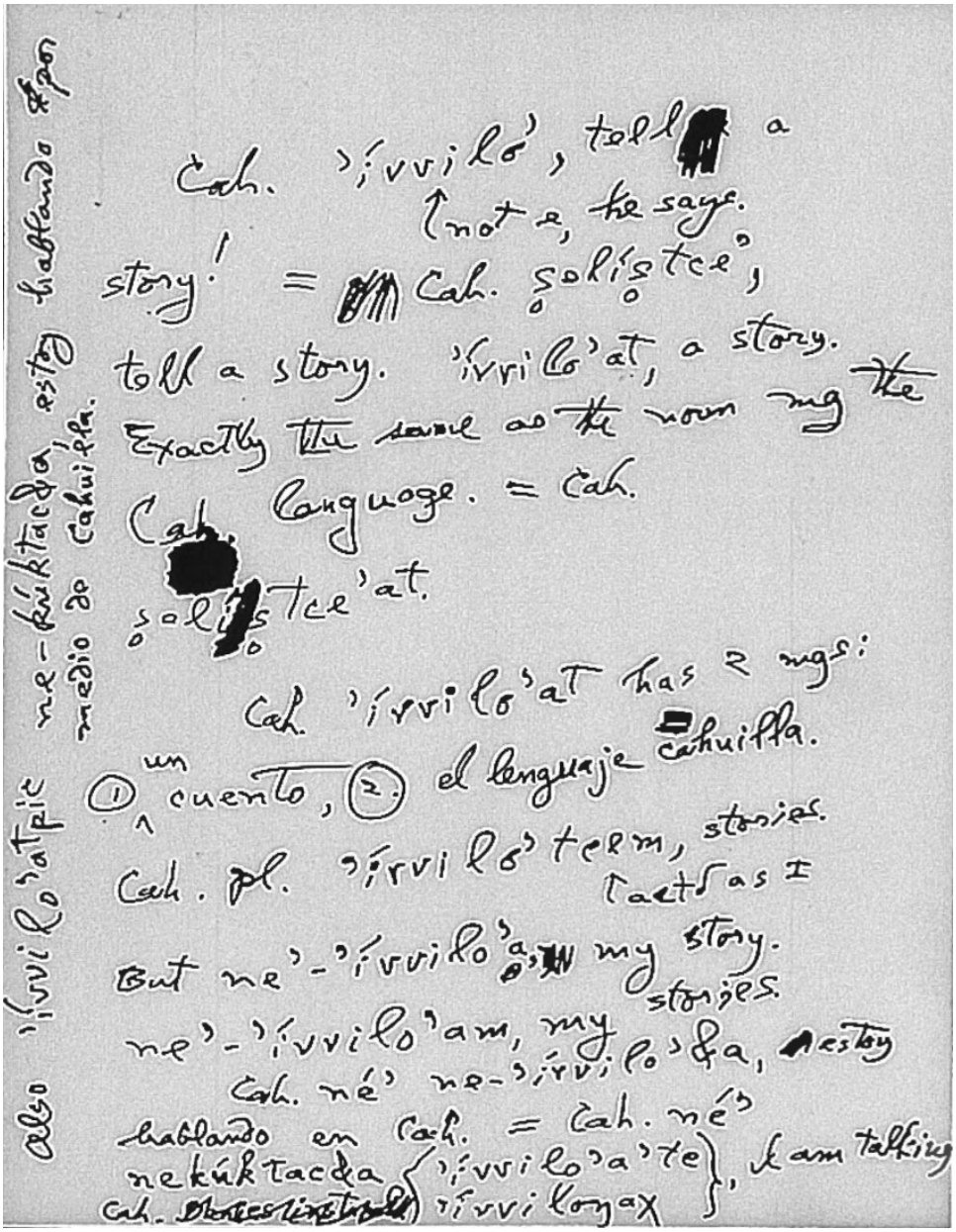

Figure 2: A page from the J.P. Harrington collection, demonstrating some of the variations in orientation, scratch-out, multilingualism, and cursive challenges unique to the data.

	CER				WER			
	Overall	English	Spanish	Cahuilla	Overall	English	Spanish	Cahuilla
Balanced								
CRNN	75.64	72.29	69.23	90	128.57	95	300	200
Pylaia	9.62	25.61	0	0	21.43	30	0	0
Jax	83.97	83.33	84.62	83.13	107.14	242.86	100	100
Unbalanced								
CRNN	80.13	78.31	84.62	86.67	103.57	95	200	114.29
Pylaia	61.69	51.22	69.23	81.67	75	48.78	100	100
Jax	96.15	95.18	92.31	98.33	100	100	100	100

Table 1: Results for each model of the character error rate (percent) and word error rate (percent) in each language.

Corpora duplication for NLP in low-resource languages: A case study of Nahuatl

Juan-José Guzmán-Landa[1], Juan-Manuel Torres-Moreno[1], Luis-Gil Moreno-Jiménez[3], Elvys Linhares Pontes[4], Miguel Figueroa-Saavedra[2], Graham Ranger[1], Martha-Lorena Avendaño-Garrido[2]

[1]Université d'Avignon (France), [2]Universidad Veracruzana (Mexico), [3]Independent Researcher (France), [4]Trading Central Labs

Correspondence: juan-manuel.torres@univ-avignon.fr

Abstract

In this paper, we aim to answer the following question: could corpus duplication be useful in Natural Language Processing (NLP) for low-resource languages? In these languages (or π-languages), corpora available for training Large Language Models are virtually non-existent. Specifically, we study the impact of corpus expansion in Nahuatl, an agglutinative and polysynthetic Amerindian π-language characterised by extensive dialectal variation. Our goal is to increase the size of Nahuatl corpora, which currently consist of a limited number of tokens, through controlled duplication techniques. Our experimental setup employs incremental duplication alongside appropriate corpus balancing, with the objective of training embeddings optimised for downstream NLP tasks. Consequently, static embeddings were trained and evaluated on a sentence-level semantic similarity task. Our results show a significant improvement in performance when incremental duplication is applied, compared to results obtained without corpus expansion. To our knowledge, this technique has not yet been explored in this field.

1 Introduction

It is well established that Large Language Models (LLMs) require training corpora comprising substantial volumes of textual data in order to acquire a deep contextual understanding of linguistic structures and usage. These amounts often run into the hundreds of millions or even billions of words. Furthermore, it has been found that performance increases logarithmically with corpus size (Kaplan et al., 2020). This massive data requirement implies a major problem for the development of LLMs trained on languages with few computational resources (π-languages), as opposed to τ-languages or languages with abundant resources (Berment, 2004; Abdillahi et al., 2006). Indeed, π-languages suffer from a severe lack of representative, large-scale textual corpora, making it impossible to train LLMs adequately. Consequently, these languages remain under-represented in Natural Language Processing (NLP), perpetuating a linguistic bias that limits their usefulness for the communities that speak them. One example of the Americas' π-languages is Nawatl (also known as Nahuatl), one of Mexico's indigenous national languages. In this country, Nawatl has been recognised as the second national language, after Spanish, with approximately 1.65 million Nawatl speakers (INEGI, 2020).

Nawatl has a significant number of dialectal varieties, with 29 recognised varieties spread across four major regions in Mexico: Western, Central, Eastern, and Huasteca[1]. This diversity represents an enormous linguistic challenge for the development of NLP tools, as it involves correctly handling significant variations in spelling and lexical choices (Zimmermann, 2019; Olko and Sullivan, 2016; Hansen, 2024). To solve this, a symbolic unifier for Nawatl spellings has recently been proposed (Guzman-Landa et al., 2025). Although the publication of digital content in Nawatl is constantly increasing, the dispersion and considerable dialectal diversity of this content mean that it cannot easily be included within the few available corpora.

The availability of digital Nawatl documents and their written application are nonetheless essential for the ongoing revitalisation of the language (Pugh et al., 2025). Our approach to addressing the scarcity of corpora involves the controlled duplication of available textual data. Combined with other techniques, this strategy could serve as a basis — in the case of π-languages — for expanding corpora on a larger scale. These corpora, in turn, could be used to train static word embeddings (Tunstall et al., 2022; Goyal et al., 2018). More specifically,

[1]See Ethnologue, 2025: https://www.ethnologue.com and (Lastra de Suárez, 1986).

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 115–127

our objective is to expand the Nawatl π-YALLI corpus[2] sufficiently to generate a positive impact on training models that produce static embeddings.

The structure of the paper is as follows: Section 2 provides a review of corpus expansion techniques in languages with few resources. Section 3 introduces the Nawatl language and the π-YALLI corpus. Section 4 introduces the strategy used in balancing the corpora, and Section 5 the duplication technique. Section 6 presents our experimental setup on a semantic similarity task. Section 7 describes the results. Finally, Section 8 concludes the paper and suggests avenues for future research.

2 Previous works

In the existing literature, research on Nawatl encompasses multiple levels of linguistic analysis. At the morphological level, a finite-state transducer has been employed to model the language's inflectional and derivational processes (Pugh et al., 2021). At the level of dialectal similarity, character-based representations have been explored using an LSTM architecture (Pugh and Tyers, 2021). Furthermore, at the syntactic level, both textual and audio modalities have been integrated to investigate syntactic structure (Pugh et al., 2024). Other models and resources for translation tasks and syntactic and speech analysis in dialects other than Nawatl have been published in (Shi et al., 2021; Gutierrez-Vasques et al., 2025)

Data duplication is found in literature primarily as a problem in τ-languages, rather than as a technique for corpora expansion. Indeed, the massive amount of data on the internet leads to significant redundancy in collected corpora, which negatively impacts model training (Lee et al., 2022; Penedo et al., 2024). Consequently, most research focuses on the detection and removal of duplicates (or deduplication), particularly in the context of corpora intended for LLM training.

This problem is particularly pronounced in τ-languages, where the volume of available text data is substantial but also highly redundant. For this reason, most research aims to produce large-scale training corpora whilst minimizing duplication as much as possible. Thus, FineWeb (Penedo et al., 2024) and CCNet (Wenzek et al., 2020) show filtering and deduplication techniques to produce high-quality, non-redundant corpora.

However, the situation is different for π-languages, where the problem is not an excess of data, but a scarcity of it. In this context, data augmentation (DA) could be an interesting strategy for expanding currently available corpora to compensate for the lack of resources (Feng et al., 2021; Chen et al., 2023). There are two main approaches proposed for DA techniques, in the literature: at the lexical level and at the syntactic level.

2.1 Lexical level DA

The EDA (Easy Data Augmentation) method (Wei and Zou, 2019) performs simple operations such as synonym substitution, as well as the insertion, deletion or random replacement of words. It has been applied exclusively in the context of text classification, and the results show performance ranging from 87.8% to 88.6%, representing improvements of less than 1%. These techniques use dictionaries to deal with synonyms.

2.2 Syntactic level DA

For languages lacking dictionaries, there are some techniques such as EDDA (Easy Distributional Data Augmentation) and TSSR (Type Specific Similar word Replacement) (Mahamud et al., 2023), which utilise distributional context and morphosyntactic labels to address this shortcoming. TSSR requires the data to be annotated with POS[3] tags. EDDA relies on the latent space generated by Word2Vec rather than a dictionary. These techniques have previously been applied to Swedish corpora.

In this article, we propose an approach to corpus expansion that utilises techniques requiring no lexical resources, such as part-of-speech taggers or dictionaries. Indeed, for Nawatl, these resources (where they exist at all) are difficult to apply directly due to the language's high degree of agglutination and polysynthesis. Furthermore, the limited dictionaries available do not cover all dialectal varieties.

Finally, we contend that EDA-type techniques and their stochastic mechanisms can introduce syntactic and semantic biases. Consequently, we aim to avoid such methods in our proposal.

[2]This corpus is available at: https://demo-lia.univ-avignon.fr/pi-yalli

[3]*Part-Of-Speech.*

3 The Nawatl language and the π-YALLI corpus

The Nawatl language is an Amerindian polysynthetic and agglutinative language. In other terms, verbal or nominal root morphemes and a range of inflexional morphemes combine productively to form new "words". At the syntactic level, Nawatl sentences follow a basic **verb–subject–object** (**VSO**) word structures, although this can be flexible. Thus, there are VO, VS, VOS and, less frequently, SV, SVO and SOV word orders (de Durand-Forest et al., 1995; Guzmán-Landa et al., 2026), depending on speakers' needs. Furthermore, the syntactic and semantic relationships between words and clauses are established through the valency of the verb and the use of conjunctive particles. These particles may also function as markers and discourse connectors.

Another distinctive feature of Nawatl is that words can be written as complete sentences, and this is particularly true of predicative words featuring verbs or verbal derivations. We therefore refer to them as phrase-words or "single-word phrases", as their morphology includes the subject and predicate, as well as information on the actants, and modal, directional and relational elements (Launey, 1978; Charles, 2016; Flores Nájera, 2019; Sasaki, 2022). Given its oral nature, there are very few written resources available for this language. Combined with the lack of standardised writing systems, this makes automated processing extremely difficult (Guzmán-Landa et al., 2025).

3.1 Some available Nawatl resources

There are very few tools and textual resources available for the Nawatl language. To our knowledge, only one machine translation tool is available for the Western Huasteca variety[4] since 2024.

In 2017, the *Instituto de Ingeniería* at the *Universidad Nacional Autónoma de México* (UNAM) published *Axolotl*, a bilingual Spanish/Nawatl corpus[5]. Furthermore, a Nawatl spelling unifier (Guzman-Landa et al., 2025) and the new π-YALLI corpus has recently been introduced. However, many dialectal varieties and texts remain inaccessible. This has a detrimental impact on the development of machine learning-based tools, thereby hindering their widespread use and adoption by Nahua-speaking communities.

3.2 The Nawatl π-YALLI text corpus

The π-YALLI corpus (Guzmán-Landa et al., 2025) is a Nawatl text resource available for machine learning and NLP algorithms. It is a heterogeneous corpus covering 16 topics and 26 dialectal varieties of Nawatl, spoken mainly in Mexico and El Salvador. It contains a limited number of words (around 6.6 million) and sentences, but it has been used successfully in various NLP tasks (Guzmán-Landa et al., 2025; Guzmán-Landa et al., 2026).

Despite its limited size, π-YALLI is, however, useful for training vector models: TF-IDF (Manning and Schütze, 1999), BM25 (Robertson et al., 2004), TF-PDF (Bun and Ishizuka, 2002) or static embedding models such as Word2Vec (Mikolov et al., 2013b), FastText (Bojanowski et al., 2017) or GloVe (Pennington et al., 2014), but clearly unsuitable for training contextualised vector models using BERT-style transformers (Devlin et al., 2019).

The acronyms used in this paper concerning the 26 dialectal varieties and the 16 topics, are listed in the Appendix A.1.

4 Statistical corpora balance

It is accepted by the scientific community that corpora must be balanced in order to avoid any bias (Arbach and Ali, 2013). However, the current π-YALLI corpus is not balanced at all (see Figure 1) either topically or dialectally.

We decided to evaluate the impact of balanced corpora on NLP tasks. For this reason, we will use two types of corpora: (i) unbalanced corpora and (ii) statistically balanced corpora (in our case, balanced only in terms of topics and dialectal varieties). Subsequently, both categories will be incrementally duplicated in order to find an optimal duplication ratio ρ that maximises the performance of the models on an NLP task.

Firstly, we proceeded to establish a statistical balancing. In this regard, we introduced two types of corpora balancing: uniform balancing and positional balancing.Both techniques can be applied on topics, dialectal varieties or others corpus categories. Corpus balancing begins by sorting the N categories in descending order, classifying them according to their number of tokens $T_i, i = 1, 2, ..., N$.

[4]See Google Translate https://translate.google.com.mx/?hl=es&sl=nhe&tl=es&op=translate

[5]The *Axolotl* corpus is also available at the following address: http://www.corpus.unam.mx/axolotl

4.1 Uniform balancing

This strategy involves balancing the varying token counts across the N categories (topical or dialectal) relative to the initial value T_1, which contains the largest number of tokens. This enables the $T_i; i = 2, 3, 4, \ldots, N$ tokens within each category to be balanced until they all match the token count of T_1.

Once the process of uniform distribution to the π-YALLI corpus is applied, in the case of the $N = 16$ topics, each one will account for 3.2 million tokens. The resulting corpus will therefore be uniformly balanced, increasing from 6.6 million to **47.9** million tokens. However, we observed that some topics—such as literature (LIT) and history (HIS)— are duplicated fewer than 5 times, while others—such as politics (POL) and music (MUS)—are duplicated more than 1,000 times. This represents a significant increase that could have a major impact on model training. In the case of $N = 26$ dialectal varieties, the new balanced corpus increases from 6.6 million to **31.3** million tokens. The number of tokens for each dialectal variety increases at a different rate, as there are fewer available topics than there are dialectal varieties.

4.2 Positional balancing

In this statistical balancing strategy, we multiply the number of tokens T_i of each topic (or dialectal variety) by their position $i = 1, 2, 3, ..., N$ in the ranking. This allows the N topics (or varieties) to be balanced positionally.

Positional balancing aims to correct the excessive uniform duplication of certain topics (or varieties). Furthermore, in $N = 16$ topic positional balancing, the corpus increases from 6.6 million to **13.9** million tokens. As there are 16 topics, no single topic will be duplicated more than 16 times. For $N = 26$ dialectal varieties, the number of tokens increases from 6.6 million to **28.5** million. A larger number of tokens is obtained because there are more dialectal varieties represented (see Figure 1).

5 Incremental corpora duplication

It has been reported that LLMs require between 10 and 100 million tokens to obtain stable embeddings (Micheli et al., 2020). We therefore decided to expand the π-YALLI corpus using balancing strategies combined with an incremental duplication technique. This study aims to assess the impact of both factors on static embedding training algorithms.

At first glance, such a strategy might appear to have no positive impact on embedding learning. Indeed, it has been found that corpus deduplication is a crucial step in achieving successful embedding learning (Lee et al., 2022). In the case of τ-languages, certain sentences are repeated 60,000 times or more; this poses a significant challenge for dense word representations, as such redundancy often leads to the overfitting of neural models.

Concerning π-languages, in addition to the lack of resources, it should be borne in mind that Nawatl is an agglutinative and polysynthetic language; consequently, the frequent use of compound words reduces the number of what we normally understand as "words", compared with other types of languages. Put differently, what would take five or six words in non-agglutinative languages is expressed in Nawatl with just one. This is very obvious in translation. Ultimately, all of this has an impact on the number of words (tokens) available in the corpora.

However, our hypothesis is that a *controlled* and *moderate* increase in the number of occurrences could facilitate the learning of textual representations in the case of π-languages, and in particular in Nawatl. We decided to empirically test our hypothesis regarding the impact of corpus expansion on learning algorithms. The aim, of course, is to seek a positive impact on the quality of static word embeddings.

6 Experimental setup

The protocol on a semantic task and the experiments concerning the incremental duplication strategy will be detailed in this section.

6.1 Similarity Semantic Task using static embeddings

Semantic similarity, a classic NLP task, involves evaluating various models (statistical, neural networks, etc.) using appropriate evaluation protocols (Francis-Landau et al., 2016). In our study, the aim is to calculate the semantic similarity between the reference sentences and the sets of candidate sentences, which may be semantically close to or distant from the references. This results in rankings of the candidate sentences, which will be compared to rankings produced by 5 native Nawatl speakers, via a statistical estimator. This is the same evaluation

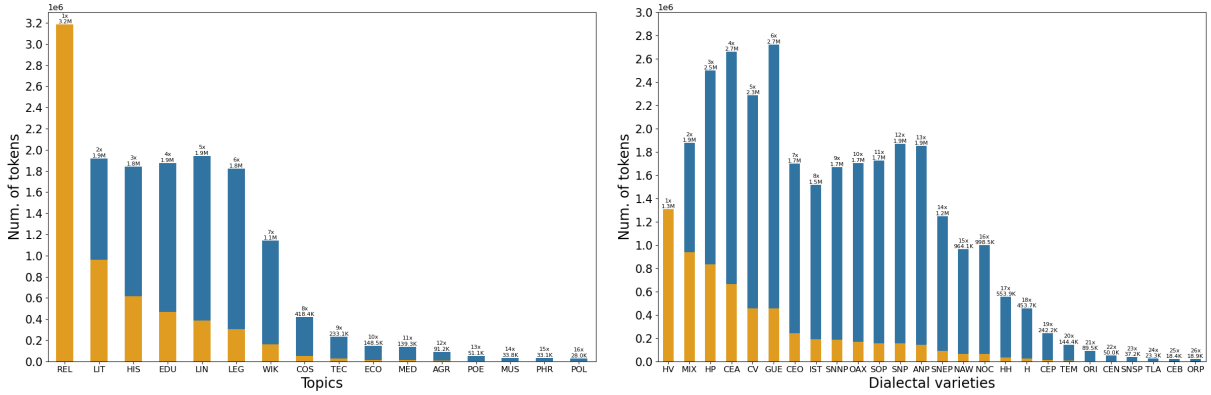

Figure 1: Corpora distribution by topics (at left) and dialectal varieties (at right). Orange bars: original unbalanced corpus. Blue bars: positional balanced corpus (numbers are in millions of tokens; see other details on the Appendix A.5).

protocol found in the recent literature (Guzmán-Landa et al., 2025): 30 reference sentences and 5 candidate sentences per reference. The final ranking of candidates allows us to estimate the impact of incremental duplication on embedding learning and also to measure their quality on a semantic proximity task. An example of sentences used in this semantic task is shown in the Appendix A.2.

Static embeddings have been widely used in NLP tasks (classification, analogies, semantic similarity, etc.), but they have been largely abandoned in favour of transformer-based contextual embedding models (such as the BERT model), whose popularity is due to their excellent performance (Devlin et al., 2019). Although transformers have shown their superiority in NLP tasks, this has only been possible in τ-languages. Indeed, this type of model requires large amounts of textual data to learn effectively. The situation changes completely when it comes to processing π-languages. In this context, non-contextual embeddings are competitive, as they can be generated from scratch, are quick to train and, most importantly, non-contextual models require small corpora to achieve meaningful learning.

Among the popular static embedding training algorithms are Word2Vec (Mikolov et al., 2013a), FastText (Bojanowski et al., 2017), and GloVe (Pennington et al., 2014). Consequently, we employed these algorithms to train word embeddings on our extended (duplicated) corpora. To operate at the vector-sentence level, we compute the average of the word vectors in the all sentences. In the Appendix A.3 we show the training hyper-parameters used for the models. The quality of the embeddings was subsequently evaluated using the aforementioned sentence semantic similarity task, establishing a ranking (Guzman-Landa et al., 2025).

The cosine similarity between each candidate phrase vector $\vec{C}_{i,j}$ and the reference phrase vector $\vec{R}_j$ for a block j, where $i = 1, 2, ..., 5$ and $j = 1, 2, ..., 30$, allows us to calculate a ranking O_j of the candidate phrases. This ranking is compared with the ranking O_j^* produced by human annotators. The correlation between the obtained O_j and ground-truth O^* rankings is evaluated by Kendall's rank correlation coefficient τ.

Kendall's coefficient τ is a non-parametric measure of correlation that assesses the ordinal association between two variables, i.e. the degree of agreement between two rankings (Kendall, 1938).

6.2 Corpus balancing and incremental duplication

In order to expand the original corpus, we incrementally duplicate ρ times the π-YALLI corpus– whether balanced or unbalanced– where $\rho = [1, 2, 4, ..., 28, 30]$ times its original size. The aim is to determine the optimal value of $\rho^\star$ that maximises the efficiency of the models.

On the one hand, in the case of unbalanced corpora, we have incrementally generated corpora ranging from: 6.6 million ($\rho = 1$), 13.2 million ($\rho = 2$), 19.8 million ($\rho = 3$), ..., to approximately 198 million words (duplicated $\rho = 30\times$ times), without regard to topical or dialectal varieties.

On the other hand, using statistically balanced corpora, we have corpora generated by progressive incrementation. Table 1 shows the corpus size in millions of words from starting point (without duplication) where $\rho = 1$, to our final experimental

point with $\rho = 30$ duplications.

Balancing	Topical		Dialectal	
	ρ=1	ρ=30	ρ=1	ρ=30
Uniform	47.9	1437.7	31.3	938.2
Positional	13.9	417.5	28.5	856.3
	Original corpus			
Unbalanced	ρ=1	ρ=30		
	6.6	198		

Table 1: Balancing and duplication impact on corpora's size in millions of words (tokens).

7 Results and Discussion

The corpus was pre-processed using a spelling Nawatl unifier (Guzman-Landa et al., 2025), followed by processes of cleaning, segmentation into paragraphs and sentences, and the removal of some stop-words (Guzmán-Landa et al., 2026).

Figure 1 presents the new statistical distributions (blue bars) obtained by applying the positional balancing techniques (orange bars) to the topical and dialectal varieties. This addresses the bias introduced by the variability and productivity of the texts selected for the corpus. This variability affects not only the quantity but also the quality of future text processing and generation.

These transformations involve considerable volumes of text, thereby significantly increasing the size of the π-YALLI corpus and reducing the underrepresentation of topic-based and speech communities.

7.1 Unbalanced corpora

Figure 2 shows the results using incremental duplication of unbalanced corpora, with an increment ratio $\rho = [1, 2, ..., 30]$. For each ratio ρ, we show the average Kendall's $\langle\tau\rangle$ over five runs and the respective standard deviation (shown as a coloured band). The FastText algorithm in skip-gram mode achieves the best $\langle\tau\rangle$ performance across most ratios ρ. However, it should be noted that Word2Vec, also in Skip-gram mode, benefits most from the unbalanced incremental duplication technique, with consistent improvements between $1\times \leq \rho \leq 16\times$. In contrast, the GloVe algorithm shows diminishing results as the number of duplications increases.

Table 2 shows further details concerning these results: the maximum average values of $\langle\tau\rangle$, the gain and the training time. Except GloVe, the unbalanced duplication yields moderate or significant benefits. We have observed that Word2Vec achieves better results when the duplication rate $\rho = [20, 22]$. On the other hand, FastText reaches its maximum values with $\rho = [8, 10]$. This shows that FastText, with a lower duplication rate, generates higher-quality static embeddings.

In order to compare our results, we have used as our baseline three well known pre-trained embedding models. These models uses the same learning algorithms but they were trained on three commonly available corpora–without duplication or balance–: (i) FastText trained on Common Crawl[6]; (ii) FastText trained on Nawatl Wikipedia[7]; and (iii) Word2Vec trained on Nawatl Common Crawl[8]. Our results outperform the three baselines, as expected.

7.2 Balanced corpora

Table 3 presents a comparison between models trained on the original unbalanced π-YALLI corpus and those trained on the newly proposed uniform and positionally balanced corpora. This study was conducted using only the Word2Vec and FastText models employing Skip-gram architectures, which yielded the best results when applying unbalanced incremental duplication to the corpora (see Figure 2). As shown in Table 3, topic positional balancing (T_{pos}) enables FastText to achieve a Kendall's $\langle\tau\rangle = 0.477$, the highest recorded across all balancing methods. Meanwhile, with uniform topic balancing (T_{+}), Word2Vec shows a **19.9%** improvement, increasing from a Kendall's $\langle\tau\rangle$ of 0.357 to 0.428. This constitutes the largest percentage gain among all balancing techniques.

We found that positional (pos) and uniform (+) balancing, when applied to topics, yield the highest scores. However, in practice, both types of balancing applied to the topic (T) or dialectal varieties (D) yield an improvement in the base Kendall's τ for Word2Vec and FastText. For this reason, we decided to apply incremental duplication to all cases: T+, T_{pos}, D+, and D_{pos}.

Figure 3 shows that topic positional balancing T_{pos} stands out significantly compared to the other cases. FastText once again achieves the highest mean Kendall's value of $\langle\tau\rangle = 0.515$ for $\rho = 12$. The percentage increase relative to the baseline τ

[6] https://commoncrawl.org

[7] FastText has been trained on 157 languages: https://fasttext.cc/docs/en/crawl-vectors.html

[8] https://sparknlp.org/2022/03/16/w2v_cc_300d_nah_3_0.html

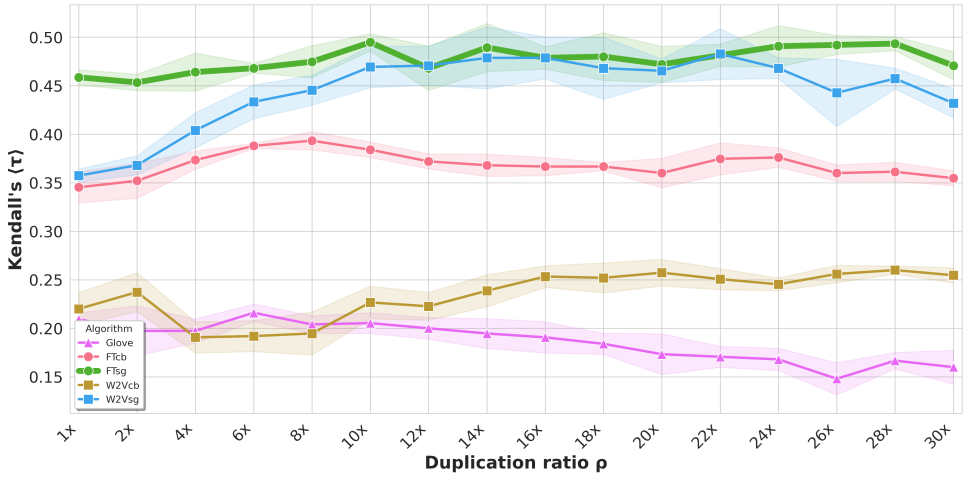

Figure 2: Unbalanced π-YALLI corpus. Kendall's coefficient $\langle\tau\rangle$ on sentence semantic similarity task. Learning on incrementally duplicated unbalanced corpus π-YALLI, using the static models GloVe, FastText (CBOW=FTcb, Skip-gram=FTsg) and Word2vec (CBOW=W2Vcb, Skip-gram=W2Vsg). There are 5 runs per duplication point ρ.

— which rises from 0.477 to 0.515 — is 8%. FastText also achieves a maximum absolute[9] value of $\tau = \mathbf{0.547}$. Word2Vec achieves its highest average $\langle\tau\rangle = 0.481$ at $\rho = 12$, and a maximum absolute $\tau = 0.500$. However, this is not the highest τ for Word2Vec, as it achieves a maximum absolute[10] $\tau = \mathbf{0.527}$ at $\rho = 18$. Therefore, $\rho^{\star} = \mathbf{12}$ can be established as a reasonably optimal duplication ratio for this NLP task.

Given the upward trend in Kendall's τ for the topic positional balancing (T_{pos}) in both Word2Vec and FastText, we decided, with purely exploratory intent, to investigate whether this improvement would persist in these algorithms. Figure 4 (see Appendix A.4) shows the performance using topic positionally balancing for values of ρ up to 50. A limit to the improvement from incremental duplication can be observed; the average $\langle\tau\rangle$ no longer exceeds the results obtained at $\rho = 12$. Indeed, for $\rho = 50$ in Word2Vec, the $\langle\tau\rangle$ obtained is lower than the initial value. In FastText, the final four average $\langle\tau\rangle$ values remain very close to the initial baseline.

Finally, a significant difference was observed between the CBOW and Skip-gram architectures of the learning algorithms. CBOW is an architecture that focuses on predicting an unknown word X based on its context — the set $C(X)$ of words surrounding X (Mikolov et al., 2013c). X is then predicted based on the information provided by its context, C. In contrast, Skip-gram predicts the words surrounding X. Word2Vec generates a single vector (embedding) for each word in the vocabulary, whereas FastText generates an embedding for each Nawatl character n-gram. This allows for the construction of vectors containing more information, by virtue of these n-grams. We confirmed experimentally that FastText's Skip-gram architecture significantly outperforms the other algorithms.

8 Conclusions and Future work

The results obtained highlight the effectiveness of the balancing and duplication techniques proposed for the corpora. Indeed, when these techniques are applied to agglutinative and polysynthetic languages having limited computational resources, they seem to facilitate the training of models that produce static representations. In this way, static representations trained with these techniques capture the language's structure more effectively.

Specifically, in the case of FastText, the topical positional balancing, combined with the $\rho = 12$ replicas of the corpus improved the Kendall coefficient $\langle\tau\rangle$ from 0.459 (using the original, unbalanced corpus) to $\langle\tau\rangle = 0.515$, representing a significant gain of **12.2%**. Although Word2Vec does not achieve the highest mean value of Kendall's $\langle\tau\rangle$, it is the model that obtained the most representative gain, illustrating the advantages of the approach presented. Their average Kendall's coefficient at the starting point (without balancing or duplication) is $\langle\tau\rangle = 0.357$, whilst topic positional balancing, combined with incremental duplication, achieves a $\langle\tau\rangle = 0.481$, i.e. a **34.7%** improvement.

[9] Not visible in the Fig. 3, at left.
[10] Not visible in the Fig. 3, at right.

Model	$\langle\tau\rangle$	$\max\langle\tau\rangle_{\rho\times}$	ρ	Gain %	Time (min)
FastText Skip-gram	0.459	**0.495**	10	7.8	46.6
Word2Vec Skip-gram	0.357	0,483	22	35.3	39.3
FastText CBOW	0.345	0.393	8	13.9	43.7
Word2Vec CBOW	0.220	0.257	20	16.8	14.9
GloVe	0.209	0.216	6	3.4	6.5
Baselines	$\langle\tau\rangle$				
FastText/Wikipedia	0.242	-	-	-	-
FastText/Common Crawl	0.240	-	-	-	-
Word2Vec/Wikipedia	0.240	-	-	-	-

Table 2: Unbalanced π-YALLI corpus. Kendall's $\langle\tau\rangle$ over five runs of the models without duplication, and $\max\langle\tau\rangle_{\rho\times}$: the maximum τ obtained with $\rho\times$ duplications. %: percentage of $\langle\tau\rangle$ improvement. The learning time is approximate for each single ρ run, executed on a cluster with a requirement of [8, 12] cores and [16, 64] GB of RAM, running under GNU/Linux in SLURM (*Simple Linux Utility for Resource Management*) mode. The baselines use pre-training models.

	Unbalanced corpus	**Balanced corpora**							
		Topical				**Dialectal**			
Skip-gram Model	$\langle\tau\rangle$	$\langle\tau\rangle$		**Gain%**		$\langle\tau\rangle$		**Gain%**	
		T_+	T_{pos}	T_+	T_{pos}	D_+	D_{pos}	D_+	D_{pos}
FastText	*0.459*	0.467	**0.477**	1.7	**3.9**	0.465	0.468	1.3	1.9
Word2Vec	*0.357*	**0.428**	0.425	**19.9**	19.0	0.413	0.381	15.7	6.7

Table 3: Starting point (without duplication) of unbalanced and balanced corpora. Kendall's $\langle\tau\rangle$ on 5 runs of Skip-gram models. +: Uniform balancing, pos: Positional balancing, T: Topical, D: Dialectal.

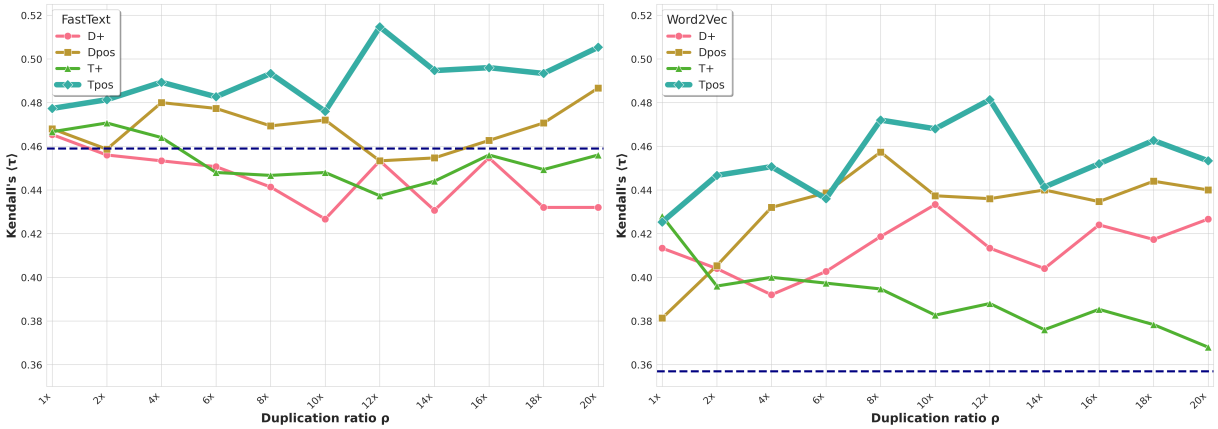

Figure 3: Balanced and incrementally duplicated corpus π-YALLI. **FastText** (left) vs. **Word2Vec** (right) in Skip-gram mode. Kendall's coefficient $\langle\tau\rangle$ of the sentence semantic similarity task. D+: Uniform balancing by dialectal variety, T+: Topic uniform balancing, Dpos: Positional balancing by dialectal variety, Tpos: Topic positional balancing. The dashed lines represent the $\langle\tau\rangle$ obtained using the original corpus (without balancing or duplication). There are 5 runs per duplication point ρ (the scale is the same on both graphics).

We found that positional balancing outperforms uniform balancing, particularly when applied to topics. We have reached the conjecture that uniform balancing tends to favour data redundancy to a greater extent. Furthermore, this is exacerbated when incremental duplication is applied; consequently, in several cases starting from $\rho = 2$, a deterioration in Kendall's τ can be observed. This suggests that while maintaining a proper balancing between classes within a corpus is crucial, it is equally important to prevent them from becoming excessively redundant.

Our research highlights the positive impact on model learning of applying a statistically balanced,

positional strategy to heavily imbalanced corpora. Furthermore, these benefits are amplified when combined with an incremental corpus-duplication technique. Even without any balancing applied, the duplication strategy alone shows positive effects on the model training. Duplication is thus an efficient and comprehensive alternative for corpora expansion, particularly when dealing with NLP of π-languages.

In future work, we will explore other ways of balancing corpora and their impact on the incremental duplication technique. Similarly, we intend to evaluate the contribution of expanded corpora to the tasks of Automatic Text Summarisation, Sentiment analysis and Named Entity recognition—such as toponyms detection—in Nawatl.

Limitations

Our results indicate that the proposed corpus balancing and duplication methods yield better results than using the original corpora.

Although these results are very promising, we recognise that further experiments are needed with other types of balancing and duplication, particularly using other NLP tasks to assess the limitations in greater detail. This is especially true for π-languages, which have very few computational resources.

Ethics Statement

We are mindful of the potential risks associated with data duplication, including the possibility of encouraging redundancy and bias, and the risk of minimizing or excluding some lects and scripts of Nawatl speech.

We therefore strongly advocate that this technology be used exclusively to promote the appreciation of Nawatl and to support the development of digital resources that facilitate its study and dissemination.

Acknowledgments

This research work has been financed by the Agorantic NAWA project and the Intermedius PhD Grant, and supported by the Laboratoire Informatique d'Avignon, from Avignon Université (France).

References

Nimaan Abdillahi, Pascal Nocera, and Juan Manuel Torres. 2006. Boites a outils TAL pour les langues peu informatisées : Le cas du Somali. In *Journées d'Analyses des Données Textuelles*, Besançon, France.

Najib Arbach and Saandia Ali. 2013. Aspects théoriques et méthodologiques de la représentativité des corpus. *Corela [En ligne], HS-13*.

Vincent Berment. 2004. *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Ph.D. thesis, Université Joseph-Fourier - Grenoble I.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL*, 5:135–146.

Khoo Khyou Bun and Mitsuru Ishizuka. 2002. Topic extraction from news archive using tf* pdf algorithm. In *3rd International Conference on Web Information Systems Engineering (WISE'02)*, pages 73–82. IEEE.

Wright-Carr. David Charles. 2016. *Lectura del náhuatl*. Instituto Nacional de Lenguas Indígenas.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in NLP. *Transactions of the ACL*, 11:191–211.

Jaqueline de Durand-Forest, Danièle Dehouve, and Eric Roulet. 1995. *Parlons Nahuatl. La langue des Aztèques*. L'Harmattan.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the ACL: Human Language Technologies, Vol 1*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the ACL: ACL-IJCNLP 2021*, pages 968–988, Online. ACL.

Lucero Flores Nájera. 2019. *La gramática de la clausula simple en el náhuatl de Tlaxcala*. Ph.D. thesis, CIESAS.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *NAACL: Human Language Technologies*, pages 1256–1261, San Diego, California. ACL.

Palash Goyal, Sumit Pandey, and Karan Jain. 2018. *Deep Learning for Natural Language Processing*. Springer.

Ximena Gutierrez-Vasques, Robert Pugh, Victor Mijangos, Diego Barriga Martínez, Paul Aguilar, Mikel Segura, Paola Innes, Javier Santillan, Cynthia Montaño, and Francis Tyers. 2025. Py-elotl: A python NLP

package for the languages of Mexico. In *5th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 38–47, Albuquerque, New Mexico. ACL.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Martha-Lorena Avendaño-Garrido, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, and Luis-Gil Moreno-Jiménez. 2025. π-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl. In *TALN, vol 1*, pages 802–816, Marseille, France. ATALA.

Juan-José Guzman-Landa, Jesús Vázquez-Osorio, Juan-Manuel Torres-Moreno, Ligia Quintana-Torres, Miguel Figueroa-Saavedra, Martha-Lorena Avendaño Garrido, Graham Ranger, Patricia Velázquez-Morales, and Gerardo Sierra-Martínez. 2025. A symbolic algorithm for the unification of nawatl word spellings. In *Advances in Soft Computing: 24th MICAI'25, Guanajuato, Mexico, 2025, Part I*, page 141–154, Berlin, Heidelberg. Springer-Verlag.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Graham Ranger, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Carlos-Emiliano González-Gallardo, Luis-Gil Moreno-Jiménez, and Martha-Lorena Avendaño-Garrido. 2026. Nawatl context-free grammars for Natural Language Processing. In *15th Language Resources and Evaluation Conference (LREC)*, pages 3333–3342, Palma, Spain. ELRA.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Miguel Figueroa-Saavedra, Carlos-Emiliano González-Gallardo, Graham Ranger, and Martha Lorena-Avendaño-Garrido. 2026. Classifying several dialectal nawatl varieties. *Preprint*, arXiv:2601.02303.

Magnus Pharao Hansen. 2024. *Nahuatl Nations: Language Revitalization and Semiotic Sovereignty in Indigenous Mexico*. Oxford University Press.

INEGI. 2020. Censo de población y vivienda 2020. In *CENSO 2020*. `https://www.inegi.org.mx/rnm/index.php/catalog/632/study-description`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Yolanda Lastra de Suárez. 1986. *Las áreas dialectales del náhuatl moderno*. UNAM, Instituto de Investigaciones Antropológicas, Mexico.

Michel Launey. 1978. *Introduction à la langue et à la littérature aztèques*, volume 1. L'Harmattan, Paris.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *60th Annual Meeting of the ACL (V1)*, pages 8424–8445, Dublin, Ireland. ACL.

Mosleh Mahamud, Zed Lee, and Isak Samsten. 2023. Distributional data augmentation methods for low resource language. *Preprint*, arXiv:2309.04862.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS - Vol 2*, NIPS, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL – HLT 2013)*, pages 746–751, Atlanta, GA, USA. ACL.

Justyna Olko and John Sullivan. 2016. Bridging gaps and empowering speakers: An inclusive, partnership-based approach to nahuatl research and revitalization. *Integral strategies for language revitalization*, pages 347–386.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. ACL.

Robert Pugh, Varun Sreedhar, and Francis Tyers. 2024. Wav2pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl. In *4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 121–126, Mexico City, Mexico. ACL.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *1st Workshop on Natural*

Language Processing for Indigenous Languages of the Americas, pages 21–27. ACL.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla Nahuatl. In *4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Vol 1*, pages 80–85. ACL.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025. Ihquin tlahtouah in tetelahtzin-cocah: An annotated, multi-purpose audio and text corpus of western sierra Puebla Nahuatl. In *Conference of the Nations of the Americas Chapter of the ACL: Human Language Technologies (Vol 1)*, pages 3549–3562, Albuquerque, New Mexico. ACL.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Thirteenth ACM international conference on Information and knowledge management*, pages 42–49.

Mitsuya Sasaki. 2022. Divide y entenderás: El papel de la polarización sintáctica en el náhuatl moderno y colonial. In *Coloquio de Investigación Lingüística, Universidad de Sonora (Mexico)*.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63. ACL.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6382–6388, Hong Kong, China. ACL.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *20th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. ELRA.

Klaus Zimmermann. 2019. Estandarización y revitalización de lenguas amerindias: funciones comunicativas e ideológicas, expectativas ilusorias y condiciones de la aceptación. *Revista de Llengua i Dret, Journal of Language and Law*, 71:111–122.

A Appendix

In this Appendix, we present: (A.1) the list of acronyms used (both dialectal and topical); (A.2) an example of Nawatl semantic similarity between a reference sentence and its candidates; (A.3) the learning hyper-parameters for the models; (A.4) the comparison of the two best-performing learning algorithms, and (A.5) the statistical distribution of topics and dialectal varieties.

A.1 Dialectal and Topical acronyms

The following acronyms will be used to designate the 26 dialectal varieties available of the π-YALLI corpus:

***CV**: Veracruz Central Nawatl, **H**: Huasteca, **HH**: Hidalgo's Huasteca, **HV**: Veracruz Huasteca, **HP**: Potosí Huasteca, **CEA**: Upper Central Region Mexican, **CEO**: Central-Western Mexican, **CEN**: Central Mexican, **CEB**: Central Low Mexican, **ORI**: Eastern Mexican, **GUE**: Guerrero's Mexican, **NOC**: Central Northwest, **IST**: Isthmus Nawatl, **NAW**: Nawat, **OAX**: Oaxaca's Nawatl, **CEP**: Puebla Center, **ANP**: Puebla's Northern Highlands, **ORP**: Eastern Puebla's Mexican, **SNP**: Puebla's Sierra Negra, **SNNP**: Puebla's Northern Sierra Negra, **SNSP**: Puebla's Southern Sierra Negra, **SNEP**: Puebla's Northeast Sierra, **SOP**: Puebla's Western Sierra, **TEM**: Temixco's Mexican, **TLA**: Tlaxcala's Mexican, and **MIX**: Mixture of dialectal varieties*[11].

The acronyms for the 16 topics are as follows[12]: ***REL**: Religion, **LIT**: Literature, **HIS**: History, **EDU**: Education, **LIN**: Linguistics, **LEG**: Legislation, **WIK**: Wikipedia, **COS**: Cosmovision, **TEC**: Technology, **ECO**: Economics, **MED**: Medicine, **AGR**: Agriculture, **POE**: Poetry, **MUS**: Music, **PHR**: Sentences without context, and **POL**: Politics.*

[11]The original INALI Spanish names for the dialectal varieties are as follows: **CV**: Nawatl Central de Veracruz, **H**: Huasteca, **HH**: Huasteca Hidalguense, **HV**: Huasteca Veracruzana, **HP**: Huasteca Potosina, **CEA**: Mexicano del Centro Alto, **CEO**: Mexicano Central de Occidente, **CEN**: Mexicano del Centro, **CEB**: Mexicano Central Bajo, **ORI**: Mexicano del Oriente, **GUE**: Mexicano de Guerrero, **NOC**: Noroeste Central, **IST:** Nawatl del Istmo, **NAW**: Nawat, **OAX**: Nawatl de Oaxaca, **CEP**: Centro de Puebla, **ANP**: Alto del Norte de Puebla, **ORP**: Mexicano del Oriente de Puebla, **SNP**: Sierra Negra de Puebla, **SNNP**: Sierra Negra Norte Puebla, **SNSP**: Sierra Negra Sur de Puebla, **SNEP**: Sierra Noreste de Puebla, **SOP**: Sierra Oeste de Puebla, **TEM**: Mexicano de Temixco, **TLA**: Mexicano de Tlaxcala, **MIX**: Mezcla de variedades.

[12]Classification of the Atlas of Languages INALI: https://atlas.inali.gob.mx/agrupaciones/info/0211

A.2 Example of a reference-candidates block 10, for the semantic similarity task

REFERENCE SENTENCE (10):
Yewehkatlahtolli momachtia ken okatka tlakayotl /
History studies the past of Humanity.

RANKED CANDIDATE SENTENCES:

1. Tikmatih tlen opanok, ken okatka tlakayotl ika yewehkatlahtolli.
   We know about humanity's past thanks to history.
2. Tlen ye wehkah otlamochih momachtia ipan weyi tlamachtiloyan.
   Historical events are studied at university.
3. Momachtilistli itechpa wehkawitl techpalewia pampa tikachtopaittaskeh yakapankawitl.
   Studying the past helps us to predict the future.
4. In wehkawitl ye wehka opanok
   The past is just that: the past.
5. Nonemilis nesi ihkin inemilis notahtzin: ohwi.
   My personal story is much like my father's: complicated.

A.3 Training Hyper-parameters used for the models

The hyper-parameters used for all models are as follows:

Number of epochs: **20**; Context window size: **5 tokens**; Embeddings' dimension: **300**; and only for GloVe algorithm: Cutoff = **100** and α = **3/4**.

A.4 Comparing mean $\langle\tau\rangle$ of FastText vs. Word2Vec with $1 \leq \rho \leq 50$

In this comparison, both algorithms use the skip-gram architecture. Kendall's $\langle\tau\rangle$ coefficient for the sentence semantic similarity task is shown in the Figure 4. The results reflect the algorithms' performance on incrementally duplicated and topic-positionally balanced corpora.

The dotted lines in the figure indicate the τ values for the trained models (FastText in green, Word2Vec in blue) using the original π-YALLI corpus (without duplication or balancing).

In this experiment there are 10 runs per duplication ratio $1 \leq \rho \leq 50$, where the standard deviation is shown as a coloured band.

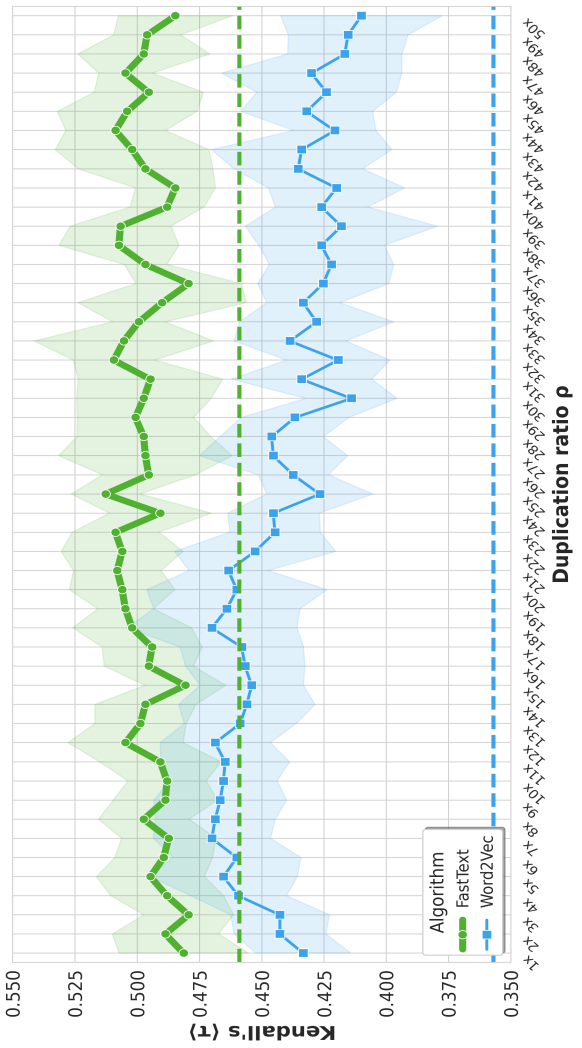

Figure 4: **FastText** (green) vs. **Word2Vec** (blue) using topic positionally balanced corpora. The dotted lines show Kendall's τ for the trained models on the original π-YALLI corpus.

A.5 Statistical distribution of tokens of the nawatl corpus

In Tables 4 and 5, we show the statistical distribution of words (tokens) by topics and dialectal varieties within the nawatl π-YALLI corpus.

Topics		
Acronym	**#**	**Num.**
REL	3 183,638	1
LIT	958,679	2
HIS	613,682	3
EDU	468,585	4
LIN	388,546	5
LEG	303,748	6
WIK	162,994	7
COS	52,304	8
TEC	25,899	9
ECO	14,854	10
MED	12,668	11
AGR	7,604	12
POE	3,934	13
MUS	2,417	14
PHR	2,208	15
POL	1,750	16

Table 4: Tokens' number (#) per topic.

Dialectal varieties		
Acronym	**#**	**Num.**
HV	1 307,351	1
MIX	938,825	2
HP	833,297	3
CEA	665,008	4
CV	457,333	5
GUE	453,994	6
CEO	242,554	7
IST	189,767	8
SNNP	185,564	9
OAX	170,483	10
SOP	156,861	11
SNP	155,944	12
ANP	142,385	13
SNEP	89,127	14
NAW	64,271	15
NOC	62,406	16
HH	32,581	17
H	25,204	18
CEP	12,746	19
TEM	7,220	20
ORI	4,263	21
CEN	2,273	22
SNSP	1,618	23
TLA	972	24
CEB	735	25
ORP	728	26

Table 5: Tokens' number (#) per dialectal variety.

On the Robustness of Morphosyntactic Transformation with Large Language Models: The Case of Quechua Collao

Pool Pocco
Chana Research Group
Pontificia Universidad Católica del Perú
Lima, Peru
pool.pocco@pucp.edu.pe

Arturo Oncevay
Chana Research Group
Pontificia Universidad Católica del Perú
Lima, Peru
arturo.oncevay@pucp.edu.pe

Abstract

Morphosyntactic transformation poses significant challenges for large language models (LLMs) in low-resource, morphologically rich languages, where multiple grammatical categories are often encoded within a single word. Applying controlled grammatical changes while preserving meaning requires both linguistic precision and robust generalization. We introduce a morphosyntactically controlled transformation dataset for Quechua Collao, built from a normalized Spanish–Quechua parallel corpus with explicit annotations and transformation labels. The dataset defines a controlled sentence-level transformation task, where models generate target sentences given a source sentence and a structured specification of grammatical changes. We evaluate multiple LLMs under varying numbers of in-context examples, selection strategies, and the use of lightweight morphological hints. Results show that performance depends strongly on prompt design and task formulation rather than model size alone. Increasing the number of examples yields model-dependent gains, benefiting smaller models more, while larger models remain relatively stable. Morphological hints provide selective improvements depending on the transformation type. These findings show that robustness arises from the interaction between model behavior, context size, and linguistic structure, highlighting the importance of controlled experimental design in low-resource settings.

1 Introduction

Morphosyntactic transformation is a structured generation task in which a system modifies specific grammatical properties of a source sentence while preserving its propositional content. Unlike word-level inflection, sentence-level transformation may require coordinated changes across multiple tokens or the insertion and deletion of functional elements. This challenge becomes particularly pronounced in morphologically rich languages, where several grammatical features are encoded within a single word through the concatenation of affixes, and even minimal modifications can entail non-trivial structural adjustments.

Southern Quechua (ISO 639-3: quz), particularly its Collao variety, provides a suitable testbed for this task. As part of the Quechuan language family, it exhibits a predominantly agglutinative morphology in which grammatical categories such as person, tense, aspect, number, and evidentiality are expressed through suffixation (Cerrón-Palomino, 2003; Chuquimamani Valer et al., 2021). This results in dense morphological structure and productive inflectional paradigms, where accurate transformation requires identifying and modifying specific morphemes while preserving the remaining structure.

In addition, Quechua Collao is situated in a low-resource setting, characterized by limited availability of digitized corpora, annotated data, and computational tools. Joshi et al. (2020) formalize this condition through a taxonomy of language resource levels, showing that most of the world's languages fall into categories with minimal computational support. Under these constraints, approaches based on large pretrained language models and in-context learning have become particularly attractive.

Recent work has explored sentence-level morphosyntactic transformation in low-resource languages, notably through shared tasks in the AmericasNLP workshop (De Gibert et al., 2025; Chiruzzo et al., 2024). In these settings, systems are required to apply explicit grammatical changes—such as person, polarity, or tense—to a source sentence and generate the corresponding modified output. Within these tasks, LLMs prompted with few-shot examples have emerged as competitive approaches, with prior studies showing that both prompt enrichment with morphosyntactic information (Vasselli et al., 2024) and example selection strategies

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 128–146

(Lupicki et al., 2025) can influence performance.

However, these efforts do not specifically address Quechua Collao, nor do they provide datasets with explicit morphosyntactic control tailored to its linguistic properties. While prior work on Quechua has primarily focused on tasks such as machine translation and related applications (Cueva Medina et al., 2024), the availability of standardized, annotated resources for controlled morphosyntactic transformation remains limited. This gap is particularly relevant for morphologically rich and low-resource languages, where the absence of linguistically structured datasets constrains both model evaluation and the systematic study of generation behavior (Camacho Caballero and Zevallos Salazar, 2020; Joshi et al., 2020).

Despite these advances, the relative contribution of different design choices remains insufficiently understood. Existing work typically reports improvements from individual techniques, but does not systematically analyze how factors such as the number of in-context examples, example selection strategies, and the inclusion of lightweight morphological hints interact under controlled conditions, particularly in relation to differences across models and transformation complexity.

In this work, we construct a morphosyntactically controlled transformation dataset for Quechua Collao, derived from a normalized Spanish–Quechua parallel corpus and enriched with explicit morphosyntactic annotations and transformation labels. This resource enables a controlled formulation of sentence-level morphosyntactic transformation, where models generate target sentences given a source sentence and a structured specification of grammatical changes.

Using this framework, we conduct a systematic study of in-context learning with multiple LLMs, varying the number of examples, example selection strategies, and the use of lightweight morphological hints. In particular, we design a linguistically motivated k-shot selection strategy that prioritizes morphosyntactic relevance while constraining spurious variation in the prompt. This setup allows us to analyze how design choices in prompt construction affect the robustness and consistency of morphosyntactic generation in a low-resource setting.

Source	Sipasqa chukuta achalan *The young woman adorns the hat*
Change	{NUMBER:PL}
Target	Sipasqa chukukunata achalan *The young woman adorns the hats*

Table 1: Example of a morphosyntactic transformation instance.

Metric	**Spanish**	**Quechua**
Sentence pairs	22,581	22,581
Total tokens	449,580	285,605
Unique tokens	52,632	73,637
Avg. tokens per sentence	19.91	12.65

Table 2: Statistics of the Spanish–Quechua parallel corpus.

2 Task and Dataset

We consider the task of sentence-level morphosyntactic transformation in Quechua Collao. Given a source sentence and a set of morphosyntactic changes, the task consists of generating a target sentence that reflects the specified transformation while preserving its propositional meaning. Table 1 illustrates the task format.

2.1 Corpus Construction

The dataset is derived from a Spanish–Quechua parallel corpus compiled from bilingual educational materials and publicly available texts. The Spanish side provides a semantic anchor and supports the initial annotation stage, while the transformation task itself is defined over Quechua sentences. The corpus was processed through text extraction, normalization[1], sentence segmentation, and heuristic alignment to obtain parallel sentence pairs. This parallel corpus serves as the basis for subsequent morphosyntactic annotation and transformation generation.

Table 2 summarizes the resulting corpus. Although the Quechua side contains fewer tokens overall, it exhibits a higher number of unique word forms and shorter average sentence length. This pattern reflects the agglutinative nature of Quechua morphology: grammatical information that often requires multiple words in Spanish can be encoded through productive suffixation attached to a single lexical root in Quechua.

[1]Orthographic normalization follows the official conventions for Southern Quechua described in Chuquimamani Valer et al. (2021).

Category	Example markers
PERSON	-ni, -nki, -n, -nchik
TENSE	-rqa, -sqa, -saq
ASPECT	-chka
NUMBER	-kuna
POSSESSION	-y, -n, -yki
EVIDENTIALITY	-mi, -si
POLARITY	mana ... -chu

Table 3: Main morphosyntactic categories and typical surface realizations in Quechua Collao.

2.2 Morphosyntactic Annotation

To enable controlled transformations, Quechua sentences were annotated with morphosyntactic features such as person, tense, aspect, number, possession, evidentiality, and polarity. Annotation is performed using a semi-automatic process based on surface morphological patterns, followed by manual verification by native speakers to ensure linguistic consistency.

Table 3 summarizes the main morphosyntactic categories represented in the dataset and their typical morphological realizations in Quechua Collao. These categories correspond to core grammatical distinctions in the language and follow the descriptive framework proposed in previous linguistic work (Chuquimamani Valer et al., 2021). A full inventory of labels and abbreviations is provided in Appendix C.

2.3 Transformation Dataset

From the annotated corpus, we construct a dataset of transformation instances, each defined as a tuple (*source sentence*, *change*, *target sentence*) (De Gibert et al., 2025). This dataset enables the systematic study of controlled morphosyntactic transformations in Quechua Collao, a setting for which no prior structured resources are available. Target sentences are obtained through controlled modification of morphosyntactic features and manually verified to preserve grammaticality and the original propositional content. Further details on the transformation pipeline are provided in Appendix B.

Table 4 provides an overview of this task-ready resource.

In addition to the minimal representation, some experimental configurations incorporate lightweight morphological hints that specify the expected surface realization of target categories. These hints are derived from the same morphosyntactic markers and are used in enriched prompting setups to provide explicit linguistic guidance during inference.

Metric	Value
Transformation instances	6328
Unique source sentences	471
Unique target sentences	474
Unique change specifications	402
Avg. transformations per source	13.44

Table 4: Statistics of the morphosyntactic transformation dataset used for in-context learning experiments.

Label	Expected realization
NUMBER:PL	noun + -kuna
POSS:3_SI	noun + -n
TENSE:PST_NEXP	verb + -sqa
TYPE:NEG	mana ... -chu
MODE:POT	verb + -man

Table 5: Examples of lightweight morphological hints used in prompting.

Table 6 presents the distribution of morphosyntactic categories in the transformation field. Person, possession, and tense transformations account for the majority of cases, reflecting their central role in Quechua verbal and nominal morphology and their high productivity in natural language usage.

The resulting dataset serves as the basis for the in-context learning experiments described in the following section.

3 Experimental Setup

3.1 In-Context Learning Setup

All experiments are conducted in an in-context learning setting using instruction-tuned large language models. Each prompt is formulated as a structured table of resolved examples followed by a final query with an empty target field, so that the model must infer the required transformation by analogy.[2] We evaluate three instruction-tuned LLMs selected to contrast parameter scale and generation behavior: Mistral 24B Instruct (Jiang et al., 2023), Qwen 14B Instruct (Bai et al., 2023), and the reasoning-oriented OpenAI GPT-OSS-20B (OpenAI et al., 2025).

To ensure comparability across runs, decoding is deterministic in all configurations, using temperature $= 0$ and top-$k = 1$. The maximum output length is fixed to 64 tokens for Mistral and Qwen, while GPT-OSS-20B is generated without

[2]A full example of the prompting format is provided in Appendix A.

Category	Frequency	Percentage
PERSON	5676	41.9%
POSS	2685	19.8%
TENSE	2385	17.6%
NUMBER	1135	8.4%
EVID	519	3.8%
MODE	397	2.9%
ASPECT	387	2.9%
TYPE	208	1.5%
PERSON_OBJ	132	1.0%
SUBTYPE	30	0.2%

Table 6: Distribution of morphosyntactic categories in the transformation field (*Change*).

Factor	Values
Dataset features	Source, Change, Target, Hints
Models	Qwen2.5-14B-Instruct Mistral-24B-Instruct GPT-OSS-20B
Shots (K)	{5, 10, 15}
Transformation size	{1, 2}
Prompting modes	w/o hints; w/ hints
Decoding	temperature = 0; top-k = 1 max tokens = 64 (Mistral/Qwen) no explicit cap (GPT-OSS)
Inference setup	LM Studio (local API)

Table 7: Experimental setup.

an explicit cap.[3] All models are executed locally through LM Studio, which also allows us to record inference latency under the same execution environment. Additional inference details are provided in Appendix E.

3.2 k-shot Selection Strategy

The quality of the in-context examples is a central factor in this task. Rather than sampling examples at random, we use a hierarchical retrieval strategy designed to preserve morphosyntactic relevance while minimizing distracting variation.

Selection proceeds in three stages:

- **Exact match.** The system first retrieves examples whose `Change` specification exactly matches the target instance, requiring identical morphosyntactic categories and values (e.g., `PERSON:1_PL_INC`, `TENSE:PST_EXP`).

- **Primary-category match.** If the required number of examples is not reached, the system backs off to instances sharing the same primary morphosyntactic category, defined as the dominant category within the `Change` specification (e.g., PERSON, TENSE, POSS), regardless of the specific feature values.

- **Similarity-based fallback.** Remaining slots are filled using TF-IDF similarity over the source sentence in order to preserve lexical and structural proximity.

[3]The 64-token limit exceeds the expected length of a single transformed sentence and prevents unnecessarily long generations for Mistral and Qwen. All models are prompted to output a `FINAL:` marker for retrieving the generated target; GPT-OSS-20B is left uncapped because it may produce intermediate reasoning before reaching that marker. This may affect latency comparability, but target retrieval remains uniform across models.

This retrieval process is further constrained in three ways. First, we apply a *subset constraint*, which prevents examples from introducing morphosyntactic categories that are absent from the target transformation. Second, we follow a *cluster-first* policy, which prioritizes examples derived from the same transformation cluster whenever possible. Third, evaluation follows a *leave-one-out* scheme, so that the target instance never appears among its own demonstrations. The full selection workflow is illustrated in Appendix D.

3.3 Experimental Factors

The experiments vary four controlled factors: model, number of in-context examples, use of morphological hints, and transformation complexity. Table 7 summarizes these settings.

We use the full transformation pool for retrieval and evaluation, rather than fixed train/dev/test partitions, since the goal is not to fine-tune model parameters but to measure controlled generalization by analogy. The number of demonstrations $K \in \{5, 10, 15\}$ allows us to test whether additional context improves performance. We also compare prompts with and without lightweight morphological hints, which provide surface-oriented cues derived from the target transformation. Finally, we distinguish between single-category transformations (*size* = 1) and compositional transformations involving two categories (*size* = 2), allowing us to evaluate how model behavior changes with structural complexity.

3.4 Evaluation Metrics

We evaluate model outputs using automatic metrics that capture both exact correctness and surface-level similarity.

Exact match accuracy measures the proportion of predictions that exactly match the reference target sentence after normalization.

chrF computes character n-gram F-scores between predicted and reference sentences, providing a more fine-grained measure of similarity that is robust to minor variations in morphologically rich languages.

In addition to automatic metrics, we perform a human-in-the-loop evaluation to assess grammatical correctness, naturalness, and semantic consistency of the generated outputs. Details on the evaluation protocol and results are provided in Appendix F.

4 Results

Model performance varies substantially across context sizes, revealing distinct sensitivities to the number of in-context examples. Table 8 summarizes corpus-level results across models and values of K.

Mistral 24B achieves strong accuracy with minimal context, indicating that it can perform robustly even with a small number of demonstrations. In contrast, Qwen 14B exhibits consistent gains as K increases, reaching the highest overall performance at larger context sizes. GPT-OSS 20B shows more gradual improvements with additional examples, but remains less competitive in exact-match accuracy.

These patterns indicate that the effect of increasing K is strongly model-dependent. Mistral remains comparatively stable across configurations, suggesting limited reliance on additional contextual support. Qwen, by contrast, benefits directly from larger context, indicating a stronger dependence on in-context signals for generalizing morphosyntactic transformations. GPT-OSS follows an intermediate trend, with moderate but less consistent gains.

Overall, additional in-context examples do not provide uniform benefits across models. Their effectiveness depends on how each model leverages contextual information during inference, motivating the more detailed analyses that follow.

4.1 Model Comparison and Stability

Robustness vs. sensitivity. Model performance reveals a clear trade-off between robustness and sensitivity to experimental conditions. While all systems achieve comparable average performance, their behavior across configurations differs substantially.

	Mistral 24B		GPT-OSS 20B		Qwen 14B	
K	**Acc.**	**chrF**	**Acc.**	**chrF**	**Acc.**	**chrF**
5	52.54	90.69	44.07	91.21	44.92	92.98
10	**55.08**	**90.28**	48.73	92.63	54.66	94.40
15	53.39	89.83	**53.81**	**93.95**	**58.90**	**94.14**

Table 8: Corpus-level results by model and number of in-context examples (K). Bold indicates the best-performing configuration for each model across values of K.

Table 9 summarizes performance variability across context sizes using mean and standard deviation as indicators of stability. Lower variance reflects more consistent behavior across configurations, while higher variance indicates sensitivity to changes in prompting conditions.

Mistral exhibits consistently low variance, indicating stable performance that is largely invariant to both the number of in-context examples and the use of hints. This suggests that the model internalizes morphosyntactic transformations without relying heavily on prompt-specific cues.

In contrast, Qwen displays markedly higher variability, reflecting strong sensitivity to contextual factors. Its performance improves as additional examples are provided, indicating a more direct reliance on in-context signals for generalizing transformation patterns, particularly in settings involving higher structural complexity.

GPT-OSS occupies an intermediate position. Although it benefits from increased context, its improvements are less consistent and do not follow a stable trend, suggesting a limited ability to systematically exploit additional examples compared to Qwen, while lacking the robustness observed in Mistral.

Accuracy vs. surface similarity. A complementary distinction emerges between exact-match correctness and surface similarity. Despite achieving high chrF scores, Qwen and GPT-OSS do not consistently match this performance in exact accuracy, indicating that morphologically similar outputs may still diverge from the correct transformation. This highlights the importance of jointly evaluating both metrics in morphologically rich languages.

Overall, these results indicate that model behavior in this task is systematic rather than random. Mistral emerges as the most robust model, Qwen

Model	Acc.	chrF
Mistral 24B	**53.67**± 1.06	90.27± 0.35
Qwen 14B	52.82± 5.85	**93.84**± 0.61
GPT-OSS 20B	48.87± 3.98	92.60± 1.12

Table 9: Performance stability across models (averaged over $K = 5, 10, 15$).
Values are reported as mean $\pm$ standard deviation over $n = 708$ instances per model.

as the most context-sensitive, and GPT-OSS as an intermediate system with moderate adaptability. These differences motivate the more fine-grained analyses of contextual effects and prompting strategies presented in the following subsections.

4.2 Effect of Context Size

The effect of the number of in-context examples (K) is not uniform across models and is better captured through statistical analysis than raw metric differences. To assess whether changes in K produce systematic variation, we apply Kruskal–Wallis tests for global comparisons across context sizes and Mann–Whitney U tests for pairwise contrasts.[4]

For Qwen, the effect of K is statistically significant across configurations, confirming a strong dependence on contextual information. Performance improves consistently as additional examples are provided, indicating that the model relies on in-context signals to infer transformation rules, particularly in settings with lower frequency or higher structural complexity. In this regime, additional examples act as explicit guidance that supports generalization by analogy.

In contrast, Mistral does not exhibit statistically significant differences across values of K, indicating that its performance remains largely invariant to the number of demonstrations. This reinforces the robustness observed in the previous subsection: the model appears to internalize morphosyntactic regularities in a way that does not depend strongly on the quantity of contextual support at inference time.

GPT-OSS shows a weaker and less consistent effect. Although performance tends to improve with larger K, the differences are comparatively mild and not systematically significant, suggesting that the model can partially benefit from additional context without exploiting it as consistently as Qwen.

[4]All statistical tests are conducted using a significance level of $\alpha = 0.05$.

Model	Metric	Test	p-value	Sig.
Mistral 24B	Accuracy	Kruskal (K)	0.853	n.s.
	chrF	Kruskal (K)	0.915	n.s.
GPT-OSS 20B	Accuracy	MW (5 vs 15)	0.034	*
	chrF	MW (5 vs 15)	0.033	*
Qwen 14B	Accuracy	Kruskal (K)	0.0077	**
	chrF	Kruskal (K)	0.0042	**
	Accuracy	MW (5 vs 15)	0.0024	**
	chrF	MW (5 vs 15)	0.0012	**

Table 10: Effect of the number of in-context examples (K) on model performance.
Significance levels: n.s. ($p \geq 0.05$, no consistent effect), * ($p < 0.05$, mild effect), ** ($p < 0.01$, strong effect).

Overall, these findings indicate that increasing K is not universally beneficial. Its effectiveness depends on the model's inference strategy: context-sensitive models derive substantial gains from additional examples, while more robust models exhibit diminishing returns.

4.3 Effect of Morphological Hints

We evaluate the impact of morphological hints through paired comparisons between configurations with and without hints, controlling for model, context size, and input instance. To determine whether observed differences are systematic, we apply the Wilcoxon signed-rank test. Table 11 summarizes the results.

For Mistral, the effect is statistically significant across metrics, indicating that hints consistently improve performance. Rather than compensating for missing knowledge, hints appear to reinforce correct morphological decisions by providing surface-level constraints aligned with the model's internal representations. This results in more stable and accurate outputs, particularly in suffix selection.

Qwen exhibits a more selective effect. While hints improve accuracy, their impact on chrF is not consistent, suggesting that their primary contribution lies in guiding the correct application of the requested transformation rather than improving surface similarity. This indicates that Qwen benefits from explicit cues, but does not fully integrate them across all aspects of generation.

In contrast, GPT-OSS does not show statistically significant differences between configurations. This suggests that the model does not consistently leverage hints, and that its predictions are influenced by factors not directly modulated by the additional information provided.

Overall, these findings indicate that morpholog-

Model	Metric	p-value	Sig.
Mistral 24B	Accuracy	$< 1 \times 10^{-6}$	***
	chrF	$< 1 \times 10^{-6}$	***
Qwen 14B	Accuracy	5.55×10^{-3}	**
	chrF	5.21×10^{-1}	n.s.
GPT-OSS 20B	Accuracy	4.54×10^{-1}	n.s.
	chrF	1.66×10^{-1}	n.s.

Table 11: Effect of morphological hints evaluated with the Wilcoxon signed-rank test (paired comparisons between configurations with and without hints).
Significance levels: n.s. ($p \geq 0.05$, no consistent effect), * ($p < 0.05$, mild effect), ** ($p < 0.01$, moderate improvement), *** ($p < 0.001$, strong improvement).

ical hints act as auxiliary signals whose effectiveness depends on the model's inference strategy. They function as a stabilizing mechanism for models with stronger internal representations, and as partial guidance for more context-sensitive systems, but do not provide universal improvements.

Qualitative examples illustrating these effects are provided in Appendix G, where hints guide correct suffix selection in some cases while having no effect in others.

4.4 Efficiency: Performance vs. Latency

Increasing the number of in-context examples (K) improves performance for some models, but also increases the amount of information processed at inference time, leading to longer end-to-end response times. This trade-off is particularly relevant in low-resource settings, where larger prompts may be required to achieve competitive performance. Latency trends across models and values of K are reported in Appendix H.

Differences across models reflect varying degrees of dependence on contextual scaling. Qwen benefits the most from increasing K, but requires longer prompts to reach strong performance. In contrast, Mistral remains comparatively stable with fewer examples, reducing its reliance on extensive context. GPT-OSS exhibits less consistent behavior, with weaker and less predictable gains from additional examples. Overall, these results suggest that improvements obtained through larger context windows should be considered jointly with their computational cost, particularly when resources are constrained.

4.5 Performance by Morphological Category

To better understand the sources of variability observed in previous sections, we analyze performance across morphosyntactic categories. While aggregate metrics suggest relatively stable behavior, Table 12 reveals that performance is not homogeneous across transformation types, but instead reflects systematic differences in linguistic complexity and data distribution. Category-wise trends across models are visualized in Appendix I.

Category	n	Model	Acc.	chrF
ASPECT	90	Mistral	52.22	88.33
	90	OSS	46.67	93.33
	90	Qwen	58.89	95.24
NUMBER	51	Mistral	32.35	84.09
	51	OSS	43.14	90.86
	51	Qwen	23.53	87.49
PERSON	120	Mistral	56.67	90.37
	120	OSS	40.83	91.87
	120	Qwen	54.17	93.77
POSS	120	Mistral	65.83	88.67
	120	OSS	65.83	94.39
	120	Qwen	75.83	95.74
TYPE	84	Mistral	47.62	92.96
	84	OSS	47.62	91.01
	84	Qwen	41.67	92.73

Table 12: Performance by morphosyntactic category and model.

Categories involving regular and localized morphological alternations—such as *PERSON* and *POSS*—are handled consistently well across models. In particular, *POSS* achieves the highest overall accuracy, while *PERSON* maintains strong performance across systems. These categories benefit from predictable suffixal patterns and higher frequency in the dataset, which generally correlates with improved performance by increasing the likelihood of retrieving relevant k-shot examples during inference.

In contrast, *NUMBER* emerges as the most challenging category across all models. Performance drops substantially, especially for Qwen, highlighting a notable exception to this trend: despite its relatively high frequency, NUMBER remains difficult due to structural ambiguity. Plural marking in Quechua can apply to multiple candidate nouns within a sentence, requiring the model to correctly identify the scope of the transformation. This introduces ambiguity that is absent in more localized morphological changes and increases sensitivity to example selection.

A similar, though less pronounced, pattern is observed for *TYPE*. These transformations often involve coordinating multiple elements (e.g., *mana* +

verb + *-chu*), making them partially syntactic rather than strictly morphological. As a result, models may preserve surface similarity while failing to fully reconstruct the intended grammatical structure, leading to moderate accuracy despite high chrF scores.

Differences across models further highlight the interaction between model-specific behavior and category-specific difficulty. Qwen achieves the highest scores in categories such as *ASPECT* and *POSS*, suggesting effective pattern extraction when sufficient contextual evidence is available, but degrades sharply in *NUMBER*, indicating higher sensitivity to ambiguity. Mistral, in contrast, exhibits more balanced performance across categories, maintaining consistently high chrF scores and moderate accuracy even in more challenging cases. GPT-OSS shows intermediate behavior, with relatively strong performance in *NUMBER* but less consistent results overall.

Taken together, these findings indicate that robustness in this task is not determined solely by model size or prompting strategy, but also by the intrinsic properties of each morphosyntactic transformation. Categories with regular structure and higher frequency yield more stable predictions, whereas structurally complex or underrepresented transformations introduce variability that cannot be fully mitigated by increasing k-shot examples alone.

5 Discussion

Across models, the most reliable behavior is not associated with the highest scores, but with consistency across configurations. Mistral 24B exhibits low variance and stable performance regardless of prompt conditions, suggesting that it internalizes morphosyntactic transformations in a way that reduces dependence on external cues. In contrast, Qwen 14B achieves competitive results but shows greater sensitivity to changes in context size and prompting configuration. This indicates that robustness is better characterized by reproducibility than by isolated metric gains.

The interaction between model size and the number of in-context examples reveals two distinct strategies. Larger models, such as Mistral, achieve strong performance with minimal context, whereas smaller models, such as Qwen, rely on increasing K to reach comparable results. However, increasing the number of examples shifts part of the modeling burden from the model itself to data availability. In low-resource settings, where curated examples are scarce and costly to obtain, this introduces an additional constraint: improvements obtained through larger context windows may not scale in practice due to limitations in data collection, annotation, and curation. As a result, robustness depends not only on performance, but on the ability to maintain it under constrained data and inference conditions.

Variation across morphosyntactic categories shows that performance is not determined solely by model architecture or prompting strategy, but also by the intrinsic properties of each transformation. Categories with regular and localized morphology, such as *PERSON* and *POSS*, are consistently easier to model, while structurally complex or ambiguous transformations, such as *NUMBER* or multi-element constructions in *TYPE*, remain challenging even under favorable prompting conditions. This suggests that certain forms of linguistic complexity introduce variability that cannot be fully mitigated through additional context or prompt engineering alone.

6 Conclusion

Our results show that performance in morphosyntactic transformation is shaped not only by model size, but by the interaction between model behavior, context size, and linguistic complexity. Larger models such as Mistral achieve stable performance with minimal contextual support, while smaller models such as Qwen benefit more from increasing the number of in-context examples. Analysis across morphosyntactic categories further reveals that structurally complex transformations remain challenging regardless of prompting strategy.

These findings highlight that robustness in low-resource settings depends on the alignment between linguistic structure and prompt design, rather than model size alone. Future work could extend morphosyntactic coverage, further analyze category-level imbalance, and explore alternative modeling approaches, including hybrid systems that integrate explicit linguistic knowledge with in-context learning.

7 Limitations

Dataset scope and coverage. The proposed dataset, while enabling controlled morphosyntactic transformations, is limited in size and coverage. Al-

though it contains 6,328 transformation instances derived from 471 source sentences, it does not exhaustively represent the full range of morphosyntactic phenomena in Quechua Collao. In particular, the distribution of transformation types is inherently imbalanced, reflecting both the availability of source data and the selective focus on specific grammatical categories.

Linguistic coverage. The dataset focuses on a subset of morphosyntactic categories and does not fully capture the richness of Quechua Collao morphology. While it includes core features such as person, tense, aspect, number, and polarity, other relevant phenomena—such as case marking, subordination, and more complex derivational processes—are not systematically represented. As a result, the task formulation reflects a simplified view of the language, which may limit the generalization of the findings to more complex or less structured linguistic contexts.

Modeling and experimental setup. Our experiments are restricted to a small set of instruction-tuned language models evaluated under a specific in-context learning setup. We do not explore alternative approaches such as fine-tuning, hybrid systems combining symbolic rules with neural models, or retrieval-augmented methods. As a result, the findings are specific to the considered models and prompting strategies, and future work could examine whether similar patterns hold under different modeling paradigms or training regimes.

Evaluation and human validation. The evaluation primarily relies on automatic metrics such as exact match accuracy and chrF, which may not fully capture linguistic acceptability or naturalness in morphologically rich languages. Although we complement this analysis with a human-in-the-loop validation involving native speakers and expert evaluators, this component is limited in scale and based on a relatively small sample of generated instances. While the results show consistent and high ratings across evaluators, the limited number of participants and examples restricts the generalizability of these findings.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Luis Camacho Caballero and Rodolfo Zevallos Salazar. 2020. Lingüística computacional para la revitalización y el poliglotismo. *Revista Letras UNMSM*, 91(134):184–198.

Rodolfo Cerrón-Palomino. 2003. *Lingüística quechua*, 2 edition. Centro de Estudios Regionales Andinos Bartolomé de Las Casas, Cusco.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Nonato Rufino Chuquimamani Valer, Oscar Chávez Gonzales, Felix Alain Riveros Paravicino, César Jara Luna, Moisés Cárdenas Guzmán, and Melquíades Quintasi Mamani. 2021. *Urin qichwa qillqay yachana mayt'u = Manual de escritura quechua sureño*. Ministerio de Educación, Lima.

Beatrice Cueva Medina, Gabriel Fabrizio Tuco Casquino, and José Alfredo Sulla-Torres. 2024. Development of a neural machine translation model optimized with bert for translation from quechua to spanish. In *Proceedings of the 22nd LACCEI International Multi-Conference for Engineering, Education, and Technology*, pages 1–7, San Jose, Costa Rica. LACCEI.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and

fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, and David Yarowsky. 2025. JHU's submission to the AmericasNLP 2025 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 105–111, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

A Prompt Example

We provide an example of the prompt format used for in-context learning. The prompt consists of a table of resolved examples followed by a final instance with an empty target field, which the model must complete.

Each row specifies a source sentence, a set of morphosyntactic changes, and the corresponding transformed target. The final instance follows the same structure but omits the target, requiring the model to infer the correct transformation based on the preceding examples.

```
Source                     | Change            | Target
---------------------------------------------------------------------------
Sipasqa chukuta achalan    | {NUMBER:PL}       | Sipasqa chukukunata achalan
Qayna wayk'urqani wasiypi  | {PERSON:1_PL_INC} | Qayna wayk'urqanchik wasiypi
...
---------------------------------------------------------------------------
<FINAL INSTANCE>
Source                      | Change            | Target
Sipasqa chukuta achalan     | {NUMBER:PL}       |
```

The model is instructed to produce a single output line in the format:

```
FINAL: <target_sentence>
```

This structured format constrains generation by explicitly aligning input transformations with output examples, enabling controlled application of morphosyntactic changes.

B Dataset Construction

B.1 Overview of the Construction Pipeline

The dataset used in this work is constructed through a multi-stage pipeline that transforms a raw parallel corpus into a structured resource for controlled morphosyntactic transformation. The process consists of three main stages: (i) parallel corpus construction, where bilingual sentence pairs are extracted, cleaned, and aligned; (ii) morphosyntactic annotation, where linguistic features are automatically identified and manually validated; and (iii) transformation dataset construction, where annotated sentence pairs are converted into controlled transformation instances. This pipeline enables the generation of high-quality training and evaluation examples in which morphosyntactic changes are explicitly specified and systematically applied, while preserving the original meaning of the source sentence. Figure 1 provides an overview of this process.

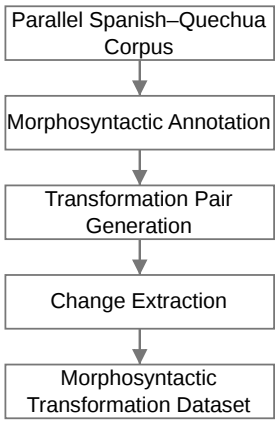

Figure 1: Pipeline for generating the morphosyntactic transformation dataset.

B.2 Parallel Corpus Construction

The parallel corpus is compiled from a combination of heterogeneous bilingual sources, including religious publications and official educational materials. These sources provide complementary properties: while religious texts contribute large-scale and diverse sentence pairs, educational materials offer more consistent orthographic and grammatical usage.

Corpus construction involves multiple preprocessing steps. First, textual data are extracted either through optical character recognition (OCR) for scanned documents or through structured parsing for digital sources. The extracted text is then cleaned to remove formatting artifacts such as line breaks, special characters, and segmentation inconsistencies.

Sentence segmentation is applied to obtain aligned units, followed by heuristic sentence alignment based on length and structural similarity. To ensure alignment quality, a subset of sentence pairs is manually reviewed and corrected.

Given the high degree of orthographic variation in the raw data, especially in non-standardized sources, an additional normalization step is applied using rule-based transformations informed by the official orthographic conventions for Southern Quechua (Chuquimamani Valer et al., 2021), which we adopt here as the reference standard for Quechua Collao (quz). Existing Quechua resources often exhibit inconsistent spelling and may mix surface variants, which introduces noise and affects downstream modeling. In our corpus, this includes forms such as “noqa” or “nuqa”, normalized to “ñuqa”, “qollqe”, normalized to “qullqi”, and “huasi”, normalized to “wasi”. Applying a consistent normalization scheme ensures uniform surface forms across the corpus, improves the reliability of morphosyntactic analysis, and reduces variability unrelated to the linguistic transformations of interest.

B.3 Morphosyntactic Annotation

Morphosyntactic annotation is performed through a semi-automatic pipeline that combines rule-based analysis with syntactic parsing and manual validation. The goal of this stage is to assign structured linguistic features to each sentence, enabling the controlled generation of morphosyntactic transformations.

In the automatic stage, candidate annotations are generated using complementary sources of linguistic evidence. For Quechua, rule-based pattern matching is applied to identify morphological markers such as person, number, tense, and evidentiality. For Spanish, syntactic parsing is used to extract grammatical information that can support disambiguation. These sources are then combined to produce a set of candidate tags along with associated evidence.

Given the ambiguity of certain morphological markers and the potential noise introduced by automatic processing, a manual validation step is performed by native speakers. This step resolves ambiguous cases, adds missing labels, corrects an-

notation errors, and ensures that the assigned tags accurately reflect the intended grammatical structure of each sentence.

For example, the suffix *-n* may mark third-person verbal agreement, as in *rima-n* 'he/she speaks', or third-person nominal possession, as in *wasi-n* 'his/her house'. Such cases are resolved during manual validation by considering the lexical category and syntactic context of the marked word. Similarly, possessive forms may require an epenthetic *-ni* after consonant-final roots, as in *yawar-ni-n* 'his/her blood', rather than the simpler vowel-final pattern *wasi-n*. These checks ensure that the final tags reflect the intended morphosyntactic function rather than only surface string matching.

The final annotated corpus includes, for each sentence, a set of validated morphosyntactic tags that serve as the basis for constructing transformation pairs. Figure 2 illustrates an example of this process, showing both automatically extracted candidates and their manually validated counterparts.

B.4 Transformation Dataset Construction

The transformation dataset is constructed from the annotated corpus by generating pairs of base and transformed sentences under controlled morphosyntactic modifications. Each transformation corresponds to a change in a subset of morphosyntactic features while preserving the underlying propositional meaning of the original sentence.

Controlled transformation generation. Starting from an annotated base sentence, one or more transformed variants are generated by modifying specific grammatical categories such as person, number, tense, or polarity. This process goes beyond surface-level rule application: transformations may require coordinated changes across multiple elements in the sentence to maintain agreement and grammaticality.

This design enables the creation of multiple valid transformation instances from a single base sentence, increasing dataset diversity without introducing uncontrolled noise. **Unlike standard data augmentation approaches**, which rely on perturbations or random variation, transformations in this dataset are linguistically grounded and explicitly defined through morphosyntactic features.

Transformation representation. Each instance is represented as a tuple of the form *(Source, Change, Target)*, where *Change* encodes the difference between the morphosyntactic features of the source and target sentences. This structured representation allows models to infer transformations from explicit grammatical specifications.

Change computation. The set of changes is computed by comparing the annotated features of the source and target sentences, retaining only those categories that differ. This ensures that each transformation is minimally specified and avoids introducing irrelevant information.

Table 13 illustrates how multiple controlled transformations can be generated from a single base sentence.

B.5 Dataset Representation and Change Computation

Each transformation instance is represented as a tuple of the form (Source, Change, Target), where the Change component encodes the morphosyntactic differences between the source and target sentences. The *Source* corresponds to the base sentence, the *Target* to the transformed variant, and the *Change* component encodes the specific morphosyntactic differences between them.

Change computation. The *Change* representation is obtained by comparing the validated morphosyntactic annotations of the source and target sentences. Each set of tags is first normalized into a dictionary of the form *CATEGORY:VALUE*. The change is then defined as the subset of categories whose values differ between the source and target, ensuring that only the relevant grammatical modifications are retained.

For example, given the following pair:

- **Source:** Sipasqa chukuta achalan {PERSON:3_SI, TENSE:PRE_SIM}
- **Target:** Sipaskunaqa chukunkuta achalanku {PERSON:3_PL, NUMBER:PL, POSS:3_PL, TENSE:PRE_SIM}

the resulting change is:

{PERSON:3_PL, NUMBER:PL, POSS:3_PL}

This procedure ensures that transformations are minimally specified and avoids including features that remain unchanged. To further improve consistency and reproducibility, the elements in the *Change* field follow a fixed ordering of categories (e.g., PERSON, NUMBER, POSS, TENSE, ASPECT, MODE, TYPE), which standardizes the representation across all instances.

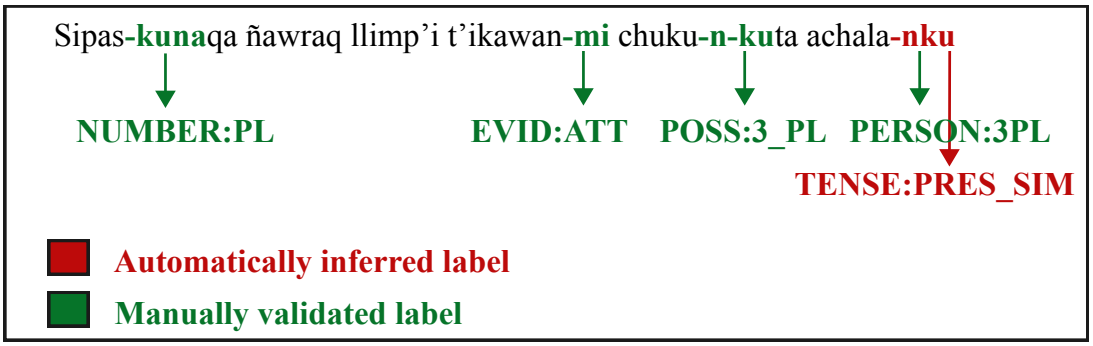

Figure 2: Example of morphosyntactic annotation showing automatically extracted candidates and manually validated tags.

Source	Change	Target
Purikuq runaqa chay wasiman chayarqan	{PERSON:3_PL, NUMBER:PL}	Purikuq runakunaqa chay wasiman chayarqanku
Purikuq runaqa chay wasiman chayarqan	{TENSE:PST_NEXP}	Purikuq runaqa chay wasiman chayasqa
Purikuq runaqa chay wasiman chayarqan	{TENSE:PRE_SIM, ASPECT:PRG}	Purikuq runaqa chay wasiman chayachkan
Purikuq runaqa chay wasiman chayarqan	{MODE:POT}	Purikuq runaqa chay wasiman chayanman

Table 13: Examples of controlled morphosyntactic transformations derived from a single base sentence.

This structured representation provides a clear interface for in-context learning, allowing models to infer transformations from explicit linguistic specifications rather than implicit patterns.

C Morphosyntactic Label Inventory

The morphosyntactic labels used in the *Change* field were selected to cover productive grammatical categories in Quechua Collao that can be modeled as controlled sentence-level transformations. The inventory includes subject person and number, verbal tense, aspect and mood, polarity, interrogation, nominal possession, nominal pluralization, object-person marking, and evidentiality. Although this inventory does not exhaust the full morphology of the language, it provides a structured set of categories for the controlled transformations studied in this work.

Labels follow a `CATEGORY:VALUE` format. For example, `PERSON:1_PL_INC` denotes a subject-person transformation whose value is first-person plural inclusive, while `TENSE:PST_NEXP` denotes a tense transformation to non-experienced past. When only the category is discussed, we use the category name alone, such as PERSON, TENSE, or NUMBER.

Table 14 summarizes the main labels, their meanings, and typical surface realizations.

Table 15 summarizes the main value abbreviations used in label values.

D k-shot Selection Workflow

Figure 3 details the retrieval workflow used to select in-context examples for each test instance. The procedure first prioritizes exact matches in the `Change` specification, then backs off to examples sharing the same primary morphosyntactic category, and finally fills remaining slots using TF-IDF similarity over the source sentence. This ordering is designed to preserve morphosyntactic relevance before relying on surface similarity.

Two additional constraints make the retrieved context auditable and comparable across runs. First, a subset constraint prevents examples from introducing morphosyntactic categories that are absent from the target transformation. Second, a leave-one-out constraint ensures that the evaluated instance never appears among its own demonstrations. Whenever possible, examples from the same transformation cluster are prioritized to preserve lexical and structural proximity.

E Experimental Conditions

This appendix provides additional implementation details for reproducibility, complementing the experimental setup summarized in the main paper.

Label	Meaning	Typical realization	Example
PERSON	Subject person/number	*-ni, -nki, -n, -nchik, -yku, -nku*	*rima-n* ‘he/she speaks’
TENSE	Verbal tense	*-rqa, -sqa, -saq, -nqa*	*wayk’u-rqa-ni* ‘I cooked’
ASPECT	Verbal aspect	*-chka*	*riku-chka-n* ‘he/she is seeing’
MODE	Verbal mood/modality	*-man, -chus*	*taki-n-man* ‘he/she could sing’
TYPE	Clause type or polarity	*mana ... -chu, ama ... -chu*	*mana yacha-ni-chu* ‘I do not know’
SUBTYPE	Clause subtype	*-chu*	*riku-n-chu?* ‘does he/she see?’
NUMBER	Nominal number	*-kuna*	*waka-kuna* ‘cows’
POSS	Nominal possession	*-y, -yki, -n, -nchik, -yku, -nku*	*wasi-y* ‘my house’
PERSON_OBJ	Subject–object person relation	*-yki, -wa-nki, -su-nki*	*qu-yki* ‘I give you’
EVID	Evidentiality	*-mi/-m, -si/-s*	*amawta-m* ‘the teacher’ (attested)

Table 14: Morphosyntactic labels used in the transformation specifications.

Abbreviation	Meaning
1, 2, 3	First, second, third person
SI	Singular
PL	Plural
INC	Inclusive
EXC	Exclusive
PRE_SIM	Present/simple or non-future simple
PST_EXP	Experienced past
PST_NEXP	Non-experienced past
FUT_SIM	Future/simple
PRG	Progressive aspect
POT	Potential mood
DUB	Dubitative mood
IMP	Imperative
NEG	Negation
PROH	Prohibitive
INT	Interrogative
ATT	Attestative evidential
REP	Reportative evidential

Table 15: Abbreviations used in morphosyntactic label values.

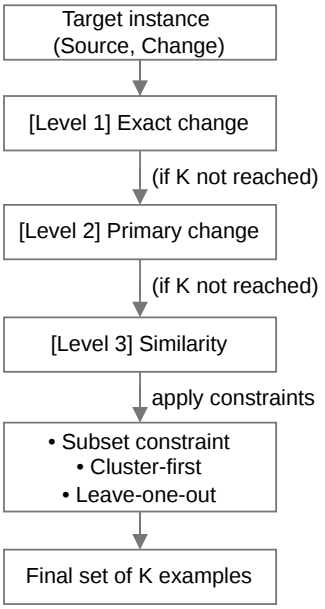

Figure 3: Hierarchical k-shot selection strategy combining exact matching, primary-category retrieval, and similarity-based fallback under structural constraints.

E.1 Local Inference Environment

All experiments were run locally through LM Studio on a Windows machine with an AMD Ryzen 5 3400G CPU, 32 GB of RAM, and AMD Radeon RX Vega 11 integrated graphics reported with 2 GB of adapter memory. No NVIDIA/CUDA device was available in this setup.

E.2 Model Files and Quantization

The evaluated models were loaded in GGUF format through LM Studio. Table 16 reports the exact local model variants used in the final experiments, including quantization format and model size as reported by LM Studio.

E.3 Model Selection Rationale

The model set was selected to provide a compact comparison among instruction-tuned LLMs with different parameter scales and generation profiles. Mistral 24B represents the larger instruction-tuned model in our comparison, Qwen 14B provides a smaller multilingual instruction-tuned alternative, and GPT-OSS 20B introduces a reasoning-oriented generation behavior. The goal was not to exhaustively benchmark all available models, but to analyze how different model profiles respond to the same controlled morphosyntactic transformation setting.

E.4 Decoding and Paired Prompting

All runs used deterministic decoding with temperature $= 0$ and top-$k = 1$, following the reproducibility setting used throughout the pipeline. Mistral and Qwen used a 64-token output cap, which was

Model	Local LM Studio identifier	Quantization	Size
Mistral 24B	`mistral-small-3.2-24b-instruct-2506`	Q4_K_M	15.21 GB
Qwen 14B	`qwen-qwen2.5-14b-instruct-1m`	Q4_K_M	8.99 GB
GPT-OSS 20B	`openai/gpt-oss-20b`	MXFP4	12.11 GB

Table 16: Local model files used in the experiments, as reported by LM Studio.

sufficient for the required single-sentence target format. GPT-OSS 20B was left uncapped only to ensure that the model could reach the required `FINAL:` output line when producing intermediate reasoning.

The experimental grid combines model, number of in-context examples ($K \in \{5, 10, 15\}$), transformation size (1 or 2 morphosyntactic categories), and prompting mode (with or without morphological hints). Paired runs with and without hints used the same selected k-shot examples, so that differences between prompting modes reflect the presence of hints rather than changes in the demonstrations.

F Human Evaluation

F.1 Evaluation Setup

To complement automatic metrics, we conduct a human-in-the-loop evaluation of model outputs. A subset of 44 transformation instances is selected to cover a diverse range of morphosyntactic categories and transformation types, including both single-category and compositional cases.

The evaluated outputs are generated using the best-performing configuration identified in the main experiments (Mistral 24B Instruct). The evaluation is carried out by two native speakers of Quechua Collao and one expert annotator, ensuring both linguistic competence and consistency in judgments.

F.2 Evaluation Protocol

Each generated instance is evaluated along three dimensions:

Grammatical correctness measures whether the output follows the morphosyntactic rules of Quechua Collao and correctly applies the specified transformation.

Naturalness assesses the fluency and acceptability of the generated sentence from a native speaker perspective.

Semantic consistency evaluates whether the output preserves the meaning of the source sentence while accurately reflecting the intended transformation.

Criterion	Average Score
Grammatical correctness	4.87 ($\pm$ 0.60)
Naturalness	4.76 ($\pm$ 0.60)
Semantic consistency	4.82 ($\pm$ 0.60)

Table 17: Human evaluation results (Likert scale from 1 to 5; $\pm$ indicates standard deviation).

Evaluators assign scores using a Likert scale from 1 (very poor) to 5 (excellent) for each dimension. This setup captures both exact-match correctness and perceived fluency, which are particularly important for morphologically rich and low-resource languages.

F.3 Results

Table 17 reports the average scores across evaluators for each evaluation dimension. The results show consistently high ratings, indicating that the generated outputs are generally well-formed, natural, and semantically coherent.

The relatively low and uniform standard deviation indicates stable judgments among evaluators, suggesting that the outputs align well with native speaker intuitions. These findings complement the automatic evaluation results reported in the main text, providing additional evidence of the practical validity of the generated transformations.

G Qualitative Comparison: Hints vs No Hints

Table 18 presents representative examples from Mistral 24B comparing predictions with and without morphological hints. These examples illustrate how hints influence the application of morphosyntactic transformations at the token level.

In particular, hints guide the model toward the correct insertion or modification of morphological markers (e.g., plural suffixes or evidential markers), reducing errors in morpheme placement. However, their effect is not uniform across categories, as some transformations (e.g., negation) are consistently handled correctly even without additional

guidance.

H Latency Analysis

Figure 4 shows response time as a function of the number of in-context examples (K) across models, with and without morphological hints, for both single-change ($size = 1$) and multi-change ($size = 2$) settings.

Latency is measured as the end-to-end response time (in milliseconds) under local inference, computed as the elapsed time between sending a prompt to the model and receiving the generated output. This measurement reflects the combined cost of processing longer prompts and generating responses, rather than an isolated estimate of model-internal inference speed.

Response time increases consistently with K for Mistral and Qwen, reflecting the expected cost of processing longer inputs. In contrast, GPT-OSS exhibits less predictable behavior, with non-monotonic trends across configurations, suggesting that response time is influenced not only by input length but also by model-specific implementation and decoding dynamics.

The effect of morphological hints on response time is comparatively small, indicating that adding structured guidance does not substantially increase computational cost. Overall, these results highlight a trade-off between context size and computational cost, as well as variability in how different models scale with longer inputs.

I Category-wise Performance

This appendix provides additional visualizations of model performance across morphosyntactic categories, disaggregated by the number of in-context examples (K).

Figure 5 shows accuracy by category and model. Performance varies substantially across categories, with transformations such as NUMBER consistently yielding lower accuracy, while categories like POSS and ASPECT exhibit more stable and higher performance. These differences reflect variation in structural complexity and the difficulty of correctly applying specific morphological operations.

The effect of increasing K is also category-dependent. For example, Qwen shows notable improvements in categories such as POSS and ASPECT as more examples are provided, while gains are less consistent for more challenging transformations like NUMBER. This supports the observation that the benefits of additional context depend on both the model and the type of transformation.

Figure 6 presents chrF scores by category. While chrF remains high across most categories, discrepancies between chrF and accuracy are evident, particularly in more complex transformations. This indicates that outputs can be lexically similar to the target while still failing to capture the correct morphological structure.

Together, these plots complement the main results by highlighting how performance varies not only across models and configurations, but also across linguistic categories with differing structural properties.

Change	Source	Target	No hints	With hints
NUMBER:PL	Yana yuraq llimp'iyuq wakataqa allqa waka nini	Yana yuraq llimp'iyuq **wakakunataqa** allqa waka nini	Yana yuraq **llimp'ikuniyuq** wakataqa allqa waka nini	Yana yuraq llimp'iyuq **wakakunataqa** allqa waka nini
TYPE:NEG	Qhipa wiñay warmiqa aqsuta riqsin	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**	Qhipa wiñay warmiqa **mana aqsuta riqsinchu**
EVID:ATT	Inkataqa amawta yana-pasqa	Inkataqa **amawtam** yanapasqa	Inkataqa **amawtami** yanapasqa	Inkataqa **amawtam** yanapasqa

Table 18: Qualitative comparison of predictions from Mistral 24B with and without morphological hints. Bold indicates the modified segment relative to the source.

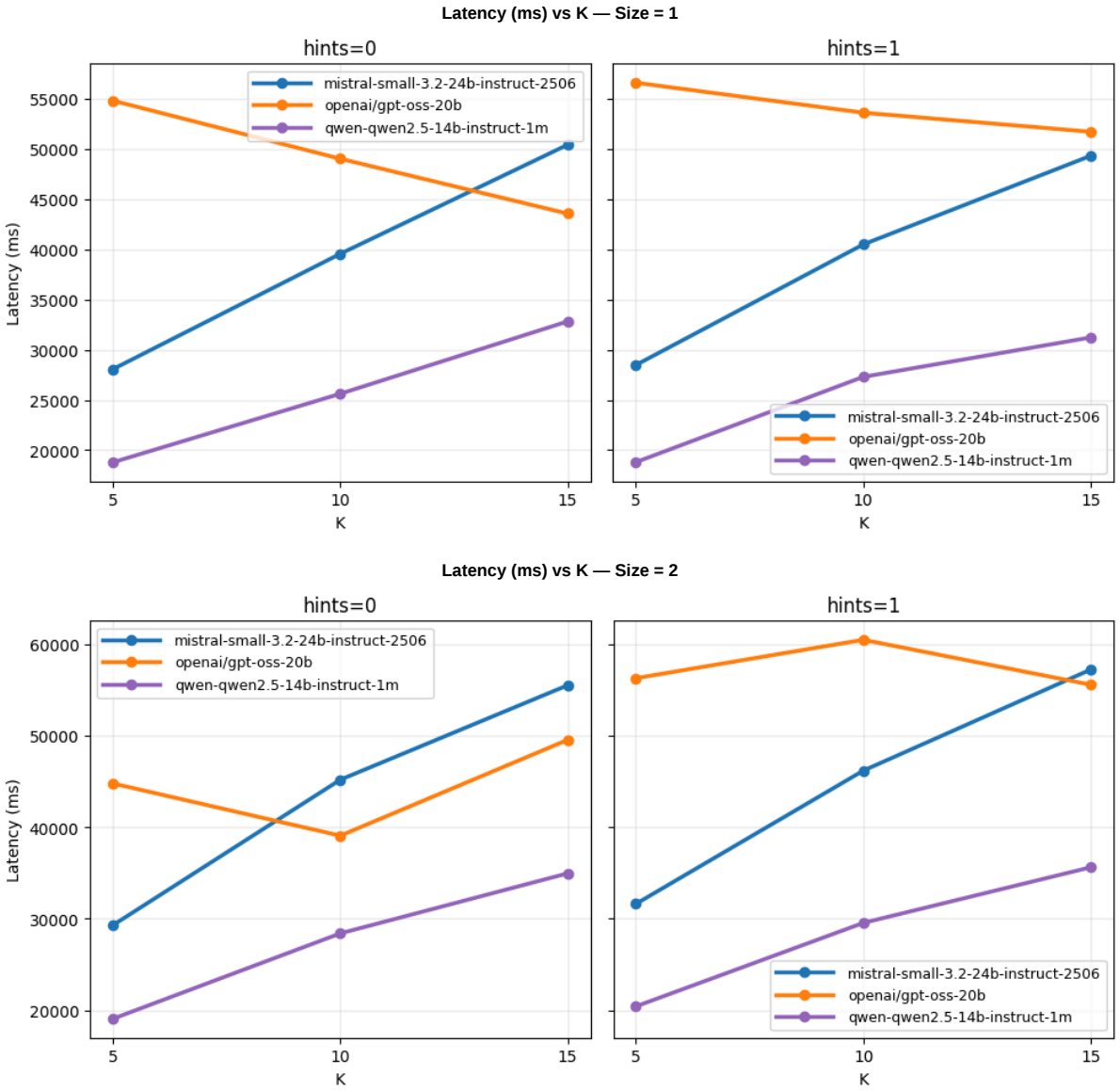

Figure 4: Inference latency as a function of the number of in-context examples (K) across models, with and without morphological hints, for $size = 1$ and $size = 2$.

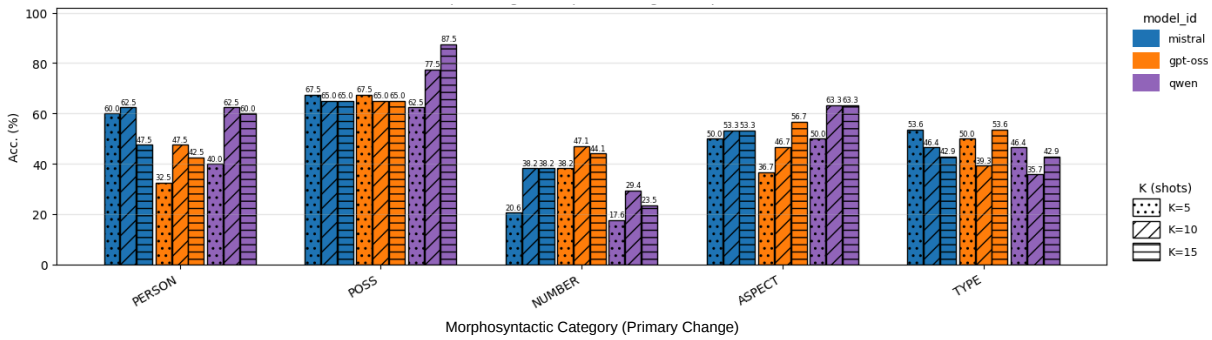

Figure 5: Accuracy by morphosyntactic category and model, disaggregated by the number of in-context examples (K).

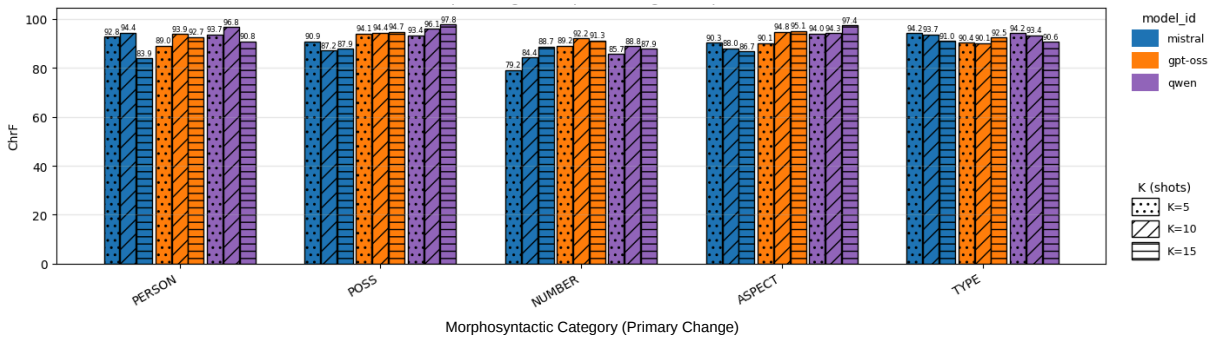

Figure 6: chrF scores by morphosyntactic category and model, disaggregated by the number of in-context examples (K).

Building Community-Centred NLP Resources for Puno Quechua

Elwin Huaman[1], **Adrian Gamarra Lafuente**[2], **Johanna Cordova**[3], **Anna Korhonen**[1]

[1]University of Cambridge (UK), [2]Stanford University (USA), [3]ERTIM - Inalco (France)

Correspondence: elh97@cam.ac.uk, agamarra@stanford.edu, johanna.cordova@inalco.fr

Abstract

The preservation of under-resourced languages requires digital tools and resources shaped by and for their speakers. We present the first dedicated ASR resources for Puno Quechua (ISO 639-3: qxp): (1) the largest speech corpus for any single Quechua variety, consisting in 66 hours of recordings for scripted and spontaneous speech (including 36 hours of manually transcribed and validated data), collected via a participatory design campaign; (2) the first systematic ASR benchmark for Puno Quechua, evaluating state-of-the-art models and fine-tuning Whisper-base, wav2vec2-base, and XLS-R-300M, with and without continued pre-training (CPT); (3) an open release of all datasets and fine-tuned models.

1 Introduction

The revitalization of indigenous languages depends increasingly on digital tools that promote language use and confer economic and social value (Galla, 2016). Puno Quechua (qxp) is spoken by approximately 465,000 people in the Puno region of Peru[1], yet formal education is conducted almost exclusively in Spanish, leaving its speakers largely illiterate in their own language. This literacy gap prevents native speakers from interacting with text-input AI applications (ChatGPT or Gemini), excluding them from the growing digital ecosystem.

An Automatic Speech Recognition (ASR) system represents a more human-centred solution: rather than forcing community members to adapt to text-heavy interfaces, ASR can adapt technology to the oral-centric reality of communities. Despite this need, no dedicated ASR resources exist for Puno Quechua. Prior work on Quechua ASR tends to treat the language family as a homogeneous entity: the variant(s) in question are sometimes not formally identified, or they are aggregated by linguistic group (e.g. Southern Quechua (Cardenas et al., 2018) or Collao Quechua (Paccotacya-Yanque et al., 2022)) without any proper examination of how their differences may affect practical applications. Furthermore, existing corpora suffer from data scarcity, restricted access (*Siminchik* by Cardenas et al., one of the largest corpora referenced in the scientific literature, has never been published in open access), and the absence of a variety-specific benchmark. Figure 1 shows an overview of the Puno Quechua ASR pipeline.

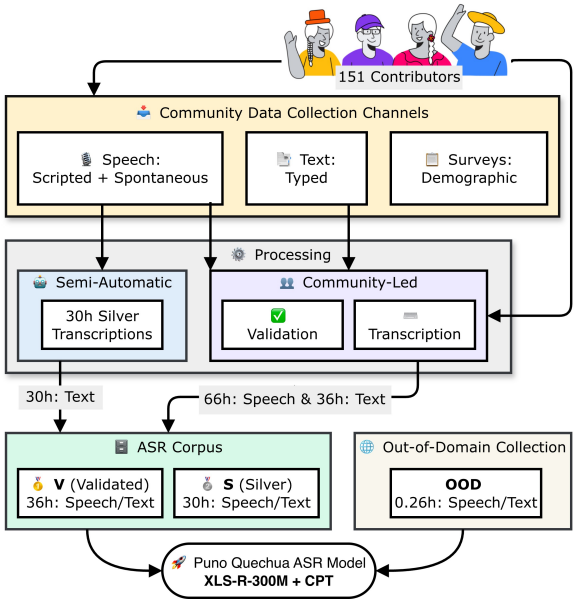

Figure 1: Puno Quechua ASR Pipeline.

This paper makes three contributions: (1) the largest ASR corpus for a Quechua variety via participatory data collection; (2) the first systematic ASR benchmark for Puno Quechua; and (3) open-sourced code and splits,[2] datasets,[3] and models.[4]

[1]Estimation based on 2017 Peruvian National Census, https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1563/

[2]https://github.com/QuechuaBase/asr-puno-quechua

[3]https://mozilladatacollective.com/datasets

[4]https://huggingface.co/QuechuaBase

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 147–152

2 Background

Puno Quechua (qxp). It belongs to the Southern Quechua branch (QIIC) (Torero, 2002) and is characterised by a rich consonant inventory of 26 phonemes, including ejective, aspirated, and uvular stops, and three vowels (/a/, /i/, /u/). It exhibits Aymara influence in its phonology, suffix inventory, and vocabulary (Adelaar, 1987).

ASR for Quechua (que). ASR development for Quechua faces several compounding challenges: (a) Data scarcity: labelled speech data are extremely scarce, and when corpora exist they often aggregate varieties under a macrolanguage label (que), conflating varieties that are not mutually intelligible; (b) Morphological complexity: Quechua is highly agglutinative, resulting in poor word-level WER metrics; (c) Low written literacy: most speakers are illiterate in Quechua, making community-driven transcription difficult, and (d) No variety-specific benchmark: to our knowledge, no prior work establishes a consistent evaluation benchmark for an individual Quechua variety. Recent work by Keren et al. (2025) covers 30+ Quechua varieties within omnilingual ASR, providing reference baselines including for qxp.

3 Participatory Design Data Collection

We collected the Puno Quechua speech corpus through a four-phase participatory design process (Huaman et al., 2025; Ulla et al., 2026; Spinuzzi, 2005; Wilson et al., 2025):

Planning. Identifying the ISO 639-3 code qxp, establishing partnerships with the National University of Altiplano Puno and the local community Illariy Ch'aska, and assessing community needs.

Preparation. Setting up data governance under CC0-1.0 licence, preparing seed sentences and questions covering agriculture, healthcare, and technology, and localising the Mozilla Common Voice platform to Puno Quechua.[5]

Collection. Voluntary, skill-based contributions such as reading, speaking, listening, or writing, with community-led validation and privacy-preserving processing data.

Deployment. Open release on Mozilla Data Collective,[6] certificates of contribution, voucher incentives for participants, and impact assessment.

The campaign ran from January to February 2026. A total of 396 volunteers registered, 292 were confirmed, 151 contributed, and 31 completed the full campaign. The resulting speech data have been aggregated and released as Common Voice Scripted Speech v25 (including v23 and v24 by Huaman et al.) with a total of 34.81 hours (30.5 validated) and Common Voice Spontaneous Speech v3 (including v1 and v2 by Huaman et al.) with a total of 35.3 hours (5.18 validated).

Table 1 summarizes the processed data that has been collected and can be used for training and evaluating models.

Dataset	Type	Validated (V)	Silver (S)
SCS-25	Scripted	30.5	-
SPS-3	Spontaneous	5.5	30.0
Total		**36.0**	**30.0**
out-of-domain (OOD) corpus			
Add_data	Radio	0.27	-

Table 1: Datasets collected for ASR, expressed in hours.

4 ASR models for Puno Quechua (QXP)

Datasets. Two primary corpora were used: i) SCS-25 with 30.5 validated hours of scripted speech; and ii) SPS-3 with 5.5 validated hours of spontaneous speech (after excluding recordings longer than 30 seconds and adding 1 hour validated by a native speaker), supplemented by 30 hours of automatically generated silver transcriptions using omniASR_LLM_7B model. A small OOD corpus (Add_data, ∼16 minutes) sourced from radio and social media, which was transcribed and validated manually by a native speaker, provides a third evaluation domain.

Foundation models. We fine-tuned three architectures: (a) Whisper-base (74M parameters), an encoder-decoder Transformer trained on 680k hours of supervised multilingual speech. We fine-tuned setting the transcript prefix to Spanish. The unbalanced setting (V) used a learning rate (LR) of 5×10^{-6} and the balanced setting (V+S) used a learning rate of 1×10^{-5}. Audio files longer than 30 seconds were excluded. (b) wav2vec2-base (95M parameters), a self-supervised convolutional-Transformer model pre-trained on 960 hours of Librispeech. Both configurations were trained with LR: 1×10^{-4}, with stronger attention dropout (0.1) for the unbalanced corpus (V) to mitigate overfitting.

[5] https://commonvoice.mozilla.org/qxp
[6] https://mozilladatacollective.com/

Audio file longer than 20 seconds were excluded. and (c) XLS-R-300M (315M parameters), a multilingual wav2vec2 model pre-trained on 436k hours across 128 languages (Babu et al., 2022), making it the strongest starting point for low-resource languages with unusual phonological features such as ejectives and uvulars, and allophonic variations. A CTC projection head over a 46-character vocabulary (Puno Quechua Latin orthography) was added. Training runs for 20,000 updates with a tri-stage scheduler and LR: 5×10^{-5}. The encoder was frozen for the first 10,000 updates to prevent the randomly initialised CTC head from corrupting pre-trained representations before it stabilised. The best checkpoint was selected by validation WER.

Continued Pre-Training (CPT). For XLS-R-300M, we additionally performed CPT on the 65 hours of unlabelled Puno Quechua audio prior to fine-tuning. CPT adapts the model's acoustic representations to the target language without requiring transcriptions, and has demonstrated consistent gains in low-resource settings (DeHaven and Billa, 2022; Mutisya and Mugane, 2026). Clips shorter than 1 second or longer than 15 seconds were excluded from training. We train for 10,000 updates (LR: 1×10^{-4}, polynomial decay, 1,000-step warmup), selecting the best checkpoint by validation loss. The best checkpoint occurs at update 9,000 with a validation loss of 2.249. Two models were fine-tuned from the CPT checkpoint (`ft_xlsr_validated` and `ft_xlsr_silver`); using the identical protocol described above for fine-tuned XLS-R-300M.

Reference baselines. We evaluated the omniASR model family (Keren et al., 2025), which combines a wav2vec2-style encoder with either CTC decoding (`CTC_300M`, `CTC_7B`; up to 6.5B parameters) or an LLM decoder (`LLM_300M`, `LLM_7B`; up to 7.8B parameters), and explicitly supports `qxp`. We also evaluated `MMS-1b-all` (1B parameters) by setting the language parameter to Cuzco Quechua (`quz`) for inference, the closest supported variety to Puno Quechua.[7]

5 Experiments and Results

5.1 Baseline with off-the-shelf models

As shown in Table 2, hybrid ASR-LLM models outperform CTC-only variants. The most balanced model across domains, omniASR LLM_7B_v2, achieves a 20.1% mean WER. Notably, all omniASR models handle spontaneous speech more accurately than scripted speech and `OOD` (the CER is significantly better), maybe because the sentences in this corpus are very short (often between 3 and 5 words), and don't provide enough context.

	Scripted		Spontaneous		OOD	
Model	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**
omniASR						
`CTC_300M_v2`	47.8	10.3	29.0	4.4	41.0	6.0
`CTC_7B_v2`	35.4	7.4	18.1	2.7	34.5	5.7
`LLM_300M_v2`	**25.9**	5.8	17.9	2.9	24.4	**3.9**
`LLM_7B_v2`	26.6	6.2	**11.1**	**1.9**	**23.7**	4.1
MMS						
`mms-1b-all`	35.0	**5.3**	36.4	6.5	38.0	6.2

Table 2: Performance of SOTA off-the-shelf models on 1,000 files samples for `qxp` (Scripted, Spontaneous) and on `OOD`.

The MMS model, despite being trained on a different Quechua variety (Cuzco Quechua, `quz`), remains competitive, reaching 5.3% CER on scripted speech.

5.2 Results with fine-tuned models

Training setup. Data were split 70/25/5 (train/dev/test). We compared two training configurations:

- **validated-only** (`V`, 36 hours) with 3× upsampling of spontaneous to compensate for class imbalance,
- **validated-plus-silver** (`V+S`, 66 hours) including silver spontaneous transcriptions.

Fine-tuning evaluation used three test sets: scripted (1.53 h), spontaneous (0.27 h), and `OOD` (0.27 h). All the experiments were conducted on a 48GB L40S single GPU.

Fine-tuned and CPT results. Table 3 reports WER and CER across all fine-tuned conditions. Three findings stand out. First, CPT yields consistent gains on scripted speech: XLS-R+CPT (`V`) achieves 1.19% WER versus 2.06% without CPT (a relative improvement of 42%). Second, silver data is the decisive factor for spontaneous speech: XLS-R+CPT trained on `V+S` reduces spontaneous WER (13.6% → 3.15%, a relative reduction of 77%); the same pattern holds without CPT (13.6% to 6.68%, a relative reduction of 51%). Third, a clear `OOD` generalisation gap persists for all fine-tuned models: silver models achieve ∼35-54% WER versus

[7] Both belong to the Collao linguistic subgroup and share a similar phoneme inventory and writing system.

		Scripted		Spontaneous		OOD		Mean	
Base model	**Dataset**	WER	CER	WER	CER	WER	CER	WER	CER
whisper-base	V	8.57	1.38	26.2	4.13	54.7	10.8	29.8	5.43
whisper-base	V+S	3.81	0.60	17.1	2.74	42.0	7.77	21.0	3.70
wav2vec2-base	V	5.84	0.77	21.6	3.06	54.2	10.3	27.2	4.71
wav2vec2-base	V+S	7.37	0.96	13.9	1.70	50.2	9.45	23.8	4.03
xls-r-300m	V	2.06	0.30	13.6	1.71	35.5	6.03	17.1	2.68
xls-r-300m	V+S	4.36	0.57	6.68	0.81	28.9	4.35	13.3	1.91
xls-r + CPT	V	**1.19**	**0.19**	13.6	1.73	35.0	6.09	16.6	2.67
xls-r + CPT	V+S	2.11	0.30	**3.15**	**0.41**	**27.4**	**4.55**	**10.9**	**1.75**

Table 3: Performance of foundation models fine-tuned on validated corpus (V) and on complete corpus (V+S). WER and CER are expressed in %.

∼27-50% for validated-only models, but the omniASR LLM_7B_v2 still outperforms all fine-tuned systems on OOD (WER: 23.7%), indicating that fine-tuning on in-domain data comes at some cost to out-of-domain robustness.

6 Discussion

Silver data as the decisive factor for spontaneous speech. Across all model architectures, including V+S silver transcriptions drastically reduces spontaneous WER. The effect is most noted for XLS-R+CPT (a relative reduction of 77%) and is consistent even for Whisper-base (26.2% → 17.1%, a relative reduction of 34.7%) and wav2vec2-base (21.6% → 13.9%, a relative reduction of 35.65%). This confirms that the low validation rate of spontaneous speech (14.7%) is a bottleneck for improving ASR system's performance, and that automatically generated silver transcriptions, despite their lower quality, provide crucial coverage of speech variation.

Effect of CPT on scripted speech. CPT on unlabelled Puno Quechua audio provides consistent gains on scripted speech regardless of data configuration. The relative improvement for XLS-R (V) (2.06% → 1.19%, a relative improvement of 42%) demonstrates that adapting the pre-trained acoustic model to the target-language's phonology characteristics is valuable even when the same data are later used for fine-tuning (Getman et al., 2024). The CPT benefit is somewhat diminished when silver data is added (4.36% vs. 2.11% without CPT), suggesting that silver data partially compensates for the lack of language-specific pre-training.

OOD generalisation gap. A clear generalisation gap exists for all fine-tuned models on out-of-domain data. The omniASR LLM_7B_v2 achieves the best OOD WER (23.7%), outperforming the best fine-tuned system (XLS-R + CPT, V+S: 27.4%). This suggests that task-specific fine-tuning on a narrow domain comes at the cost of robustness to unseen acoustic conditions and speaking styles. However, there is a substantial resource disparity involved. The omniASR LLM_7B_v2 operates at ∼7.8B parameters and requires ∼30 GB of VRAM at inference[8], while our XLS-R+CPT competitive model operates at just 317M parameters and ∼2GB of VRAM, making it deployable on commodity hardware. Closing the OOD gap for our model will therefore require not simply more training data, but exploration of lightweight strategies that preserve cross-lingual generalisation while remaining deployable in low-resource settings.

7 Conclusions and Future Work

This paper makes three concrete contributions.

Largest Puno Quechua ASR corpus. We have constructed, to the best of our knowledge, the largest corpus ever prepared for ASR in a single Quechua variety. The corpus comprises 66 hours of recordings for scripted and spontaneous speech (including 36 hours of manually transcribed and validated data), supplemented by 30 hours of automatically transcribed silver spontaneous speech, and 0.27 hours of out-of-domain annotated data. The data were collected through a four-phase participatory design process involving 151 native speakers and released openly via Mozilla Data Collective under a CC0-1.0 licence. The participatory methodology ensured that the corpus reflects domains directly relevant to the community.

[8] https://huggingface.co/facebook/omniASR-LLM-300M

Systematic ASR benchmark for Puno Quechua. We establish the first variety-specific ASR benchmark for Puno Quechua, evaluating SOTA models (omniASR CTC and LLM variants up to 7B parameters; MMS-1b-all) and fine-tuned foundation models (Whisper-base, wav2vec2-base, XLS-R-300M, with and without CPT) across scripted, spontaneous, and out-of-domain test sets. Key findings are: (a) silver transcriptions are the decisive factor for spontaneous speech performance, reducing WER by up to 77% relative; (b) continued pretraining on unlabelled Puno Quechua audio yields consistent gains on scripted speech; (c) omniASR models outperform all fine-tuned systems on out-of-domain data, revealing a generalisation gap that remains an open challenge.

Open release of fine-tuned models. We release all fine-tuned model variants for Puno Quechua, including Whisper-base, wav2vec2-base, and XLS-R-300M under both `V` and `V+S` configurations, as well as the CPT checkpoint and the CPT-based fine-tuned models. The best-performing system, XLS-R-300M with CPT, fine-tuned on `V+S`, achieves 2.11% WER and 0.30% CER on scripted speech, and 3.15% WER and 0.41% CER on spontaneous.

Future research follows from these results. The enrichment of the corpus, in accordance with the quality requirements outlined above, must be continued. Another objective is to incorporate data representative of a wider range of domains, as we have begun to do through the annotation of data crawled from media sources. With the ongoing goal of developing tools that genuinely meet users' needs, we also aim to design more resource-efficient models (for example through quantization) that can be integrated into everyday applications, particularly mobile voice input systems.

Limitations

Despite the contributions presented in this paper, some limitations should be acknowledged.

Corpus size. Although the corpus is the largest for any single Quechua variety, only 14.7% of the 35.3 recorded hours of spontaneous speech have been validated and transcribed, reflecting the difficulty of manual annotation in a community with low written literacy rates in Quechua.

Out-Of-Domain. All fine-tuned models exhibit a clear generalisation gap relative to the off-the-shelf models. Expanding the diversity of training and evaluation domains, e.g., radio, television, and social media, will be necessary to close this gap without sacrificing the parameter efficiency that makes our models deployable on commodity hardware.

Acknowledgements

This research was supported by UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 under the UK government's funding guarantee for ERC Advanced Grants for the project entitled "Towards Globally Equitable Language Technologies (EQUATE)"; by netidee Förderung (`www.netidee.at`); by SILICON Stanford (`silicon.stanford.edu`); and by the French National Research Agency and Ministry of Higher Education, Research and Innovation (MESR).

References

Willem F. H. Adelaar. 1987. Aymarismos en el quechua de puno. *Indiana*, 11:223–231. Published by Gebr. Mann Verlag.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2278–2282. ISCA.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP*, 2:21.

Mitchell DeHaven and Jayadev Billa. 2022. Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training. *arXiv preprint arXiv:2207.00659*.

Candace Galla. 2016. Indigenous language revitalization, promotion, and education: function of digital technology. *Computer Assisted Language Learning*, 29:1137 – 1151.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sámi. In *Interspeech*.

Elwin Huaman, Wendi Huaman, and Jorge Luis Huaman. 2025. Making an under-resourced language available on the wikidata knowledge graph: Quechua language. In *Information Management and Big Data*, pages 212–224, Cham. Springer Nature Switzerland.

Elwin Huaman, Wendi Huaman, Jorge Luis Huaman, and Ninfa Quispe. 2026. Quechua speech datasets in common voice: The case of puno quechua. In *Information Management and Big Data*, pages 184–193, Cham. Springer Nature Switzerland.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. Omnilingual ASR: open-source multilingual speech recognition for 1600+ languages. *CoRR*, abs/2511.09690.

Hillary Mutisya and John Mugane. 2026. Continued pretraining for low-resource swahili asr: Achieving state-of-the-art performance with minimal labeled data. *arXiv preprint arXiv:2603.11378*.

Rosa Y. G. Paccotacya-Yanque, Candy A. Huanca-Anquise, Judith Escalante-Calcina, Wilber R. Ramos-Lovón, and Álvaro E. Cuno-Parari. 2022. A speech corpus of quechua collao for automatic dimensional emotion recognition. *Scientific Data*, 9(1):778.

Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication*, 52(2):163–174.

Alfredo Torero. 2002. *Idiomas de los Andes. Lingüística e historia*. Editorial horizonte.

Petti Ulla, M. Claus Hannah, Barford Anna, Sadek Malak, Reichart Roi, and Korhonen Anna. 2026. COACT – a community-centered, participatory and actionable roadmap for equitable language ai. In *PrePrint*.

Marianne Wilson, David M. Howcroft, Ioannis Konstas, Dimitra Gkatzia, and Gavin Abercrombie. 2025. Participatory design for positive impact: Behind the scenes of three NLP projects. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 252–263, Vienna, Austria. Association for Computational Linguistics.

The Power of Simplicity: N-Grams and Transformers in Nahuatl Language Identification

Luis Armando Mercado-Campos[1] and **Robert Pugh**[2] and **Alexis Palmer**[1]

[1]University of Colorado Boulder, Department of Linguistics

[2]Indiana University Bloomington, Department of Linguistics

{lmercadocampos, alexis.palmer}@colorado.edu, pughrob@iu.edu

Abstract

In the context of real-world language technology applications, the language or variety in which a given text is written is often unknown or uncertain. Yet, this information is crucial in order to adequately select and apply appropriate models or resources. Language identification (LID), or the process of determining the language or variety of a text sample, is thus often an important fundamental task in natural language processing. LID can be particularly challenging when: (1) there are not many labeled texts for training; and (2) similar or related languages are involved, since these may share a number of surface-level features. In this paper, we present an LID system for Nahuatl, a group of closely-related language varieties spoken in Mexico and Central America. Nahuatl LID involves both of the aforementioned challenges: Nahuatl varieties can be quite similar, sharing morphemes and even many lexical items, and there is a relative paucity of representative, variant-labeled Nahuatl text. We describe LID experiments for a total of 11 Nahuatl varieties, achieving generally good results (90.59% $\pm$0.09% in 5-fold cross-validation experiments). Many of the outstanding errors are the result of confusion between three highly-similar Huasteca variants.

1 Introduction

Nahuatl is a group of endangered, closely-related, morphologically rich languages[1] in the Nahuan branch of the Uto-Aztecan family, spoken primarily in Mexico. There are 30 formally recognized variants of Nahuatl (INALI, 2009), which can have substantial phonological, morphological, syntactic, and lexical differences (Gruda et al., 2023). There is also significant variant-internal diversity.

Figure 1: Map showing the geographic distribution of Nahuatl varieties in Mexico. Green, blue, and orange regions correspond to Central, Eastern, and Western varieties, respectively. Adapted from Pugh and Tyers (2024a).

Research on Nahuatl dialectology dates back to at least Lehmann (1920). Since then, researchers generally agree with the variant sub-classifications presented in Lastra (1986), Canger (1988), and Kaufman (2001). While not identical, these three agree on the existence of Eastern varieties, corresponding to one wave of early migration; Central varieties, corresponding to the Nahuatl spoken in the central valley of Mexico; and Western varieties, including Nayarit/Durango Nahuatl. Pharao Hansen (2014) provides additional recommendations for the classification of Eastern and Central/Western varieties based on a survey of linguistic evidence.

Importantly, while Nahuatl languages share a non-trivial number of grammatical features and lexical items, attempts to homogenize or downplay the diversity of Nahuatl varieties can negatively impact revitalization efforts (Hansen, 2013), and also fail to acknowledge the real and often substantial mu-

[1]We use "language," "variant," and "variety" interchangeably, but avoid the use of the term "dialect." In Mexico, this distinction carries real social, educational, and political consequences, as historically the term "dialect" has been used as a tool to marginalize indigenous language speakers (Aguayo Rousell and Piña Osorio, 2016). In government documentation, schooling, and public discourse, referring to these languages as "dialects" can contribute to marginalization and under-resourcing (Parodi, 2011).

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 153–167

tual unintelligibility between speakers of different regional varieties. For example, Huasteca Nahuatl and Isthmus Nahuatl speakers often cannot communicate effectively without prior exposure (de Suárez et al., 1986).

We note that this system is designed as a variety classifier; it assumes the input is already known to be Nahuatl, and should be understood as operating downstream of a general language identification stage. Integrating it with a broader LID system (e.g., one that first filters for Nahuatl before variety classification) is a natural direction for deployment.

In a linguistic context like that of Nahuatl, a reliable variety identifier can serve many purposes. It can help organize and search web-scraped text or mixed-variant corpora such as Gutierrez-Vasques et al. (2016a), facilitating access for speakers, educators, students, and researchers. In applied settings, variety identification can support corpus building and documentation workflows by speeding up metadata assignment and flagging potential mislabels for manual review. Furthermore, for applications like automated dialog systems, being able to determine a user's variant can help ensure that any generated content is aligned with the user's language variety.

1.1 Related work

LID for text is a well-established and widely-researched task in NLP (Jauhiainen et al., 2019). The Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) has heavily focused on distinguishing between similar languages and varieties, including it in numerous shared tasks over the years (Zampieri et al., 2014). Early approaches to this task involved leveraging lexical and character features with linear classifiers, often combined with multi-stage classification (e.g. first identifying the language, then classifying the variety) and ensemble methods (Zampieri et al., 2014, 2015). More recently, approaches typically rely on large, pretrained transformer language models like BERT (Devlin, 2018) or XLM (Lample and Conneau, 2019), though uses of, e.g. SVM or XGBoost classifiers, are still common (Aepli et al., 2023).

In the context of LID for Indigenous languages of the Americas, there has been less focus. Pugh et al. (2025a) explore the task focusing specifically on the languages of Mexico, finding that character n-gram features with Linear SVM provide a strong baseline.

Computational work on Nahuatl is limited but growing. Gutiérrez-Vasques (2018) explores automatic bilingual lexicon extraction, demonstrating alignment potential despite data sparsity. Other work on the language has explored computational approaches to orthographic variation (Gutierrez-Vasques et al., 2025; Guzman-Landa et al., 2025), morphological analysis (Maxwell, 2005; Pugh et al., 2021; Tona et al., 2023; Tyers and Pugh, 2023), syntactic analysis (Pugh et al., 2022; Pugh and Tyers, 2024b), and corpus creation (Gutierrez-Vasques et al., 2016a; Pugh et al., 2025b; Guzmán-Landa et al., 2025). With respect to computational Nahuatl dialectology, Farfan (2019) developed a finite-state morphological analyzer built from a grammar of Classical Nahuatl to explore points of convergence among contemporary written Nahuatl variants. Character language models have also been shown to be effective at measuring the similarity between Nahuatl variants and for language detection (Pugh and Tyers, 2021).

The most relevant work to the present paper is Guzmán-Landa et al. (2026), which investigates methods of Nahuatl variant detection on a custom corpus. Their corpus is not publicly available.

A note on data availability: Data sovereignty is an essential concern for many language communities, and especially many Indigenous communities where people speak endangered languages. The CARE framework (Carroll et al., 2020) encodes ethical data governance practices to support scientific advancement while both safeguarding and honoring Indigenous data sovereignty. The data we have collected includes a mix of publicly-available data and data collected and/or built through direct personal collaborations with speakers and speaker communities. We intend to release at least a portion of this corpus, and we are still in the process of discussing with our collaborators how best to do this, while respecting CARE and general principles of fair use.

1.2 Research question and contributions

The core of this work is the evaluation of models that learn variety-discriminating features directly from naturally occurring text, without relying on hand-crafted linguistic rules. Specifically, we investigate how effectively architectures like XLM-R and supervised machine learning models trained on character n-gram features (as well as ensembles of various models) can capture regional linguis-

tic patterns and orthographic variations to achieve high-accuracy language identification across a diverse linguistic continuum.

We introduce a system for automatic classification of Nahuatl variants, using a weighted soft-voting ensemble combining a linear SVM, a logistic regression model, and XLM-R. We show that this combination is effective for LID, achieving 91% accuracy on a held-out test set. Importantly, our model retains the orthographical variation seen in naturally-occurring data.

2 Linguistic properties of Nahuatl languages

Nahuatl languages are morphologically agglutinative, with extensive use of derivational and inflectional morphology. Canonical word order is typically SOV (Subject-Object-Verb), though constituent order can vary depending on pragmatic or regional factors (Olko et al., 2018; Gruda et al., 2023). Despite their shared history, Nahuatl varieties diverge significantly in phonology, lexicon, and syntax, which reflects both internal change and external influences from Spanish and neighboring Indigenous languages (Parodi, 2011).

Like other Uto-Aztecan languages, Nahuatl has a relatively small but distinct phonemic inventory. The overview in Table 1 represents a synthesized core inventory based on our analysis of both Classical Nahuatl (*nci*) as a stable reference system and documented modern regional innovations. While this work uses text rather than speech, phonological distinctions between varieties are directly relevant because they are often systematically reflected in orthography. For example, the 'tl-t-l' isogloss and the presence or absence of the saltillo (glottal stop) manifest as consistent spelling differences across varieties, making them useful discriminative features for test-based classification. This aggregated inventory draws specifically from the surveys of Canger (2001) and de Suárez et al. (1986) to establish the segments that participate in systematic regional mappings. Rather than reflecting a single 'standard', this inventory serves as a comparative work to ensure that phonological differences (e.g., 'saltillo' ʔ or the /t͡ɬ/ variants) remain available as discriminative features for the classification models.

Place of Articulation	Stop	Affricate	Fricative	Nasal	Liquid	Glottal
Bilabial	p	-	-	m	-	-
Alveolar	t	ts, t͡ɬ	s	n	l, r	-
Palatal	-	t͡ʃ (ch)	ʃ (x)	-	-	-
Velar	k, k^{w}	-	-	-	-	-
Glottal	-	-	h	-	-	ʔ

Table 1: Aggregated consonant inventory across Nahuatl varieties, serving as a reference for understanding the phonological distinctions that may surface as discriminative orthographic features in text. Not all contrasts are present in every variety. This inventory represents the union of documented segments across varieties, with Classical Nahuatl (nci) as a reference point.

3 Data

We begin by assembling a corpus of over 1.1M sentence-level entries from eleven regionally distinct Nahuatl languages. he distribution of this data across varieties, sources, and domains is shown in Table 2, with additional details in Table 10 in the Appendix.

3.1 Corpus Sources and Description

Data is curated from a variety of sources, both publicly available and private. Only some of the private-source data will be made available, by agreement with those sources.

(i) Existing Machine Translation Corpus. Mercado-Campos (2023) compiled a corpus of approximately 31,000 regionally labeled sentences from openly-licensed datasets. This subcorpus includes texts from multiple genres such as dialogues, narratives, sermons, and health pamphlets. Although not structurally parallel, many of these documents address similar domains, particularly education and religious instruction.

(ii) New Collaborator Contributions (2024-2025). In ongoing collaborations, we have collected new data from several Nahuatl varieties. Most of this data has been shared informally through personal interactions, primarily via social media or local archives. This subcorpus includes a mix of narrative fragments, conversational transcripts, and educational material. Consent to use and share this data was granted by the language communities for this research project only.

(iii) Personal Language Learning Collection. As part of an ongoing language learning effort, new examples of one variety were collected from annotated lessons, beginner notes, and other class-related materials. Familiarity with the variety through active language study informed quality

Lang	Name	# of Sentences	Domain
nhw	Western Huasteca Nahuatl	369,830	Daily Life, Bible
nhe	Eastern Huasteca Nahuatl	168,799	Legal, Stories, Bible
azz	Highland Puebla Nahuatl	130,151	Daily Life, Stories, Bible
ncj	Northern Puebla Nahuatl	129,829	Transcripts, Legal, Bible
nch	Huasteca Nahuatl	105,856	Educational, Myths, TikTok Stories, Bible
nhi	Western Puebla Sierra Nahuatl	90,157	Transcripts, Bible
nhg	Tetelcingo Nahuatl	87,859	Daily Life, Bible
nlv	Orizaba Nahuatl	14,636	Educational, Government, Bible
nhm	Morelos Nahuatl	7,515	Educational, History, Bible
nci	Classical Nahuatl	5,466	Poetry
nhn	Central Nahuatl	3,062	Stories, Daily Life, Children Stories
Total		**1,113,160**	

Table 2: Language variety statistics (ordered by # of sentences)

judgments during curation. The informal nature of this data required extra curation to be suitable for use in this project. The subcorpus consists of approximately 43,400 sentences, after filtering out low-quality, duplicate, or dictionary-style entries. We additionally exclude many examples extracted from images due to OCR errors or poor formatting.

(iv) AmericasNLP Shared Task Data. The AmericasNLP 2025[2] shared task competitions (De Gibert et al., 2025) include a small amount of non-parallel Nahuatl text (e.g., daily conversations) from 2 varieties.

(v) Bible Corpus. The massively parallel Bible corpus (Christodouloupoulos and Steedman, 2015) contains the New Testament (NT) for some Nahuatl languages. Between this source and a scripture website,[3] we collect NT Bibles for 16 different Nahuatl languages. There is a significant history of research using the Bible as a data source for very low-resource languages, to varying effect (among others, Chew et al., 2006; Mayer and Cysouw, 2014; Agić et al., 2015; Nicolai and Yarowsky, 2019; Ebrahimi and Kann, 2021; Liu et al., 2021; Kann, 2024; Marashian et al., 2025; Le Ferrand et al., 2025). Although the Bible is a readily-accessible source of parallel data for many languages, it must be used with caution. Researchers have raised concerns about the very particular domain of the Bible, the frequent use of archaic expressions, and – crucially – the fact that many early Bible translations were performed by colonial missionaries with only partial knowledge of the target languages.

(vi) Axolotl Parallel Corpus. Axolotl (Gutierrez-Vasques et al., 2016b; Gutiérrez-Vasques, 2018) is a freely-available parallel Spanish-Nahuatl corpus, accessible online and free to use either through it website or as a Python package (pyElotl Gutierrez-Vasques et al., 2025). Axolotl is also the Nahuatl name for the animal currently known as "ajolote" in Spanish or "axolotl" in English. Axolotl includes Nahuatl data from a range of variants and domains. Using pyElotl, we retrieve a subcorpus of about 13.5K sentences from 6 language varieties, summarized in Table 3.[4]

Lang	# of Sentences
azz	2,884
nci	5,421
nhe	149
nhm	1,938
nhn	1,757
nhw	1,449
Total	**13,526**

Table 3: Portion of Axolotl Corpus used for this study: # of sentences by variant

3.2 Variation within Variants: Orthographies, Registers, and Domains

Nahuatl materials span multiple orthographies and writing habits, adding to the variability seen even

[2]https://github.com/AmericasNLP/americasnlp2025/

[3]https://scriptureearth.org

[4]More detailed statistics in Table 11 in the Appendix.

within one variety. As an example, one important difference is the encoding of vowel length. Vowel length is phonemically contrastive in Nahuatl, meaning that short and long vowels distinguish meaning. Many of the varieties in our corpus use orthography to indicate vowel length, but there is no standardization across varieties in how specifically to write the contrast. Some orthographies mark long vowels with macrons (e.g., ā, ē, ī, ō, ū), others duplicate the vowels (e.g., aa, ee, ii, oo, uu), and others do not represent the contrast at all.

Our system is designed to perform variety classification without any orthographic normalization. We are, however, interested in understanding to what extent the model relies on orthographic differences for classification. To investigate that, we additionally produce a normalized version of the corpus using the `py-elotl` normalizer (Gutierrez-Vasques et al., 2025) set to the INALI standard, which maps orthographic variants to a single standardized form.

Additionally, our corpus includes a range of domains, as well as registers ranging from highly formal Bible translations and historical documents to contemporary daily-life conversations, social media data, and educational transcripts.

Using Bible data. To experimentally assess the effect of including Bible translations in datasets for this task, we create two additional subcorpora (see Section 5.2 for results and discussion). These datasets are restricted to the six languages for which we have both a Bible translation and non-Bible data, and each is balanced across languages. **Setting A** uses a balanced corpus of naturally occurring text, and **Setting B** adds a balanced selection of Bible data across the same languages.

3.3 Data Processing

Unit of Analysis. The goal in this project is language identification at the sentence level. We adopt a sentence-level approach for several reasons. First, many sources in our corpus do not have preserved document boundaries, making the document-level classification impractical. Second, some of the primary practical applications of this system (e.g., flagging mislabeled sentences in mixed-variety corpora) are inherently sentence-level tasks. Finally, in translated and parallel corpora, sentence-level correspondences across varieties are not guaranteed, as translators may convey the same meaning through different constructions entirely, making document-level context potentially misleading rather than helpful. Intuitively, we define "sentence" as a self-contained text unit conveying a single propositional idea, whether a standalone utterance or narrative fragment. This includes full constructions, short utterances common in dialogue, lines of poetry, and transcribed speech segments that may be grammatically incomplete but pragmatically whole. Some examples appear in Table 4.

Category	Nahuatl	English
Retained		
One-word sentence	*nimitstlasohtla*	I love you
Short utterance	*ninomachtia*	I study
Short utterance	*xinichpalewe*	Help me
Poetic line	*in xochitl, in kwikkatl*	the flower, the song
Filtered Out		
Repetitive/Noise	*tla tla tla*	if if if
Unbound prefix	*nino-*	(reflexive prefix) I, me

Table 4: Examples of sentence types retained and removed from the dataset.

We operationalize sentence segmentation utilizing line breaks, punctuation patterns, and formatting markers (e.g., chapter headings) as heuristics, followed by manual validation of the segmentation heuristics and spot-checking a random sample of sentences to verify data quality. Entries are cleaned of duplicate lines and dictionary-style lists, and highly fragmented tokens with insufficient context are removed to ensure data quality.

Sampling and Balancing. Distribution of sentences across language varieties is highly unbalanced (Table 2). While skew of this degree may reflect the prevalence of different varieties in accessible written media, it can obscure the true effectiveness of supervised classification systems. We thus create two versions of the corpus, one unbalanced and one balanced. Instances in the unbalanced version are split using stratified sampling to preserve the empirical regional proportions found in the raw corpus. To address potential domain and quantity bias, the balanced dataset is created by downsampling all varieties to match the smallest class (3062 sentences).

For experimentation, we reserve a held-out evaluation split using stratified shuffle sampling by language. Approximately 20% of the sentences for each Nahuatl variety are set aside as a test set, and the remaining 80% are used for training (and, where relevant, internal validation). This is applied

both to the unbalanced corpus and the downsampled balanced version, so that the label and domain distributions in the test set mirror those of the training data. All results reported are computed on these held-out test splits.

4 Models and Methodology

We train and compare several different classification models. We are specifically interested in combining the interpretability of non-neural models with the robust performance of pretrained multilingual models. To that end, we combine several models into an ensemble. All classification is performed at the sentence level.

4.1 Classification Models

We investigate three different models. First, we train a standard **logistic regression** model using TF-IDF weighted character n-gram features (n=2–5) to predict language variety given an input sentence. Second, using the same features, we train a **linear SVM** for the same task. Logistic regression is effective for sparse input features (Vimal and Anupama Kumar, 2020), while linear SVMs are particularly well-suited to capturing the orthographic and phonological cues highly predictive for Nahuatl (Çöltekin and Rama, 2016). Both model types have been shown to be strong, interpretable models for Indigenous language identification and classification. Together, these models serve as strong, interpretable baselines established for indigenous language identification and classification (Pugh et al., 2025a).

We compare XLM-RoBERTa Large (Conneau and Khandelwal, 2019) and mBERT (Devlin, 2018) as the multilingual baseline. Because of a mismatch between available sizes for these models, our comparison is between models of substantially different scales. XLM-RoBERTa Large (550M parameters), trained on 2.5TB of CommonCrawl data across 100 languages, is a more capable multilingual model for our task, and we find it outperforms mBERT-Large (330M parameters).

4.2 Model Ensemble

We combine the three models in a weighted soft-voting ensemble.

To determine relative weighting of models in the ensemble, we use a two-stage approach. First, we **optimize ensemble weights** using an 80/20 split of the training data. A grid search on this configuration determines the following weights: Linear SVM (0.60), Logistic Regression (0.20), and XLM-R (0.20).

Next, we perform a **robustness check** on the proposed ensemble weights using 5-fold cross-validation across the complete dataset. Models are retrained for each fold, using the ensemble weights above. Accuracies across the five folds range from 90.43 to 90.67. The stability of performance across folds verifies the stability of the proposed weights. NOTE: We performed the cross-validation experiments simply to validate our ensemble weights. No other results reported in this paper come from the cross-validation set up.

4.3 Tokenization

We compare SentencePiece (Kudo and Richardson, 2018) (Unigram LM (Kudo, 2018)), Byte-Pair Encoding (BPE (Sennrich and Haddow, 2015)), and character-level approaches. The custom SentencePiece Unigram tokenizer is selected based on classification perfomance in preliminary experiments (Macro F1 0.76, Accuracy 90%), as it best segments words into meaningful subwords while preserving orthographic patterns.

4.4 Implementation Details

Implemented using a single NVIDIA A100 GPU and Hugging Face Transformers. Maximum sequence length of 128. Training used AdamW optimizer, batch size 32, learning rate 2e-5, and early stopping based on validation loss with a patience of 3 epochs.

5 Results and Analysis

The results of the three individual models and the ensemble system on the held-out data are listed in Table 5. The system achieves an F1 score of 0.91.[5] It is worth highlighting that the ensemble system only slightly outperforms the much simpler Logistic Regression and Support Vector Machine models, a finding that supports the power of simple statistical modeling of character features for text-based LID.

The classification report shows strong performance for several more isolated languages (often approaching 0.99 recall), while the most challenging boundary is within the "Huasteca cluster" (*nch*, *nhw*, and *nhe*). In particular, *nhw* has the lowest

[5]This performance is consistent with the stable average found from the 5-fold cross-validation experiment.

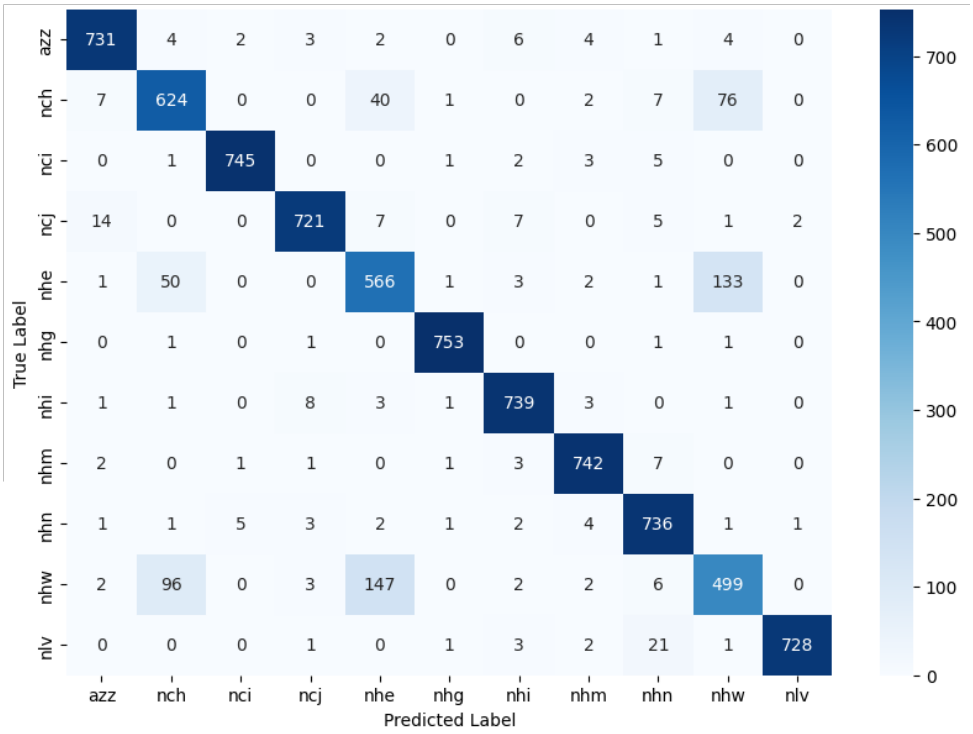

Figure 2: Confusion matrix of the ensemble model's predictions. The most frequently confused variants are the three Huasteca varieties (nhe, nhw, nch), which share a large number of phonological, morphological, and vocabulary features, and are typically considered members of a single Eastern Nahuatl subgroup Canger (2001).

Lang.	LR	SVM	XLMR	Ens.
azz	0.95	**0.96**	0.95	**0.96**
nch	0.79	0.78	0.78	**0.82**
nci	0.98	0.98	0.98	0.98
ncj	0.94	0.95	0.94	**0.97**
nhe	0.72	0.72	0.72	**0.75**
nhg	0.99	0.99	0.99	0.99
nhi	0.96	0.96	0.95	**0.97**
nhm	0.95	0.96	0.96	**0.97**
nhn	0.94	0.94	0.94	**0.96**
nhw	0.65	0.64	0.59	**0.70**
nlv	**0.98**	**0.98**	0.97	**0.98**
Avg	**0.90**	**0.90**	**0.89**	**0.91**

Table 5: Performance (F1 score) of the three component systems (Logistic Regression LR, Support Vector Machine SVM, and XLMR language-model) and the ensemble of the three.

Class	Top 10 discriminative features (LR)
azz	ta, w, ne, j, k, eju, ten, no, de, ejua
nch	ta, t, cati, ti, j, hu, inta, tal, ten, tali
nci	y, uh, yn, in, z, au, auh, l, v, ll
ncj	in, n, ahmo, ahm, iya, huan, ehh, ehhu, ehhua, aki
nhe	ax, tlen, j, hu, len, queja, ueja, tl, eja, imo
nhg	ie, e, que, tie, hua, ua, nu, tlo, tli, o
nhi	h, u, ua, eh, queh, ueh, uan, aua, ou, iu
nhm	k, non, a, o, h, nin, z, i, ce, kej
nhn	k, w, l, wa, ce, in, ka, tl, ll, ts
nhw	catli, atli, catl, j, hu, ij, an, se, tli, quen
nlv	k, h, w, ki, eh, iwa, iw, iwan, ka, lli

Table 6: Top 10 discriminative character n-grams per class from the logistic regression model trained on TF–IDF features.

recall (0.66), with persistent confusion against *nch* (approximately 15% of confusions), which supports a continuum interpretation for closely related regional languages rather than a sharply separable boundary. The complete confusion matrix heatmap can be seen in Figure 2.

5.1 Linguistic vs. Orthographic Features

The model appears to rely heavily on surface cues, including some well-established isoglosses as well as orthographic tendencies (artifacts of the specific texts in the corpus vs. dialectological features).

In Table 6, we show the top-most discriminative features in the Logistic Regression model. For example, one well-known isogloss among Nahuatl varieties is the "tl-t-l" distribution, where some varieties use "tl", others "t", and still others "l" for corresponding lexical items and morphemes. We see that this is valuable for distinguishing between varieties, since two "t-varieties" in our corpus, azz and nch, have sequences like *ta* and *ten* (*tla* and *tlen* in other varieties) in their top discriminative features. Other useful discriminative linguistic features are the lexical item *auh* for nci, a particle that is prevalent in Classical Nahuatl texts, and *ax*, a verbal negation prefix in the Huasteca region, as the top discriminative feature for nhe.

Language	Precision	Recall	F1-score
azz	0.96	0.94	0.95
nch	0.79	0.84	0.82
nci	0.97	0.97	0.97
ncj	0.95	0.95	0.95
nhe	0.72	0.77	0.75
nhg	0.99	1.00	0.99
nhi	0.95	0.94	0.94
nhm	0.94	0.94	0.94
nhn	0.93	0.91	0.92
nhw	0.70	0.65	0.67
nlv	0.98	0.96	0.97
Accuracy			0.89
Macro Avg	0.89	0.89	0.89

Table 7: Classification performance, final ensemble model, on orthographically normalized version of the balanced dataset. cf. rightmost column of Table 5.

Additionally, the list of top discriminative features highlights the value of orthographic cues in our experiment, such as *k* and *w* for nlv and nhn, substrings containing *y* for nci, and *h* and *u* (followed by a vowel) for nhi. It is important to note that, while some orthographic practices may be adopted by specific communities (e.g. the so-called "Tenango" orthography described in Pugh et al. (2025b)), orthographic patterns are largely an artifact of text/author rather than the Nahuatl variety. In order to evaluate the extent to which our model learned to perform LID via document-specific rather than language-specific features, we train and evaluate the ensemble model on an orthographically-normalized version of the corpus, produced using the py-elotl normalizer (Gutierrez-Vasques et al., 2025) set to the INALI standard. These results, shown in Table 7, show that indeed, to some extent, orthographic patterns alone contribute to some of the performance, since the normalized experiment shows small but consistent drops for all varieties.

5.2 The effect of domain

We also briefly explore the impact of the source text domain. Specifically, we compare performance on two small subcorpora (see Section 3.2) containing only the six languages where we have both Bible and non-Bible data. Table 8 shows results for Settings A (only non-Bible data) and B (mixed Bible and non-Bible data) for those six languages, in a very small data setting. Table 9 shows the same for five of those six languages, increasing the dataset size by removing the smallest language.

Although inclusion of Bible data doubles the amount of data available, performance in Setting B either stays the same as Setting A or decreases. This suggests that the Biblical register may introduce domain-specific noise that obscures authentic regional features. Variety-specific scores, though, fluctuate significantly with the inclusion of Bible data. For instance, in Table 8, the *nhw* variety drops from 0.84 to 0.77, while *nlv* increases from 0.82 to 0.91. This suggests that the model may be learning the standardized, formal voice of the Bible rather than capturing authentic regional cues. The domain of religious translations introduces a stylistic similarity that can distort performance metrics for specific variants, motivating our decision to prioritize naturally occurring text in our final model iterations.

	Setting A (No Bible)		Setting B (With Bible)	
Variety	**F1-Score**	**Support**	**F1-Score**	**Support**
azz	0.95	74	0.90	147
nch	0.82	74	0.76	148
nhi	0.89	73	0.88	147
nhm	0.89	73	0.91	147
nhw	0.84	74	0.77	147
nlv	0.82	74	0.91	148
Macro Avg	**0.87**	442	**0.86**	884
Accuracy	**0.87**	442	**0.86**	884

Table 8: Performance comparison between non-Bible data (Setting A) and mixed data (Setting B), for 6 languages. Balanced settings, with 370 instances per language in Setting A and 735 in Setting B.

Variety	Setting A (No Bible) F1-Score	Support	Setting B (With Bible) F1-Score	Support
azz	0.93	162	0.95	325
nch	0.83	163	0.83	326
nhi	0.93	163	0.95	325
nhm	0.89	163	0.93	325
nhw	0.90	162	0.86	325
Macro Avg	**0.90**	813	**0.90**	1626
Accuracy	**0.90**	813	**0.90**	1626

Table 9: Performance comparison between non-Bible data (Setting A) and mixed data (Setting B), for 5 languages. Balanced setting, with 810 instances per language in Setting A and 1625 in Setting B.

6 Concluding remarks and future work

This paper introduces an ensembled text classification model to address the challenges posed by the lack of labeled corpus data for varieties of Nahuatl, a minoritized language family often treated incorrectly as a single language. By leveraging a weighted soft-voting ensemble combining Linear SVM (0.60), Logistic Regression (0.20) and XLM-R (0.20), the best system achieves a robust classification accuracy of 91% on naturally occurring test data.

Results indicate that transformers and character n-gram models can implicitly learn cross-variant correspondences robustly from naturally occurring text, effectively navigating the linguistic diversity of Nahuatl.

In future work, we plan to incorporate more conversational, social media, and community-generated content, to expand the scope of the system and to avoid formal and/or translated Bible data. We also plan to investigate the effectiveness of moving beyond the surface-level subwords we get from tokenization to a morphologically-motivated approach, which may capture useful morphological structures. This sort of approach may be especially useful for the agglutinative nature of Nahuatl languages.

Other future directions highlight the importance of engaging with speaker communities. First, we hope to integrate feedback from speaker communities to refine the classification system, with a particular focus on determining meaningful units of analysis. Finally, we plan to incorporate this classifier to develop variety-appropriate language processing tools such as spell-checkers, OCR post-correction, and machine translation.

Limitations

This work has several limitations that should be considered when interpreting the results.

First, although we control for data imbalance through downsampling in our experimental design, there is a tradeoff between building balanced datasets and using all available data. Training on the pre-sampling corpus, which is uneven in both size and domain distribution, may improve performance for many varieties, but it may also bias the models toward higher-resource varieties and more represented genres.

Second, as shown in our orthographic normalization experiments, a portion of model performance appears to rely on surface-level orthographic conventions rather than deeper linguistic distinctions. This could potentially limit generalization to unseen writing styles or communities with different conventions. Some varieties in our corpus are heavily skewed toward a single domain; most notably Classical Nahuatl (nci), which appears exclusively in the poetry domain. This raises the possibility that the model may be learning domain or register signal rather than variety-specific linguistic features for these classes. Future work should evaluate performance on held-out domains to disentangle these effects.

Third, the inclusion of domain-specific data, particularly Biblical text, introduces stylistic regularities that may not reflect everyday language use, affecting the classifier's flexibility. Additionally, our sentence-level formulation may not fully capture broader discourse-level cues relevant for variety identification.

Fourth, the system does not perform open-set language identification: it will assign one of the 11 Nahuatl variety labels to any input, including non-Nahuatl text. In practice, this classifier should be composed with a general-purpose LID system that first identifies the input as Nahuatl.

Finally, given the limited availability of labeled data and focus on 11 varieties, the generalizability of our findings to the full diversity of Nahuatl variants remains an open question.

Ethical considerations

This work uses data from an Indigenous language, where questions of data ownership, representation, and use are especially important. In line with CARE principles for indigenous data governance,

we recognize that not all data used in this study can or even should be freely redistributed.

Portions of the corpus were collected through direct collaboration with speakers and are used here with permission for research purposes only. Any future data release will be done in accordance with this to ensure that community preferences, access restrictions, and appropriate attributions are respected.

We also acknowledge that automatic language or variety identification systems may have unintended consequences if used without care. For example, misclassifications could affect downstream applications such as educational tools or language technologies, potentially reinforcing incorrect associations or privileging certain variants over others. Also, modeling decisions that treat language varieties as fixed and separable categories may not align with speakers' own linguistic identities or practices.

For these reasons, we emphasize that such systems should be developed and applied in collaboration with communities, with attention to their goals, expectations, and concerns.

7 Acknowledgements

Thanks to the anonymous reviewers for helpful and interesting suggestions. Thanks also to those who helped with gathering additional corpus data, without whom this project would have been even more challenging. Specifically, we extend our thanks to Rodrigo Ortega Acoltzi, who translated "The Birth of the Fifth Sun," Lydia Leija, who translated "Theft of Music," and Chicome Itzcuintli Amatlapalli, who authored both books. And thanks to members of the LECS Lab at CU Boulder for comments and suggestions. This work was supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another."

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial Evaluation Campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. "If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages". In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Hilda Berenice Aguayo Rousell and Juan Manuel Piña Osorio. 2016. Expressions of racism in a sample of university students in Mexico. *Sinéctica*, (46).

Una Canger. 1988. Subgrupos De Los Dialectos Nahuas (1988). In J. Kathryn Josserand and Karen Dakin, editors, *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan.Part. Oxford: BAR International Series 402 (Ii)*, volume 402 of *BAR International*, pages 473–98. BAR, Oxford.

Una Canger. 2001. Nahuatl dialectology: A survey and some suggestions. *Tonos: Revista de Estudios Filológicos*.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita C. Holbrook, Raymond Lovett, Simeon Materechera, Mark A. Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE Principles for Indigenous Data Governance. *Data Sci. J.*, 19:43.

Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. 2006. "Evaluation of the Bible as a Resource for Cross-Language Information Retrieval". In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74, Sydney, Australia. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Çağrı Çöltekin and Taraka Rama. 2016. "Discriminating Similar Languages with Linear SVMs and Neural Networks". In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexis Conneau and Kartikay Khandelwal. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, Shruti Rijhwani, Katharina Von Der Wense, and Manuel Mager. 2025. "Findings of the AmericasNLP

2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas". In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Lastra de Suárez et al. 1986. Las áreas dialectales del náhuatl moderno.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abteen Ebrahimi and Katharina Kann. 2021. "How to Adapt Your Pretrained Multilingual Model to 1600 Languages". In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

J.I.E. Farfan. 2019. *Nahuatl Contemporary Writing: Studying Convergence in the Absence of a Written Norm*. University of Sheffield.

Szymon Gruda, Gregory Haimovich, and John Sullivan. 2023. Lexical creativity in modern Nahuatl: An analysis of multidialectal data. *Lingua*, 285:103488.

María Ximena Gutiérrez-Vasques. 2018. EXTRACCIÓN LÉXICA BILINGÜE AUTOMÁTICA PARA LENGUAS DE BAJOS RECURSOS DIGITALES.

Ximena Gutierrez-Vasques, Robert Pugh, Victor Mijangos, Diego Barriga Martínez, Paul Aguilar, Mikel Segura, Paola Innes, Javier Santillan, Cynthia Montaño, and Francis Tyers. 2025. "Py-Elotl: A Python NLP package for the languages of Mexico". In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 38–47, Albuquerque, New Mexico. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016a. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016b. "Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Miguel Figueroa-Saavedra, Carlos-Emiliano González-Gallardo, Graham Ranger, and Martha Lorena-Avendaño-Garrido. 2026. Classifying several dialectal Nawatl varieties. *arXiv preprint arXiv:2601.02303*.

Juan-José Guzmán-Landa, Juan-Manuel Torres-Moreno, Martha Lorena Avendaño Garrido, Miguel Figueroa-Saavedra, Ligia Quintana-Torres, Graham Ranger, Carlos-Emiliano González-Gallardo, Elvys Linhares-Pontes, Patricia Velázquez-Morales, and Luis-Gil Moreno-Jiménez. 2025. "π-YALLI : un nouveau corpus pour des modèles de langue nahuatl / Yankuik nawatlahtolkorpus pampa tlahtolmachiotl". In *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, pages 802–816, Marseille, France. ATALA \\& ARIA.

Juan-José Guzman-Landa, Jesús Vázquez-Osorio, Juan-Manuel Torres-Moreno, Ligia Quintana Torres, Miguel Figueroa-Saavedra, Martha-Lorena Avendaño-Garrido, Graham Ranger, Patricia Velázquez-Morales, and Gerardo Sierra-Martínez. 2025. A symbolic algorithm for the unification of nawatl word spellings. In *Mexican International Conference on Artificial Intelligence*, pages 141–154. Springer.

Magnus Pharao Hansen. 2013. Nahuatl in the plural: Dialectology and activism in Mexico. In *Proceedings of the American Anthropological Association, Annual Meeting*.

INALI. 2009. *Catálogo de las lenguas indígenas nacionales: Variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas*. Instituto Nacional de Lenguas Indigenas, México, D.F.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Amanda Kann. 2024. "Massively Multilingual Token-Based Typology Using the Parallel Bible Corpus". In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.

Terrence Kaufman. 2001. The history of the Nawa language group from the earliest times to the sixteenth century: Some initial results. *Paper posted online at http://www. albany. edu/anthro/maldp/Nawa. pdf. University of Pittsburgh*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yolanda Lastra. 1986. *Las areas dialectales del nahuatl moderno*. Universidad Nacional Autonoma de Mexico, Instituto de Investigaciones Antropologicas.

Eric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. "Faithful Transcription: Leveraging Bible Recordings to Improve ASR for Endangered Languages". In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 333–342, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Walter Lehmann. 1920. Die Sprachen Zentral-Amerikas in ihren Beziehungen zueinander sowie zu Sud-Amerika und Mexiko, 1/2. *Zentral-Amerika, Teil I*.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. "The Usefulness of Bibles in Low-Resource Machine Translation". In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. "From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation". In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Mike Maxwell. 2005. "Language Documentation: The Nahuatl Grammar". In *Computational Linguistics and Intelligent Text Processing*, pages 474–485, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thomas Mayer and Michael Cysouw. 2014. "Creating a massively parallel Bible corpus". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Luis Armando Mercado-Campos. 2023. Design and implementation of an NMT system for Spanish-Nahuatl. Master's thesis, Universidad del País Vasco / Euskal Herriko Unibertsitatea, Donostia-San Sebastián, España, June.

Garrett Nicolai and David Yarowsky. 2019. "Learning Morphosyntactic Analyzers from the Bible via Iterative Annotation Projection across 26 Languages". In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.

Justyna Olko, R. Borges, and John Sullivan. 2018. Convergence as the driving force of typological change in Nahuatl. *STUF - Language Typology and Universals*, 71:467 – 507.

Claudia Parodi. 2011. Multiglosia virreinal novohispana: el náhuatl. *Cuadernos de la ALFAL*, 2:89–101.

Magnus Pharao Hansen. 2014. The East-West split in Nahuan Dialectology: Reviewing the Evidence and Consolidating the Grouping. In *Friends of Uto-Aztecan Workshop*.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. "Universal Dependencies for Western Sierra Puebla Nahuatl". In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27.

Robert Pugh and Francis Tyers. 2024a. Experiments in multi-variant natural language processing for Nahuatl. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 140–151, Mexico City, Mexico. Association for Computational Linguistics.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards and Open Source Finite-State Morphological Analyzer for Zacatlán-Ahuacatlán-Tepetzintla Nahuatl. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 80–85.

Robert Pugh, Francis Tyers, and Brian OConnor. 2025a. 8. Implementación de Identificación de Idiomas para las Lenguas Indígenas de México. *UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO*, page 88.

Robert Pugh and Francis M. Tyers. 2024b. A Universal Dependencies Treebank for Highland Puebla Nahuatl. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Robert Pugh, Cheyenne Wing, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, and Francis Tyers. 2025b. "Ihquin tlahtouah in Tetelahtzincocah: An annotated, multi-purpose audio and text corpus of Western Sierra Puebla Nahuatl". In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3549–3562,

Albuquerque, New Mexico. Association for Computational Linguistics.

Rico Sennrich and Alexandra Haddow. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ana Tona, Guillaume Thomas, and Ewan Dunbar. 2023. "A morphological analyzer for Huasteca Nahuatl". In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 112–116, Remote. Association for Computational Linguistics.

Francis Tyers and Robert Pugh. 2023. "A finite-state morphological analyser for Highland Puebla Nahuatl". In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 103–108, Toronto, Canada. Association for Computational Linguistics.

Bhartendoo Vimal and S Anupama Kumar. 2020. Application of logistic regression in natural language processing. *Int J Eng Res*, 9(06).

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. "A Report on the DSL Shared Task 2014". In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. "Overview of the DSL Shared Task 2015". In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

A Appendix

A.1 Corpus details

Table 10 shows in more detail the distribution of text types per language in our corpus.

Table 11 shows detailed information about language varieties and sources in the Axolotl corpus (Gutierrez-Vasques et al., 2016b,a), from which we use a subset.

Lang	Source	# of Sentences	Avg. Sent. Length (tokens)
nhn	El Ajolote de Xochimilco (Story)	133	3.04
	Classes (Daily life)	1,700	13.12
	Historias para Niños (Stories)	2,300	2.39
nhi	UD (Transcripts)	813	8.24
	UD (Transcripts)	90,200	5.48
nhg	Conversaciones (Daily life)	30	4.10
	Diccionario de Frases (Daily life)	87,800	4.10
nch	UTexas (Educ.)	1,000	2.30
	Mitos en Náhuatl (Myths)	104,800	5.33
ncj	Bible (Religion)	88,900	3.92
nci	4 Poemas Cortos (Poetry)	121	3.26
	Poemas en Náhuatl (Poetry)	5,300	23.47
nlv	La lengua de los Aztecas (Educ.)	28	9.82
	Guiones Bilingües (Legal)	340	1.17
	Bible (Religion)	14,300	11.62
azz	Conversaciones (Transcripts)	2,900	6.31
	OpenSLR (Mixed)	40,900	8.66
	Bible (Religion)	127,300	4.53
nhe	Conversaciones (Mixed)	149	161.07
	Bible (Religion)	168,600	5.02
nhm	Various Books (Mixed)	1,900	10.26
	Bible (Religion)	5,600	7.46
nhw	Constitución (Legal)	1,400	3.71
	Translated songs (Daily life)	2,100	1.14
	Bible (Religion)	366,300	5.11

Table 10: Dataset Statistics: Variant, Source (Domain)

Corpus	Sentences	Variant	Domain
Mitos y cuentos nahuas de la Sierra Madre Occidental	93,466	azz	Short stories
Los cuentos en náhuatl de Doña Luz Jiménez	34,801	azz	Short stories
Primer Axotli Libro	15,794	nci	Didactic
López Austin, Alfredo. Augurios y Abusiones	32,392	nci	Historical
Documentos nauas de la Ciudad de México del siglo XVI	123,473	nci	Historical
Testimonios de la Antigua Palabra	43,444	nci	Historical
Chimalpain Cuauhtlehuanitzi	34,203	nci	Historical
Historia de México narrada en náhuatl y español	10,881	nci	Historical
Anales de Tepeteopan	8,162	nci	Historical
Nican Mopohua	9,176	nci	Literature
Veinte Himnos Sacros de los Nahuas	9,296	nci	Literature
Trece Poetas del Mundo Azteca	8,405	nci	Literature
La tina negra y roja	28,657	nci	Literature
La llave del náhuatl	26,957	nci	Literature
La tierra nos escucha	7,146	nci	Literature
Teatro náhuatl II Selección y estudio Crítico	38,939	nci	Literature
Recetario Nahua el Norte de Veracruz	24,040	nci	Recipes
Recetario Nahua de Milpa Alta	18,836	nci	Recipes
Antología del cuento náhuatl	36,469	nci	Short stories
Adivinanzas	326	nci	Short stories
Lo que relatan de antes. Kuentos tének y nahuas de la Huasteca	14,134	nci	Short stories
Reyes García, Luis y Christensen, Dieter. El anillo de Tlalocan	23,102	nci	Short stories
Garibay, vida económica de Tenochtitlan	30,794	nhe	Historical
Revista La lengua y cultura Nahuatl	33,800	nhe	Magazine
Untitled	46,868	nhm	Historical
Yancuitlalpan, tradicion y discurso ritual	16,527	nhm	Short stories
El Náhuatl de Tetzcoco en la Actualidad	32,825	nhn	Didactic
Método autodidáctico español-náhuatl náhuatl-español	29,166	nhn	Didactic
La voz profunda	6,778	nhn	Literature
Revista Amerindia	12,710	nhn	Magazine
Cuéntos Indígenas de México	496	nhw	Musical
TOTAL	**851,563**		

Table 11: Axolotl Corpus Overview (Sorted by Variant and Domain)

RAN: Resource Abundance Notation for Languages in NLP

Jared R. Coleman[1] **Tainã G.D. Coleman**[2] **Bhaskar Krishnamachari**[3]
[1]Loyola Marymount University [2]San Diego Supercomputer Center
[3]University of Southern California
jared.coleman@lmu.edu t1coleman@sdsc.edu bkrishna@usc.edu

Abstract

The term "low-resource" is used pervasively in NLP but communicates almost nothing precise. We propose **RAN (Resource Abundance Notation)**, a compact, multi-dimensional notation for quantifying a language's NLP resource profile. A RAN score is written as S/M/L_1-B_1/L_2-B_2/. . . , where $S = \lfloor \log_{10}(\text{speakers}) \rfloor$, $M = \lfloor \log_{10}(\text{monolingual sentences}) \rfloor$, and each L_i-B_i pair records a bilingual partner and $\lfloor \log_{10}(\text{parallel sentences}) \rfloor$. Values derive from canonical sources: Wikidata for speakers, OSCAR 23.01 for monolingual corpora, and (where available) OPUS for parallel corpora. We score 20 typologically diverse languages (including Quechua, Guarani, Cherokee, and Owens Valley Paiute) and correlate each profile against published benchmarks for machine translation (MT, via NLLB-200 chrF++), named entity recognition (NER, via XTREME XLM-R WikiANN F1), and part-of-speech tagging (POS, via XTREME XLM-R UD accuracy). The RAN components carry complementary information: a linear model using all three explains 52% of MT variance, 76% of NER variance, and 72% of POS variance. Among single predictors, $B_{\max}$ (the largest bilingual corpus, regardless of partner) is strongest for the cross-lingual transfer tasks (NER, POS), while M and B_{en} are strongest for MT. RAN is designed first as a *communication* tool, not a predictive model.

1 Introduction

"Low-resource" appears in thousands of NLP papers with no shared definition. A paper may use the term whether a language has 500 parallel sentences or 500,000, whether it has 10 fluent speakers or 10 million. This makes it difficult to compare results across papers, assess whether a technique might transfer to a new language, or communicate the difficulty of a task. Existing classifications, such as Joshi et al. (2020) provide coarse, single-dimensional categories that are hard to reproduce from first principles and too imprecise for quantitative reasoning.

We propose **RAN (Resource Abundance Notation)**, designed first as a *communication* tool. When a paper reports evaluation on a `4/2/en-4` language (Cherokee) versus a `7/4/en-7/fr-6` language (Swahili), a reader instantly grasps the difference across multiple dimensions, without looking up corpus statistics or consulting a classification table. The notation is **compact** (fits in an abstract), **multi-dimensional** (separating speakers, monolingual data, and bilingual data), **reproducible** (derived from canonical, citable sources), and **interpretable** (each integer is an order of magnitude).

This matters especially for Indigenous languages, where the heterogeneity of "low-resource" is extreme: Quechua (`6/0/en-6/es-2`), Guarani (`6/1/en-6/es-2`), Cherokee (`4/2/en-4`), and Owens Valley Paiute (`1/3/en-3`) differ by orders of magnitude in speaker vitality and corpus availability yet are often bundled under a single label.

2 Notation

A RAN score is written as:

$$\text{S/M/}L_1\text{-}B_1\text{/}L_2\text{-}B_2\text{/. . .} \quad (1)$$

where $S = \lfloor \log_{10}(\#\text{ fluent speakers}) \rfloor$, $M = \lfloor \log_{10}(\#\text{ monolingual sentences}) \rfloor$, and $B_i = \lfloor \log_{10}(\#\text{ parallel sentences with language } L_i) \rfloor$, with B_i listed in descending order.

Each dimension corresponds to a different kind of NLP training or downstream use. S measures the *community*: language vitality, the realistic ceiling on future annotation or participatory data work, and the downstream population any tool will serve. It is the only dimension that cannot be grown by scraping, and it motivates the ethical and sovereignty considerations that corpus counts alone do not capture (we expand on this in §5). M measures the raw

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 168–172

material for self-supervised pretraining: the text a self-supervised language model (e.g., BERT, GPT, XLM-R) can consume, and therefore the ceiling on what any monolingual or multilingual encoder can learn about the language without alignment. B_i is the supervised counterpart: parallel sentences are the direct training data for MT, and they enable cross-lingual transfer from a high-resource partner. L_i records the partner identity, because a language paired with English behaves differently from one paired only with a regional neighbor. Listing the pairs in descending B_i order puts the strongest connection ($B_{\max}$) first. This is the quantity most relevant for cross-lingual transfer (§4), and the ordered list itself reads as a pivot map (§8).

We intentionally report M and the B_i as counts in their *source corpora* (OSCAR and OPUS respectively) rather than as a single combined figure. This keeps each component independently citable and reproducible from a canonical, dated snapshot. As a consequence, a configuration like $M = 1$, $B_{\text{en}} = 6$ (e.g. Guarani in our data) is internally consistent: it means OSCAR 23.01 contains $\sim$10 Guarani sentences while OPUS lists $\sim 10^6$ Guarani–English pairs. A reader who needs the *total* monolingual text available, ignoring source provenance and duplication, can read $\max(M, \max_i B_i)$ off the notation directly as a lower bound.

The $\lfloor \log_{10} \rfloor$ quantization is primarily a communication choice, but it also matches what is known about how NLP performance responds to data: cross-entropy loss scales as a power law in corpus size (Kaplan et al., 2020; Bansal et al., 2022), so meaningful differences between languages appear at the order-of-magnitude level rather than in raw counts. The gap between 100 and 1,000 sentences is qualitative, while the gap between 100,000 and 100,900 is invisible. Integer values capture this directly, fit in an abstract, and are comparable across languages at a glance. A useful side-effect is robustness to upstream noise: deduplication policy, sentence-splitter choice, and domain coverage can shift raw counts by tens of percent without changing the floored log. When a task calls for finer precision, a decimal form ($M = 3.4$ for $\sim$2.5K sentences) coexists with the integer notation.

3 Data

We scored 20 typologically diverse languages spanning the full RAN range, from English (`9/10/...`) to Owens Valley Paiute (`1/3/en-3`). **Speaker counts** are the maximum P1098 value across statements on each language's Wikidata entity (Q-ids in `languages.csv`), typically the L1+L2 total. **Monolingual corpus sizes** are estimated from OSCAR 23.01 deduplicated word counts (OSCAR Project, 2023), converted to sentences at 15 words/sentence (5 for logographic scripts). Hausa, absent from OSCAR 23.01, falls back to CC-100. **Bilingual data** comes from OPUS (Tiedemann, 2012), queried with `preprocessing=moses`. For each pair we record the size of the *largest single corpus* OPUS lists, not the sum: this avoids double-counting derived/aliased releases (e.g. NLLB derives from CCMatrix; HPLT/MultiHPLT report identical counts) and corresponds to the realistic ceiling for a single-corpus model. Queries use the ISO 639-1 macrocode (e.g. `ar`, `zh`); sub-variant tags (`arz`, `cmn`, `swh`, etc.) are not aggregated by OPUS and including them would not change any integer bin here. The released `languages.csv` records the OPUS corpus that produced the maximum for every pair, so each value is reproducible. For Owens Valley Paiute (`mnr`), effectively absent from OPUS, both M and B_{en} derive from the community-curated Kubishi Dictionary[1] (4,484 `mnr`–en sentence pairs).

We compare against published benchmark numbers (no models were trained for this work): **machine translation (MT)** via NLLB-200 on FLORES-200 xx→eng, chrF++ (NLLB Team et al., 2024) (17/20 languages); **named entity recognition (NER)** via XLM-R Large zero-shot on WikiANN F1 (Conneau et al., 2020; Hu et al., 2020) (10/20); and **part-of-speech tagging (POS)** via XLM-R Large zero-shot on Universal Dependencies accuracy (9/20). Table 1 gives the full inventory.

4 Correlation with Benchmark Performance

We fit linear regressions predicting benchmark scores from RAN components and compare across a small set of models (Table 2). The linear form is used as a coarse summary of how the integer-log components track benchmark performance. We do not claim that quality is literally linear in $\log(\text{data})$, only that log-scale components capture the order-of-magnitude effects discussed in §2. Our aim is twofold: (a) test whether the three components carry complementary information by comparing

[1] https://dictionary.kubishi.com/

Language	ISO	RAN	sum
English	eng	9/10/es-8/fr-8/de-8/zh-7	27
Spanish	spa	8/9/en-8/fr-8/pt-8	25
Chinese†	zho	9/9/en-7/ja-7	25
Hindi	hin	8/8/en-7/ur-6	23
Arabic†	ara	8/8/en-7/fr-7	23
Vietnamese	vie	7/9/en-7/zh-6	23
Turkish	tur	7/8/en-7/de-7	22
Korean	kor	7/8/en-7/ja-6	22
Sinhala	sin	7/7/en-7	21
Nepali†	nep	7/7/en-7/hi-6	21
Mongolian†	mon	6/7/en-7/zh-5	20
Hausa	hau	7/6/en-6/fr-6	19
Swahili†	swa	7/4/en-7/fr-6	18
Welsh	cym	5/6/en-7	18
Yoruba	yor	7/2/fr-6/en-6	15
Maltese	mlt	5/3/en-7/it-6	15
Guarani†	grn	6/1/en-6/es-2	13
Quechua†	que	6/0/en-6/es-2	12
Cherokee	chr	4/2/en-4	10
OVP‡	mnr	1/3/en-3	7

Table 1: RAN components for the 20 languages, sorted by RAN_{sum}. OVP = Owens Valley Paiute. English–English "bilingual" is reported as $B_{\text{en}} = 0$ by convention. For English we use $B_{\max}$ as its largest partner. † Macro-language ISO 639-3 code (zho, ara, swa, mon, que, nep, grn). The numbers shown are dominated by one variant in practice (e.g. Arabic by Modern Standard *arb*; Quechua by *quy* Ayacucho Southern, the variant used in NLLB-200's `quy_Latn` FLORES split). Sub-variants (e.g. *quz* Cuzco, *qub* Huallaga) would receive distinct RAN scores when scored individually. We recommend reporting the ISO 639-3 code alongside RAN whenever a macro-language label could be ambiguous. ‡ OVP is not in OSCAR or OPUS, so both M and B_{en} derive from the community-maintained Kubishi dictionary (`https://dictionary.kubishi.com/`).

the full model against single-predictor baselines, and (b) identify which single component is the strongest predictor for each task.

Model	MT	NER	POS
$S + M + B_{\max}$	**0.52**	**0.76**	**0.72**
$S + M + B_{\text{en}}$	0.52	0.53	0.47
$B_{\max}$ only	0.41	0.48	0.64
M only	0.26	0.22	0.35
B_{en} only	0.41	0.06	0.12
RAN_{sum}	0.20	0.19	0.37
S only	0.00	0.00	0.05

Table 2: Training R^2 for each predictor model across tasks. MT: NLLB-200 chrF++ (n=17); NER: XTREME WikiANN F1 (n=10); POS: XTREME UD accuracy (n=9). For MT, $B_{\text{en}} = B_{\max}$ for all 17 languages.

Three observations follow from the regression experiment (Figure 1 visualizes the aggregate trend). First, the full three-component model dominates every task: at $R^2 = 0.76$ for NER, no single component is within 0.28, so dropping any of S, M, or $B_{\max}$ leaves real variance on the table. The R^2 gap between the full model and any single component is itself the evidence that the components carry complementary information rather than restating each other. Second, $B_{\max}$ is the strongest single predictor for NER ($R^2 = 0.48$) and POS ($R^2 = 0.64$), while B_{en} alone explains almost nothing on those tasks ($R^2 = 0.06$ and 0.12), confirming that cross-lingual transfer benefits from connections to *any* high-resource partner, not just English. Third, S alone is uninformative of benchmark performance. We argue in Section 5 that this is a feature of the notation, not a defect. Because samples are small ($n \in [9, 17]$), we treat these R^2 values as descriptive rather than as evidence of out-of-sample predictive power.

Chinese sits well below its RAN-predicted score on NER and POS, large enough to ask whether the gap reflects RAN missing something Chinese-specific or a property of logographic, non-segmented scripts more broadly. The other non-Latin-script languages in our set are consistent with the latter reading: Korean (alphabetic syllabary, space-segmented) lands close to its RAN-predicted NER/POS scores, while Vietnamese (Latin, segmented, tonal) shows the same pattern. We therefore attribute Chinese's gap to script and tokenization in the zero-shot benchmark setting (XLM-R transfer from English WikiANN/UD) rather than to a corpus signal RAN fails to capture. Confirming this would benefit from adding a second logographic language (e.g. Japanese) and a non-segmented non-logographic language (e.g. Thai). We leave this to future work.

5 Why Keep Speaker Count?

S's non-predictiveness should not be read as an argument to drop it from the notation. Corpus size tells you what a model *can do today*, while S tells you why we should care and what resources could plausibly become available. It captures language vitality (Swahili `7/4/...` and Maltese `5/3/...` have near-identical corpus profiles but very different endangerment status), the realistic ceiling on future annotation and participatory data work, the scale of downstream impact, and the community capacity and sovereignty considerations that shape whether a dataset should exist at all. A notation

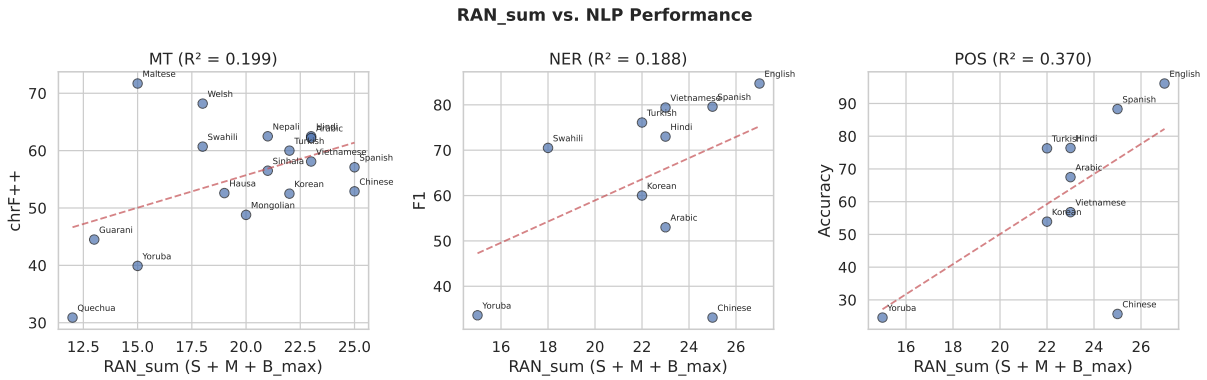

Figure 1: RAN_{sum} vs. benchmark performance across the three tasks, with linear fits.

that omitted S would be a poorer communication tool even if it remained an equally good predictor of benchmark scores.

6 A Living RAN Database

A notation is only as useful as the shared numbers it references. We deploy a community-maintained web database at `https://ran.kubishi.com` where anyone can submit an update (ISO 639-3 code, any subset of S, M, or L_i-B_i pairs with citations). Every submission is queued for human review of source reproducibility and correct denomination (deduplicated sentences, not bytes), and submissions and revisions retain stable IDs so a cited RAN score can always be traced back to the exact revision it was drawn from.

7 Related Work

Joshi et al. (2020) introduced a six-class taxonomy (0–5) for language resource levels, widely adopted but offering only a single dimension. Blasi et al. (2022) demonstrated systematic inequalities in NLP performance across languages, motivating more precise quantification of resource gaps. NLLB (NLLB Team et al., 2024) and XTREME (Hu et al., 2020) provide the benchmark data enabling our empirical validation.

8 Conclusion and Possible Extensions

RAN provides a compact, reproducible, multi-dimensional notation for communicating language resource profiles. Across 20 languages and three tasks, the components carry complementary information ($R^2 = 0.52$ for MT, 0.76 for NER, 0.72 for POS), and $B_{\max}$ outperforms B_{en} for cross-lingual transfer. We frame RAN as a *communication* tool: `6/0/en-6/es-2` (Quechua) and `1/3/en-3` (Owens Valley Paiute) name resource profiles a single "low-resource" label would flatten. We encourage adoption in abstracts, dataset and model cards, and shared-task descriptions (e.g., resource envelopes like "all targets satisfy $M \leq 4$"). Natural extensions of the core notation include rendering the L_i-B_i graph to surface pivot paths, a decimal form for finer precision, and optional dimensions (script status, lexicon size, typological distance, quality flags) appended per task.

This paper deliberately scopes RAN to written text and text-based benchmarks (MT, NER, POS), because the data sources it draws on (OSCAR, OPUS) and the benchmarks it validates against (NLLB-200, XTREME) are themselves text-only. But many of the languages RAN is most useful for (particularly Indigenous languages of the Americas like Owens Valley Paiute, Cherokee, and many Quechuan and Tupian variants) are predominantly or exclusively spoken, and a text-only profile understates their actual resource picture. A natural extension is an optional speech component $H = \lfloor \log_{10}(\text{hours of recorded speech}) \rfloor$, populated from sources such as Common Voice, FLEURS, and community archives, and paired with bilingual H_i values for aligned speech–text resources. This would let RAN describe ASR/TTS resource profiles alongside the text-based one. We leave the design of H and its empirical validation against speech benchmarks to future work, but flag it here as the most important near-term direction for languages of the Americas.

Limitations

Our sample is small ($n \in [9, 17]$ per task) and English-centric: NER and POS benchmarks rely on XLM-R zero-shot transfer from English, and B_{en} coincides with $B_{\max}$ for every MT language.

The floor-of-log formulation compresses very different corpus sizes (e.g. 1K vs. 9K sentences) into the same integer. RAN does not encode data *quality* (domain, noise, script-register match), which can shift effective performance by 10+ chrF++. Counting in *sentences* is also an approximation: corpora differ in average sentence length and complexity, and our words÷15 (or ÷5 for logographic scripts) heuristic only coarsely normalises this. A corpus of short, simple sentences and one of long, syntactically rich sentences with the same M are not equivalent training material. We accept this loss of precision in exchange for a unit that is meaningful across scripts (unlike raw token counts, which are tokenizer-dependent) and that fits in an abstract. The decimal form $M = 3.4$ and per-task quality flags can recover precision where needed. RAN also currently assumes catalogued parallel sentences are human-produced: it does not yet distinguish machine-translated or back-translated pairs, which can now be generated cheaply and would otherwise inflate B_i without a commensurate quality gain, nor does it account for silver/synthetic data more broadly. Speaker counts inherit Wikidata's heterogeneity (different sources, years, L1/L2 conventions). Finally, RAN reports only text resources: until the H extension sketched above is in place, languages with substantial speech corpora but little text will be understated. RAN should complement, not replace, qualitative community-facing context.

Ethical Considerations

Several languages in our dataset (Cherokee, Quechua, Guarani, and Owens Valley Paiute) are spoken by Indigenous communities, some critically endangered. RAN is descriptive and uses only aggregate, publicly cited statistics. It is not intended to rank languages by "worth" or to justify deprioritizing any language. We emphasize that low S (few speakers) is precisely where language-sovereignty considerations and community partnership matter most. Decisions about whether and how to build NLP tools for an Indigenous language must rest with the speaker community, not with corpus counts.

References

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, and 1 others. 2022. Data scaling laws in NMT: The effect of noise and architecture. In *Proceedings of the 39th International Conference on Machine Learning*.

Damian E. Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5486–5505. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, and 1 others. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OSCAR Project. 2023. OSCAR 23.01. `https://oscar-project.github.io/documentation/versions/oscar-2301/`. Open Super-large Crawled Aggregated coRpus, version 23.01.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association.

Bringing Mapudungun into the Modern MT Ecosystem: Morphology-Aware Tokenization for NLLB-200 Fine-Tuning

Isaac M. Thompson[1], **Brandon M. A. Rogers**[2], **Eric K. Ringger**[1]
[1]Department of Computer Science, Brigham Young University
[2]Department of Spanish & Portuguese, Brigham Young University
{it238, brandon.rogers, ringger}@byu.edu

Abstract

For Mapudungun arn→es translation, morphology-aware tokenization can substitute for a 5× increase in model parameters. We fine-tune three sizes of Meta's NLLB-200 on Mapudungun–Spanish translation across eight tokenization strategies, including our novel Morfessor-VC method, which constrains Morfessor morpheme segmentation to tokens already present in NLLB's pretrained vocabulary. Our 600M Morfessor-VC model is competitive with our own fine-tuned 3.3B Standard BPE model on arn→es (43.2 vs. 42.9 chrF++, $\Delta = +0.3$, $p = 0.039$, 95% CI [0.02, 0.60]) while using five times fewer parameters, and all fine-tuned conditions surpass frontier LLMs by over 27 chrF++. Mapudungun is an indigenous polysynthetic language spoken by 200,000+ Mapuche people in Chile and Argentina, absent from NLLB-200 and not supported by major commercial MT providers; prior work predates large-scale multilingual models and does not address the tokenization challenges posed by its agglutinative morphology. These results establish new state-of-the-art baselines for Mapudungun MT and provide a practical foundation for community language tools in pedagogy, social media, and language revitalization.

1 Introduction

Mapudungun (also *Mapuzugun*, *Mapudungu*, or *Mapuche*; historically *Araucanian*, the colonial designation reflected in ISO 639-3 code `arn`) is spoken by more than 200,000 Mapuche people across southern Chile and Argentina (Duan et al., 2020), making it one of the largest indigenous languages in South America by speaker count. Despite this, Mapudungun remains absent from nearly all commercial and open-source machine translation (MT) systems—a gap with direct consequences for community language vitality (Ahumada et al., 2022). The language's polysynthetic, agglutinative morphology—where a single verb form can encode tense, aspect, subject, object, evidentiality, and spatial deixis—poses well-documented challenges for standard subword tokenization (Rust et al., 2021; Mielke et al., 2021; Petrov et al., 2023). Byte-pair encoding (BPE) (Sennrich et al., 2016), the dominant tokenization strategy in multilingual models, is not designed to respect morpheme boundaries, and high fertility rates on agglutinative languages have been shown to correlate with degraded translation quality (Ahia et al., 2023; Banerjee and Bhattacharyya, 2018).

Prior work on Mapudungun NLP is sparse. Levin et al. (2002) collected early bilingual data and explored foundational language technologies for Mapudungun. Duan et al. (2020) introduced the AVENUE corpus and established sequence-to-sequence baselines. Pendas et al. (2023) and Lira et al. (2025) have since explored neural approaches, and Chandiá (2022) developed morphological analysis tools. None of this prior work has fine-tuned large-scale multilingual models on Mapudungun, nor systematically studied how tokenization choices interact with model capacity for this language.

We address both gaps. Our contributions are:

- The first fine-tuning study of Meta's NLLB-200 (NLLB Team et al., 2022) on Mapudungun–Spanish translation, across three model sizes (600M, 1.3B, 3.3B parameters) and both translation directions.
- A systematic ablation of eight tokenization strategies, ranging from standard NLLB BPE to Morfessor-based segmentation (Virpioja et al., 2013) and SentencePiece UnigramLM (Kudo, 2018).
- **Morfessor-VC**, a novel tokenization method that runs Morfessor segmentation and then constrains the resulting vocabulary to tokens

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 173–185

already present in NLLB's pretrained vocabulary, preserving embedding alignment while improving morpheme boundary coverage.

- A comprehensive evaluation showing that Morfessor-VC with the 600M model achieves 43.2 chrF++ on Mapudungun→Spanish, matching our standard BPE 3.3B baseline (42.9 chrF++), and outperforming frontier LLMs (Aya Expanse 8B: 15.9 chrF++) by over 27 points.

Code and evaluation scripts are available at https://github.com/byu-matrix-lab/mapudungun-nllb; fine-tuned model weights are released as a HuggingFace collection at https://huggingface.co/collections/byumatrixlab/mapudungun-nllb.

2 Related Work

Mapudungun NLP. Levin et al. (2002) collected early bilingual data and explored foundational language technologies for Mapudungun at CMU. Duan et al. (2020) introduced the AVENUE corpus and established sequence-to-sequence baselines, reporting plain chrF (character n-gram F-score, 0–1 scale; not directly comparable to chrF++) of 0.50 arn→es and 0.40 es→arn on 220k training pairs using a custom Transformer. Ahumada et al. (2022) developed educational NLP tools for Mapudungun. Pendas et al. (2023) applied active learning to Mapudungun MT; their BLEU scores are not directly comparable to standard held-out evaluation due to test-train overlap in the active learning simulation. Lira et al. (2025) explored transfer learning from high-resource language pairs (Spanish–English, Spanish–Finnish) and reported 30.30 chrF on arn→es on a separate 1,250-pair test set. Chandiá (2022) developed a finite-state morphological analyser for Mapudungun, providing linguistic groundwork relevant to our tokenization study.

Morphology-aware tokenization. The interaction between subword tokenization and morphologically rich languages is well studied. Rust et al. (2021) showed that multilingual models produce unequal tokenization quality across languages, with agglutinative languages suffering highest fertility. Ahia et al. (2023) and Petrov et al. (2023) further quantified how tokenization inequity translates to downstream performance gaps. Ataman and Federico (2018) and Banerjee and Bhattacharyya (2018) proposed combining morphological segmentation with neural MT, motivating our Morfessor-BPE condition. Mager et al. (2022) conducted a direct comparison of BPE and morphological segmentation for four polysynthetic Amerindian languages, finding that morphological segmentation can outperform BPE in low-resource settings—consistent with our findings for Mapudungun. Gowda and May (2020) established that vocabulary size has diminishing returns beyond a language-specific threshold, informing our Mono BPE and Optuna BPE conditions. Morfessor 2.0 (Virpioja et al., 2013), the unsupervised morpheme segmenter underlying three of our conditions, uses minimum description length to learn morpheme boundaries without linguistic supervision. Subword regularization via UnigramLM (Kudo, 2018) has shown robustness benefits in low-resource settings, motivating our inclusion of that condition.

Low-resource MT with multilingual models. NLLB-200 (NLLB Team et al., 2022) supports 200 languages but does not include Mapudungun. Downey et al. (2023) studied methods for adapting multilingual vocabularies to new languages, directly relevant to our tokenization approach. Fine-tuning NLLB on new languages has shown strong results in AmericasNLP shared tasks (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025). Gow-Smith and Sánchez Villegas (2023) and DeGenaro and Lupicki (2024) demonstrated competitive NLLB fine-tuning results across indigenous languages at the 2023 and 2024 shared tasks respectively; our work is the first to systematically study tokenization strategies—rather than data augmentation or model selection—for NLLB fine-tuning on a polysynthetic language. Mager et al. (2023) provides a comprehensive overview of MT for indigenous languages of the Americas.

3 Data

We use the AVENUE corpus (Duan et al., 2020), a Mapudungun–Spanish parallel resource derived from the Mapudungun Speech Corpus (Caniupil et al., 2019)—oral history interviews recorded by Mapuche community members and linguists, transcribed and aligned using ELAN (Wittenburg et al., 2006). The corpus was developed through an academic collaboration (CMU, Universidad de La Frontera, Chilean government partners) and uses

the Alfabeto Mapuche Unificado (AMU) orthography. The corpus covers domains such as traditional medicine, cultural practices, and personal narrative.

Extraction and cleaning. The AVENUE data is structured as parallel ELAN annotation blocks. We extract sentence-level pairs by treating each ELAN block as a translation unit. We then apply a cleaning pass that strips incomplete ELAN tags, CHAT transcription codes, overlap and uncertainty markers, and degenerate segments. Two standard defaults required language-specific overrides: the minimum word count was set to 1 (single-word Mapudungun clauses are grammatically complete in a polysynthetic language) and long-word filtering was disabled (Mapudungun compounds regularly exceed 30 characters). The resulting corpus is split into train / dev / test sets of 55,452 / 1,581 / 9,382 sentence pairs, respectively.

Code-switching. A notable property of the AVENUE corpus is widespread code-switching: approximately 24% of Mapudungun utterances contain embedded Spanish words or phrases, as marked by ELAN <SPA> tags in the original transcriptions. This reflects natural bilingual speech patterns in contemporary Mapuche communities rather than transcription artifacts. Our cleaning pipeline retains these mixed-language utterances; we analyze their effect on translation quality in Section 7.3.

Data statistics. Table 1 summarizes corpus statistics. The Mapudungun training side contains 656,969 whitespace-delimited tokens (avg. 11.8 tokens/sentence); the Spanish side contains 888,476 tokens (avg. 16.0 tokens/sentence). The higher Spanish token count reflects that Mapudungun's polysynthetic morphology encodes in a single word what Spanish expresses across multiple words.

Split	Lang	Sents	Tokens	Types	Avg len
Train	arn	55,452	656,969	105,375	11.8
	es	55,452	888,476	40,059	16.0
Dev	arn	1,581	23,519	6,863	14.9
	es	1,581	29,314	4,381	18.5
Test	arn	9,382	108,361	26,264	11.5
	es	9,382	144,870	13,624	15.4

Table 1: Corpus statistics after cleaning. Tokens and types are whitespace-delimited. Mapudungun has substantially more types per token than Spanish, reflecting its richer morphology.

4 Tokenization Methods

A core challenge in fine-tuning NLLB-200 on Mapudungun is the mismatch between NLLB's pretrained BPE vocabulary—optimized for 200 languages with Spanish heavily represented—and Mapudungun's polysynthetic morphology. Naive fine-tuning feeds the model unsegmented Mapudungun text, which NLLB's internal tokenizer fragments unpredictably. We compare eight tokenization strategies that differ in how they pre-segment Mapudungun before fine-tuning. All conditions use NLLB's standard tokenizer for the Spanish side; only the Mapudungun side is varied. Pre-segmented text uses the `@@` boundary marker convention (e.g., *mapu@ dun@ gun*) so that desegmentation is applied at inference time before scoring.

4.1 Standard BPE (NLLB)

The baseline condition feeds raw, unsegmented Mapudungun text directly to NLLB's built-in SentencePiece tokenizer without any pre-processing. This is the standard fine-tuning setup for NLLB and serves as our primary comparison point.

4.2 Joint BPE

Following Duan et al. (2020), we train a shared SentencePiece BPE model (Kudo and Richardson, 2018) on the concatenated Mapudungun and Spanish training data with a vocabulary size of 5,000. Joint training encourages shared subword representations across languages, which can aid cross-lingual transfer but may produce poor segmentations for the morphologically richer language.

4.3 Mono BPE

We train a Mapudungun-only SentencePiece BPE model, setting the vocabulary size via the 95%-character-coverage heuristic of Gowda and May (2020): we find the number of character types needed to cover 95% of character occurrences, then multiply by 10 to estimate the subword vocabulary size. This yields a vocabulary focused on Mapudungun morphology without interference from Spanish.

4.4 Optuna BPE

To address the objection that our results might be sensitive to vocabulary size selection, we use Optuna (Akiba et al., 2019) to tune the BPE vocabulary size over 15 trials, optimizing chrF++ on the development set using a 600M NLLB model

fine-tuned for 3 epochs per trial. The optimal vocabulary size found was 10,954. Notably, chrF++ was largely flat across a wide range of vocabulary sizes (1,431–14,517 all achieved 45.72 chrF++), suggesting BPE vocabulary size is not a critical hyperparameter for this language pair.

4.5 Morfessor

We apply Morfessor 2.0 (Virpioja et al., 2013), an unsupervised morpheme segmenter based on minimum description length, to the Mapudungun training corpus. Morfessor learns morpheme boundaries without linguistic supervision, producing segmentations that reflect statistical regularities in the character sequences. Boundaries are marked with `@@`. The Spanish side is left unsegmented.

4.6 Morfessor-VC (Our Method)

Standard Morfessor segmentation introduces boundaries that do not align with NLLB's pretrained vocabulary, causing *double tokenization*: NLLB's internal tokenizer re-splits the already-segmented morphemes, producing fragmented representations that were not seen during pretraining.

Morfessor-VC (Vocabulary-Constrained) is our proposed solution. Starting from Morfessor segmentation, we greedily merge adjacent morphemes whose concatenation corresponds to a single token in NLLB's SentencePiece vocabulary. The algorithm processes each word left-to-right:

1. Given morphemes $[m_0, m_1, \ldots, m_k]$ within a word, consider the candidate merge $m_i \oplus m_{i+1}$.
2. If $\_(m_i \oplus m_{i+1})$ or $(m_i \oplus m_{i+1})$ is a single piece in NLLB's vocabulary, merge and repeat from m_i.
3. Otherwise, retain the boundary and advance to m_{i+1}.

The result preserves morpheme boundaries only where the split is both linguistically motivated (by Morfessor) *and* corresponds to a genuine subword boundary in NLLB's vocabulary. Boundaries that span a single NLLB vocabulary item are removed. This reduces double-tokenization while retaining meaningful morphological signal. Unlike prior work applying standard Morfessor to polysynthetic languages (Mager et al., 2022), Morfessor-VC does not introduce segments that the pretrained model's tokenizer will re-split; the VC step is specific to fine-tuning scenarios where a fixed pretrained vocabulary must be respected.

Table 2 illustrates the difference on *kutrankautulay* ("does not become sick"; kutran = illness, kautu = to become, -lay = negation).

Method	Segmentation
Standard BPE	`_ku tran anka ut ulay`
Morfessor	`kutran@@ ka@@ u@@ tulay`
Morfessor-VC	`kutran@@ kau@@ tulay`

Table 2: Tokenization of *kutrankautulay* under three conditions. Standard BPE (top) shows NLLB's internal SentencePiece output on the raw word, ignoring morpheme structure entirely. Morfessor (middle) adds pre-segmentation boundaries, but splits *kau* into *ka* + *u*; when each pre-segment is fed to NLLB, *ka* and *u* receive word-initial _ sentinels—embeddings pretrained as independent Spanish function words, not as morpheme-internal pieces. Morfessor-VC (bottom) merges *ka* + *u* → *kau*, a single item in NLLB's vocabulary (id 22,381), eliminating this spurious split.

4.7 Morfessor-BPE

Following Banerjee and Bhattacharyya (2018), we apply BPE *within* Morfessor morphemes. First, Morfessor segments the corpus; then a BPE model (vocabulary size 4,000) is trained on the resulting morpheme tokens and applied within morpheme boundaries. This combines morphological segmentation with BPE's data-driven compression.

4.8 UnigramLM

We train a SentencePiece Unigram language model (Kudo, 2018) on the Mapudungun training data using the same 95%-character-coverage vocabulary size heuristic as Mono BPE. Unlike BPE, UnigramLM directly optimizes a probabilistic segmentation model and can produce multiple segmentations for the same word, which has been shown to improve robustness in low-resource settings (Kudo, 2018).

4.9 Fertility Comparison

Figure 1 shows the fertility (pre-segmented tokens per whitespace-delimited word) for each condition on the Mapudungun training side. Morfessor and Morfessor-VC have the lowest fertility (1.13), close to one-to-one morpheme–word mapping. BPE-based methods are more aggressive (1.55–2.04), and UnigramLM (1.98) is comparable to Mono BPE. The near-identical fertility of Morfessor and Morfessor-VC confirms that the VC merging step removes boundaries without adding

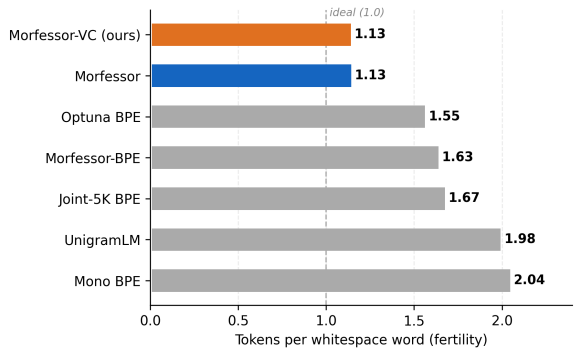

Figure 1: Pre-segmentation fertility (tokens per whitespace word) for each tokenization condition on the Mapudungun training set, sorted in ascending order. Morfessor and Morfessor-VC achieve the lowest fertility (1.13); BPE-based methods and UnigramLM are more aggressive (1.55–2.04).

new ones. Standard BPE is excluded from this comparison because its segmentation is applied internally by NLLB's SentencePiece tokenizer and produces incomparable fertility values.

5 Experimental Setup

Models. We fine-tune three sizes of NLLB-200-distilled (NLLB Team et al., 2022): 600M, 1.3B, and 3.3B parameters. All models are initialized from Meta's publicly released checkpoints. We fine-tune each model independently for each tokenization condition and both translation directions (arn→es and es→arn), for a total of 48 fine-tuning runs.

Training. All models are fine-tuned for up to 10 epochs with early stopping (patience 3, dev chrF++), AdamW ($lr = 3{\times}10^{-5}$, 1K warmup steps, weight decay 0.01), effective batch size 128, max length 128, beam size 4, on a single NVIDIA A100.

Evaluation. Our primary metric is chrF++ (Popović, 2017), a character n-gram F-score that has shown correlations with human judgments across language pairs. Our pilot human evaluation in Section 6.1 found no consistent preference between systems despite a ∼3-point chrF++ gap at 3.3B, though the study's single-annotator design limits the strength of that conclusion. We report BLEU (Papineni et al., 2002) as a secondary metric. Both metrics are computed with sacreBLEU (Post, 2018) against the held-out test set (9,382 sentence pairs). We also report COMET (wmt22-comet-da) in Appendix B for completeness; because Mapudungun is absent from XLM-R's training data, COMET scores on the Mapudungun side are unreliable and we do not use them to draw conclusions. For pre-segmented conditions, output is desegmented (removing `@@` markers) before scoring. We assess statistical significance for headline comparisons using paired bootstrap resampling (10,000 iterations) over per-sentence chrF++ scores (Koehn, 2004).

Human evaluation. We conduct a pilot preference ranking of Standard BPE (3.3B) vs. Morfessor-VC (3.3B) outputs on 50 stratified sentences per direction; see Section 6.1.

6 Results

Table 3 reports chrF++ for all 8 tokenization conditions across 3 model sizes and both translation directions, alongside zero-shot NLLB baselines and prompted LLM comparisons.

Fine-tuning vs. baselines. Zero-shot NLLB and prompted LLM scores serve as pre-fine-tuning references: both reflect performance without any Mapudungun-specific training, not directly comparable competitors to our fine-tuned systems. Fine-tuning yields dramatic gains over both in both directions (Figure 2). Even the smallest fine-tuned model under Standard BPE (34.9 chrF++) substantially outperforms Aya Expanse 8B (15.9 chrF++) on arn→es, and all morphology-aware conditions at 600M exceed 27 points above the best LLM baseline.

Tokenization effects. Figure 3 shows chrF++ by condition and model size. Standard BPE lags at 600M (34.9 arn→es) but catches up at 3.3B (42.9), while all morphology-aware methods cluster between 41.7–43.2 at 600M. Morfessor-VC leads arn→es at both 600M and 3.3B. For es→arn, tokenization effects are substantially smaller, with Standard BPE competitive at 3.3B. This asymmetry is important: es→arn produces Mapudungun output, the direction most relevant for community applications, and morphology-aware tokenization does not deliver a clear advantage there. This contrasts with Mager et al. (2022), who found the largest morphological segmentation gains on the polysynthetic-output direction; we attribute the difference to NLLB's fixed decoder-side tokenizer, which limits the benefit of pre-segmented targets. Claims about morphology-aware tokenization as

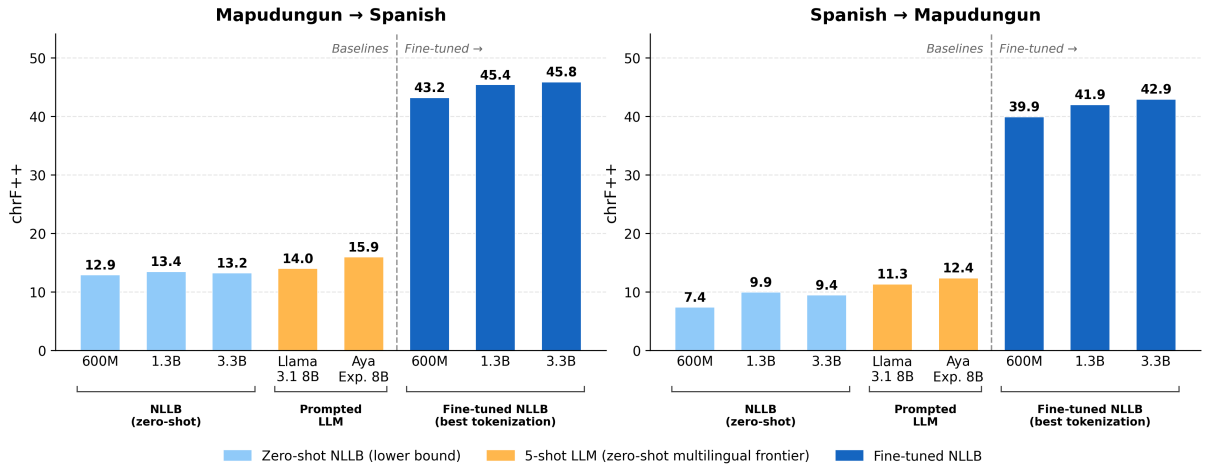

Figure 2: chrF++ by system and translation direction. Fine-tuned NLLB bars show the best-performing tokenization condition per direction (Morfessor-VC for arn→es, Standard BPE for es→arn). Dashed line separates baselines from fine-tuned systems. Zero-shot NLLB and prompted LLM bars establish lower bounds and the frontier of zero-shot multilingual capability before any Mapudungun-specific training; they are not fine-tuned competitors. Duan et al. (2020) and Pendas et al. (2023) are omitted: Duan reported plain chrF (0–1 scale), not directly comparable to chrF++; Pendas reported BLEU under an active learning simulation protocol. Lira et al. (2025) are omitted because their test set differs from ours; see Section 2.

Condition	arn→es				es→arn			
	600M	1.3B	3.3B	8B†	600M	1.3B	3.3B	8B†
Baselines (not fine-tuned)								
Zero-shot NLLB	12.86	13.43	13.19	—	7.37	9.91	9.42	—
Llama 3.1 8B	—	—	—	13.96	—	—	—	11.32
Aya Expanse 8B	—	—	—	15.92	—	—	—	12.36
Fine-tuned NLLB-200								
Standard BPE	34.90	40.64	42.85	—	39.87	41.94	**42.89**	—
Joint-5K BPE	42.36	44.68	45.25	—	38.17	41.06	42.30	—
Mono BPE	41.72	43.99	44.85	—	38.04	40.62	42.02	—
Optuna BPE	42.41	44.67	45.21	—	38.79	41.34	42.33	—
Morfessor	43.09	**45.40**	45.51	—	**39.90**	42.00	42.76	—
Morfessor-VC	**43.16**	45.37	**45.84**	—	39.89	**42.04**	42.77	—
Morfessor-BPE	42.80	44.97	45.56	—	38.74	41.25	42.40	—
UnigramLM	42.18	44.64	45.07	—	38.20	41.05	42.30	—

Table 3: chrF++ on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. †LLM baselines evaluated with 5-shot prompting.

an efficiency lever should therefore be understood as specific to the arn→es direction.

Scaling efficiency. Morfessor-VC at 600M (43.2 chrF++) matches Standard BPE at 3.3B (42.9 chrF++) on arn→es ($\Delta = +0.31$, $p = 0.039$, 95% CI $[0.02, 0.60]$; significant but with a narrow margin), using five times fewer parameters. Morfessor-VC 3.3B also significantly outperforms Standard BPE 3.3B ($\Delta = +2.99$, $p < 0.001$) and vanilla Morfessor 3.3B ($\Delta = +0.34$, $p < 0.001$, 95% CI $[0.14, 0.53]$).

6.1 Human Evaluation

We conducted a pilot preference ranking evaluation comparing Standard BPE (3.3B) and Morfessor-VC (3.3B) outputs across 50 sentences per direction, stratified by per-sentence chrF++ quartile. A single evaluator—a second-language learner of both Spanish and Mapudungun (native English speaker), not a native speaker of either language—indicated which system output was preferred for each sentence, or whether the two were of equal quality. No inter-annotator agreement was measured, as only one evaluator was available for this pilot study.

Results are shown in Table 4. Preferences are evenly split in both directions, with no clear advantage for either system. This is itself a finding: a 3-point chrF++ advantage for Morfessor-VC on arn→es does not translate to a perceptible quality

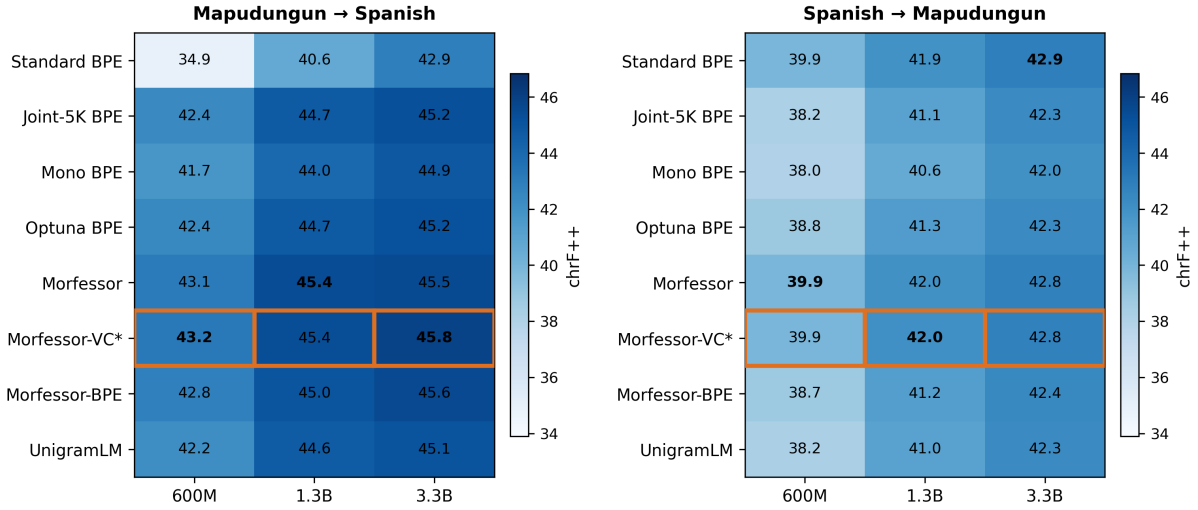

Figure 3: chrF++ by tokenization condition and model size for both translation directions. Darker cells indicate higher chrF++. Morfessor-VC achieves the highest arn→es scores at 600M and 3.3B; for es→arn, differences between conditions are smaller and Standard BPE is competitive at 3.3B.

Direction	Std BPE	Morfessor-VC	Tied
arn→es	19	19	12
es→arn	18	17	15

Table 4: Human preference rankings (out of 50 sentences per direction). Tied = annotator found the two outputs equally acceptable.

difference at the sentence level. The high tie rate (24% arn→es, 30% es→arn) further reflects that the two systems frequently produce equivalently adequate outputs. In this pilot study, chrF++ differences between systems did not translate to consistent human preference at the 3.3B scale, suggesting a possible ceiling effect or reference translation noise in the corpus; a larger study with multiple native-speaker annotators would be needed to draw firm conclusions about perceptual thresholds.

7 Analysis

7.1 Why Morfessor-VC Works

Figure 1 shows that Morfessor-VC has identical fertility to vanilla Morfessor (1.13), confirming that the VC step only removes boundaries—it never adds new ones.

The key mechanism is *double tokenization avoidance* (Table 2). When Morfessor splits two character sequences that together form a single NLLB vocabulary item, that boundary is spurious: NLLB's tokenizer will never produce it, so the corresponding embedding pair was never pretrained. By merging such adjacent morphemes, Morfessor-VC aligns pre-segmented tokens with embeddings NLLB actually learned.

Of the 87,382 boundaries introduced by Morfessor, the VC step removes 2,215 (2.5%)—confirming it is conservative. The VC advantage over vanilla Morfessor is 0.07 chrF++ at 600M ($p = 0.23$, n.s.) and 0.34 at 3.3B ($p < 0.001$). The primary driver at 600M is Morfessor-style morpheme segmentation; the VC refinement pays off more at larger scale where the model can exploit the improved embedding alignment. The broader finding—that any morphology-aware segmentation dramatically outperforms Standard BPE at 600M—is more robust than the VC-specific advantage.

Together with the Optuna BPE flatness across vocabulary sizes (1,431–14,517) and weak tokenization sensitivity in es→arn, these findings suggest that tokenization may not be the binding constraint at the current data scale—corpus size and domain narrowness may constitute harder ceilings.

7.2 Linguistic Analysis

A Mapudungun linguist annotated 20 Morfessor-VC outputs (10 per direction) for adequacy and naturalness in this pilot study. No systematic morphological assembly errors were observed in either direction, suggesting (on this small sample) that Morfessor-VC segmentation does not introduce boundary artifacts into the output.

arn→es. Outputs were generally adequate; in one case the model appeared to outperform the reference, retaining an adverb the transcriber omitted. A preliminary observation was that the model

tended toward narrow lexical interpretations of the polysemous noun *dungu* (language / thing / matter / way-of-being). Aspect errors (habitual vs. perfective) occurred occasionally.

es→arn. Several outputs were judged equivalent to or more concise than the reference, and the model sometimes preferred native vocabulary (e.g., *üytulafiñ*, "I will not name it") over calqued reference forms. One failure was severe: the model reproduced the Spanish source verbatim in Mapudungun orthography, a source-copy error likely triggered by high lexical overlap with mixed-language training sentences (Section 7.3). Repeated misuse of *chumuechi/chumngechi* ("how/in what way") suggests an overfit collocation pattern for this adverb.

7.3 Code-switching and es→arn Quality

Approximately 24% of Mapudungun utterances contain embedded Spanish words (ELAN <SPA> tags). This asymmetry explains why Standard BPE is relatively competitive for es→arn (Spanish input is well-covered by NLLB pretraining) and why the source-copy failure in Section 7.2 occurs only in that direction.

We used a word-level language identification model to correlate Spanish token proportion with per-sentence chrF++ from Morfessor-VC 3.3B on arn→es. The correlation is negligible (Pearson $r = +0.039$, $r^2 \approx 0.002$; Spearman $r = +0.051$); the $p < 0.001$ result reflects the large test set ($n = 9{,}382$) rather than practical effect size. Stratifying by code-switching level: no Spanish words averages 45.75 chrF++, light CS (1–20%) averages 45.11, heavy CS (>20%) averages 46.77. The uptick at heavy CS likely reflects Spanish loanwords and proper nouns that pass through unchanged, inflating character overlap with the reference. These results are post hoc and correlational; the model does not appear to substantially degrade on mixed-language arn→es input. A controlled future experiment—training conditions that tag or strip code-switched tokens—would directly measure the effect on es→arn quality, particularly the source-copy failure mode described in Section 7.2.

8 Conclusion

We presented the first systematic fine-tuning study of NLLB-200 on Mapudungun–Spanish translation, comparing eight tokenization strategies across three model sizes and both translation directions. Our proposed Morfessor-VC method—which constrains Morfessor segmentation to tokens present in NLLB's pretrained vocabulary—achieves the highest arn→es chrF++ at both 600M (43.2) and 3.3B (45.8), while matching Standard BPE at 3.3B using only a 600M model; the VC refinement over vanilla Morfessor is itself small ($\Delta \leq 0.34$), and the primary driver is morpheme-boundary presegmentation broadly. This 5× parameter efficiency gain—specific to the arn→es direction and significant but with a narrow margin ($p = 0.039$, 95% CI [0.02, 0.60])—suggests that morphology-aware tokenization can be a practical lever for compute-constrained deployment when translating from polysynthetic languages. For es→arn, tokenization choice has little impact and improving that direction—which produces Mapudungun output required for community use—is the direct priority for future work.

All fine-tuned conditions substantially outperform zero-shot NLLB and frontier LLMs, surpassing Aya Expanse 8B by over 27 chrF++ points and establishing new state-of-the-art baselines for Mapudungun MT. Linguistic analysis confirmed that Morfessor-VC does not introduce morpheme boundary artifacts, with the principal remaining error types being lexical (polysemy, collocation) rather than structural.

We release all models, tokenization code, and evaluation scripts to support future work on Mapudungun NLP and low-resource MT more broadly.

Several directions remain open. Improving es→arn—the direction that produces Mapudungun output, required for community use cases—is a direct priority; concrete next steps include back-translation augmentation, prefix-control for code-switching, and investigating whether corpus size or domain narrowness constitute harder ceilings than tokenization. Morfessor-VC is language-agnostic and applies to other polysynthetic languages in NLLB's coverage gap. Mapudungun is now included in the BOUQuET benchmark (Alastruey et al., 2026); we hope this work provides a technical foundation for future efforts pursued in coordination with Mapuche community priorities (Ahumada et al., 2022).

Limitations

Domain. All data comes from the AVENUE corpus, which consists of oral history interviews.

Translation quality on other domains (news, legal, educational texts) is unknown and may differ substantially.

Code-switching. The 24% code-switching rate in the AVENUE corpus is a property of this specific community's speech style. Our models may be poorly calibrated for monolingual Mapudungun text, and we have not evaluated on such data.

Reference quality. AVENUE translations were produced by human transcribers and may contain errors, as noted by our linguistic annotator in several cases. Automatic metrics evaluated against imperfect references will underestimate true translation quality.

Dialect variation. Mapudungun has significant dialectal variation across Chile and Argentina. The AVENUE corpus reflects a subset of Mapuche communities; performance on other dialects has not been assessed.

Evaluation metric. chrF++ is our primary metric, but it is a surface-level measure that does not capture morphological correctness or semantic adequacy directly.

Human and linguistic evaluation. Both our human preference evaluation (Section 6.1) and our linguistic analysis (Section 7.2) are pilot studies. The human evaluation was conducted by a single second-language learner of Spanish and Mapudungun (not a native speaker of either language); no inter-annotator agreement was measured. The linguistic annotation covers only 20 sentences. The divergence between chrF++ gains and human preferences may reflect a chrF++ ceiling effect at this performance level, reference translation noise in the AVENUE corpus, or both (see also the reference quality limitation above). A cleaner reference set would be needed to reliably discriminate between systems at the 3.3B scale. A stronger design would compare our best system against prior-work output, use multiple annotators including native Mapudungun speakers, and measure inter-annotator agreement. Preliminary observations in Section 7.2 (e.g., *dungu* polysemy errors, *chumuechi* overfitting) warrant confirmation on a larger annotated sample.

Missing ablations. We do not evaluate vocabulary-adaptation baselines such as WECHSEL (Minixhofer et al., 2022), which initializes new subword embeddings for cross-lingual transfer; this represents a complementary approach to Morfessor-VC and is a natural direction for future work. We also do not compare to parameter-efficient fine-tuning methods (e.g., LoRA) or to backtranslation-augmented training, both of which have shown benefits in indigenous-language MT (Ebrahimi et al., 2023).

Ethics Statement

This work develops machine translation technology for Mapudungun, an endangered indigenous language. We are attentive to the risk that MT systems can homogenize dialectal variation or be used to extract or commodify indigenous linguistic knowledge without community consent. The AVENUE corpus was collected through an academic collaboration involving Carnegie Mellon University, the Universidad de La Frontera (Temuco, Chile), and Chilean government partners. It is not missionary or religious data; it was collected with informed consent from Mapuche speakers and made publicly available by its creators (Duan et al., 2020).

Orthography. Mapudungun has multiple competing orthographic systems, including the Alfabeto Mapuche Unificado (AMU), Ragileo, and Azümchefe, each associated with different community and political positions. The AVENUE corpus uses the AMU orthography. Our models implicitly privilege AMU; community deployment should attend to which orthographic communities the models will serve and involve speakers of those communities in evaluation and adaptation.

Community engagement. We follow the CARE Principles for Indigenous Data Governance (Carroll et al., 2020) in spirit: data comes from a community-controlled corpus, we do not introduce new surveillance or extraction, and we release all models and code to support community-driven future work. We acknowledge that we did not establish direct partnerships with Mapuche governance bodies for this study, and encourage future work to do so. The ethical norms articulated by Mager et al. (2023) for indigenous-language MT research informed our approach.

Human evaluation. Human evaluation and linguistic annotation was conducted by a specialist collaborator who is a Mapudungun linguist.

We do not release any new primary data in this work.

Acknowledgments

We are grateful to the Mapuche speakers who contributed to the original corpus recordings, including Luis Caniupil, Flor Caniupil, Héctor Painequeo, Rosendo Huisca, and Hugo Carrasco. Compute resources were provided by the BYU Office of Research Computing. This work was supported by the BYU Computer Science Graduate Fellowship and generous donations to the BYU MAchine Translation Research and Interlingual eXperimentation (MATRIX) Lab.

References

Orevaoghene Ahia, Luke Bandarkar, Ife Henshaw, Sabrina Schneider, Sachin Kumar Singh, Antonios Anastasopoulos, David Mortensen, Graham Neubig, and Yulia Tsvetkov. 2023. All languages are NOT created (tokenized) equal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14614–14627, Singapore. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for Mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196, Seattle, Washington. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM.

Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Éric Saint-James, Ioannis Tsiamas, and 1 others. 2026. OmniLingual MT: Machine translation for 1,600 languages. *arXiv preprint arXiv:2603.16309*.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Sreejit Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the 2nd Workshop on Subword/Character LEvels Models*, pages 55–60, New Orleans, Louisiana. Association for Computational Linguistics.

Luis Caniupil, Flor Caniupil, Héctor Painequeo, Rosendo Huisca, Hugo Carrasco, Rodolfo M. Vega, Lori Levin, and Jaime Carbonell. 2019. Mapudungun speech corpus.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Andrés Chandiá. 2022. A Mapudüngun FST morphological analyser and its web interface. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6540–6547, Marseille, France. European Language Resources Association.

Ona de Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael DeGenaro and Luke Lupicki. 2024. Low-resource machine translation for indigenous languages of the Americas. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–7, Mexico City, Mexico. Association for Computational Linguistics.

C.M. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore. Association for Computational Linguistics.

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo M. Vega, Antonios Anastasopoulos, Lori Levin, and Alan W. Black. 2020. A resource for computational experiments on Mapudungun. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2872–2877, Marseille, France. European Language Resources Association.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages.

In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Peutêtre, Katharina Kann, and Alexis Palmer. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–17, Toronto, Canada. Association for Computational Linguistics.

Edward Gow-Smith and Danaé Sánchez Villegas. 2023. Low-resource machine translation for low-resource languages: Corpus augmentation and pretrained model leverage. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 74–86, Toronto, Canada. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lori Levin, Rodolfo Vega, Jaime Carbonell, Ralf Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2002. Data collection and language technologies for Mapudungun. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Gran Canaria, Spain. European Language Resources Association.

Hernan Lira, Luis Martí, and Nayat Sanchez-Pi. 2025. Spanish–Mapudungun translation using transfer learning for low-resource languages. In *2025 15th IEEE International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4007, Seattle, United States. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

María Begoña Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of Spanish to Mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11, Toronto, Canada. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Aalto University.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559, Genoa, Italy. European Language Resources Association.

A Full BLEU Results

Table 5 reports BLEU (sacreBLEU, `13a` tokenizer) for all fine-tuned conditions across three model sizes and both translation directions.

B Full COMET Results

Table 6 reports COMET scores (wmt22-comet-da) for all fine-tuned conditions across three model sizes and both translation directions. *Note: the wmt22-comet-da model has not been validated on Mapudungun, which is absent from its training languages (based on XLM-R pretraining data). COMET scores for the Mapudungun side should be interpreted cautiously; we report them for completeness alongside the primary chrF++ results in the main text.*

Condition	arn→es			es→arn		
	600M	**1.3B**	**3.3B**	**600M**	**1.3B**	**3.3B**
Baselines (not fine-tuned)						
Zero-shot NLLB	1.68	1.94	2.16	0.32	0.64	0.68
Llama 3.1 8B		1.55			0.20	
Aya Expanse 8B		2.13			0.39	
Fine-tuned NLLB-200						
Standard BPE	10.24	15.83	16.78	8.53	**9.93**	**10.57**
Joint-5K BPE	14.20	15.91	17.20	7.92	9.02	9.73
Mono BPE	11.93	12.97	15.45	7.24	8.30	9.16
Optuna BPE	14.50	16.33	17.97	8.12	9.26	9.93
Morfessor	16.15	18.21	18.72	**8.61**	9.83	10.32
Morfessor-VC	**16.45**	**18.35**	**18.94**	8.50	9.84	10.32
Morfessor-BPE	14.74	15.98	17.79	8.00	9.12	9.93
UnigramLM	12.61	14.11	15.77	7.40	8.69	9.33

Table 5: BLEU on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. LLM baselines were 8B models evaluated with 5-shot prompting and are shown in a single merged column for clarity.

Condition	arn→es			es→arn		
	600M	**1.3B**	**3.3B**	**600M**	**1.3B**	**3.3B**
Baselines (not fine-tuned)						
Zero-shot NLLB	0.390	0.392	0.385	0.408	0.489	0.440
Llama 3.1 8B		0.483			0.484	
Aya Expanse 8B		0.509			0.497	
Fine-tuned NLLB-200						
Standard BPE	0.592	0.631	0.648	0.666	0.677	**0.681**
Joint-5K BPE	0.636	0.654	0.659	0.658	0.672	0.679
Mono BPE	0.628	0.644	0.653	0.658	0.670	0.677
Optuna BPE	0.638	0.654	0.662	0.660	0.672	0.679
Morfessor	0.644	0.660	0.664	0.666	0.677	0.680
Morfessor-VC	**0.644**	**0.661**	**0.665**	**0.667**	**0.677**	0.680
Morfessor-BPE	0.640	0.655	0.662	0.660	0.673	0.679
UnigramLM	0.633	0.650	0.657	0.657	0.672	0.677

Table 6: COMET (wmt22-comet-da) on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. LLM baselines were 8B models evaluated with 5-shot prompting and are shown in a single merged column for clarity.

QomL'aqtaqa: A Qom–Spanish Parallel Corpus for Natural Language Processing with Machine Translation Evaluation

Viviana Cotik[1,2], **Aleksei Korablev**[3], **Paola Cúneo**[4,5] , **Pablo Laciana**[2]

[1] Universidad de Buenos Aires. FCEN. Departamento de Computación. Argentina.
[2] CONICET-UBA. Instituto de Ciencias de la Computación (ICC). Buenos Aires, Argentina.
[3] Minerva University
[4] CONICET, Argentina
[5]Universidad de Buenos Aires. FFyL. Instituto de Lingüística, Argentina

vcotik@dc.uba.ar

Abstract

Qom, a language of the Guaycuruan family, is a low-resource language for NLP and speech processing. We present the first parallel Qom–Spanish corpus in a computationally usable format, comprising 33,392 parallel segments, totaling 1,469,905 Qom tokens and 891,344 Spanish tokens. A subset of 2,943 segments excludes Bible-derived content. It includes alignments at different levels: sentences, sentence fragments, and paragraphs, and is compiled from multiple sources, both previously available and newly collected. We also present bidirectional neural machine translation baselines based on NLLB-200, achieving competitive performance in both translation directions on the full dataset, and lower performance on the non-Bible subset. An ablation study shows that training exclusively on biblical data reduces performance on non-biblical text, highlighting the importance of domain diversity in low-resource machine translation.

1 Introduction

Qom (also known as Toba) is an under-resourced Indigenous language of the Guaycuruan family spoken by one of the Indigenous peoples of Argentina, who number over 80,000 and mainly inhabit the Gran Chaco region, although migration to urban centers has diversified their sociolinguistic situation. While Qom remains vital in rural areas, language shift toward Spanish is increasing in urban contexts, particularly in the Buenos Aires metropolitan area.

In this context, the development of parallel corpora for machine translation is relevant not only for advancing low-resource NLP research but also for supporting language revitalization efforts, as computational resources can contribute to documentation, visibility and intergenerational transmission of ancestral languages.

However, to the best of our knowledge, parallel Qom–Spanish corpora suitable for machine translation are not yet available.

As a traditionally oral language with relatively recent written forms, Qom lacks a universally adopted orthographic standard, resulting in variation across communities that introduces additional challenges for computational processing. Furthermore, its morphological complexity and syntactic divergence from Spanish make it a particularly challenging and valuable case study for machine translation research.

Our key contributions are the following:

- the creation of Qom l'aqtaqa,[1] a parallel Qom–Spanish corpus by converting existing resources, originally available only as PDFs, into a machine-readable format and aligning additional Qom and Spanish texts that were previously unpaired,
- bidirectional machine translation baselines for Qom–Spanish and Spanish–Qom.

The corpus is accompanied by a comprehensive Data Statement that provides detailed documentation of its composition, creation process, and limitations (Bender and Friedman, 2018; Gebru et al., 2021). The resulting dataset is ready for computational use. Its public release is planned for a future version, which will incorporate additional data collected through ongoing recording, transcription, and translation efforts. The code will be available on GitHub upon publication.

The rest of the paper is organized as follows: Section 2 provides linguistic background on Qom. Section 3 reviews prior work on data and low-resource machine translation for Latin American Indigenous

[1] *Qom l'aqtaqa* (lit. "the people's words") is often translated as "the Qom language" but its semantic scope extends beyond a strictly linguistic or grammatical notion. More broadly, speakers understand *qom l'aqtaqa* as encompassing Qom ways of speaking, including not only referential meaning but also the intentional and agentive force attributed to words, and to play a central role in valued practices such as oratory and healing (Messineo, 2014).

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 186–202

languages, with a focus on Qom. Section 4 presents the construction and characteristics of the Qom–Spanish parallel corpus. Section 5 describes the machine translation baseline and reports experimental results. Section 6 concludes the paper and outlines future work. Additional details are provided in Appendices A.1 (Qom phonological system) , A.2 (Bible corpus coverage), A.3 (source-specific processing decisions), A.4 (alignment quality), A.5 (lexical analysis), A.6 (translation examples), and A.7 (Data Statement).

2 The Qom Language

The Qom language[2] presents several typological features that are relevant for machine translation. It is a morphologically rich language with polysynthetic and agglutinative tendencies, such that single words may encode information corresponding to an entire clause in English or Spanish. Morphemes are relatively segmentable and tend to be formally invariant, as illustrated in example (1):[3]

Example (1)

N'axaỹaxangui
n-ʔaGay-aGan-gi
3M-listen-CAUS-DIR

añi	*lquiyaqte*	*so*	*ỹaxaiquiolec*
a-ɲi	l-kizaqte	so	yaGaiki-ole-k
F-D:trid	3POSS-heart	D:DIST	old.man-DIM-MASC

'(He/she) was listening to the little old man's heartbeat.'

Both nominal and verbal morphology are highly complex: nouns inflect for number and gender and combine with deictic classifiers encoding positional and spatial distinctions, while possessed nouns distinguish between alienable and inalienable possession through different morphological marking strategies; for instance, in (1), the prefix *l-* marks a third-person possessor in *lquiyaqte* 'heart', yielding a possessive construction equivalent to that 'little old man's heartbeat'. In the same example, *ñi* (tridimensional) occurs with 'heart', while *so* (distal) modifies 'little old man' in a past narrative context. The verbal morphology is particularly elaborate, including three distinct sets of person prefixes conditioned by semantic roles and participant affectedness (active, middle, and inactive paradigms). For instance, in (1) the prefix *n-* indexes a middle person paradigm, marking an affected agent (i.e., the participant who listens). Verbs also encode aspectual distinctions and may take suffixes expressing direction, position, reflexivity, and reciprocity, and causation, as seen in (1) (*-aGan* 'CAUS', *-gi* 'DIR'). In addition, Qom lacks copular verbs. Basic constituent order differs from Spanish and English: transitive clauses generally follow Subject–Verb–Object order, whereas intransitive clauses tend toward Verb–Subject, and certain pronominal objects precede the verb.

These typological characteristics pose well-known challenges for machine translation. Differences in constituent order require substantial structural reordering during translation, while rich morphology necessitates modeling below the word level, typically through subword segmentation approaches (Jurafsky and Martin, 2025). Furthermore, the high degree of morphological synthesis complicates alignment between Qom lexical units and those of less synthetic target languages.

As a traditionally oral language, Qom has only relatively recently developed written forms, and no single orthographic standard has been universally adopted across communities. As a result, variation exists not only across dialects but also across domains of literacy use, including schools, churches, health services, and community organizations, each of which may promote distinct writing conventions. Consequently, multiple orthographic variants coexist in contemporary written Qom, reflecting different community-level agreements regarding the representation of particular phonemes. This variation affects several phonological segments. For instance, the palatal glide may be represented by multiple equivalent graphemes, including ỹ, ŷ, ý, ȳ, and ÿ. Other frequent alternations include d/r (e.g., *doqshe/roqshe* 'non-Indigenous person'), h/j (e.g., *hec/jec* 's/he goes'), and e/i (e.g., *nache/nachi* 'then'). Special mention should be made of the apostrophe ('), which represents the glottal stop, a phoneme absent in Spanish and inconsistently represented in writing (e.g., *do'onataxan/d'onataxan*).

A complete overview of the Qom phonological system and its orthographic variants is provided in Table 5 in Appendix A.1.

[2]ISO-639-3: tob. Glottocode: toba1269 https://glottolog.org/resource/languoid/id/toba1269)

[3] The first line gives the Qom text, the second its phonological representation, the third the morphological glosses, and the final line the English translation. Abbreviations in the glosses conform to the Leipzig Glossing Rules: 3 (third person), CAUS (causative), D (deictic classifier), DIM (diminutive), DIR (directional), DIST (distal), F (feminine), M (middle person marker), MASC (masculine), POSS (possessive), TRID (tridimensional).

3 Related Work

While efforts to develop computational resources for Latin American Indigenous languages have increased in recent years (Huamán-Águila et al., 2024; Ortiz Coronel et al., 2024; Tonja et al., 2024), these languages are still widely considered underrepresented. These efforts include initiatives such as the AmericasNLP shared tasks on machine translation (MT)[4] between Indigenous languages and Spanish, organized from 2021 to 2025, that have progressively expanded to 14 languages in recent editions and cover both translation directions. However, most MT benchmarks and corpora focus on a limited set of languages, many of which are primarily represented by varieties from countries such as Mexico, Paraguay, Bolivia, and Peru. In contrast, Indigenous languages spoken in Argentina remain largely underrepresented (Ticona et al., 2025)[5].

Several parallel resources for Qom–Spanish exist, covering a range of linguistic domains and text types (e.g. Messineo (2014); Martínez et al. (2013)). However, these materials are not available in formats readily usable for computational processing and require substantial effort for cleaning, alignment, or normalization.

To our knowledge, there are no parallel or text-based corpora available for machine translation in Qom that are directly suitable for NLP applications. Computational work on Qom is very limited, with prior studies restricted to spoken language identification (Garber and Riera, 2022) and a morphological description based on a linear context-free grammar (Porta, 2010). A small spontaneous speech dataset has been released through Mozilla Common Voice[6], but it is designed for Automatic Speech Recognition and is not suitable for MT. More broadly, computational resources for other Guaycuruan languages such as Pilagá and Mocoví are, to our knowledge, also absent.

Low-resource machine translation studies rely on relatively small parallel datasets, although sizes vary considerably across languages and tasks, ranging from a few hundred to over 100k sentence pairs, with a median of approximately 14.6k in AmericasNLP (de Gibert et al., 2025).

MT systems commonly evaluate performance using BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017). These metrics are widely adopted in low-resource settings, including shared tasks such as AmericasNLP (de Gibert et al., 2025), which also explore additional evaluation metrics. In extremely low-resource scenarios, Biblical or religious corpora are sometimes used as training data, although their impact and prevalence vary across languages and regions.

Recent work on low-resource machine translation increasingly relies on pretrained multilingual models such as NLLB-200, where joint multilingual fine-tuning across language pairs has been shown to outperform per-language approaches (Gow-Smith and Sánchez Villegas, 2023; de Gibert et al., 2025; NLLB Team et al., 2022).

Prior work suggests that transfer learning benefits from typological similarity between languages, supporting the use of related languages as proxies for low-resource settings (Moreno et al., 2024).

4 Corpus Collection and Curation

The corpus used in this work consists of a collection of bilingual Qom–Spanish texts spanning multiple discourse genres, including oral narratives, educational materials, and literary and religious translations. The dataset combines pre-existing parallel resources with texts that were aligned specifically for this study. For the former, we converted already aligned bilingual materials into a machine-readable format. For the latter, we performed the alignment of previously unpaired Qom–Spanish texts.

In this section, we present an overview of the corpus (4.1), describe the data collection and processing steps (4.2), the quality assessment of the alignments (4.3), and finally detail the data filtering and deduplication steps (4.4) and corpus statistics (4.5). A full Data Statement documenting the Qom–Spanish parallel corpus is provided in Appendix A.7.

4.1 Overview of the Corpus

We first present the pre-existing parallel resources, and then those aligned in the present study.

Pre-existing parallel resources

Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco)[7] (Messineo, 2014) (here-

[4] AmericasNLP: https://americasnlp.org/, last accessed Apr. 2026.

[5] Although some languages spoken in Argentina—such as Quechua, Aymara, and Guaraní—are included in initiatives such as AmericasNLP, available datasets for these and other languages spoken in the country, such as Mapudungun, generally correspond to varieties from other countries and do not reflect those used within Argentina.

[6] https://mozilladatacollective.com/datasets/cmn1ppks200u7o1078xxw5izl

[7] [Qom Verbal Art: Advice, Prayers, and Narratives from

after, ***Arte Verbal***) is a compilation of 85 bilingual Qom–Spanish texts comprising advice, ritual speech, and traditional narratives. It represents a valuable source of formalized oral discourse, exhibiting significant pragmatic and stylistic diversity. The texts were originally produced orally in Qom, recorded, and later transcribed and translated into Spanish through collaborative fieldwork. The Spanish translations are subordinate to the Qom source and aim to remain closely aligned in structure and meaning.

Educación Sanitaria Intercultural: Manual de promoción de la salud entre los tobas (qom) del Chaco Central – Comunidades Tobas del Río Bermejito, Chaco (Argentina)[8] (Martínez et al., 2013) (hereafter, ***Manual de Salud***) is a bilingual manual for health promotion among the Qom communities of the Central Chaco. This resource provides expository health content and domain-specific terminology in both Qom and Spanish.

Las aventuras de Copaic, el gato montés.[9] (Haddad, 2022) (hereafter, ***Copaic***) is a short illustrated bilingual story intended for children. It provides examples of simple narrative structures and relatively controlled syntax.

Materiales del Taller de Lengua y Cultura Toba[10] (Messineo and Dell'Arciprete, 2005) (hereafter, ***Taller Derqui***) is a collection of materials for language and cultural outreach, including 10 texts, 6 songs, and Article 75 of the Argentine National Constitution. The materials were developed over four years of community-based workshops held in an urban Indigenous neighborhood, with the aim of strengthening the Qom language and fostering its transmission and visibility within the community.

Resources aligned in this work

La ***Declaración Universal de los Derechos Humanos***[11] (hereafter, ***UDHR***) (OHCHR, 1948a,b) is an internationally recognized legal instrument available in bilingual Qom–Spanish format, containing 4,037 words in Qom and 2,085 words in Spanish.

El Principito[12] is the short novel (96 pages) by Saint-Exupéry, which was manually aligned primarily at the paragraph level, with some alignments at the sentence level between the Spanish version (Saint-Exupéry, 1943) and its Qom translation (Saint-Exupéry, 2005).

La Biblia[13] (hereafter, ***The Bible***) is a collection of texts from both the Old and New Testament, available in bilingual Qom–Spanish format. The Qom version (Sociedad Bíblica Argentina, 2013) and Spanish version (Sociedades Bíblicas Unidas, 1992) were aligned through a custom-developed program at the verse level, where each verse typically corresponds to at least one full sentence, but may sometimes extend to a paragraph.

The corpus preserves the orthographic variation present in the original sources, as it reflects naturally occurring written materials produced by speakers from different communities and sociolinguistic backgrounds rather than a normalized standard (see Section 2). It comprises heterogeneous textual units, including sentence fragments, full sentences, and paragraphs, which may co-occur within the same text. This variation in granularity reflects the original parallelization of the materials. For further details, see Table 1.

4.2 Data Creation and Processing

Both the PDF-aligned text and the version aligned for this work were produced by two authors with a background in computer science, who are Spanish speakers (one native and one non-native) and do not speak Qom. The final alignments were subsequently reviewed by a linguist, a native Spanish speaker and Qom specialist, with over twenty years of experience in the documentation and linguistic description of this Indigenous language (see Section 4.3). The entire process was overseen by a PhD in computer science with experience in NLP, corpus creation, and annotation, who is a native Spanish speaker and does not speak Qom. For further details, see Appendix A.7.

The four originally bilingual sources were available only as PDFs, so the first step was to convert them into a machine-readable tabular format. Extraction strategies varied depending on PDF quality and document structure, combining automated methods (including large language model (LLM)-assisted text reconstruction (e.g., GPT-4) in cases

El Espinillo (Chaco)]

[8] *Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito* [*Intercultural Health Education: A Health Promotion Manual among the Toba (Qom) of the Central Chaco – Toba Communities of the Bermejito River, Chaco (Argentina)*]

[9] *Lmitaxamaxac so copaic* [*The Adventures of the Wildcat*]

[10] *Lo'onatacpi na qom Derqui l'ecpi* [*Qom Language and Culture Workshop Materials (Derqui)*]

[11] *Na nqataxacpi na ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua* [*Universal Declaration of Human Rights*] (United Nations)

[12] *So shiỹaxauolec nta'a* [*The Little Prince*]

[13] *La'aqtaqa Ñim Lo'onatac 'Enauacna* [*The Bible*]

of corrupted encodings) with manual correction where necessary.

Across sources, we filtered out non-parallel or document-structural elements, such as tables without complete sentences, captions, speaker labels, cross-references, and other metadata not corresponding to translational content. Bracketed content was also excluded, as it typically marks neologisms without direct Spanish counterparts.

Texts were then aligned at the paragraph, sentence, or verse level, depending on the structure of each source. Alignment relied on punctuation cues and document structure, and was complemented by manual inspection.

While some sources required additional preprocessing—such as reconstruction from corrupted encodings, exclusion of non-bilingual sections, or normalization of orthographic inconsistencies—others presented relatively clean parallel structures. Detailed, source-specific processing decisions are provided in Appendix A.3.

For texts obtained from web sources (e.g., *UDHR*), Qom and Spanish versions were aligned manually from the outset. In the case of *El principito*, minor discrepancies in layout required paragraph-level restructuring prior to sentence alignment.

We used the Qom version of *the Bible* Sociedad Bíblica Argentina (2013), available at Bible.com (YouVersion)[14], under the code LÑLE13 (*La'aqtaqa Ñim Lo'onatac 'Enauacna*, 2013). The platform allows parallel visualization at the verse level, although the provenance of the Spanish source text is not specified. Among the more than 25 Spanish versions available, we selected *Dios Habla Hoy, Standard Edition (DHHS94)* (Sociedades Bíblicas Unidas, 1992), as it most closely matches the Qom text in style and content. There is a clear correspondence between the two languages at the book, chapter, and verse levels,[15] which allowed us to systematically extract each verse from every chapter and book and store them in parallel format in a CSV file for subsequent analysis. The corpus presents a high level of alignment, but we identified differences in completeness between the two versions, that are discussed in Appendix A.2.

4.3 Alignment Quality Assessment

A full revision of the corpus was conducted by a native Spanish-speaking linguist with expertise in Qom, comparing the original Qom texts with the aligned CSV version. This process included correcting orthographic inconsistencies, errors introduced during PDF-to-CSV conversion (e.g., extra spaces, character confusions), and occasional errors in the source texts. Potentially problematic segments identified during the parallelization stage were also reviewed.

Spelling was standardized across the dataset to better reflect the original forms. Ambiguous cases identified during preprocessing, including misalignments, were carefully checked against both the Qom and Spanish originals. For more details, refer to Appendix A.4.

4.4 Data Preprocessing

After extraction and alignment, all sources underwent a common automated filtering pipeline. Filtering proceeded in three steps: (i) exact duplicate pairs, matched on both the Qom and Spanish sides within the same source document, were removed; (ii) pairs with an empty side were discarded; and (iii) pairs were retained only if the token-length ratio $(|\text{ES}|+1)/(|\text{QOM}|+1)$ fell within $[0.15,\ 6.0]$, ensuring that neither side is more than roughly six times longer than the other, and if both sides contained at least two tokens, to remove extreme length mismatches likely to reflect alignment errors. Apostrophe-like characters were also normalized to U+0027 across all sources, since the glottal stop phoneme is represented by multiple visually similar Unicode codepoints. The segment counts in Table 1 reflect the corpus after these steps.

4.5 Corpus Statistics

We define three corpus variants: *QomL'aqtaqa-Base* (all sources except UDHR[16] and the Bible), *QomL'aqtaqa-Bible*, and *QomL'aqtaqa-Base+Bible*. We will sometimes use *QomL* as an abbreviation of *QomL'aqtaqa*. Table 1 shows the statistics of our corpus.

To account for variability in textual granularity, we report dataset size using segment-based statistics. These segments include full sentences, sentence fragments, and paragraphs. The reported

[14] Bible.com (YouVersion) is a free digital platform to read, listen to, and study the Bible. It offers multiple translations and audio versions, among others: https://www.bible.com/bible.

[15] See https://www.bible.com/bible/574/GEN.1.DHHS94?parallel=128

[16] This material was not included, as it became available after processing had begun.

statistics include the total number of segments, total token counts for each language, average segment length, and the standard deviation of segment lengths. The high standard deviation in some sources reflects heterogeneous segmentation granularity within those documents.

Token counts are based on orthographic segmentation (i.e., units separated by whitespace and punctuation).[17]

5 Machine Translation Baseline

We fine-tuned a pretrained neural machine translation model based on NLLB-200 on the Qom–Spanish parallel corpus. This system serves as a baseline for future work.

Since NLLB-200 does not include Qom (ISO 639-3: `tob`), we used the Guaraní tag (`grn_Latn`) as a proxy, following expert validation in Chacoan typology, which identified it as the most suitable available option. This choice enables a warm-start initialization from a typologically related language, allowing the model to adapt pretrained representations during fine-tuning (Moreno et al., 2024).

5.1 Model and Training Setup

We fine-tuned `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2022), a 600M-parameter distilled variant of the NLLB-200 model covering 200 languages. For fine-tuning, we used the Hugging Face `Seq2SeqTrainer` with the Adafactor optimizer (learning rate 5×10^{-4}), an effective batch size of 16 (2 per device with 8 gradient accumulation steps), fp16 precision, and 10 epochs. Experiments were run on a single NVIDIA T4 GPU (16 GB) via Kaggle (12-hour session limit).

5.2 Corpus Configurations and Data Splits

We experimented with two corpus configurations: **Base** (*QomL'aqtaqa-Base*) and **Base+Bible** (*QomL'aqtaqa-Base+Bible*). The latter incorporates Biblical data, making the corpus roughly 11 times larger (from 2,943[18] to 33,392 pairs). This mirrors a common practice in low-resource machine translation. For instance, Chiruzzo et al. (2022) augmented their Guaraní–Spanish system with Bible data and reported consistent improvements, even though the text differs substantially from contemporary usage.

For each configuration, we applied two split strategies. The **random** split assigns pairs to train/development/test sets uniformly at random (approximately 66/17/17% for *QomL-Base*, and 80/10/10% for *QomL-Base+Bible*). The **stratified** split assigns pairs by source document, ensuring that each document contributes proportionally to every partition and thus maintaining a balanced representation of sources across splits; this results in an approximate 86/7/7% split for *QomL-Base* and 80/10/10% for *QomL-Base+Bible*. The smaller test set in the stratified *QomL-Base* configuration reflects the limited corpus size while retaining a representative evaluation sample. In both cases, the splits are disjoint: no sentence pair appears in more than one partition. Table 2 shows the resulting sizes of the dataset splits.

5.3 Results and Discussion

We report ChrF++ as the primary metric and BLEU as a secondary reference for comparability with prior work, following the evaluation protocol of the AmericasNLP shared tasks (de Gibert et al., 2025), with both metrics computed using SacreBLEU (Post, 2018). Results are shown in Table 3.

QomL-Base+Bible yields substantially higher scores. This is likely due to the fact that the Bible constitutes approximately 91% of its training data and is highly represented in the test sets under both split strategies. Therefore, these scores primarily reflect Bible-register performance rather than general Qom–Spanish translation quality. Given that the test sets of both configurations differ in size and composition—particularly since *QomL-Base+Bible* is dominated by biblical text—absolute scores are not directly comparable across configurations and should therefore be interpreted cautiously.

Under *QomL-Base*, stratified split scores are uniformly higher than random split scores across both directions (e.g., ChrF++ 30.89 vs. 28.81 for ES→QOM). This is the opposite of what is sometimes expected and requires explanation. Under the random split, the test set is dominated by *Arte Verbal* sentences (the largest *QomL-Base* source), which also dominate training; the model is thus largely evaluated on in-distribution examples. The stratified split ensures that each source document contributes proportionally to every partition, including *Taller Derqui* and *Manual de Salud*, which

[17] While Spanish shows moderate morphological complexity, Qom displays a higher degree of synthesis, with single tokens often encoding information that corresponds to multiple words in Spanish. Accordingly, "tokens" are treated here as practical units for corpus comparison rather than as strict morpholexical units.

[18] 2,882 without UDHR.

Title	Reference	Segs	Seg. unit	Tokens		Avg seg. len. (tokens)	
				Qom	ES	Qom (avg ± std)	ES (avg ± std)
Arte verbal Qom	Messineo (2014)	1,831	fragment	16,242	15,009	8.87 ± 5.44	8.20 ± 4.67
Educación Sanitaria Intercultural	Martínez et al. (2013)	212	paragraph; sentence; fragment	7,438	6,343	35.08 ± 35.18	29.92 ± 30.58
Materiales del Taller de Lengua y Cultura Toba	Messineo and Dell'Arciprete (2005)	273	sentence; fragment	2,299	1,946	8.42 ± 4.97	7.13 ± 3.42
Las Aventuras de Copaic	Haddad (2022)	16	paragraph; sentence; fragment	560	484	35.00 ± 26.95	30.25 ± 20.95
El Principito	Saint-Exupéry (1943, 2005)	550	paragraph; sentence; fragment	16,817	15,793	30.58 ± 25.98	28.71 ± 26.44
La Declaración Universal de los Derechos Humanos	OHCHR (1948b)	61	paragraph; sentence	4,037	2,085	66.18 ± 43.44	34.18 ± 22.47
Biblia†	Sociedad Bíblica Argentina (2013); Sociedades Bíblicas Unidas (1992)	30,449	paragraph; sentence	1,422,512	849,684	46.72 ± 21.14	27.91 ± 12.10
Subtotal QomL-Base		*2,943*		*47,393*	*41,660*		
Total QomL-Base+Bible		*33,392*		*1,469,905*	*891,344*		

Table 1: *QomL'aqtaqa* parallel corpus sources and statistics. † Source added in the extended (*QomL-Base+Bible*) configuration. *Segs* counts parallel segment pairs; each segment may correspond to a fragment, a full sentence, or a paragraph, depending on the granularity of the source text. *Fragments* are sub-sentential units reflecting the prosodic segmentation of the original texts, where a single utterance may span multiple lines. Token counts are based on orthographic segmentation. Average segment length and standard deviation are reported in tokens.

Config	Split	Train	Dev	Test
QomL-Base	Random	1,912	485	485
	Stratified	2,488	197	197
QomL-Base+Bible	Random	26,802	3,256	3,273
	Stratified	26,582	3,335	3,414

Table 2: Dataset split sizes for each corpus configuration.

Config	Direction	Split	BLEU	ChrF++
QomL-Base	ES→QOM	Random	4.42	28.81
		Stratified	4.61	30.89
	QOM→ES	Random	4.02	23.05
		Stratified	5.50	23.88
QomL-Base+Bible	ES→QOM	Random	23.55	53.97
		Stratified	23.37	53.45
	QOM→ES	Random	24.73	46.42
		Stratified	22.80	45.02

Table 3: MT results for both corpus configurations and split strategies.

exhibit distinct vocabulary, register, and sentence structure. The slightly higher stratified scores likely reflect the contribution of *El Principito* to the test set under stratification, whose narrative register aligns well with *Arte Verbal* training examples. The QOM→ES direction shows a smaller gap (23.05 vs. 23.88 ChrF++) because generating Spanish benefits from the model's strong Spanish decoder regardless of the Qom source domain.

Although our scores are not directly comparable to those reported in the AmericasNLP 2025 shared task for low-resource Indigenous language MT (de Gibert et al., 2025) due to differences in languages, test sets, and, in our case, a Bible-dominated evaluation set, we provide the shared-task results for contextualization. The best-performing reference systems for each translation direction achieved 47.81 ChrF++ for Indigenous-to-Spanish translation and 36.76 for Spanish-to-Indigenous translation, while our models obtain 46.42 and 53.97 ChrF++, respectively.

5.4 Ablation: Data Composition vs Domain Specificity

The improvements observed in *QomL-Base+Bible* are likely driven by the substantially larger amount of training data. However, its strong performance may be influenced by the high proportion of Bible data (91% of the training data), raising questions about generalization beyond this specific domain.

To investigate the effect of training data composition, we train a **Bible-only** (with *QomL-Bible*) model and compare it against the *QomL-Base* model. Both models are evaluated on a shared held-out test set: the 158-pair *QomL-Base* strati-

fied test set. This provides a controlled setting to compare a model trained exclusively on a single domain (Bible) against a model trained on more diverse non-Bible sources under the same evaluation conditions.

Table 4 reports ChrF++ scores for both models on this test set.

Model	Direction	ChrF++
QomL-Base	ES→QOM	38.25
QomL-Bible	ES→QOM	33.78
QomL-Base	QOM→ES	30.88
QomL-Bible	QOM→ES	20.31

Table 4: Ablation comparing a Bible-only model (*QomL-Bible*) and a model trained on non-Bible data (*QomL-Base*), evaluated on the same 158-pair *QomL-Base* stratified test set.

The *QomL-Bible* model performs consistently worse than the *QomL-Base* model when evaluated on non-Bible content. The larger degradation in the QOM→ES direction (−10.57 ChrF++ points) suggests that the Spanish decoder overfits to a narrow biblical register, which does not transfer well to general Qom–Spanish translation. The smaller gap in the ES→QOM direction (−4.47 ChrF++ points) indicates that lexical knowledge from the Bible may still be partially useful for Qom generation, likely due to shared morphological patterns across registers.

A stratified set of 28 translations per direction from the test set of *QomL-Base*, produced by the *QomL-Base+Bible* model, was evaluated by a linguist with expertise in Qom. Results show 82% correct for Qom→Spanish, and for Spanish→Qom, 89% correct, 7% incorrect, and the remainder uncertain. Evaluation focused on preserving semantic and syntactic integrity with respect to the source, assessing model outputs independently of the references, which were sometimes less accurate. A brief descriptive analysis of translation errors is presented in Appendix A.6.

5.4.1 Lexical Analysis of Generated Output

To characterize the register difference qualitatively, we analyzed the most frequent content words (excluding Spanish stopwords) in the Spanish outputs generated by each model on the *QomL-Base* test set. See Appendix A.5 for word clouds and top-10 content words for each model.

A clear contrast is observed. The *QomL-Base* model produces domain-appropriate vocabulary: *paredes* (walls)[19], *indicador* (indicator), *parásito* (parasite), *árboles* (trees), *agua* (water), and *tobas* reflect the health and educational sources in the corpus. The Bible-only model's output is markedly sparser—its top content word appears only 10 times, compared to 93 for the *QomL-Base* model—and its vocabulary (*comen* (eat), *puede* (can), *lugar* (place), and *lengua* (language)). This confirms that training exclusively on the Bible leads to a model that generates narrow, domain-inappropriate Spanish for general Qom–Spanish translation.

6 Conclusion

We have presented *QomL'aqtaqa*, the first parallel Qom–Spanish corpus in a computationally usable format, along with bidirectional MT baselines fine-tuned from pretrained NLLB-200 models. The corpus combines seven sources spanning oral narratives, educational materials, and literary and religious translations, totaling 33,392 aligned segment pairs in its full configuration. Our ablation confirms that the Bible, while a valuable source of parallel data, introduces a strong register bias: a model trained exclusively on biblical text shows a drop of 10.57 ChrF++ points for QOM→ES and 4.47 for ES→QOM when evaluated on non-biblical content, and generates lexically narrow Spanish dominated by nature and cosmological vocabulary. Overall, the results indicate that the proposed approach is effective for bidirectional Qom–Spanish machine translation.

Several directions remain for future work. We plan to compare the use of the Guaraní proxy code (`grn_Latn`) with adding a new `tob_Latn` embedding, evaluate *QomL-Base+Bible* models on *QomL-Base* test sets to better assess out-of-domain generalization, and explore a two-stage training strategy with multilingual fine-tuning followed by Qom–Spanish adaptation. Future work should also revisit evaluation metrics for Qom given its polysynthetic morphology, and incorporate additional resources such as the *Vocabulario Toba* (Buckwalter and Litwiller de Buckwalter, 2013). Finally, we will expand the corpus with ongoing data collection efforts and develop a lightweight web interface for interactive translation.

[19]The high frequency of *paredes* (walls) reflects its domain-specific use in Chagas prevention materials, where walls are a key site for detecting and controlling vector infestation (e.g., cracks, stains, and hiding places of triatomine bugs).

Limitations

Several limitations should be considered. First, segment heterogeneity (including full sentences, fragments, and paragraphs) may affect both training and evaluation of machine translation systems; however, prior work suggests that sub-sentential fragments can still provide useful translational signal in low-resource settings (Steingrímsson et al., 2023). Second, due to resource constraints, alignment quality was not comprehensively assessed across the entire corpus. Third, the evaluation relies on automatic metrics such as BLEU and chrF++, which may not fully capture translation quality in low-resource, morphologically rich languages. Fourth, we did not evaluate the Bible-only model on a held-out Bible test set, limiting direct comparison between in-domain and out-of-domain performance; we leave this controlled comparison to future work.

Regarding data construction, the splits are group-disjoint at the discourse-unit level, meaning that train, development, and test partitions are defined over complete discourse units (i.e., paragraphs, sentence fragments, or sentences), rather than individual sentence pairs. All sentences belonging to the same discourse unit are assigned to the same partition, ensuring that no unit is split across different subsets and that no exact duplicate pairs appear across partitions. However, substring overlap across units was not explicitly checked, although it is structurally unlikely. In addition, results for *QomL-Base+Bible* are heavily influenced by Bible data (approximately 91% of the training set), and thus primarily reflect performance on this specific register rather than general Qom–Spanish translation quality, and should be interpreted accordingly. Finally, the use of Guaraní (`grn_Latn`) as a proxy language tag for Qom constitutes a practical workaround in the absence of a dedicated code, but introduces a confound in the learned representations. Future work should address this limitation by introducing a dedicated `tob_Latn` embedding.

Ethical considerations

This work builds on prior efforts to collect and parallelize Qom–Spanish data, in which native speakers and community members have been actively involved. In line with established guidelines for working with Indigenous language communities (Bird, 2020), we maintain communication with collaborators and speakers regarding the present work and aim to respect community perspectives and cultural context. Although the primary source materials are mostly publicly available, the processed resources reported in this study are not yet released. We plan to make them available in formats that are accessible and useful to the community in future work.

Acknowledgments

This work was partially supported by the Lacuna Fund Natural Language Processing 2024 grant "Corpus Lengua y Cultura Qom" (Grantee 109).

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6:587–604.

Steven Bird. 2020. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519.

Alberto S. Buckwalter and Lois Litwiller de Buckwalter. 2013. Vocabulario Toba–Castellano y Castellano–Toba. Equipo Menonita. `https://chacoindigena.net/materiales/`. Accessed: 2026-04-07.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2022. Jojajovai: A parallel guaraní–Spanish corpus for MT evaluated with humans and automatic metrics. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2098–2107. European Language Resources Association.

Ona de Gibert, Robert Pugh, Ali Marashian, Raúl Vázquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 134–152. Association for Computational Linguistics.

Leandro Martín Garber and Pablo Ernesto Riera. 2022. Sistema de identificación de idioma (LID) para grabaciones de entornos naturales bilingües en comunidades qom. Ph.D. thesis, Master thesis, Universidad de Buenos Aires.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Communications of the ACM, 64(12):86–92.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 192–199. Association for Computational Linguistics.

María del Rosario Haddad. 2022. Las aventuras de Copaic, el gato montés: ¿Dónde está mi música? Number 1 in Cuentos originarios ilustrados. María del Rosario Haddad and Instituto de Investigación en Etnomusicología, Ciudad Autónoma de Buenos Aires. Illustrated by Carlus Rodríguez. Translated to Qom by Amada Farías et al. Bilingual edition: Spanish–Qom.

Óscar Huamán-Águila, C. E. Fernández-García, and C. R. Gonzales García. 2024. Uso de la transcripción fonética para lograr que avatares de inteligencia artificial pronuncien discurso en lengua quechua: caso illariy. Lengua y Sociedad, 23(2):833–851.

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript draft released August 24, 2025.

Gustavo J. Martínez, Paola Cúneo, and Mauricio Maidana. 2013. Educación Sanitaria Intercultural. Manual de promoción de la salud entre los tobas (qom) del Chaco Central - Comunidades Tobas del Río Bermejito, Chaco (Argentina). Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito, Chaco (Argentina). Museo de Antropología, Universidad Nacional de Córdoba, Córdoba, Argentina. Edición bilingüe Qom–Español.

Cristina Messineo. 2003. Lengua toba (guaycurú): aspectos gramaticales y discursivos. Number 48 in LINCOM Studies in Native American Linguistics. LINCOM Europa, Munich, Germany.

Cristina Messineo. 2014. Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco). Textos y comentarios de Mauricio Maidana. Asociación Civil Rumbo Sur / Ethnographica, Buenos Aires, Argentina. Contains 85 bilingual Qom–Spanish texts compiled with contributions from Mauricio Maidana.

Cristina Messineo and Ana Dell'Arciprete. 2005. Lo'onatacpi na qom Derquil'ecpi: materiales del taller de lengua y cultura toba, 1 edition. Comunidad Toba Daviaxaiqui, Buenos Aires.

Oscar Moreno, Yanua Atamain, and Arturo Oncevay. 2024. Awajun-OP: Multi-domain dataset for Spanish–Awajun machine translation. In Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024), pages 112–120, Mexico City, Mexico. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

OHCHR. 1948a. La Declaración Universal de los Derechos Humanos. Proclamada por la Asamblea General de las Naciones Unidas, Resolución 217 A (III), 10 de diciembre de 1948. Accessed: 2026-04-07.

OHCHR. 1948b. Na nqataxacpi na Ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua [Universal Declaration of Human Rights in Toba/Qom]. Toba (Qom) translation of the 1948 declaration; translation date not recorded by OHCHR. Accessed: 2026-04-07.

D. Ortiz Coronel, R. Lacerda de Sá, L. Trigo, and J. R. Pichel Campos. 2024. Proyecto guaná ia: innovación en la enseñanza de la lengua en riesgo. Lengua y Sociedad, 23(2):945–961.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pages 612–618. Association for Computational Linguistics.

Andrés Osvaldo Porta. 2010. The use of formal language models in the typology of the morphology of amerindian languages. In Proceedings of the ACL 2010 Student Research Workshop, pages 109–114.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191. Association for Computational Linguistics.

Antoine de Saint-Exupéry. 1943. El Principito. Reynal & Hitchcock; Gallimard, Estados Unidos / Francia. Publicada originalmente en francés; existen muchas traducciones posteriores, incluyendo al español y al qom.

Antoine de Saint-Exupéry. 2005. So Shiyaxauolec Nta'a. AEAC Editores, Argentina. Traducción de El Principito al idioma Qom (Toba).

Sociedad Bíblica Argentina. 2013. La'aqtaqa Ñim Lo'onatac 'Enauacna: Qom (Toba) Bible. Accessed: 2026-04-07.

Sociedades Bíblicas Unidas. 1992. Dios Habla Hoy: Versión Española. Spanish Bible (DHH, Versión Española). Accessed: 2026-04-07.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2023. Do not discard – extracting useful fragments from low-quality parallel data to improve machine translation. In Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation, pages 1–13, Macau SAR, China.

Belu Ticona, Fernando Martín Carranza, and Viviana Cotik. 2025. Indigenous languages spoken in argentina: a survey of nlp and speech resources. In Proceedings of the 31st International Conference on Computational Linguistics, pages 8449–8461.

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP Progress in Indigenous Latin American Languages. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 6972–6987.

A Appendix

A.1 Qom phonological system

Below, we present the Qom phonological inventory, based on Messineo (2003). Table 5 lists the consonant phonemes along with their corresponding graphemic representations, while Table 6 shows the vowel phonemes.

A.2 Details on Coverage and Missing Data in the Parallel Bible Corpus

The corpus comprises 35,173 aligned verses from the LNLE13 (Qom) and DHHS94 (Spanish) versions, with high correspondence at the book, chapter, and verse levels. However, completeness differs: LNLE13 reaches 87.9% coverage (4,247 missing verses), mainly due to the absence of the seven deuterocanonical books (Sirach, 1–2 Maccabees, Wisdom, Judith, Tobit, and Baruch), as well as additional partial gaps in several books (e.g., Matthew, Mark, Psalms, Romans, and Luke) and 139 entirely empty chapters. In contrast, DHHS94 is nearly complete (99.0% coverage, 346 missing verses), with only minor, distributed gaps and two empty chapters (Psalms 42–43).

A.3 Source-Specific Processing Details

This appendix provides detailed, source-level descriptions of the extraction, cleaning, and alignment procedures summarized in Section 4.2.

Arte Verbal. This source presented significant challenges due to the use of a legacy font encoding, which resulted in systematically corrupted text under standard extraction methods. To address this, the document was processed in segments of up to six pages. Non-parallel introductory and commentary sections were removed prior to processing. Each segment was then reconstructed using a large language model (GPT-4), producing a machine-readable version of the text. Two page ranges could not be recovered in usable form and were excluded (pp. 84–86 and pp. 95–96 following section Ro.15). Additional post-processing included deduplication and removal of bracketed content marking neologisms without Spanish equivalents.

Taller Derqui. The PDF quality of this source was very low, making automated extraction unreliable. As a result, the full text was manually reviewed and corrected. The document contains word lists presented as grammatical examples; these were excluded, and only the parallel narrative texts were retained.

Manual de Salud. This source required selective extraction due to its heterogeneous structure. Chapters 1 and 5 contained no bilingual content and were excluded. Within the remaining chapters, the following elements were removed: tables lacking complete sentences, figure captions, cross-references to figures, and speaker-role indicators (e.g., *Lta'a (Padre)*, *Qa'ñole (Jovencita)*), as these correspond to document structure rather than translational content. Bracketed content was also excluded, as it marks Qom neologisms without direct Spanish counterparts.

Copaic. This source required minimal processing. The text is short, the layout is clean, and the parallel structure is explicit. No major structural filtering was necessary beyond normalization of apostrophes.

UDHR. The Qom and Spanish versions were obtained from web sources and aligned manually at the article level. Due to the clean and consistent structure of the texts, no additional preprocessing was required.

El Principito. Both the Qom and Spanish versions were processed. Although the placement of illustrations is largely consistent across versions,

Consonants	Labial	Alveolar	Palatal	Velar	Uvular	Laryngeal / Glottal
Plosive	/p/ *p*	/t/ /d/ *t* *d, r*	/tʃ/ *ch*	/k/ /g/ *c, qu* *g, gu*	/q/ /G/ *q* *x*	/ʔ/ *'*
Fricative		/s/ *s*	/ʃ/ /ʒ/ *sh* *y*			/h/ *j*
Nasal	/m/ *m*	/n/ *n*	/ɲ/ *ñ*			
Lateral		/l/ *l*	/ʎ/ *ll*			
Glide	/w/ *hu, u, v*		/y/ *ỹ*			

Table 5: Qom consonant inventory with corresponding grapheme representations, adapted from Messineo (2003).

Vowels	Front	Central	Back
Close	*i*		*o*
Mid	*e*		
Open		*a*	

Table 6: Qom vowel inventory, adapted from Messineo (2003).

minor discrepancies were observed. The text was first reorganized to achieve paragraph-level alignment, guided by illustration placement and manual inspection. Subsequently, sentence- or paragraph-level alignment was performed using punctuation cues, such as full stops and colons.

A.4 Details on Alignment Quality

As mentioned in the body of the paper, the alignment was reviewed by a native Spanish-speaking linguist with extensive expertise working on Qom. While a subset of potentially problematic segments had been identified during the parallelization stage, the revision was performed over the entire corpus.

The review process involved a systematic comparison between the original Qom texts and the aligned CSV version. Several types of issues were identified and corrected:

- **Errors introduced during format conversion (PDF to CSV)**, including:
  - extra or missing whitespace,
  - character confusions (e.g., i instead of l),
  - incorrect handling of special characters representing the glottal stop (phoneme /ʔ/), written with an apostrophe-like symbol (') (see Table 6), and
  - segmentation inconsistencies.
- **Orthographic inconsistencies in the original texts**, which were revised to improve internal consistency while preserving the original forms as much as possible.
- **Occasional errors in the source material**, which were corrected when clearly identifiable.
- **Alignment issues**, such as shifted or mismatched segments, which required cross-checking both the original and processed Qom and Spanish versions.

In addition, previously flagged ambiguous cases were carefully re-examined in context. For *Taller Derqui*, the low PDF resolution meant that the two orthographic variants $\tilde{y}$ and $\bar{y}$, which are equivalent in Qom (see Table 6), could not always be distinguished visually; all instances were unified to $\tilde{y}$ to ensure internal consistency.

The Bible corpus was not reviewed in its entirety; instead, attention was restricted to previously identified doubtful segments.

A.5 Lexical Analysis

Figure 1 and Table 7 present word clouds and the top-10 content words from Spanish outputs generated by the *QomL-Base* and *QomL-Bible* models, excluding stopwords. Analysis is provided in Section 5.4.1.

A.6 Translation Examples

A brief overview of selected translation examples (see Table 8) from both directions of the *QomL-Base+Bible* corpus follows, illustrating the types

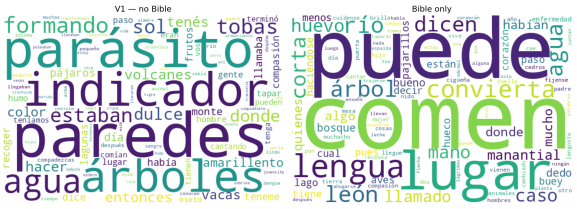

Figure 1: Word clouds of content words in Spanish outputs on the *QomL-Base* test set. Left: *QomL-Base* model (no Bible). Right: *QomL-Bible* model. Word size reflects frequency. For *QomL-Base* the most frequent words are: *paredes* (walls), *indicador* (indicator), *parásito* (parasite), *árboles* (trees), *agua* (water), and textittobas (Tobas). For *QomL-Bible* the most frequent words are: *comen* (eat), *puede* (can), *lugar* (place), *lengua* (language).

QomL-Base		***QomL-Bible***	
Word (EN)	**Count**	**Word (EN)**	**Count**
paredes (walls)	93	comen (eat)	10
parásito (parasite)	66	puede (can)	10
indicador (indicator)	52	lugar (place)	8
árboles (trees)	48	lengua (tongue)	8
agua (water)	46	convierta (turn)	8
formando (forming)	35	león (lion)	7
tobas (Tobas)	30	árbol (tree)	7
estaban (were)	27	río (river)	7
sol (sun)	26	agua (water)	7
donde (where)	20	dicen (say)	6

Table 7: Top-10 content words (stopwords removed) in Spanish outputs generated by each model on the *QomL-Base* stratified test set. English translations are provided for readability only. The *QomL-Base* model produces domain-specific vocabulary from health and educational texts; the *QomL-Bible* model generates sparse output dominated by lower-frequency biblical vocabulary.

of errors identified in the evaluation. The model produces recognizable Qom morphological structure in some cases (example 4, ES→QOM, where the output is verbatim correct), but also exhibits characteristic low-resource errors grounded in specific generated translations.

The Qom→Spanish outputs reveal two distinct types of errors. On the one hand, some cases can be attributed to relatively shallow **lexical substitutions or hallucination-like behavior** (see example 2), where the system selects semantically inappropriate equivalents despite otherwise well-formed structure—for instance, producing *nido de guazuncho y cuero* ('nest of deer and leather') in place of *gallineros y corrales* ('chicken coops and corrals').

On the other hand, more problematic cases involve a deeper **misanalysis of syntax and information structure** (see example 1). In particular, the model fails to correctly interpret discourse-fronted constituents such as *para los qom* ('for the Qom people'), reanalyzing them as canonical subjects (*esa gente toba*, 'those Toba people'). This results in translations that are not only pragmatically infelicitous but also syntactically distorted, suggesting that the system does not adequately capture the interaction between word order and discourse functions in Qom.

In the Spanish→Qom direction, errors tend to cluster around **morphosyntactic constraints** rather than lexical choice. For instance, in example 4, the system produces *Ndoteec da ayem 'auchoxoden*, where *ndoteec* ('only/just') appears in a position that is not grammatically licensed, even though its meaning is appropriate. This indicates partial knowledge of lexical semantics without full control over distributional restrictions. At the same time, there are encouraging signs at the morphological level: the model appears capable of identifying and generating derivational morphology such as diminutives, as in *-olec* (DIM) in *nsoqolec* ('young man', see example 3).

A.7 Data Statement

We present a data statement for *QomL'aqtaqa* corpus following Bender and Friedman (2018)'s data statements and incorporating elements from Gebru et al. (2021)'s Datasheets for datasets.

A. **Curation Rationale**

The primary criterion for constructing the corpus was the inclusion of texts that are publicly available, and accessible in both Qom and Spanish. The corpus includes the following materials:

(a) Arte verbal Qom: consejos, rogativas y relatos de El Espinillo (Chaco) [Qom Verbal Art: Advice, Prayers, and Narratives from El Espinillo (Chaco)] Messineo (2014).

(b) Educación Sanitaria Intercultural: Manual de promoción de la salud entre los tobas (qom) del Chaco Central. Comunidades Tobas del Río Bermejito, Chaco (Argentina) / *Paxaguenaxac da qantela'a da chalataxac yalexa't da nataxac. Lma' na qom tala Bermejito* [Intercultural Health Education: A Health Promotion Manual among the Toba (Qom) of the Central Chaco. Toba Communities of the Bermejito River, Chaco (Argentina)] Martínez et al. (2013).

(c) Las aventuras de Copaic, el gato montés /

	Corpus	Source	Reference	Hypothesis
QOM→ES				
1	*Arte Verbal*	qaq ne'ena qom nache ñi'ipiaqa' taqa'en yi'iyi	y para los qom ese era un lugar de marisca.	y esa gente toba es un lugar de marisca
2	*Manual de Salud*	Da qaỹo'ot ca lauo' ana 'olegaxa qataq nlo laỹi' qoỹoqta que'eca noỹic. Ỹaqto saishet da taigui ca noỹic aca chiincha qataq ca llalec huetaigui ca 'olegaxa lauo' qataq nlo laỹi'.	Poner los gallineros y corrales lo más lejos posible de la vivienda, para que no entren en la casa las vinchucas y crías que estén instaladas en estos lugares.	Haz un nido de guazuncho y cuero lejos de la vivienda, para que no entren en la vivienda las vinchucas y sus crías.
ES→QOM				
3	*Arte Verbal*	"Vos, jovencito, cuando vas al monte	am nsoq da 'anqo'ona ada aviaq	Am nsoqolec da huo'o ca taigui ada'ada aviaq
4	*Taller Derqui*	Solamente te pido que me tengas compasión	Chaqdata da aỹem auchoxonnec qome	Ndoteec da aỹem 'auchoxoden

Table 8: Selected translation examples from the *QomL-Base+Bible* non-stratified model. Examples 1–2 are QOM→ES; examples 3–4 are ES→QOM.

Lmitaxamaxac so copaic [The Adventures of the Wildcat] (Haddad, 2022).

(d) Materiales del Taller de Lengua y Cultura Toba / *Lo'onatacpi na qom Derqui l'ecpi* [Qom Language and Culture Workshop Materials (Derqui)] (Messineo and Dell'Arciprete, 2005).

(e) Declaración Universal de los Derechos Humanos / *Na nqataxacpi na ỹotta'a't shiỹaxauapi mayi netalec ana 'alhua* [Universal Declaration of Human Rights] (OHCHR, 1948a,b).

(f) El Principito / *So shiỹaxauolec nta'a* [The Little Prince] (Saint-Exupéry, 1943, 2005).

(g) La Biblia / *La'aqtaqa Ñim Lo'onatac 'Enauacna* [The Bible] (Sociedades Bíblicas Unidas, 1992; Sociedad Bíblica Argentina, 2013).

Some of these texts *(a-d)* were already aligned, but in PDF format and were converted into a machine-readable format as part of this work, other texts *(e-g)* were previously unpaired. In these cases the alignment constitutes a contribution of our team. The selection strategy combined two complementary types of texts: First, a subset of the corpus (notably *a, c* and *d*) is of particular relevance because these texts were originally produced in the Qom language. They represent culturally grounded genres and styles, reflecting Qom verbal art, worldview, and communicative practices. These materials provide valuable insight into endogenous linguistic structures and discourse patterns. Within this group, *(c) Copaic* is especially notable as it targets a child audience, thus contributing variation in register and genre. Second, the corpus also includes translations from Spanish into Qom, such as widely disseminated texts like *(e) UDHR*, *(f) El Principito*, and *(g) The Bible*. These texts are commonly available across multiple languages and contribute to cross-linguistic comparability. Importantly, they expand the lexical and thematic coverage of the corpus, incorporating domains beyond poetic expression, including educational, health-related, legal, and institutional discourse.

Overall, the curation rationale balances cultural representativeness (through original Qom texts) and domain diversity and comparability (through translated materials). This combination supports broader generalization potential for systems trained on the dataset, while maintaining a strong grounding in authentic Qom language use.

B. **Language Varieties**

Qom, also known as Toba. Guaycuruan family. Area: Gran Chaco, South America. Typological features: polysynthetic and agglutinative tendencies; rich morphology; nouns inflect for number and gender; possessed nouns distinguish between alienable and inalienable; deictic classifiers; verbal morphology includes three distinct sets of person prefixes and suffixes enconding aspectual distinctions, direction, position, reflexivity, and reciprocity; no

copular verb; no adpositions; basic word order is SVO and VS. ISO-639-3: tob. Glottocode: toba1269 (https://glottolog.org/resource/languoid/id/toba1269). Language tag: tob_Latn.

Spanish. Rioplatense Spanish as used in Argentina. Standard Indo-European Romance language with relatively analytic morphology and SVO word order. Used as the target/source language in all translation pairs. Language tag: es_Latn.

C. **Speaker Demographic**

We do not have complete sociolinguistic metadata for all contributors involved in the production of the Qom texts, as such information is not consistently documented in the original materials. Below we report the available information:

(a) Arte Verbal and *(b) Manual de Salud*: Produced by a native Qom speaker (male, 60+ years old) from the western region of Chaco (*dapigueml'ec* variety). The speaker has extensive experience in linguistic–anthropological work, as well as in teaching and translation related to Qom language and culture.

(c) Copaic: Produced collaboratively by a group of Qom speakers, including both men and women, aged 25-45. The contributors are native speakers with strong knowledge of their cultural traditions and oral narratives.

(d) Taller Derqui: Developed collaboratively by adult Qom speakers of diverse ages, genders, and regional origins, all residing in an Indigenous neighborhood in Buenos Aires. All Qom participants are native speakers. The process also involved researchers and students (primarily from linguistics and anthropology) affiliated with the University of Buenos Aires.

(f) El Principito: Translated collectively by multiple adult Qom speakers from the provinces of Chaco and Formosa. The work was carried out across different stages, workshops, and versions. All contributors are native speakers.

(e) UDHR and *(g) The Bible*: No speaker-level metadata available.

Integrated Sociolinguistic Overview:

Age: When available, contributors range from approximately 25 to 60+ years old.

Gender: Both male and female contributors are represented in the subset of texts with available metadata.

Race/ethnicity: Contributors include Qom (Indigenous) speakers and non-Indigenous collaborators.

Native language: All identified Qom contributors are native speakers of Qom; materials also involve non-native collaborators.

Socioeconomic status: All Qom contributors are characterized as low socioeconomic status (low SES).

Number of speakers: The corpus includes both single-author texts and collaboratively produced materials involving multiple contributors; however, the total number of distinct speakers cannot be fully determined due to incomplete records.

Disordered speech: There is no evidence or documentation of disordered speech (e.g., dysarthria) in any part of the corpus.

D. **Annotator Demographic**

Annotation and data processing (including PDF-to-CSV conversion) were carried out by two authors with a background in computer science. Both are Spanish speakers (one native, one non-native) and do not speak Qom. The final alignments and evaluation were reviewed by a trained linguist, a native Spanish speaker with in-depth knowledge of Qom, with over 25 years of experience working with Qom communities. The entire process was overseen by a PhD in computer science with experience in NLP, corpus creation, and annotation, who is a native Spanish speaker and does not speak Qom.

All contributors are adult, non-Indigenous researchers (two men and two women) with higher education.

We acknowledge that the exclusively non-Indigenous composition of the team may influence both the annotation process and the interpretation of the data, representing a potential limitation. At the same time, the long-term collaborative experience of the linguist with Qom communities provides sustained sociolinguistic expertise.

E. **Speech Situation**

We have partial information regarding the speech situation and production context of the materials included in the corpus. The available details are summarized below.

Time and place:

A subset of the materials —specifically *(a) Arte Verbal*, *(c) Copaic*, and *(d) Taller Derqui*—

were collected from the early 2000s onward. These texts originate from Qom-speaking communities primarily in the Chaco region, as well as from collaborative workshop settings in Buenos Aires.

For other materials, such as *(e) UDHR* and *(g) The Bible*, precise information about the time and place of translation into Qom is not available. The health material *(b) Manual de Salud* and *(f) El Principito* were produced/translated during the early 2000s in Buenos Aires, Chaco and Formosa.

Modality (spoken vs. written):

Materials *(a)*, *(c)*, and *(d)* originate in oral production in Qom, recorded in audio, and later transcribed and translated into Spanish through collaborative fieldwork, thus preserving features of naturally occurring speech while providing aligned bilingual material. The transcription is segmented into discourse lines based on prosodic units rather than strictly syntactic or semantic sentence boundaries, reflecting the organization of the original oral performance. Moreover, the Spanish version is fully subsidiary to the Qom source text, offering a free translation intended to remain closely aligned with the structure and meaning of the original rather than functioning as an independent adaptation.

In contrast, *(b)*, *(e)*, *(f)*, and *(g)* are written texts, either originally composed or translated in written form.

Scripted/edited vs. spontaneous:

The orally based materials *(a, c, d)* are relatively spontaneous, although later subject to transcription and little degree of editing.

The translated and institutional materials *(b, e, f, g)* are scripted and edited, as they derive from established written sources and underwent controlled translation processes.

Synchronous vs. asynchronous interaction:

The original oral productions *(a, c, d)* were generated in synchronous communicative contexts (e.g., storytelling, workshops), though the corpus itself represents their later, asynchronous textual form. The written and translated materials *(b, e, f, g)* are asynchronous, as they were produced without real-time interaction between participants.

Intended audience:

(a) Arte verbal: primarily community-internal audiences, with cultural and narrative functions, as well as the explicit aim of contributing to the transmission and dissemination of Qom language and culture.

(c) Copaic: oriented toward children, reflecting an educational and narrative purpose.

(d) Taller Derqui: mixed audience, including Qom community members and learners in educational contexts, and a role in the maintenance, transmission, and dissemination of Qom language and cultural practices.

(b) Manual de Salud: aimed at community health communication, targeting Qom-speaking populations.

(e) UDHR and *(g) The Bible:* broad and general audiences, as part of widely disseminated multilingual texts; in the case of the Bible, it has also historically been used as a tool for evangelization in Indigenous communities.

(f) El Principito: literary audience, including both educational and general readership.

Overall, the corpus reflects a combination of oral, culturally grounded productions and written, translated materials.

F. **Text Characteristics**

The corpus encompasses a broad and diverse range of genres and discourse types, which directly influence both lexical selection and structural patterns.

First, it includes genres that are intrinsic to Qom verbal tradition, particularly represented in *(a) Arte Verbal*, as well as in *(c) Copaic* and *(d) Taller Derqui*. These comprise narrative genres (e.g., traditional stories), as well as persuasive and ritual forms such as advice, exhortations, prayers, and chants. These texts reflect culturally specific communicative practices and exhibit distinctive stylistic and discourse features.

Second, the corpus incorporates health-related genres in *(b) Manual de Salud*, including descriptions of diseases and treatments, as well as simulated or reported medical interactions. These texts introduce domain-specific terminology and explanatory discourse structures.

Third, legal and institutional genres are represented in *(e) UDHR*, as well as in parts of *(d)*, which include the translation of an article of the Argentine Constitution. These materials are characterized by highly specialized vocabulary and formal, technical registers.

In addition, the corpus includes Western literary narrative genre, specifically the short fic-

tional work *(f) El Principito*, which reflects stylistic conventions of European literary tradition.

Finally, *(g) The Bible* represents biblical/religious genre, with its own highly conventionalized structures and specialized vocabulary.

Overall, this diversity of genres and topics results in a corpus that captures a wide spectrum of linguistic variation, from culturally embedded oral discourse to formal, technical, and literary registers. This heterogeneity should be taken into account when interpreting linguistic patterns and evaluating generalization capacity.

G. **Recording Quality**

N/A

H. **Other**

We have obtained permission to use the following resources for building a computationally usable corpus, to be released in the future, and for supporting the development of the translation system: *(a) Arte verbal*, *(b) Manual de Salud*, *(c) Las aventuras de Copaic*, *(d) Taller Derqui*, and *(f) El Principito*.

Toward a Coarse-Labeled Spoken Language Identification Dataset for Central Alaskan Yup'ik and Samoan from US Broadcast Archives

Yangyang Chen **Kyeongmin Rim** **James Pustejovsky**
Department of Computer Science
Brandeis University
{yangyangchen, krim, jamesp}@brandeis.edu

Abstract

Publicly available spoken language identification (LID) systems give sparse and inconsistent coverage of the languages spoken in US communities beyond the contiguous mainland — Alaska Native languages and the languages of the US Pacific Island territories. No system on HuggingFace covers Central Alaskan Yup'ik except the largest variant of Meta's MMS-LID family, and only three MMS-LID variants cover Samoan, while Whisper and VoxLingua107-based models lack both despite including other Polynesian languages. We describe an ongoing effort to build a coarse-labeled LID dataset for Yup'ik and Samoan from US public broadcast archives, benchmark publicly available LID systems on it, and train a simple MLP classifier on top of two frozen pretrained speech encoders as prototypes. We report preliminary corpus statistics, off-the-shelf model performance, and prototype results. Guided by the distinctive phonological typology of the target languages, we outline a phonologically-informed fine-tuning direction as future work.

1 Introduction

Spoken language identification is a prerequisite for nearly every downstream speech-processing task in multilingual settings, including automatic speech recognition, transcription, search, and indexing of audio archives, yet off-the-shelf LID coverage of the languages spoken in US communities beyond the contiguous mainland — Alaska Native languages and those of the US Pacific Island territories — is sparse and inconsistent. Table 1 surveys the publicly available spoken LID systems on HuggingFace against our two target languages — Samoan (ISO 639-3: `smo`) and Central Alaskan Yup'ik (hereafter Yup'ik;[1] ISO 639-3: `esu`) — together with related languages from the same families (Hawaiian and Maori for Polynesian; Central Siberian Yupik for Eskimo-Aleut) that illustrate the inconsistency of existing coverage. Whisper (Radford et al., 2023), VoxLingua107-based models (Valk and Alumäe, 2021), and ESPnet OWSM (Peng et al., 2024) include Hawaiian and Maori but neither Samoan nor any Yupik variety. Meta's Massively Multilingual Speech (MMS) LID family (Pratap et al., 2023) covers Samoan only at its 1024-label size and above, and covers Yup'ik only at the largest 4017-label variant. At 4017 labels, the per-class prior is approximately 2.5×10^{-4}, and zero-shot performance on under-represented languages is accordingly poor. To our knowledge, `mms-lid-4017` is the only publicly available spoken LID model that covers both target languages.

		Target		Related		
System	#langs	`smo`	`esu`	`haw`	`mi`	`ess`
Whisper large-v3	99	—	—	✓	✓	—
VoxLingua107	107	—	—	✓	✓	—
ESPnet OWSM v4	∼150	—	—	✓	✓	—
mms-lid-126/256/512	126–512	—	—	—	—	—
mms-lid-1024	1024	✓	—	—	—	—
mms-lid-2048	2048	✓	—	—	—	—
mms-lid-4017	4017	✓	✓	✓	✓	✓

Table 1: Coverage of the two target languages (Samoan `smo`, Yup'ik `esu`) and three related languages from the same families (Hawaiian `haw`, Maori `mi` for Polynesian; Central Siberian Yupik `ess` for Eskimo-Aleut) across publicly available spoken LID systems. Only `mms-lid-4017` covers both target languages. That it also covers `ess` (the primary source of Yup'ik confusion in Section 5) means the model has both labels in its vocabulary yet cannot reliably distinguish them.

Against this backdrop, we describe an ongoing effort to build LID capability for Yup'ik and Samoan using archival broadcast recordings from US public media (Section 4). This setting differs from standard LID corpora in three important

[1] We follow the standard orthographic convention in which the apostrophe in Yup'ik marks the geminate /pː/ specific to the Central Alaskan variety; *Yupik* without the apostrophe refers to the broader language family.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 203–211

ways: (i) speech is interspersed with English in code-switching and bilingual programming contexts, (ii) audio quality varies across decades from analog tape to digital capture, and (iii) a large fraction of the raw material is non-speech (music, ambient sound, silence) that must be filtered before any language labeling.

The pairing of these two languages is driven by the archive and the communities it serves, not chosen arbitrarily. The American Archive of Public Broadcasting (AAPB) — a national collection of digitized US public radio and television — holds decades of non-English programming in many under-resourced languages; Yup'ik and Samoan are two for which the holdings are both substantial and tied to a concrete access need, comprising long-running Yup'ik public-television programming from Alaska and Samoan-language broadcasting from American Samoa. For the speaker communities that produced this material, language is the primary axis of access: without language-level metadata the recordings remain effectively unfindable to the people they most concern. Supplying that metadata reliably is the role of LID here, and doing so responsibly requires sustained outreach with those communities, who are to lead the larger-scale annotation effort that this pilot is designed to bootstrap (Section 7). This effort is one part of a broader program to make US public-broadcast collections more accessible to the communities they document; a parallel line of work prompts vision–language models to extract structured entities, including Hawaiian personal names, from on-screen television chyrons (Lynch et al., 2026). Neither language is served by existing tooling: each is absent from or marginal in every off-the-shelf LID system we surveyed (Table 1). The pairing is also methodologically useful, because the two languages sit at near-opposite poles of phonological typology (Section 2): a single system that must separate both from English and from their respective family neighbors is a demanding testbed for the phonologically-informed direction we pursue.

This paper makes four contributions. First, we describe a pipeline for coarse segment-level LID annotation over long-form broadcast material, designed to be executable by non-native researchers on a short timeline without sacrificing label reliability. Second, we report preliminary corpus statistics for a small in-house annotated set covering both target languages. Third, we benchmark publicly available LID systems on this set, documenting how off-the-shelf models perform when the target language is either absent from or severely under-represented in the training label space. Fourth, we train a simple MLP head on top of two frozen pretrained speech encoders (XLS-R-300M and Whisper-medium) as prototype classifiers, showing that even minimal in-domain supervision substantially improves Yup'ik recall over zero-shot MMS-LID, and that the choice of pretrained encoder is itself a major lever at this scale. The results motivate a phonologically-informed fine-tuning direction, which we sketch in Section 7.

2 Target Languages

Central Alaskan Yup'ik and Samoan occupy maximally distant points in several dimensions of phonological typology. Table 2 summarizes the contrasts we return to in Section 7.

Yup'ik, an Eskimo-Aleut language spoken in western Alaska, exhibits a relatively rich consonantal inventory including velar and uvular consonants, voiceless sonorants, consonant gemination, and permissive consonant clustering (Jacobson, 1995). Its phonology additionally includes a four-vowel system with contrastive vowel length and an iambic prosodic system associated with systematic consonant lengthening (Woodbury, 1987). These properties produce dense consonantal transitions and distinctive durational patterns in running speech.

In contrast, Samoan, a Polynesian language of the Austronesian family, has a much smaller consonant inventory and highly restrictive syllable structure. Samoan syllables are strictly of the form (C)V, disallowing consonant clusters and codas entirely. The language additionally exhibits phonemic vowel length and a phonemic glottal stop, yielding speech with high vocalic density and relatively regular rhythmic structure (Mosel and Hovdhaugen, 1992).

These typological contrasts are directly relevant to LID. Yup'ik contains several acoustically salient but cross-linguistically rare features, including velar and uvular consonants as well as voiceless nasals, while Samoan is characterized by highly regular vowel-rich phonotactics and minimal consonant complexity. The two languages also differ substantially from English along multiple dimensions of syllable structure and segment inventory, suggesting that phonologically-informed representations may provide useful inductive bias for low-

Feature	Yup'ik	Samoan	English
Vowel inventory	4	5+length	11+
Consonant inventory	∼25	∼13	∼24
Velar and uvular consonants	✓		
Voiceless nasals	✓		
Consonant clusters	✓		✓
Gemination	✓		
Syllable structure	CVC	(C)V	complex
Glottal stop phonemic		✓	

Table 2: Contrastive phonological features across the three languages in our LID task.

resource spoken LID.

3 Related Work

Our paper follows a line of AmericasNLP work that documents dataset construction for an individual under-resourced language as the central contribution, with modest or no experimental results. Reyes Pérez and García Zuñiga (2024)'s description of the curated Amuzgo dataset is a direct precedent in this tradition, pairing a detailed linguistic description with a multi-phase data collection and annotation pipeline. A long line of work on St. Lawrence Island Yupik (Schwartz et al., 2021; Chen et al., 2020) has produced textual corpora and morphological analyzers but no speech resources for LID evaluation.

On the methods side, phonologically-informed LID is an active area. Liu et al. (2022) unify acoustic-phonetic and phonotactic information into a single encoder (PHO-LID) and report over 40% relative improvement on AP17-OLR. Shahin et al. (2023) show that a two-stage pipeline (pretrain on manner and place of articulation detection, then fine-tune on LID) improves code-switching LID by 5–6% relative. Universal phone recognizers such as Allosaurus (Li et al., 2020) and Allophant (Glocker et al., 2023) provide a pathway to obtain phonetic pseudo-labels in any language, which can be mapped to articulatory feature vectors via PanPhon (Mortensen et al., 2016) and PHOIBLE (Moran and McCloy, 2019). We leverage these directions in Section 7.

4 Data and Annotation

4.1 Source material

We take data from the AAPB Online Reading Room,[2] a publicly available collection of digitized public radio and television content from stations across the United States. The Yup'ik material originates with KYUK, a public broadcasting station in western Alaska that has produced Yup'ik television news and programming since the 1970s; the recordings used here are video, and contain a mixture of fully Yup'ik segments, English segments, and code-switching within single broadcasts. The Samoan material comes from KVZK, the public television broadcaster in American Samoa, and covers Samoan-language news and cultural programming. The archive spans several decades and includes both edited broadcast output and raw field recordings, with substantial variation in audio quality. All materials used in this pilot study are accessed under the AAPB's existing access and use policies; we do not redistribute the source audio.

4.2 Annotation protocol

Our annotation protocol is designed for a short timeline with in-house researcher annotators rather than expert native speakers. Two observations justify this choice. First, programming from Alaska in particular has a high base rate of Yup'ik in its non-English content: the overwhelming majority of non-English speech in this broadcast context is Yup'ik rather than another language. Similarly, the American Samoa-source material is predominantly Samoan and English. In both cases, program metadata (titles, descriptions) and audiovisual contexts provide additional disambiguation. Second, our target label set is deliberately coarse. We annotate at the segment level with five labels: `english`, `esu`, `smo`, `mixed`, and `other`. The `other` category includes non-speech material and unintelligible audio. This coarse label set is sufficient to train and evaluate a LID system while being reliably producible by a non-native researcher annotator working from a video source with context.

The annotation pipeline is as follows:

1. Voice activity detection and segmentation to strip non-speech regions and produce speech chunks suitable for annotation.

2. Manual segment-level labeling using an in-house annotation interface, with video context available to the annotator throughout.

3. *Dual annotation*: each segment in the current evaluation set is labeled independently by two authors, and inter-annotator agreement is reported in Section 4.4.

[2] https://americanarchive.org/

4.2.1 Labeling guidelines

Broadcast material mixes languages in structured ways that our coarse label set must handle consistently across annotators. We adopt the following rules, finalized after an initial annotation pilot:

- **Dominant-language rule.** If a segment contains speech in more than one language but one language clearly dominates, the segment receives the dominant language's label rather than `mixed`. A segment is labeled `mixed` only when the two languages are roughly balanced or alternate in genuine code-switching.
- **News broadcast convention.** News segments in which a target-language anchor delivers content first and an English broadcaster restates the same content afterwards are labeled with the target language (`esu` or `smo`), not `mixed`, since the English portion is a repetition rather than parallel code-switching.
- **Proper nouns.** Proper nouns (personal names, place names, organizations) embedded in otherwise-monolingual speech do not trigger the `mixed` label; the segment takes the label of the surrounding discourse.
- **Numbers and borrowed English vocabulary.** Yup'ik broadcasters frequently read numbers, weather figures, and commercial catch report data in English inside otherwise-Yup'ik speech. We treat such stretches as `esu` under the dominant-language rule, because the English tokens are routine lexical borrowings rather than genuine code-switching.

These rules are encoded in the annotator guidelines we plan to hand off to native-speaker annotators in subsequent work.

4.3 Corpus statistics

Table 3 reports corpus statistics for the current annotated set, using the primary annotator's labels as gold.

4.4 Inter-annotator agreement

A second author independently annotated 10 of the 15 videos (6 Yup'ik, 4 Samoan), producing 1,342 dual-annotated segments covering 9.3 hours of audio. Table 4 reports the confusion matrix between the two annotators. Raw agreement is 75.3% and Cohen's κ is 0.67, indicating substantial agreement.

	Yup'ik	Samoan
Videos annotated	9	6
Total duration	5.1 h	5.1 h
Total segments	589	1,189
labeled `esu`/`smo`	283	685
labeled `eng`	238	308
labeled `mixed`	53	122
labeled `other`	15	74

Table 3: Corpus statistics for the in-house annotated set (primary annotator). Segments labeled `other` (non-speech) are excluded from all evaluations.

	Annotator 2				
Annotator 1	`eng`	`esu`	`mix`	`oth`	`smo`
`eng`	249	2	39	6	11
`esu`	10	112	6	5	0
`mixed`	39	17	113	0	29
`other`	6	4	0	83	26
`smo`	35	0	49	43	458

Table 4: Inter-annotator confusion matrix over 1,342 dual-annotated segments (raw agreement 75.3%, Cohen's $\kappa = 0.67$).

The primary source of disagreement is the `mixed` label: 179 of 327 disagreements (55%) involve one annotator choosing `mixed` and the other choosing a specific language or `other`. This is consistent with the inherent subjectivity of the dominant-language rule (Section 4.2.1), which requires a judgment call on whether code-switching is "genuine" or incidental. Agreement on the two target-language labels is considerably higher: Yup'ik segments agree 84% of the time (112/133) and Samoan segments agree 78% (458/585). These per-class agreement rates support the reliability of the gold labels used in the MMS-LID and prototype evaluations (Sections 5–6).

5 Benchmarking Off-the-Shelf LID

We evaluate the three MMS-LID variants that cover at least one target language (Table 1) on our annotated evaluation set, excluding segments labeled `other` (non-speech and unintelligible audio). Whisper, VoxLingua107, and ESPnet OWSM cover neither target language and are omitted from this evaluation. Table 5 reports per-class precision, recall, and F1 for the target-language and English labels within each language group. The `mixed` class is never predicted by any MMS-LID model (which outputs a single language label) and therefore has zero recall across the board; we omit it from the table.

System	Class	Yup'ik videos P	R	F1	Samoan videos P	R	F1
mms-lid-1024	eng	.96	.96	.96	.94	.65	.77
	target	—	—	—	.92	.57	.71
mms-lid-2048	eng	.96	.93	.94	.95	.52	.67
	target	—	—	—	.93	.58	.72
mms-lid-4017	eng	.98	.92	.95	.97	.50	.66
	target	.85	.41	.55	.98	.61	.75

Table 5: Zero-shot per-class precision (P), recall (R), and F1 for the MMS-LID model family on our annotated evaluation set, broken down by language group. "target" = esu for Yup'ik, smo for Samoan. Dashes mark models that cannot emit the target label. Segments labeled other and mixed are excluded.

Gold Label	Predicted Label	Count
esu	ess	111
esu	esu	110
esu	esk	16
esu	esi	7
smo	smo	390
smo	haw	128
smo	ton	93
smo	jav	6

Table 6: Top predicted labels for gold Yup'ik (esu) and Samoan (smo) clips under MMS-LID-4017. Most errors remain within related language families.

Finding 1: Samoan recall is moderate but precision is high. All three MMS-LID variants that cover Samoan achieve high precision on smo (0.92–0.98) but only moderate recall (0.57–0.61): the models recognize Samoan when they predict it, but frequently assign related Polynesian labels, primarily Hawaiian (haw) and Tongan (ton), to segments that are in fact Samoan. This Polynesian confusion is expected given that these languages share much of their phonological profile (small inventories, strict CV syllable structure, phonemic vowel length). Performance does not improve with model size, suggesting the confusion is driven by acoustic similarity rather than label-space sparsity.

Finding 2: Yup'ik is systematically confused with other Yupik varieties. Only mms-lid-4017 can emit the esu label. It achieves high precision (0.85) but low recall (0.41): the majority of Yup'ik segments are assigned to Central Siberian Yupik (ess) or Northwest Alaska Inupiatun (esk) instead. This is not a general-purpose model failure: Central Alaskan and Central Siberian Yupik are close relatives within the Yupik subgroup of Eskimo-Aleut, sharing most of their segmental inventory. Table 6 shows that the most frequent errors concentrate on genealogically related languages. The confusion is best read as a linguistically principled limitation that the aggregate MMS-LID label space does not resolve. It also constitutes direct motivation for phonologically-informed modeling: the two varieties differ in specific features (e.g., certain uvular realizations, subsets of the voiceless sonorant series, and prosodic detail) that an articulatory-aware encoder could in principle exploit. We revisit this in Section 7.

Finding 3: English recall degrades in Samoan-source videos. English is recognized with 92–96% recall in Yup'ik-source videos but only 50–65% in Samoan-source videos. We attribute this to the Samoan material containing more code-switching and borrowed English vocabulary, which MMS-LID tends to assign a Polynesian label rather than English.

6 Prototype Classifier

To test how much of the off-the-shelf performance gap can be closed without fine-tuning the speech encoder, we train a lightweight MLP classification head on top of a *frozen* encoder and repeat the experiment with a second frozen encoder of comparable scale. We compare XLS-R-300M (Conneau et al., 2020), a self-supervised wav2vec 2.0 model, against the encoder of Whisper-medium (Radford et al., 2023), a multilingual ASR model whose encoder has about 307M parameters. Both produce 1024-dimensional hidden states. For each segment we extract mean-pooled final-layer activations (over the audio-valid frames in the Whisper case, whose input is padded to 30 seconds) and train a two-layer MLP (1024→256→4, with ReLU and dropout) on those fixed features. Holding the head, the data, and the recipe constant lets the comparison isolate two effects: the value of in-domain supervision relative to zero-shot MMS-LID, and the choice of pretrained encoder.

The classifier is trained with cross-entropy over four labels {eng, esu, smo, other}: segments labeled mixed in the annotation are collapsed into the target language of their source video (Section 4.2.1). We evaluate with 15-fold leave-one-video-out cross-validation across all Yup'ik and Samoan videos combined; each fold trains on 14

Encoder	Class	P	R	F1
XLS-R-300M	eng	.68	.75	.71
	esu	.79	.85	.82
	smo	.80	.69	.74
Whisper-medium	eng	.93	.83	.88
	esu	.96	.92	.94
	smo	.91	.91	.91

Table 7: Prototype classifier with a trainable MLP head on top of two frozen encoders, evaluated with 15-fold leave-one-video-out cross-validation. Overall accuracy / macro F1 are .74 / .76 for XLS-R-300M and .89 / .91 for Whisper-medium. Segments labeled `other` and `mixed` are excluded from the reported metrics; `mixed` is collapsed into the video's target language during training.

videos and evaluates on the held-out video. Table 7 reports per-class precision, recall, and F1 for both encoders, excluding the `other` class to match the MMS-LID evaluation in Table 5.

Comparison with MMS-LID, and across encoders. Despite using only frozen encoders, both prototypes substantially outperform zero-shot MMS-LID-4017 on Yup'ik: XLS-R-300M lifts `esu` recall from 0.41 to 0.85 (F1 0.55→0.82), and Whisper-medium lifts it further to 0.92 (F1 0.94). MMS-LID's principal Yup'ik failure mode — the `esu`/`ess` confusion in Table 6 — is absent here by construction, since our head emits only the four annotation labels; the relevant question is how cleanly each encoder's representations separate those four classes, and the gap between XLS-R and Whisper answers it. For Samoan, where MMS-LID was already moderately strong (F1 0.75), XLS-R offers parity (F1 0.74); Whisper-medium reaches F1 0.91, with the off-diagonal `smo`→`eng` count dropping from 162 under XLS-R to 19. English precision follows the same pattern: 0.68 with XLS-R, recovering to 0.93 with Whisper, comparable to MMS-LID's 0.97–0.98. At this parameter scale, the choice of pretrained encoder dominates the supervision recipe: a model whose pretraining objective rewards phonetic discriminability across many languages (Whisper, multilingual ASR) yields markedly more LID-discriminable representations than one trained by self-supervised reconstruction (XLS-R).

Encoder scale. We also ran the same head on top of Whisper-large-v3, roughly twice the encoder parameter count of Whisper-medium: overall macro F1 is essentially identical (0.91 vs. 0.91), and per-class scores move within noise. We speculate that at this scale of supervision and label granularity — about 1.6k labeled clips and four coarse classes — the bottleneck has shifted away from encoder capacity, and that the additional parameters in large-v3 chiefly refine fine-grained acoustic detail that a coarse four-class LID head cannot exploit.

Per-video variance. Two Yup'ik folds stand out as hard cases under XLS-R-300M (Folds 4 and 8 in the LOO-CV, accuracies 0.52 and 0.60 against the Yup'ik-group average of 0.87). Switching to Whisper-medium resolves one of them: Fold 8, a raw field recording with markedly noisier audio than the surrounding broadcast segments, climbs to 0.94, while Fold 4, an edited broadcast clip set, remains the hardest at 0.59 even with the stronger encoder. The split points to two distinct sources of difficulty — acoustic distance from the encoder's pretraining distribution, which a stronger encoder can absorb, and fold-specific factors that persist across encoders and warrant further investigation — and further motivates encoder adaptation (Section 7).

Pilot-scale caveat. The headline macro F1 of 0.91 under Whisper-medium should be read against the scale of the supporting evidence: 15 broadcast videos, about 1.6k labeled segments, and one held-out video per fold. The picture is likely to shift once the larger native-speaker-led annotation effort (Section 7) yields a more domain-diverse training and evaluation set, and the numbers reported here should be treated as pilot-scale evidence rather than settled levels.

7 Discussion and Future Work

7.1 Discussion

The benchmark results in Section 5 reveal a shared failure mode across both target languages: MMS-LID achieves high precision but low recall, because it systematically confuses target-language segments with closely related languages within the same family: Hawaiian and Tongan for Samoan, Central Siberian Yupik and Northwest Alaska Inupiatun for Yup'ik. These are not random misclassifications: they reflect genuine acoustic similarity within language subgroups that a model trained on broad typological coverage cannot fully resolve.

To probe whether these confusions align with broad phonological tendencies, we ran the universal phone recognizer Allosaurus over the annotated clips and computed two simple phone-level proxies

Gold label	Vowel ratio	Uvular-like ratio
eng	0.409	0.018
esu	0.441	0.051
smo	0.516	0.023

Table 8: Exploratory phone-level statistics computed from Allosaurus pseudo-phone sequences. Samoan exhibits the highest vowel-token ratio, consistent with its strict (C)V syllable structure, while Yup'ik shows the highest rate of uvular-like phones, consistent with its documented velar and uvular inventory.

from the resulting pseudo-phone sequences. These outputs are not treated as gold phonetic transcriptions, but as approximate acoustic-phonetic representations suitable for exploratory analysis. Table 8 summarizes two aggregate measures by gold label: vowel-token ratio and uvular-like phone ratio.

The resulting distributions align with the typological contrasts outlined in Section 2. Samoan clips exhibit the highest vowel-token ratio, consistent with the language's strict (C)V syllable structure and high vocalic density, while Yup'ik clips show the highest rate of uvular-like phones, consistent with its documented velar and uvular inventory. These exploratory results support the interpretation that the MMS-LID confusions are phonologically structured rather than random.

This pattern directly motivates a phonologically-informed approach to LID of closely related varieties. For example, Central Alaskan Yup'ik shares substantial similarities with other Yupik varieties, but comparative descriptions between Central Alaskan and Central Siberian Yupik also note several systematic differences, such as retroflex segments reported for Central Siberian Yupik that are not typically reported in Yup'ik, and differences in gemination and prosodic timing (Section 2) (Jacobson, 1990). Features like these may help reduce the frequent confusions observed in our experiments. Similarly, although Samoan, Hawaiian, and Tongan share many phonological properties, finer-grained differences in segment inventories and specific prosodic patterns such as tonal marking of absolutive case (Yu, 2021) may support improved discrimination within the language family. These are precisely the dimensions that a loss designed around articulatory targets can steer an encoder toward. The prototype results in Section 6 already provide indirect empirical support: under the same MLP head, Whisper-medium (multilingual ASR pretraining) substantially outperforms XLS-R-300M (self-supervised reconstruction), suggesting that pretraining objectives rewarding phonetic discriminability across languages already carry much of the inductive bias LID benefits from.

We sketch a two-stage phonologically-informed fine-tuning pipeline as the direction of our continuing work. Stage 1 follows Shahin et al. (2023): the wav2vec 2.0 encoder is first fine-tuned on articulatory feature classification (manner and place of articulation) with pseudo-labels obtained from universal phone recognizers such as Allosaurus or Allophant, mapped to feature vectors via PanPhon. This requires no target-language transcription. Stage 2 fine-tunes the resulting articulatory-aware encoder for LID on the annotated broadcast data, optionally augmenting the classification loss with a regularization term that encourages representations whose inter-language distances correlate with PHOIBLE inventory distances. Recent work (Choi et al., 2026) has shown that self-supervised speech models already encode distinctive features as linear subspaces in their hidden states; the proposed fine-tuning should sharpen exactly the dimensions that are discriminative within the Yupik subgroup, and should help most in the short-segment regime where aggregate statistical LID degrades.

7.2 Ongoing Annotation and Future Directions

Beyond the modeling direction, immediate next steps include transitioning primary annotation to trained native-speaker annotators, who will expand the dataset over the coming months; expanding the label set to include code-switching boundaries and speaker turns; testing the MMS-LID Samoan result under noisier, non-studio recording conditions; and adding speaker and recording-era metadata to support fairness evaluation across the multi-decade archive.

Since the initial submission of this paper, a larger-scale annotation effort has begun with community language fellows working directly on Yup'ik and Samoan broadcast materials. These annotators are contributing segment-level review, correction, and expansion of the coarse-label dataset using the same annotation interface and guidelines described in Section 4. Their ongoing work will substantially expand both the scale and linguistic reliability of the corpus, and will enable future evaluation under native-speaker supervision.

8 Conclusion

We present a pilot spoken language identification benchmark for Central Alaskan Yup'ik and Samoan using archival broadcast recordings. Experiments with off-the-shelf multilingual LID systems show that exact-label recognition remains challenging, with many errors concentrated within the related language families. Lightweight supervised adaptation substantially improves performance, and the choice of frozen pretrained encoder is itself a major lever: under an otherwise identical head, an ASR-pretrained encoder yields substantially fewer errors than a self-supervised one of comparable scale. Future work will expand annotation, incorporate native-speaker review and explore phonologically informed modeling approaches.

Limitations

The annotated set reported in this paper is small (9 Yup'ik and 6 Samoan videos) and was produced by in-house researcher annotators rather than native speakers. A deliberately coarse five-label scheme and high base rate of Yup'ik and Samoan in non-English content mitigate annotator noise, and dual annotation yields substantial agreement ($\kappa = 0.67$), but native-speaker validation is required before strong claims can be made about absolute label quality. The dominant source of annotator disagreement is the `mixed` label, which is inherently subjective; the target-language labels themselves agree at 78–84%. The prototype classifiers are trained and evaluated on a dataset far smaller than typical low-resource LID settings; the reported numbers, including the Whisper-medium macro F1 of 0.91, should be read as pilot-scale evidence and may not generalize to a larger or more domain-diverse set. The multi-decade time span of the broadcast source introduces acoustic variability (analog vs. digital capture, channel and codec differences) that we do not yet explicitly model. Finally, the phonologically-informed approach in Section 7 is proposed, not yet implemented or evaluated.

Acknowledgments

We thank KYUK in western Alaska and KVZK in American Samoa for providing access to the broadcast recordings that made this study possible. These recordings were provided through the American Archive of Public Broadcasting [3], a collaboration between GBH and the U.S. Library of Congress. This research was supported in part by Andrew W. Mellon Foundation.

[3] https://americanarchive.org/

References

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved finite-state morphological analysis for St. Lawrence Island Yupik using paradigm function morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2676–2684.

Kwanghee Choi and 1 others. 2026. Self-supervised speech models discover phonological vector arithmetic. *Computing Research Repository*, arXiv:2602.18899.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. volume arXiv:2006.13979.

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-lingual phoneme recognition with articulatory attributes. In *Proceedings of Interspeech 2023*.

Steven A. Jacobson. 1990. Comparison of central alaskan yup'ik eskimo and central siberian yupik eskimo. *International Journal of American Linguistics*, 56(2):264–286.

Steven A. Jacobson. 1995. *A Practical Grammar of the Central Alaskan Yup'ik Eskimo Language*. Alaska Native Language Center, University of Alaska Fairbanks.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of ICASSP 2020*.

Hexin Liu, Leibny Paola Garcia Perera, Andy W. H. Khong, Suzy J. Styles, and Sanjeev Khudanpur. 2022. PHO-LID: A unified model incorporating acoustic-phonetic and phonotactic information for language identification. In *Proceedings of Interspeech 2022*.

Kelley Lynch, Owen King, Kyeongmin Rim, Gabrielle Keen, Yangyang Chen, and James Pustejovsky. 2026. Structured entity extraction from Hawaiian television chyrons using vision-language models. In *Proceedings of the SIGUL 2026 Workshop: Towards Inclusivity and Equality — Language Resources and Technologies for Under-Resourced and Endangered Languages*, Palma, Mallorca, Spain. Joint workshop with ELE, EURALI, and DCLRL, co-located with LREC 2026.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016*.

Ulrike Mosel and Even Hovdhaugen. 1992. *Samoan Reference Grammar*. Scandinavian University Press, Oslo.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-weon Jung, Soumi Maiti, and Shinji Watanabe. 2024. OWSM v3.1: Better and faster open Whisper-style speech models based on E-Branchformer. In *Proceedings of Interspeech 2024*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. volume arXiv:2305.13516.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Antonio Reyes Pérez and Hamlet Antonio García Zuñiga. 2024. From field linguistics to NLP: Creating a curated dataset in Amuzgo language. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 127–131.

Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn, and Sylvia L.R. Schreiner. 2021. A digital corpus of St. Lawrence Island Yupik. In *Proceedings of the Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Mostafa Shahin, Zheng Nan, Vidhyasaharan Sethu, and Beena Ahmed. 2023. Improving wav2vec2-based spoken language identification by learning phonological features. In *Proceedings of Interspeech 2023*.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: A dataset for spoken language recognition. In *Proceedings of IEEE SLT 2021*.

Anthony C. Woodbury. 1987. Meaningful phonological processes: A consideration of Central Alaskan Yupik Eskimo prosody. *Language*, 63(4):685–740.

Kristine M. Yu. 2021. Tonal marking of absolutive case in Samoan. *Natural Language & Linguistic Theory*, 39:291–365.

Retrieval-Augmented Long-Context Translation for Cultural Image Captioning: Gators submission for AmericasNLP 2026 shared task

Aashish Dhawan
University of Florida
aashish.dhawan@ufl.edu

Christopher Driggers-Ellis
University of Florida
driggersellis.cw@ufl.edu

Dzmitry Kasinets
University of Florida
dkasinets@ufl.edu

Christan Grant
University of Florida
christan@ufl.edu

Daisy Wang
University of Florida
daisyw@cise.ufl.edu

Abstract

We present the University of Florida Gators submission to the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. Our two-stage pipeline generates a Spanish intermediate caption with Qwen2.5-VL, then produces the target-language caption using retrieval-augmented many-shot prompting with Gemini 2.5 Flash. We achieve 164.1%, 131.7%, and 122.6% improvements over the shared task baseline for Bribri, Guaraní, and Orizaba Nahuatl captioning, respectively, in our dev set evaluation and maintain >150% improvements for the Bribri and Orizaba Nahuatl languages in the test set evaluation. We find retrieval is highly language-dependent, beneficial only for large, in-domain corpora, and that synthetic data augmentation accounts for around 28 chrF++ of the dev set Guaraní performance gain. Our submission is the overall winner of the shared task, placing second out of five finalist submissions in human evaluations of target-language captions. Code and prompts are available on GitHub.[1]

1 Introduction

The AmericasNLP 2026 (Bui et al., 2026) consists of generating culturally grounded captions for images in Indigenous languages of the Americas. We identify three major challenges. First, the target languages are low-resource. Second, the captions are culturally specific rather than generic visual descriptions. Third, the task requires not only lexical transfer, but also stylistic control. Successful systems must produce short, natural captions that match the reference register expected by the organizers.

[1] https://github.com/dhawan98/AmericasNLP2026-Gators-Submission

While the cultural image captioning task is new to AmericasNLP 2026, we find that current vision-language models (VLMs) are not able to directly caption images in the target languages. This is unsurprising given the limited training data available for the target languages. We therefore formulate the problem as a compound of image captioning in a high-resource language followed by machine translation from that language into one of the shared task targets, which mirrors the shared task's baseline approach.

We first attempt to build on the work of Dhawan et al. (2026) by providing intermediate Spanish (Es) VLM captions to an mBART-based machine translation (MT) model. Dhawan et al. (2026) show that synthetic parallel data and language-specific preprocessing improve low-resource Indigenous MT, including Es-Guaraní (Grn) translation. In the 2026 shared task, however, a standard neural machine translation pipeline proves insufficient. When we apply the existing Es-Grn translation model to Spanish intermediate captions produced from the dev images, the resulting captions are often fluent enough at the sentence level but do not match the target caption register. In other words, the performance bottleneck shifts from generic translation quality to domain adaptation and culturally grounded caption style.

To address this mismatch, we move from sequence-to-sequence translation toward retrieval-augmented in-context translation with Large Language Models (LLMs). The core idea is simple. Instead of relying on a single model fine-tuned on mixed-domain low-resource language corpora, we retrieve Es-low-resource (LoRes) examples that are similar to the current caption and provide them as in-context exemplars. This design lets the decoder

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 212–223

adapt its lexical choices and stylistic register at inference time. We evaluate multiple OpenAI and Gemini models, vary the number of retrieved training examples and development exemplars, and test prompt variants and retrieval heuristics.

Our experiments lead to three main findings. First, among OpenAI models, many-shot direct translation outperforms both our mBART baseline and LLM post-correction, but gains are modest and highly sensitive to prompt composition. Second, Gemini 2.5 Flash (Comanici et al., 2025) is dramatically stronger in Es-LoRes in-context translation than the GPT-family models. Third, development references are useful as in-context exemplars for matching the expected caption register, but they must be interpreted carefully because same-pool development prompting can inflate development-set scores.

Our primary contributions are as follows. We present a retrieval-augmented LLM caption translation pipeline for low-resource cultural image captioning. We document an extensive set of negative and positive ablations, including model substitutions, prompt revisions, and development-exemplar hyperparameter searches. Finally, we highlight evaluation design as a central methodological issue for dev-conditioned in-context learning in low-resource captioning.

2 Background

Machine translation for Indigenous languages of the Americas has advanced largely through the AmericasNLP shared tasks, which have established evaluation benchmarks and made parallel corpora available for many low-resource languages (Mager et al., 2021; Ebrahimi et al., 2023, 2024). Strong systems in these shared tasks have typically relied on multilingual pretrained models such as mBART, M2M-100, or NLLB-200, often combined with synthetic data generation, multilingual transfer, and additional corpus collection (Gow-Smith and Sánchez Villegas, 2023; Tonja et al., 2023; Costa-Jussà et al., 2022). A recurring pattern in this literature is that performance improvements come not only from larger models, but also from better alignment between model capacity, augmentation strategy, and the target domain.

Character-based evaluation metrics are also especially relevant in this area. chrF and chrF++ are widely used for morphologically rich languages because they are more robust than BLEU to inflectional variation and spelling differences (Popović, 2015, 2017). The AmericasNLP 2026 organizers likewise use chrF++ as the first-stage ranking metric for the shared task.

Finally, our work relates to the broader use of in-context learning for low-resource generation. Instead of relying exclusively on fixed model parameters after fine-tuning, retrieval-augmented prompting can adapt the decoder to the current example at inference time. In our setting, this is especially attractive because the target outputs are short and stylistically constrained: parallel demonstrations can serve not only as semantic guides, but also as direct evidence of the desired caption register.

3 Dataset and Methodology

3.1 Baseline System

We compare our retrieval-augmented LLM pipeline against the Qwen3VL-8B (Qwen Team, 2025) Captioning and Sheffield 2023 (Gow-Smith and Sánchez Villegas, 2023) MT method, which is the stated baseline for the current shared task on image captioning. We apply this baseline to our submission results for dev set captioning in each of the applicable target languages and report percent improvement over the baseline scores.

For MT in Guaraní captioning, we also compare against the mBART-based Es-Grn model from Dhawan et al. (2026), trained on curated and synthetic parallel data. This serves as the strongest conventional translation baseline in our pipeline and allows us to assess whether retrieval-augmented in-context generation improves over standard sequence-to-sequence translation. We evaluate several retrieval-augmented GPT-4-family (OpenAI et al., 2024) model variants, summarized in Table 2: GPT-4o-mini, GPT-4.1-mini, GPT-4.1, along with the competing Gemini 2.5 Flash.

Across these models, we vary the number of retrieved training examples r, development exemplars d, and prompting strategy.

3.2 Task Data

The official development set $\mathcal{D}$ (dev set) contains 50 examples, and the organizers release the data in JSONL format paired with images. The pilot set includes Spanish captions for reference, but the organizers explicitly note that these are pilot-only and will not be present in development or test.

Our final submission pipeline is two-stage. Stage 1 produces a Spanish caption from the image using

Qwen 2.5B (Bai et al., 2025b). Stage 2 utilizes Gemini 2.5 Flash (Comanici et al., 2025) for Es-LoRes translation.

3.3 Datasets by Language

The shared task covers five target languages: Guaraní, Yucatec Maya, Orizaba Nahuatl, Bribri, and Wixárika. As shown in Table 1, the available retrieval data $\mathcal{R}$ differs substantially across languages, which motivates treating the retrieval size r and the number of development exemplars d as inference-time hyperparameters that vary across target language submissions.

Guaraní We use the largest retrieval bank in our setup. The retrieval Es-Grn bank of 53,183 pairs contains AmericasNLP 2023 (Ebrahimi et al., 2023) training data augmented with synthetic examples from the MultiScript30k project (Driggers-Ellis et al., 2025). The Guaraní retrieval dataset is relatively well aligned with the captioning task, as it contains culturally specific terms, proper nouns, and short descriptive examples useful for visual caption generation.

Yucatec Maya We do not have a comparable parallel training corpus for retrieval. The development set $\mathcal{D}$ is the only retrieval source, so we rely on dev exemplars and the pretrained knowledge of the LLM and we fix $r = 0$ in all experiments.

Other Target Languages We use the available Es-LoRes parallel data as retrieval banks. Their retrieval banks vary in size and domain match: Orizaba Nahuatl has 16,145 pairs, Bribri has 7,508 pairs, and Wixárika has 8,966 pairs. The Wixárika retrieval data is notably less caption-like, since much of the available corpus is narrative or literary rather than visual-description oriented. These differences motivate the language-specific hyperparameter choices reported in Table 1.

3.4 Retrieval Bank and Prompt Construction

For each target language, we construct a language-specific Es-LoRes retrieval bank $\mathcal{R}$ from the available parallel data described above. The Spanish side of each retrieval bank is indexed with BM25 (Robertson and Zaragoza, 2009), a TF-IDF-style lexical retrieval method that ranks candidate examples by query-term overlap while accounting for term importance and document length normalization. At inference time, the Spanish caption generated in Stage 1 is used as a query q, and the top-r retrieved pairs are selected as $\mathcal{R}_r(q) \subset \mathcal{R}$.

In addition to retrieved training pairs, some configurations include development exemplars. Let $\mathcal{D}$ denote the shared-task development set for a target language, and let $\mathcal{D}_d \subset \mathcal{D}$ denote the d development examples included in the prompt. For a query caption q, the full prompt context is defined as

$$P(q) = \mathcal{D}_d \cup \mathcal{R}_r(q),$$

where $\mathcal{R}_r(q)$ provides retrieval-based semantic and lexical grounding, and $\mathcal{D}_d$ provides examples of the caption style expected by the task. The model then generates the target-language caption conditioned on $P(q)$.

We sweep r and d where applicable. The values selected for submissions are in Table 1, and detailed grid-search results appear in our appendix.

3.5 Prompting Strategy

The MT prompt structure is deliberately simple. The system prompt instructs the model to translate from Spanish into the target language, match the example style, stay concise, preserve culturally specific nouns when appropriate, and produce exactly one line. The user prompt contains two evidence blocks, development exemplars $\mathcal{D}_d$ and retrieved Spanish–target-language pairs $\mathcal{R}_r(q)$, followed by the current Spanish caption.

We use the same general prompt structure across target languages, adding language-specific modifications only when required. After several prompt-engineering attempts, we found that more aggressive instructions, such as suppressing generic lead-ins or forcing noun-first phrasing, reduced performance. The final prompts therefore keep the system instruction minimal and rely on in-context examples for lexical and stylistic guidance. We summarize the final and ablation prompt files in Table 5 of the appendix.

3.6 University of Florida Gators Submission

Our submission system features a retrieval-augmented long-context translator embedded in a two-stage image captioning pipeline. Figure 1 illustrates the full system, which is organized into five steps. Stage 1 corresponds to Step 1 in the diagram. A vision-language model generates one Spanish caption q for each target image. We use either Qwen2.5-VL-72B-Instruct in 4-bit precision or Qwen3-VL-8B (Bai et al., 2025a) for this stage.

The prompt is culturally aware; follows a noun-first style; and encourages concise descriptions of visible entities, objects, clothing, actions, and scene context. We treat this stage as fixed and do not optimize it extensively in this paper.

Stage 2 corresponds to Steps 2–5 in Figure 1. It transforms the generated Spanish caption q into the final target-language caption using retrieval-augmented many-shot MT. The q is used as a BM25 query over the Spanish side of the available Es-LoRes retrieval bank $\mathcal{R}$. In Step 3, the retrieved subset $\mathcal{R}_r(q)$ and development subset $\mathcal{D}_d$ are assembled into a many-shot prompt with an instruction to translate from Spanish to the target language while matching the example style. Each prompt contains approximately 3K–5K tokens. The r nearest Es-LoRes training pairs provide semantic and lexical grounding, while the d development pairs provide direct evidence of the target caption register. In Step 4 of Figure 1, Gemini 2.5 Flash performs the final target-language generation with temperature set to 0.0, thinking disabled, and a maximum output length of 120 tokens. We then strip prefixes in Step 5 and normalize whitespace to produce the final JSONL submission. The system is evaluated using chrF++, followed by human judgment for the top-ranked submissions.

4 Results

Table 1 gives our final submission's performance for each of the target languages in the shared task while Table 2 summarizes the main progression of Es-Grn dev set experiments from an mBART translation model to in-context MT with Gemini 2.5 Flash, which is competent in all target languages. Several trends emerge immediately. First, direct many-shot translation is better than post-correction, confirming that it is more effective to generate the caption in one step than to repair the mBART output after the fact. Second, among the OpenAI models we tested, GPT-4o-mini remains the strongest, but the gains over the mBART baseline are modest. Third, Gemini 2.5 Flash yields much larger improvements, even before adding any development exemplars.

In our final dev set results, the Guaraní target achieves the highest absolute chrF++ score among the five target languages by at least 10 chrF++ and more than doubles the Bribri and Wixárika target performances. For Guaraní we achieve a remarkable 131.7% improvement over the baseline

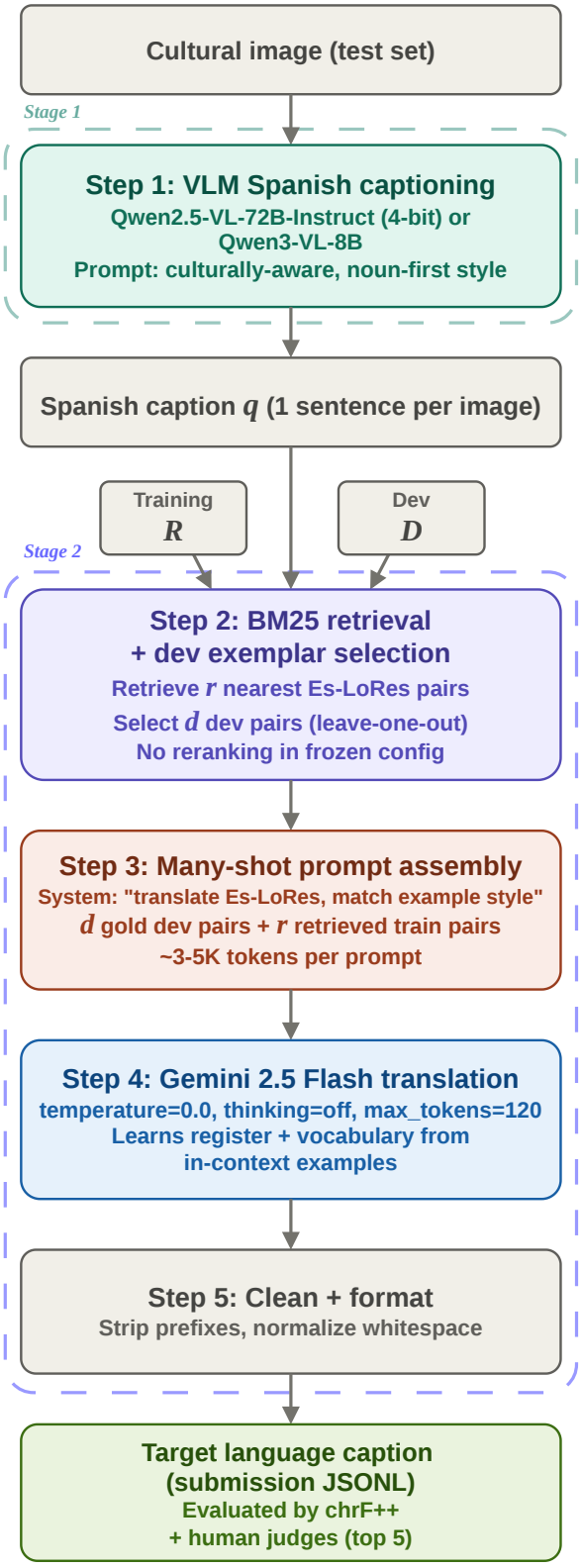

Figure 1: Overview of the proposed two-stage image captioning pipeline. Here, $\mathcal{R}$ denotes the full Es-LoRes retrieval bank, $\mathcal{R}_r(q) \subset \mathcal{R}$ denotes the r nearest Es-LoRes training pairs retrieved for Spanish caption q, $\mathcal{D}$ denotes the full development set, and $\mathcal{D}_d \subset \mathcal{D}$ denotes the d gold dev pairs used in the many-shot prompt.

Configuration					test		dev	
Language	r	d	Test Examples	Retrieval Pairs (R)	Test chrF++	% Imp.	Dev chrF++	% Imp.
Bribri	80	20	267	7,508	17.90	155.2%	19.99	**164.1%**
Guaraní	80	49	101	53,183*	23.10	14.72%	**48.24**	131.7%
Orizaba Nahuatl	40	20	200	16,145	**25.42**	**166.9%**	25.67	122.6%
Wixárika	40	20	201	8,966	17.58	3.952%	18.99	6.866%
Yucatec Maya	–	49	212	0†	21.11	–	26.29	–

Table 1: Submission chrF++ results across languages for the proposed system. Here, r is the number of retrieved training pairs included in each prompt, d is the number of development exemplars included in each prompt, and R denotes the total size of the available retrieval bank for each language. Baseline results are those provided for the shared task (Gow-Smith and Sánchez Villegas, 2023). **Bold** and underlined entries in the right-side columns indicate the best and second-best results, respectively. * Includes parallel Es-Grn MultiScript30k (Driggers-Ellis et al., 2025) synthetic exemplars. † No external retrieval bank is used; only development exemplars are included.

System	r	d	Notes	Setting type	chrF++
mBART Baseline	–	–	Inherited MT Baseline	Baseline	22.40
GPT-4o-mini Post-Correction	–	8	Revise mBART Draft	Post-Editing	19.81
GPT-4o-mini	24	4	Direct Many-Shot Translation	Direct Prompting	23.41
GPT-4o-mini	28	4	Best OpenAI Setting	Direct Prompting	23.60
GPT-4.1-mini	28	4	Model Swap	Direct Prompting	23.23
GPT-4.1	28	4	Model Swap	Direct Prompting	23.49
Gemini 2.5 Flash	28	0	No dev Exemplars	Train-Only Prompt	32.14
Gemini 2.5 Flash	28	4	Same-Pool dev Prompting	Dev-Assisted	39.25
Gemini 2.5 Flash	28	20	Same-Pool dev Prompting	Dev-Assisted	42.90
Gemini 2.5 Flash	28	49	Same-Pool dev Prompting	Dev-Assisted	42.19
Gemini 2.5 Flash	80	49	Same-Pool dev Prompting	Dev-Assisted	**48.24**

Table 2: Dev chrF++ Guaraní captioning results across baseline machine translation and RAG-based prompting configurations. **Bold** indicates the best result, and Underline indicates the second-best result.

method (Gow-Smith and Sánchez Villegas, 2023). In Section 5, we investigate the effect that synthetic exemplars may have had on Guaraní target language performance versus the other languages for which no MultiScript30k data exists. However, as the final column in Table 1 attests, each language target outperforms the baseline method whenever the baseline is available. In particular, for the Bribri and Orizaba Nahuatl targets, respectively, we achieve 164.1% and 122.6% improvements over the baseline. As we highlight in Table 1, our improvement for Bribri is the greatest relative improvements over the baseline method for any target language.

The test results are slightly different. Bribri is replaced by Orizaba Nahuatl as most improved language. The languages achieve 155.2% and 166.9% improvement over the testing baseline, respectively (Bui et al., 2026). Guaraní performance drops by more than half in the test evaluation and most of the dev set performance gain is not reproduced in the official evaluation as a result. Wixárika performance improvement falls by a similar amount.

5 Ablations

In addition to the results we list in the previous section, we provide a number of ablations to demonstrate the superiority of the final submission over numerous alternative approaches to various facets of the captioning pipeline. In the experiments reported here, we hold the Spanish captions fixed and optimize only the Guaraní translation stage. This lets us analyze ablation effects independent of the image captioner and makes the ablations directly comparable to our dev set results. We focus on the dev set evaluation here because ablations are all performed on the dev set before the release of final shared task results.

5.1 Machine Translation Architecture

We allude in the first section to how the initial approach for the machine translation step in Guaraní utilizes an mBART based MT model (Dhawan et al., 2026). Table 2 shows Guaraní captioning performance across several MT architectures, including mBART, GPT-4o, and Gemini 2.5 Flash. Where appropriate, we adopt the retrieval-augmented approach in our final submission, and

we vary the values of r and d within architectures.

The results show that the best configuration for Guaraní translation is the Gemini 2.5 Flash LLM prompted with $r = 80$ and $d = 49$ exemplars. We traverse much of the r, d search grid with OpenAI models. The best OpenAI setting used GPT-4o-mini with 28 retrieved examples and 4 development exemplars, reaching 23.60 chrF++. This performance is reflected in Table 2. Removing or increasing development exemplars reduced performance. GPT-4.1-mini and GPT-4.1 were both competitive but did not surpass GPT-4o-mini. This pattern suggests that, for the OpenAI models we tested, the in-context regime has a narrow optimum. Too little context leaves the model underconstrained; too much context appears to add noise or dilute the style signal.

For its superior performance over GPT models and its positive receptivity to context, we adopt Gemini 2.5 Flash for Es-LoRes MT in our final submission.

5.2 Hyperparameter Search

For each language, we sweep values of r and d that we list in Section 3. Table 1 of final submission results shows performance at the submittal configuration featuring its r, d pair. To be thorough, Table 6 reports the results of a partial grid search for each target language in our appendix.

Though this ablation shows that the r and d exemplar counts from Table 1 are optimal within our search grid, there is no accounting for values outside of it. Additionally, performance changes as one scans the grid indicate differing impacts of the r and d hyperparameters for different languages and data sources. Thus, we stress in this ablation the importance of a thorough search for the optimal r, d pair for any new target language or data configuration in retrieval-augmented Es-LoRes MT.

5.3 Synthetic Exemplars

We notice in Section 4 that our pipeline performs Guaraní image captioning much more effectively in absolute terms versus any of the other target languages, even though we achieve similar percent improvement over the applicable baseline for the Bribri and Orizaba Nahuatl targets. We also note that Guaraní is the only target language for which we include synthetic exemplars from MultiScript30k. Testing for the effect of this synthetic data, we ablate the original Guaraní captioning submission by only using AmericasNLP 2023 (Ebrahimi et al., 2023) training data for retrieval exemplars (r). Table 3 compares performance with and without MultiScript30k (Driggers-Ellis et al., 2025) synthetic exemplars for three r, d pairs and provides the greatest performance overall without synthetic exemplars.

Data	r	d	Ret. Pairs	chrF++
ANLP2023 + MS30k	40	49	53,183	51.34
ANLP2023 + MS30k	40	20	53,183	48.38
ANLP2023 + MS30k	80	49	53,183	48.24
ANLP2023	80	0	26,032	22.65
ANLP2023	40	49	26,032	21.03
ANLP2023	40	20	26,032	20.50
ANLP2023	80	49	26,032	20.75

Table 3: Dev chrF++ results for Guaraní captioning with differing retrieval exemplar sets. Data includes AmericasNLP2023 (ANLP2023) training data (Ebrahimi et al., 2023) and/or MultiScript30k (MS30k) (Driggers-Ellis et al., 2025) synthetic exemplars as noted in the column *Data*.

The results are clear. Controlling for our retrieval hyperparameters by fixing r and d to three pairs, we observe that the original configuration with both genuine AmericasNLP 2023 training data and synthetic MultiScript30k exemplars outperforms the ablation with AmericasNLP 2023 alone by more than 100% relative improvement in chrF++ in each case. Additionally, the best Guaraní captioning performance without synthetic exemplars is 55.9% less than the best performance with them. These results mirror the comparison of our submission's Guaraní performance to the other target languages in the shared task and attribute much of our improvement in Guaraní captioning to synthetic exemplars in the retrieval-augmented Es-Grn translation step.

5.4 Alternative Prompting and Reranking

We also test several prompting ablations that looked reasonable but are consistently negative in their impact on performance, with the exception of a specific strategy we give additional attention to in Section 5.5. In this section, we quickly summarize other alternative prompts, and in Table 5, we produce all of the relevant prompts for completeness.

A prompt rewrite aimed at suppressing generic scene-introduction phrases reduces performance substantially, and a retrieval reranker that attempts to prefer short caption-like pairs also reduces performance, both for OpenAI and Gemini. Likewise, using 49 same-pool development exemplars with

GPT-4o-mini degraded performance instead of improving it. These results matter because they show that this task does not respond well to aggressive heuristic control. The most effective systems are built by keeping the instruction stable and varying only model choice and amount of context supplied.

For Wixárika in particular, we devise special prompts with cultural context in the form of a glossary. These glossaries contain the names of common objects from the Wixárika culture and their definitions in Spanish, with the hope that this additional context will help the LLM MT architectures translate intermediate Spanish captions into the Wixárika target language. We utilize two versions of the cultural glossary, but the results do not improve for either one. We include the additional prompts in Table 5 to document this alternative prompting.

5.5 Morphological Considerations for Bribri

We ablate our submission for the Bribri (Bzd) language by considering performance with and without morphological prompting and post-processing for the Es-Bzd MT. Bribri has complex tonal marking in orthography (circumflexes, underlines, multiple diacritics) which makes the deduplication logic harder (Coto-Solano, 2021). The tokenizer splits differently on tonal characters.

In our final submission, we utilize a complex postprocessing and additional prompting for the Gemini MT model to improve performance. Table 4 shows the evolution of our strategies to account for the morphological complexity of the Bribri language. First, careful examination of our pipeline's output reveals that Bribri captions are initially in NFC encoding, which combines base letter and diacritic encodings into one character. This does not match the dev set examples, which use NFD encoding. For instance, our model outputs *ë* as a single unit, but the dev set's NFD encoding would separate the letter *e* from the umlaut above it. Because chrF++ is a character-level lexical metric, this mismatch depresses the score. We therefore perform NFD-Normalization (NFD-Norm.) to account for the difference. Secondly, we believe the morphological considerations of Es-Bzd translation significant enough that they deserve special prompting. To this end, we formulate a Morphological Prompt (Morph.) for Bribri that includes special instructions about Subject-Object-Verb (SOV) word order, tonal diacritics, common consonant clusters, verb-final clauses, and possessive noun prefixes.

Language	Prompting	Postprocessing	chrF++
Bribri	Standard	Standard	11.50
Bribri	Standard	NFD-Norm.	17.02
Bribri	Morph.	NFD-Norm.	19.99

Table 4: Morphological ablations for our submission in the Bribri captioning task. We either utilize NFD-Normalization (NFD-Norm.) as a postprocessing step, both NFD-Norm. and Morphological Prompting (Morph.), or we apply the same prompting and postprocessing as other target languages (Standard). Rightmost column shows Dev ChrF++.

As in previous ablations, the results in Table 4 largely speak for themselves. NFD-normalization overcomes the encoding mismatch, and morphological prompting provides additional improvements.

6 Discussion

We now proceed to a discussion of our method and the results achieved. We analyze the available data and investigate the potential sources of performance improvement over the baseline method.

6.1 In-Context Retrieval

Though limited to the Guaraní target, we show in our ablation of MT architectures that long-context retrieval-based MT using state-of-the-art LLMs greatly improves image captioning performance at the MT stage versus the dedicated mBART model (Dhawan et al., 2026). While we cannot say for certain whether this relationship holds for other language targets, from comparisons to baseline performances, it appears that additional context from real and synthetic exemplars significantly boosts LLM translation in Es-LoRes tasks.

6.2 Synthetic Exemplars

For the Guaraní language, we include approximately 30k synthetic exemplars for in-context retrieval from the MultiScript30k (Driggers-Ellis et al., 2025) dataset. We observe in Section 4 that Guaraní achieves substantially higher absolute dev chrF++ than the other target languages.

Remembering that synthetic exemplars apply only to Guaraní and its significantly positive effect on Es-Grn MT in related work (Dhawan et al., 2026), we ablate for the synthetic exemplars' effect on Guaraní captioning.

For Guaraní captioning without the MultiScript30k synthetic retrieval pairs, the results we elaborate in Section 5.3 and Table 3 show a comparison similar to Guaraní captioning versus the other target languages. Results improve over 100%, approximately 28 total chrF++ or more, for Guaraní captioning with the synthetic exemplars for the same r, d pair. These results also mirror observations from (Dhawan et al., 2026) on the effect of synthetic exemplars on Es-Grn MT. We therefore conclude that synthetic exemplars drive Guaraní performance gains and speculate that synthetic retrieval pairs may further improve captioning performance for other target languages.

6.3 Morphology

Despite favorable results in the Guaraní target, we acknowledge in Section 4 that our greatest relative improvement is for the Bribri language. The ablation in Section 5.5 clarifies that much of this performance improvement stems from the morphological considerations we take for the Bribri language target via specialized prompting. We observe that explicitly prompting for Bribri's morphological features accounts for over 10% of the performance gain in our final submission over the shared task's baseline (Gow-Smith and Sánchez Villegas, 2023). We conclude morphological prompting has potential for LoRes MT in morphologically complex languages.

7 Future Work

The most impactful next step is improving the visual captioning stage. Error analysis indicates that approximately 54% of remaining Guaraní errors originate in the vision model rather than the translator. A stronger VLM will likely yield larger gains than further translation tuning. Beyond captioning, we may extend retrieval to incorporate visual similarity rather than Spanish text overlap alone, which would help when the intermediate caption is noisy or culturally ambiguous. The lowest-resource languages (Wixárika and Bribri languages) are bottlenecked by corpus domain mismatch rather than model choice. Even small caption-style parallel data corpora for these languages would likely produce larger gains than any prompting improvement.

8 Conclusion

We present the University of Florida Gators team system for the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages (Bui et al., 2026). Our two-stage pipeline, VLM captioning in Spanish followed by retrieval-augmented many-shot translation with Gemini 2.5 Flash, substantially outperforms fine-tuned baseline models across all five target languages, culminating at 48.24 chrF++ for Guaraní in the submission. We find that retrieval behavior is highly language-dependent. Large retrieval windows help Guaraní but hurt Yucatec Maya, where Gemini's pre-training knowledge is sufficient and BM25 retrieval adds noise. We also find that development exemplars are useful for matching caption register, but their use requires careful interpretation because they can inflate development-set scores when drawn from the same evaluation pool. For the lowest-resource languages, the ceiling is domain mismatch, not model capacity.

Limitations

Our system is a cascade: errors in the Spanish captions propagate into translation with no recovery mechanism. Because most ablations hold the Spanish captions fixed and vary only the translation stage, they likely underestimate the contribution of the visual captioning model to final performance.

A second limitation is the use of development examples as in-context exemplars. These examples are useful for matching the expected caption register, especially when little caption-style target-language data is available, but they can also inflate development-set scores when exemplar selection and evaluation draw from the same small pool. We therefore use dev-assisted results primarily primarily model-selection and submission-configuration. A more rigorous and comprehensive evaluation would report held-out or cross-split dev results.

Finally, evaluation relies primarily on chrF++, which is useful for low-resource and morphologically rich languages but cannot fully capture fluency, cultural appropriateness, or naturalness for native speakers. Although the shared task includes human evaluation for finalist systems, our ablations include no native-speaker evaluation.

Acknowledgments

The authors gratefully acknowledge Arnold and Lisa Goldberg, whose financial support helped to make this work possible. We also acknowledge the Gatorade, whose support provided the compute resources necessary for this research.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, Silvia Fernandez Sabido, Luis Samuel Santiago Melchor, Sotero Silverio, Robert Pugh, Raúl Vázquez, John E. Ortega, Arturo Oncevay, Rubén Manrique, Luis Chiruzzo, Rolando Coto-Solano, Elisabeth Mager, Shruti Rijhwani, David Ifeoluwa Adelani, Manuel Mager, and Katharina von der Wense. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Aashish Dhawan, Christopher Driggers-Ellis, Christan Grant, and Daisy Zhe Wang. 2026. Improving indigenous language machine translation with synthetic data and language-specific preprocessing. In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 119–126, Rabat, Morocco. Association for Computational Linguistics.

Christopher Driggers-Ellis, Detravious Brinkley, Ray Chen, Aashish Dhawan, Daisy Zhe Wang, and Christan Grant. 2025. Multiscript30k: Leveraging multilingual embeddings to extend cross script parallel data.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi et al. 2024. Findings of the americasnlp 2024 shared task on machine translation into indigenous languages. In *Proceedings of the AmericasNLP Workshop*.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the americasnlp shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Manuel Mager, Arturo Oncevay, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the AmericasNLP Workshop*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, Alexander Gelbukh, and Jugal Kalita. 2023. Enhancing translation for indigenous languages: Experiments with multilingual models. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Appendix

Here, we provide additional data for validation of our method versus various ablations with particular focus on alternative prompts for various target languages and the r, d grid search. In the following sections, we give tables of alternative prompts and performance at different r, d pairs for each target language.

Alternative Prompts

Table 5 yields the prompts that we utilize in our ablations and the prompts present in our final submission for each target language as indicated in the *Standing* column. The *Language(s)* column shows which languages the prompt applies to.

r, d Hyperparameter Search

We frequently refer to a search for optimal r and d retrieval hyperparameters in the main body of this paper but reserve detailed communication of the sweep for this appendix due to the number of configurations we consider. For the Gemini 2.5 Flash MT architecture and for each target language, we partially sweep a grid consisting of r, d pairs such that $r \in \{0, 10, 20, 40, 80\}$ and $d \in \{0, 10, 20, 30, 40, 49\}$. Table 6 reports the chrF++ scores for the final captioning pipeline for each combination of r and d tested for each target language.

Short Desc.	Description	Prompt URL(s)	Standing	Language(s)
Captioning	Prompts for VLM in first stage of captioning pipeline.	v1_submission/guarani_caption_prompt.txt	**Final**	All
Many-Shot	Includes r and d in-context exemplars from previous parallel text corpora and from the dev set, respectively.	v1_submission/<target>_system_prompt.txt	**Final**	All
Morph. Bribri	Accounts for Bribri target's morphological complexity. Portion shown appended to general prompt.	v1_submission/bribri_system_prompt.txt	**Final**	Bribri
Wixárika Gloss. *v2*	Provides definitions for culturally situated nouns in the Wixárika language.	wixarika/caption_es_wixarika_v2.txt	Ablation	Wixárika
Wixárika Gloss. *v3*	Provides definitions for culturally situated nouns in the Wixárika language.	wixarika/caption_es_wixarika_v3.txt	Ablation	Wixárika

Table 5: Prompts we utilize in our ablations and final submission for each of the target languages. The *Standing* column indicates whether a prompt is an Ablation or part of our **Final** submission. The final *Language(s)* column indicates what languages the prompt applies to.

Bribri				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Bribri	80	20	Uses Morphological Bribri prompting and NFD-normalization.	**19.99**
Bribri	80	20	Uses Regular Prompting.	11.50
Bribri	40	20	...	11.41
Bribri	10	10	...	11.41
Bribri	20	20	...	11.16
Bribri	20	10	...	10.95
Bribri	40	10	...	10.95
Bribri	10	10	...	10.63
Bribri	80	10	...	10.17
Bribri	80	0	...	6.53
Bribri	40	0	...	5.93
Bribri	20	0	...	5.30
Bribri	10	0	...	4.75
Guaraní				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Guaraní	40	49	Includes synthetic exemplars. Not submitted because ablation was incomplete at submission deadline.	51.34
Guaraní	40	20	...	48.38
Guaraní	80	49	Includes synthetic exemplars.	**48.24**
Guaraní	80	20	...	42.61
Guaraní	0	49	...	20.80
Orizaba Nahuatl				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Orizaba Nahuatl	40	20	–	**25.67**
Orizaba Nahuatl	40	10	–	25.59
Orizaba Nahuatl	80	20	–	25.25
Orizaba Nahuatl	80	10	–	25.16
Orizaba Nahuatl	20	20	–	24.16
Orizaba Nahuatl	20	10	–	23.91
Orizaba Nahuatl	40	0	–	16.56
Orizaba Nahuatl	80	0	–	16.37
Orizaba Nahuatl	20	0	–	15.61
Wixárika				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Wixárika	40	20	–	**18.99**
Wixárika	40	10	–	17.74
Wixárika	20	20	–	17.56
Wixárika	80	10	–	17.48
Wixárika	80	20	–	17.13
Wixárika	20	10	–	16.81
Wixárika	40	0	–	16.59
Wixárika	80	0	–	16.25
Wixárika	20	0	–	13.80
Yucatec Maya				
Language	Retrieval (r)	Dev Exemplars (d)	Notes	chrF++
Yucatec Maya	–	49	Fixes $r = 0$ for lack of retrieval exemplars.	**26.29**
Yucatec Maya	–	20	...	26.29
Yucatec Maya	–	40	...	25.07
Yucatec Maya	–	30	...	25.05
Yucatec Maya	–	0	Fixes $r = 0$ for lack of retrieval exemplars. At $d = 0$, the model receives no signal from the target language.	20.25

Table 6: Dev chrF++ results across languages for the proposed system. (...) Indicates previous *Notes* column entry applies. (–) Indicates no *Notes*. **Bold** entries *ChrF++* column are submission scores for each target language.

From Machine Translation to Image Captioning: Training Vision-Language Models for Indigenous Languages of the Americas

Luis Lara[1], **Param Raval**[1]
[1]Mila – Quebec AI Institute
luis.lara@mila.quebec

Abstract

We describe our system for the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages of the Americas. Our post-training pipeline starts from Aya Vision 32B: the vision-language model is first fine-tuned on machine translation data from prior AmericasNLP shared tasks and then further fine-tuned on the cultural Image Captioning data. This approach uses translation as an intermediate training task, while the final system produces captions directly in the requested Indigenous language rather than translating a Spanish caption afterward. Our experiments show that machine translation fine-tuning is an important initialization step. The resulting fine-tuned vision-language model also shows translation capabilities for the languages considered in this work. In addition, our zero-shot GPT-5.5 submission ranks first in the Maya language track under the official human-evaluation stage. [1]

1 Introduction

The AmericasNLP 2026 shared task studies cultural Image Captioning for Indigenous languages of the Americas (Bui et al., 2026). Given an image and a requested language, a system must produce a caption that is both visually grounded and appropriate for the linguistic and cultural setting. This is challenging because many of these languages have little or no representation in the data used to pretrain current large language models (LLMs) and vision-language models (VLMs). It is also challenging because the shared task provides only a small set of image-caption pairs, so caption-only fine-tuning has relatively few examples to learn from.

At the same time, the AmericasNLP community has built machine translation resources for several Indigenous languages of the Americas across previous shared tasks (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024; De Gibert et al., 2025). These resources do not teach a model to ground language in images, but they do provide substantially more examples of how to generate text in the relevant languages. This suggests a transfer strategy: before asking a VLM to caption images, first fine-tune it on bilingual translation data so it learns to produce text in the relevant Indigenous languages.

We implement this strategy with a multi-stage supervised fine-tuning pipeline. We use Aya Vision 32B, a VLM, as the backbone (Dash et al., 2025). We first perform supervised fine-tuning for machine translation, SFT (MT), on Spanish–Indigenous bilingual text. We then further fine-tune the same model on the cultural Image Captioning (IC) data, SFT (IC), so that the final system generates captions directly in the requested Indigenous language instead of first producing a Spanish caption and translating it afterward. All stages use Low-Rank Adaptation (LoRA) (Hu et al., 2022). We also test reinforcement learning with verifiable rewards for machine translation, RLVR (MT), after SFT (MT), but find that it provides smaller and less consistent gains than the supervised transfer step.

Our experiments support the main motivation for the pipeline: caption fine-tuning works better when it starts from the translation-tuned model. In addition, the same fine-tuned VLM can produce useful translation candidates for these languages under reference-based oracle candidate selection. We also evaluate a zero-shot direct Image Captioning submission with GPT-5.5 (OpenAI, 2026) as a frontier-model comparison, which ranks first in the Maya track under the official human-evaluation stage. Figure 1 summarizes our two submission routes: the fine-tuned Aya Vision pipeline and the separate GPT-5.5 zero-shot direct Image Captioning submission.

[1]Project code is available at https://github.com/ludolara/americasnlp2026

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 224–235

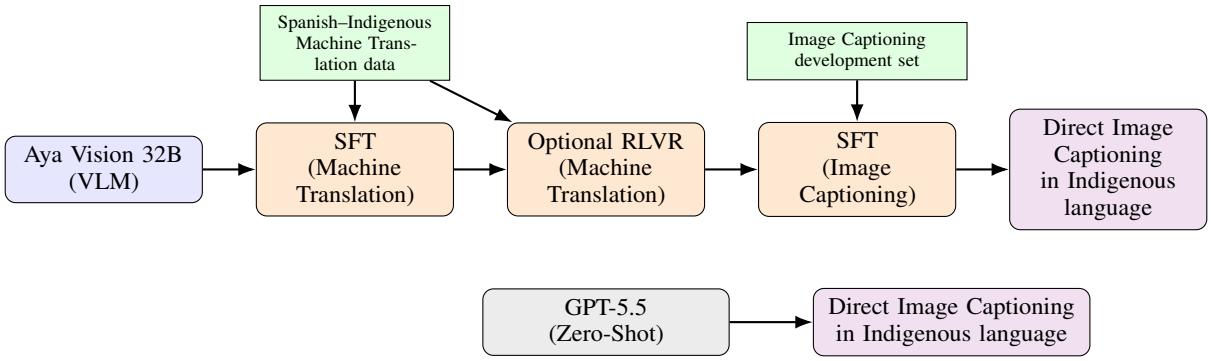

Figure 1: Overview of our Image Captioning systems. The main pipeline adapts Aya Vision 32B through supervised fine-tuning on Spanish–Indigenous machine translation data, optional RLVR on the same task, and supervised fine-tuning on the Image Captioning development set. The lower branch shows our separate GPT-5.5 zero-shot submission. Both systems generate captions directly in the target Indigenous language.

Our contributions are as follows:

1. We present a multi-stage post-training pipeline for cultural Image Captioning in Indigenous languages of the Americas, using machine translation fine-tuning as an intermediate training stage before image-captioning fine-tuning.

2. We show that the same translation-tuned VLM can produce useful translation candidates for the overlapping Indigenous languages under reference-based oracle candidate selection.

3. We include a separate zero-shot GPT-5.5 direct Image Captioning submission, which reaches the strongest human-evaluation result for the Maya track under the official shared-task evaluation.

2 Related Work

AmericasNLP shared tasks. Across the 2021, 2022, 2023, 2024, and 2025 editions, the AmericasNLP shared tasks (STs) evolve from an open machine translation benchmark into a broader suite of tasks for Indigenous languages of the Americas (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024; De Gibert et al., 2025). The 2021 ST established the benchmark with two tracks, ten languages, official baselines, and large improvements over baseline for many languages (Mager et al., 2021). The 2022 competition expanded the scope to speech by adding automatic speech recognition and speech-to-text translation alongside text-based machine translation, covering Bribri, Guaraní, Kotiria, Wa'ikhana, and Quechua (Ebrahimi et al., 2022). The 2023 edition returned the focus to machine translation but expanded the benchmark to eleven language pairs, added a new Chatino–Spanish evaluation set from the legal domain, and complemented automatic ranking with human evaluation (Ebrahimi et al., 2023). The 2024 findings emphasize a harder competitive setting: organizers released strong Sheffield and Helsinki baselines from 2023 and a repository of prior shared task training data, yet improvements over the best baseline were observed for only a subset of languages (Ebrahimi et al., 2024). The 2025 STs broadened the scope beyond machine translation to include educational-material creation and translation metrics, showing a shift from pure text-to-text translation toward a wider range of community-relevant language technologies (De Gibert et al., 2025).

This trajectory is directly relevant to our setting. First, the findings papers repeatedly show that progress depends on assembling, reusing, and extending scarce bilingual and speech resources (Mager et al., 2021; Ebrahimi et al., 2022, 2023, 2024). Second, they show that stronger baselines and human evaluation matter, because automatic gains alone do not fully characterize quality for low-resource Indigenous-language generation (Ebrahimi et al., 2023, 2024; De Gibert et al., 2025). The AmericasNLP 2026 shared task on cultural image captioning extends this line of work from text-to-text translation to visually grounded generation.

LLMs for Indigenous-language translation. Results from the AmericasNLP 2025 ST system papers do not provide strong evidence that off-the-shelf LLM-based methods are strong direct generators in the Spanish-to-Indigenous direction. Yahan and Islam (2025) report that NLLB-200

Table 1: Image Captioning test chrF++ results using the automatic evaluation metric. Maya was included only in v3, which corresponds to the GPT-5.5 zero-shot submission and was not produced by our fine-tuned image-captioning method. Versions v0 and v2 use the same training recipe, but correspond to different checkpoints.

Version	Submission	Overall	Bribri	Guaraní	Nahuatl	Wixárika	Maya
v0	SFT (MT) + SFT (IC)	17.398	**11.728**	19.634	19.421	18.811	–
v2	SFT (MT) + SFT (IC)	**17.599**	11.309	19.417	**20.655**	**19.013**	–
v1	SFT (MT) + RLVR (MT) + SFT (IC)	17.341	10.886	**19.772**	19.853	18.853	–
v3	GPT-5.5 (Zero-Shot)	12.812	4.555	12.804	19.725	10.984	**15.993**

outperforms LLaMA 3.1 and XGLM in Track 1, while Hus et al. (2025) use an LLM as a post-correction component and report stronger gains in the Indigenous-to-Spanish direction. Together with the official findings, these results motivate our setting: rather than treating a large generative model as a strong zero-shot translator, we study whether targeted supervised fine-tuning can make it a better generator for low-resource Indigenous languages (De Gibert et al., 2025) without intermediate Spanish translations.

Intermediate-task fine-tuning from translation to Image Captioning. Transfer learning in NLP commonly uses knowledge from one task, domain, or dataset to improve adaptation to another setting (Ruder et al., 2019). A more specific version of this idea is intermediate-task fine-tuning, where a pretrained model is first fine-tuned on an additional supervised task before being adapted to the final target task (Phang et al., 2018). We use this framework as methodological motivation rather than as direct evidence for our specific task combination. In our setting, machine translation serves as the intermediate task: it exposes the model to substantially more Spanish–Indigenous text and target-language generation examples before the model is fine-tuned on the much smaller Image Captioning data. This stage does not teach visual grounding. Instead, it is intended to improve the model's ability to generate text in the target Indigenous languages, while the subsequent Image Captioning fine-tuning stage teaches the model to connect visual inputs with concise target-language descriptions.

3 Shared Task and Data

3.1 Shared Task and Baseline

The shared task requires generating one caption for each input image in the requested Indigenous language. We treat this as direct image-to-Indigenous-text generation: the model is trained to produce the final caption itself, rather than first generating a Spanish caption and translating it afterward. Compared with generic image captioning, this task combines visual grounding with low-resource language generation. Further details on the shared task evaluation protocol are provided in the official findings paper (Bui et al., 2026).

The baseline for the ST follows the *generate-then-translate* pipeline where a caption is generated in Spanish using the Qwen3-VL-8B-Instruct VLM and then translated into the target language using a NLLB-200 model trained for that language. The latter stage uses the approach proposed by the winner of the AmericasNLP 2023 ST on MT (Gow-Smith and Sánchez Villegas, 2023).

3.2 Image Captioning Data

The development set consists of 250 labeled images, with 50 examples for each of Bribri, Guaraní, Yucatec Maya, Nahuatl, and Wixárika. We refer to Yucatec Maya as Maya in the result tables for compactness. We also use Nahuatl as a compact label for the shared-task Nahuatl track; where relevant, Appendix C preserves the Orizaba Nahuatl label used by the example IDs. Teams were allowed to train with the development set. The test set used to rank the submissions of the ST contains 267 Bribri images, 101 Guaraní images, 212 Yucatec Maya images, 200 Nahuatl images, and 201 Wixárika images.

3.3 Machine Translation Data

We restrict the translation data to the four overlapping languages: Bribri (`bzd`), Guaraní (`grn`), Nahuatl (`nah`), and Wixárika (`hch`). For Nahuatl, we use the available data from previous AmericasNLP shared tasks. We note that this data is not specifically labeled as *Orizaba Nahuatl*. Each training example contains Spanish text, Indigenous-language text, a language name, and a language code. The dataset is prepared by labeling these pairs bidirectionally so the model sees both Spanish-to-Indigenous and Indigenous-to-Spanish translation prompts.

The primary resources we use include Axolotl

Table 2: Per-language chrF++ on the internal development subset for Image Captioning systems.

System	Overall	Bribri	Guaraní	Nahuatl	Wixárika
None	6.053	2.077	7.688	6.836	6.687
SFT (IC)	14.589	11.570	15.714	15.444	14.848
SFT (MT) + SFT (IC)	19.313	11.457	22.143	**17.149**	**22.602**
SFT (MT) + RLVR (MT) + SFT (IC)	**19.805**	**11.508**	**23.767**	16.982	22.496

for Spanish–Nahuatl (Gutierrez-Vasques et al., 2016), a Wixárika resource derived from work on morphological segmentation (Mager et al., 2018), a Guaraní–Spanish parallel corpus (Chiruzzo et al., 2020), and a Bribri back-translation resource (Feldman and Coto-Solano, 2020). We also use additional data from the AmericasNLP 2025 ST1 language pairs, following the shared-task setup described by De Gibert et al. (2025); these additions are linked to Helsinki-NLP's earlier shared-task work (De Gibert et al., 2023).

4 System Overview

Our system is based on Aya Vision 32B (Dash et al., 2025), an open-weight multilingual multimodal model designed for image understanding, image captioning, visual question answering, text generation, and translation across 23 languages. Its open weights make it suitable for efficient fine-tuning on vision-language alignment tasks.

We use the same base model for translation and image captioning. The complete pipeline has three possible stages: SFT (MT), supervised fine-tuning for machine translation; optional RLVR (MT), reinforcement learning with verifiable rewards for machine translation; and SFT (IC), supervised fine-tuning for image captioning. The strongest and simplest configuration in our development experiments is SFT (MT) followed by SFT (IC).

4.1 Supervised Fine-Tuning (Machine Translation)

The first stage uses supervised fine-tuning on bilingual text with Aya Vision 32B. Each example is formatted as an instruction-style chat prompt, with the target sentence used as the assistant response. Training is bidirectional, so the model learns both Spanish-to-Indigenous and Indigenous-to-Spanish generation. The machine translation prompt template is listed in Appendix B.

4.2 Supervised Fine-Tuning (Image Captioning)

The second stage fine-tunes the model using the image captioning development set. Each sample provides an image and a target language, and the model learns to produce a caption directly in that language. Our approach skips intermediate translation and generates the final caption directly in the requested Indigenous language.

The image captioning prompt is intentionally simple. We avoid long explanations and multi-step instructions to minimize prompt engineering. The Spanish prompt asks for a single culturally appropriate caption in the target language and instructs the model not to include explanations. The image captioning prompt template is listed in Appendix B.

4.3 Reinforcement Learning with Verifiable Rewards (Machine Translation)

We also evaluate RLVR (MT) after SFT (MT). In this stage, we use the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) for optimization. For the automated candidate-selection procedure used in these experiments, we use sentence-level chrF++ to select high-scoring translations from multiple sampled candidates. In our development experiments, this stage gives a small overall improvement but the gains are not consistent across languages. We therefore treat RLVR (MT) as an exploratory ablation rather than as the central contribution of the system.

4.4 Zero-Shot Direct Image Captioning with GPT-5.5

In addition to the fine-tuned Aya Vision systems, we submitted a zero-shot direct Image Captioning run with GPT-5.5 (OpenAI, 2026). For each example, we provided the image and requested target language, and asked the model to generate the final caption directly in that language. This run does not use our machine-translation fine-tuning stage, RLVR, or image captioning fine-tuning, and it does not follow a generate-then-translate pipeline. We include it as an exploratory frontier-VLM com-

Table 3: Official human-evaluation results for our submissions. Mean Rating is the official average human-evaluation score reported by the organizers, based on 1–5 human ratings where higher is better. When multiple annotators rated the same example, their scores were averaged per example before computing the final mean. N Ratings and N Images reproduce the counts reported in the official results. Points follow the shared-task overall ranking rule: 5 points for first place, 4 for second, 3 for third, 2 for fourth, and 1 for fifth.

Language	Version	Submission	Rank	N Ratings	N Images	Mean Rating
Bribri	v0	SFT (MT) + SFT (IC)	4	320	267	1.994
Guaraní	v1	SFT (MT) + RLVR (MT) + SFT (IC)	5	228	101	1.764
Maya	v3	GPT-5.5 (Zero-Shot)	**1**	212	212	**3.203**
Nahuatl	v2	SFT (MT) + SFT (IC)	3	200	200	1.560
Wixárika	v2	SFT (MT) + SFT (IC)	5	201	201	2.210
Overall human-evaluation ranking			4		Total points: 12	

parison, especially for Yucatec Maya, which is included in the shared task but not covered by our translation-initialized Aya Vision submissions. The exact prompt and inference details are listed in Appendix B.2.

5 Experimental Setup

5.1 Image Captioning Evaluation

We evaluate Image Captioning with chrF++ (Popović, 2017). Following the shared-task setting, we train SFT (IC) with the development set. To select the best candidate system before submission, we hold out a 20-example subset from the development data for internal model selection. This subset contains 5 examples per evaluated language: Bribri, Guaraní, Nahuatl, and Wixárika, corresponding to 10% of the 50 development examples available for each language. Because this model-selection subset is small, scores on it should be interpreted cautiously. We list the selected example IDs in Appendix C.

5.2 Machine Translation Evaluation

For translation, we evaluate the SFT (MT) and SFT (MT) + RLVR (MT) models on the AmericasNLP machine translation test split originally introduced in the 2021 shared task. Based on the later findings papers and shared task data releases, we believe this is the same test set that continued to be used in subsequent AmericasNLP machine translation evaluations for the overlapping languages (Mager et al., 2021; Ebrahimi et al., 2024; De Gibert et al., 2025). We report chrF++ for Spanish-to-Indigenous and Indigenous-to-Spanish directions.

All translation results use best-of-100 candidate selection under sentence-level chrF++ against the reference. We report chrF++ on the shared test set, but because candidate selection uses the reference, this setting should be interpreted as a reference-based oracle candidate-quality analysis rather than as single-output interactive translation.

5.3 Baselines and Comparisons

For image captioning, our main internal comparison is a step-by-step ablation of the training pipeline. We begin with the simplest setting, where Aya Vision is trained only on the image-captioning data. We then add translation fine-tuning before image-captioning training, so the model enters the captioning stage with stronger target-language generation ability. Finally, we test whether inserting the reward-based translation stage between translation fine-tuning and captioning provides any additional benefit.

For translation, we compare against reported AmericasNLP 2025 systems where the language and test-set overlap is available. Because our results use reference-based best-of-100 candidate selection, this comparison should be read as contextual rather than as a strict interactive MT leaderboard comparison. In particular, we are interested in whether SFT (MT) can move a large generative model from weak off-the-shelf behavior toward useful candidate generation for these languages.

6 Results

6.1 Image Captioning Results

Table 1 reports the main shared task results on the test set. Under the official automatic metric, v2 is the strongest fine-tuned Aya Vision submission overall, with 17.599 chrF++ averaged over Bribri, Guaraní, Nahuatl, and Wixárika. The RLVR-based v1 submission is slightly lower overall, which is consistent with our interpretation that RLVR (MT) is exploratory and does not provide consistent gains in this setting.

Table 2 reports chrF++ on the 20-example development subset. On this subset, the two systems

are close overall: SFT (MT) + SFT (IC) reaches 19.313 chrF++, while SFT (MT) + RLVR (MT) + SFT (IC) reaches 19.805. The per-language pattern is mixed: RLVR (MT) improves Bribri and Guaraní slightly, while Nahuatl and Wixárika are slightly lower.

Table 3 reports our official human-evaluation results. Our fine-tuned Aya Vision submissions reach the human-evaluation stage for all four languages covered by our translation-initialized pipeline: Mila ranks fourth for Bribri, fifth for Guaraní, third for Nahuatl, and fifth for Wixárika. Our separate zero-shot GPT-5.5 submission ranks first for Maya, which is the only language for which v3 was used in the official human-evaluation stage. Under the shared-task point system, Mila ranks fourth overall with 12 points. This result suggests that a frontier VLM can be a strong direct caption generator for Maya in this shared-task setting, even without our task-specific fine-tuning pipeline. However, this result should be interpreted alongside the automatic chrF++ results, where the GPT-5.5 Maya submission does not rank first. Since GPT-5.5 is a closed model, we cannot determine whether its pretraining or instruction tuning included Maya data.

6.2 Machine Translation Results

Table 4 compares our SFT (MT) model with selected AmericasNLP 2025 systems for overlapping languages. Because our SFT rows use reference-based best-of-100 candidate selection, they should be interpreted as reference-based oracle-selected candidate-quality scores rather than as single-output interactive translation results. In the Spanish-to-Indigenous direction, SFT (MT) remains below the reported multilingual baseline for most overlapping languages. However, it reaches the same broad range as several submitted systems, especially for Bribri, Wixárika, and Nahuatl. In the Indigenous-to-Spanish direction, the reference-based oracle-selected SFT candidates obtain strong chrF++ scores across all four languages. These results suggest that targeted supervised fine-tuning can make a large generative model produce useful candidate translations for these languages, while also showing that fairer comparisons require direct single-candidate generation rather than oracle selection. Appendix A reports qualitative translation examples for both translation directions (Tables 7 and 8) and sentence-level chrF++ score distributions (Figure 2).

Table 4: Comparison with selected AmericasNLP 2025 translation systems using chrF++ for overlapping languages. Our SFT rows use best-of-100 reference-based candidate selection and should be interpreted as reference-based oracle-selected candidate-quality scores, not as single-output interactive translation results. Bold values indicate the best performance, while underlined values indicate the second-best performance.

System	BZD	GRN	HCH	NAH
SPA→XXX				
Baseline	**25.52**	**35.68**	**28.26**	22.42
GMU	22.51	<u>29.95</u>	26.14	20.33
Syntax Squad	22.77	16.21	26.77	12.64
SFT (MT)	24.22	29.39	27.69	<u>26.39</u>
SFT (MT) + RLVR (MT)	<u>24.31</u>	29.46	<u>27.83</u>	**26.52**
XXX→SPA				
Baseline	30.14	35.91	26.33	26.36
GMU	27.86	33.84	24.37	25.58
Syntax Squad	26.22	24.70	22.02	13.88
SFT (MT)	**33.45**	<u>38.17</u>	<u>30.14</u>	**31.95**
SFT (MT) + RLVR (MT)	<u>33.42</u>	**38.29**	**30.18**	<u>31.73</u>

Table 5: Image Captioning pipeline ablation on the 20-example internal development subset. Rows incrementally add SFT (IC), SFT (MT), and RLVR (MT). Δ is measured against the previous row.

System	chrF++	Δ
No SFT	6.053	–
SFT (IC)	14.589	**+8.536**
SFT (MT) + SFT (IC)	19.313	+4.724
SFT (MT) + RLVR (MT) + SFT (IC)	**19.805**	+0.492

6.3 Pipeline Ablation

Table 5 makes the main pipeline ablation explicit. Starting from no supervised fine-tuning, SFT (IC) gives the largest improvement in chrF++ on the internal development subset, reaching 14.589. Adding SFT (MT) before SFT (IC) further improves the score to 19.313, and adding RLVR (MT) gives a smaller additional gain to 19.805.

This ablation clarifies the role of each component. The largest single contribution comes from SFT (IC), which teaches the model the image captioning task. SFT (MT) remains useful as a language-generation initialization for the Indigenous languages, while RLVR (MT) has the smallest incremental effect in this setup.

7 Conclusion

We presented a multi-stage supervised fine-tuning approach to cultural Image Captioning for Indige-

nous languages of the Americas. Instead of relying only on scarce image captioning supervision, we first run SFT (MT) with Spanish–Indigenous translation data and then run SFT (IC). Development experiments show that SFT (MT) provides a useful initialization in our internal model-selection setting: SFT (IC) from the base model is much weaker than SFT (IC) from the SFT (MT) model. We also show that the same fine-tuned VLM can produce useful translation candidates for the overlapping languages under reference-based oracle candidate selection.

Under the official human-evaluation stage, our submissions rank in the top five for each submitted language, and our zero-shot GPT-5.5 run ranks first in the Maya language track under the same protocol. More broadly, our results suggest that low-resource cultural image captioning in this setting is not only a multimodal grounding problem but also a target-language generation problem. Translation data can therefore serve as a practical bridge toward multimodal generation in low-resource settings: SFT (MT) improves sample efficiency for image captioning because it teaches the model to produce stable text in the target Indigenous language, while SFT (IC) teaches the model to connect that language to the image. When image-caption supervision is scarce but bilingual text exists, translation fine-tuning can be an effective initialization.

Limitations

Our image captioning evaluation is limited by the small size of the internal development subset. Twenty examples are useful for internal model selection, but they cannot fully characterize performance across images, cultures, and linguistic forms. The shared task description allowed participants to use the development set for training, and we follow that setting.

Our translation results use best-of-100 candidate selection with reference-based chrF++ scoring. Although this uses the machine translation test set and chrF++, it is not the same as interactive translation: in real use, the system would not have the reference translation available when choosing an output. These scores should therefore be read as reference-based oracle-selected candidate-quality results.

The system is also limited by automatic evaluation. chrF++ is useful for low-resource translation and image captioning, but it cannot determine whether a caption is culturally appropriate or whether a translation is acceptable to speakers. Future work should include evaluation by language experts and community members.

Finally, the model may produce incorrect, offensive, or culturally inappropriate outputs. We do not recommend deployment without careful human review, community consultation, and safety procedures appropriate for each language context.

Ethical Considerations

Work on Indigenous languages requires care because language data is connected to communities, identity, and cultural knowledge. A model that generates text in an Indigenous language should not be presented as a replacement for speakers, translators, teachers, or community institutions. Our goal is to study supervised fine-tuning and reinforcement learning methods and provide a technical system for shared task evaluation, not to claim authority over language use.

The cultural image captioning setting also raises questions about what visual content should be described and how. Captions may encode cultural assumptions, and automatic systems can easily produce descriptions that are linguistically plausible but culturally wrong. Any real use of such a model should involve community-centered evaluation and control over data, outputs, and deployment contexts.

Acknowledgments

We thank the AmericasNLP organizers for creating the shared task and for supporting evaluation on Indigenous languages of the Americas.

References

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani-Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second AmericasNLP competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jonathan Hus, Nathaniel Krasner, and Antonios Anastasopoulos. 2025. Machine translation using grammar materials for LLM post-correction. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99, Albuquerque, New Mexico. Association for Computational Linguistics.

Manuel Mager, Dionico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the*

Americas, pages 202–217, Online. Association for Computational Linguistics.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-19.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mahshar Yahan and Mohammad Islam. 2025. Leveraging large language models for Spanish-indigenous language machine translation at AmericasNLP 2025. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 126–133, Albuquerque, New Mexico. Association for Computational Linguistics.

A Machine Translation Examples

Tables 7 and 8 show qualitative examples for both translation directions. Within each direction and language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++. Figure 2 shows the corresponding sentence-level chrF++ score distributions for both translation directions and the four overlapping languages.

B Training Details and Prompts

All local experiments use Aya Vision 32B as the base model. We keep the base model frozen and train only LoRA parameters, using LoRA rank 32 and LoRA alpha 64. Both SFT (MT) and SFT (IC) use a learning rate of 2×10^{-5}, a linear scheduler, and a warmup ratio of 0.03. SFT (MT) is configured for 60 epochs with batch size 32 and gradient accumulation 1, but early stopping terminates training after 6.80 epochs. SFT (IC) is configured for 10 epochs with batch size 1 and gradient accumulation 1, with the selected checkpoints obtained after 9 epochs, initialized either from Aya Vision 32B or from the SFT (MT) checkpoint depending on the experiment.

B.1 Fine-Tuning Prompt Templates

For SFT (MT), each bilingual training example uses the following instruction-style prompt, and the target sentence is used as the assistant response:

```
Traduce del <source language> al <target language>.
```

For SFT (IC), each image-captioning example uses the following Spanish prompt:

```
Escribe un solo pie de foto en {language} para esta imagen. Debe ser una descripción culturalmente adecuada de la imagen. Responde solo con el pie de foto en {language}, sin explicaciones.
```

B.2 GPT-5.5 Zero-Shot Prompt and Technical Details

For the GPT-5.5 submission (v3), we used direct zero-shot Image Captioning through the OpenAI Responses API. Each input consisted of one image and a target language. We did not provide in-context examples, reference captions, intermediate Spanish captions, or translation outputs. The instruction string was:

```
You create image captions for an AmericasNLP submission. The target language is specified in the user prompt. Answer with exactly one caption and no surrounding text.
```

The user prompt template was:

```
Look at the image and write one concise, culturally appropriate caption directly in {language} (ISO {iso_lang}). Do not write in Spanish or English. Do not include labels, explanations, markdown, quotation marks, or translations. Return only the final caption in {language}.
```

We used model `gpt-5.5`, reasoning effort `medium`, image detail `high`, and maximum output tokens 4096. We generated one caption per image. No supervised fine-tuning, RLVR, or multi-candidate reranking was used for this run.

C Internal Development Subset IDs

The internal Image Captioning development subset contains 20 examples, selected as 10% of the development set with 5 examples per language and seed 42. The IDs are listed in Table 6 for reproducibility.

Table 6: Internal development subset example IDs.

Language	IDs
Bribri	bzd_040, bzd_007, bzd_001, bzd_047, bzd_017
Guaraní	grn_016, grn_015, grn_009, grn_048, grn_007
Orizaba Nahuatl	nlv_045, nlv_049, nlv_036, nlv_006, nlv_039
Wixárika	hch_048, hch_023, hch_022, hch_026, hch_034

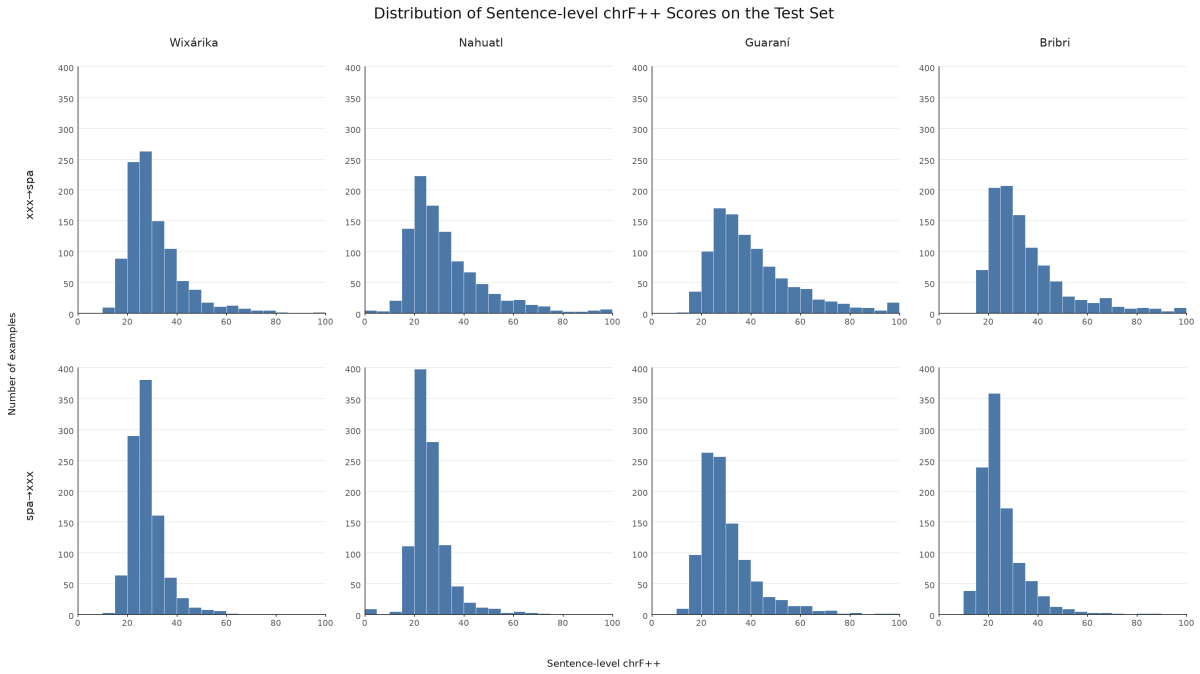

Figure 2: Distribution of sentence-level chrF++ scores on the machine translation test set for both translation directions and the four overlapping languages. These distributions complement the aggregate scores in Table 4 by showing the spread of reference-based oracle-selected candidate quality across examples.

Table 7: Qualitative translation examples from the SFT (MT) + RLVR (MT) model for the XXX→SPA direction. Within each language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++.

chrF++	Language	Source	Reference	Model translation
100.00	Bribri	Sa' ù bak tkër kãmîx káx íxẽ tã.	Nuestra casa estaba lejos de este lugar.	Nuestra casa estaba lejos de este lugar.
35.26	Bribri	Îxñẽ káxm ye' dö, e' wa tã káxm se' i sãũ.	Hoy no he ido, así que no lo he visto.	Pues por ahora yo todavía no lo he visto, después algún momento lo veré.
17.88	Bribri	Ie' bak kèq klòqnẽ tã irir sãwîx ie.	Ella era increíblemente vulnerable.	Ella no había venido desde ese entonces.
100.00	Guaraní	Kennedy oñe'ê umi pilotokuéra ndive.	Kennedy habló con los pilotos.	Kennedy habló con los pilotos.
40.79	Guaraní	Ndaikuaaái ha'épa oipota umimarandu ñesê rehegua aravo.	No sabía si él quería información sobre los horarios de viaje.	No está claro si quería la información sobre el plazo de presentación.
18.84	Guaraní	Nameméi añemonde porâ.	No me pongo ropa de vestir a menudo.	No me dejes en la oscuridad.
100.00	Wixárika	waik+, waik+ Locust Hill.	No, no Locust Hill.	No, no Locust Hill.
30.66	Wixárika	m+k+ peuku yeix+a, p+reuyepi kitenie nepa eriwa nemana kunierix+ metá nepei xei kem+raka eriekai, tsiere nepeixei mey+ h+awekai.	Él vino, abrió la puerta y recuerdo mirar atrás y ver la expresión de su rostro, y pude ver que estaba decepcionado.	El joven se puso de pie, lo pegó en la puerta de su habitación, y fue a ver que estaba allí, y fue a ver que estaba en un sueño.
17.54	Wixárika	aix+ nuereu erie ena aimieme kwanetsi nake.	Me siento bastante seguro aquí y por eso me gusta.	En las orillas estaba una serpiente inmensa y negra.
100.00	Nahuatl	Fuerza Aerea quicohqui 18 aviones	La Fuerza Aérea compró 18 aviones.	La Fuerza Aérea compró 18 aviones.
32.37	Nahuatl	Coneme quiiqueh ce achi campán	Los niños bebieron un poco de champán.	Ellos quieren tocar un poco de campana.
12.40	Nahuatl	Ya nelpano eltoya yolquentzi	Ella era increíblemente vulnerable.	Pero ya viene mi madre, no los molestes.

Table 8: Qualitative translation examples from the SFT (MT) + RLVR (MT) model for the SPA→XXX direction. Within each language block, rows are ordered from a high-scoring example to an approximately average example and then a low-scoring example according to sentence-level chrF++.

chrF++	Language	Source	Reference	Model translation
86.70	Bribri	Viví en Texas.	Ye' sèxnẽ Texas.	Ye' sèxnẽ Texas árô.
25.00	Bribri	Siempre disfrutaba escuchando las historias de mi abuela.	Wake' ye' wîxke káxpake e' tsèke yö bua'ë.	Ye' ãmìx tso' e' tö ye' wã míxköla i chèke e' tsáx.
11.81	Bribri	El diseño era horrible	I dióqköl bak sulû ichóqk.	I mîxk sa' dör kóqyoq.
100.00	Guaraní	Vivían en Atlanta.	Oikova'ekue Atlanta-pe.	Oikova'ekue Atlanta-pe.
30.27	Guaraní	Me dijo que era urgente.	He'i chéve pya'eva'erâha.	He'i chupe ha'e péicha ha'erãjeýma.
14.05	Guaraní	Está bien si cometimos algunos errores.	Oî porâ, rojavy mimi.	Oĩmi jepémo, opavavévo oiko mba'e oporomongusúvo ko tendápe.
61.15	Wixárika	Ella dijo que su madre tiene el mismo caminar.	m+k+ petay+ kename mamaya yaxeik+a reyeikakai.	m+k+ petay+ kename mama payeikakai kename y+k+
27.78	Wixárika	Tu tiempo sigue aumentando cada año.	atukari p+t+ +rir+me witari manye yeika.	kuinie timaye yeitait+ mepeku mieme m+k+k+ ne hamikiekame muwa tukari wahiawariya.
15.85	Wixárika	No sé cuánto tiempo durará.	neukwara maté kem+ reutere,	ni nekai ukaratsi eriwaniya kemariyariyari,
61.06	Nahuatl	Augusta no es una gran ciudad.	Augusta amo hueyiatepetl.	Augusta niman amo hueyi altepetl.
26.13	Nahuatl	Pon un anuncio de refresco.	xihtlali tlamachiltilli cececatl.	Tlamaquetza un tomachiliztli ic in zan tlatlamacoz.
16.07	Nahuatl	Ella era increíblemente vulnerable.	Ya nelpano eltoya yolquentzi	Quipiya in itlacenpaj tlalquetzaloyan.

Culturally-Aware Image Captioning for Guaraní with Multimodal Prompting: IUHoosiers at AmericasNLP 2026

Wenchen Shi*, **Phakphum Artkaew***, **Luke Gessler**
Indiana University Bloomington
{wencshi, partkaew, lgessler}@iu.edu
*Equal contribution

Abstract

The AmericasNLP 2026 Shared Task challenges systems to generate culturally grounded image captions in indigenous languages of the Americas, a setting that demands both cultural awareness and linguistic accuracy for severely under-resourced languages. We present IUHoosiers, Indiana University's system for the Guaraní track. Rather than fine-tuning, our approach centers on inference-time knowledge injection: for each test image, we retrieve relevant Guaraní grammatical and cultural resources using BM25 and inject them into a large vision-language model's prompt alongside the image, enabling language-specific cultural and linguistic grounding without any parameter updates. IUHoosiers placed first for Guaraní in both automatic evaluation (24.67 chrF++) and human evaluation (3.45/5), outperforming all other participating systems.

1 Introduction

The AmericasNLP 2026 Shared Task challenges systems to generate culturally grounded image captions in indigenous languages of the Americas (Bui et al., 2026). We focus on Guaraní, a Tupi-Guaraní language that is spoken widely across Paraguay, Brazil, Bolivia, and Argentina and is the co-official language along with Spanish in Paraguay (∼6.5M speakers). However, it is still considered an under-resourced language across all NLP resources (Chiruzzo et al., 2020).

Cultural captioning for Guaraní presents two intertwined challenges. First, captions must be *culturally accurate*: correctly distinguishing closely related cultural items, such as *mate* from its cold counterpart *tereré*, requires culturally grounded knowledge that generic vision-language models often lack. Second, the target language itself is challenging: Guaraní is agglutinative with active/inactive voice morphology and pervasive Spanish borrowing in everyday *jopara* speech (Estigarribia, 2020), making linguistically correct output non-trivial even for large language models.

Since the dev set contains only 50 image-caption pairs and the gold labels follow a specific format, parameter-efficient fine-tuning (PEFT) methods (Mangrulkar et al., 2022) such as low-rank adaptation (LoRA; Hu et al., 2022) may be prone to overfitting without substantial data augmentation. We therefore focus on inference-time methods instead. Recent work suggests that grammar-book and parallel-text context can improve low-resource language generation without gradient updates (Tanzer et al., 2023; Aycock et al., 2025; Zhang et al., 2025), and we build on this finding with **retrieval-augmented prompting**. For each test image, we generate a text description, query four Guaraní knowledge pools with BM25 (Robertson and Zaragoza, 2009), and inject the top-k retrieved items into the system prompt of Gemma 4 31B (Google DeepMind, 2025) to generate captions in a single forward pass. Additionally, we explore a *visual few-shot* mode that injects cultural images from Diccionario audiovisual multilingüe del Paraguay (DAMPY)[1] as multimodal context. Figure 1 gives an overview of the full pipeline.

2 Approach

Why in-context learning? The AmericasNLP 2026 Shared Task spans five indigenous languages (Guaraní, Bribri, Yucatec Maya, Wixárika, and Nahuatl), each with very distinct typological properties, morphological systems, and cultural contexts. We believe the linguistic and cultural distances among these languages are too vast for any single unified system to address all of them effectively, and we take the view that in low-resource settings, building one dedicated system per language is the more principled approach. A language-specific pipeline allows much more targeted injec-

[1] https://spl.gov.py/dampy/index.html

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 236–242

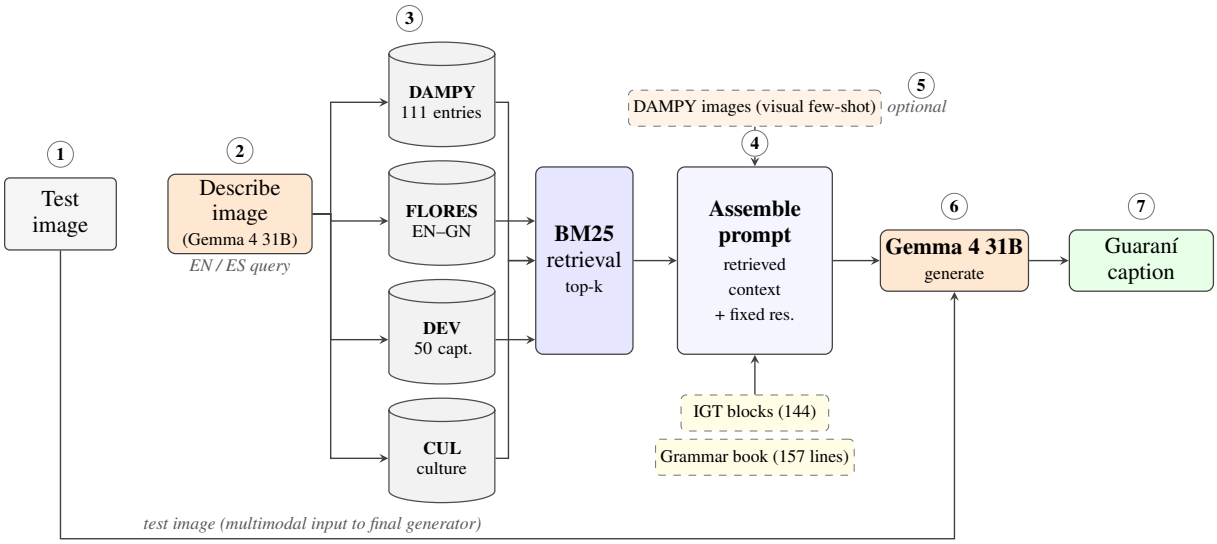

Figure 1: Overview of the IUHoosiers pipeline for Guaraní image captioning. Given a test image **(1)**, Gemma 4 31B generates an EN/ES description **(2)** that is used as a BM25 query over four Guaraní knowledge pools **(3)**: cultural entries (DAMPY), parallel sentences (FLORES), dev-set gold captions (DEV), and a curated cultural knowledge base (CUL). Step **(3)** operates in one of two modes: *RAG* dynamically selects the top-k entries per image via BM25, while *static* mode takes the first n examples from the dataset regardless of image content (used in v6 and v8; see Table 1). The retrieved context is combined with fixed resources (a 157-line grammar excerpt and 144 interlinear-glossed example blocks) into the system prompt **(4)**. In visual few-shot mode **(5)**, DAMPY images with their captions are additionally injected as multimodal shots (dashed). Gemma 4 31B **(6)** consumes the assembled prompt together with the test image and produces a Guaraní caption **(7)**. EN: English, ES: Spanish, GN: Guaraní.

tion of relevant grammatical structure and cultural knowledge than a one-size-fits-all system could provide. We therefore build a Guaraní-specific pipeline and adopt in-context learning (ICL) as our core method. Recent work on extremely low-resource languages shows that ICL, especially when paired with explicit linguistic signals, can outperform parameter-efficient fine-tuning when the target language is poorly represented in the base model (Li et al., 2025). The feasibility of injecting many resources simultaneously is further supported by recent long-context language models (Gemini Team et al., 2024), which make it practical to include large amounts of in-context material within one context window. At the same time, work on culturally aware vision-language modeling has shown that standard VLMs often miss fine-grained cultural distinctions, and that improving cultural grounding may require either specialized data construction or model adaptation (Huang et al., 2025). Our approach therefore tests whether carefully selected grammatical and cultural knowledge, injected at inference time, can close this gap without any parameter updates. [2]

Model. We use Gemma 4 31B (Google DeepMind, 2025), a 31-billion-parameter vision-language model with a 128K-token multimodal context window. The original baseline prompt (Spanish-language Guaraní captioning guidelines) is used for all nine submitted configurations.

Fixed resources. Two resources are injected into every prompt regardless of the test image: a 157-line condensed grammar-book excerpt from Estigarribia (2020) covering Guaraní morphosyntax, voice alternations, and postpositional structure, and 144 interlinear glossed example blocks (surface form / morpheme break / gloss / translation). Both resources were adapted from Aycock et al. (2025), who study grammar-book prompting across multiple low-resource languages including Guaraní. Their preprocessed materials served as our starting point.

We note that these distilled resources were generated with the assistance of Claude-Opus-4.7 in a two-step process. First, we provided Claude with the raw resource files together with the competition setting and asked it to reason about the most effective format for injecting such resources into a language model for in-context learning. Through iterative discussion, Claude then produced prompts

[2]Code is available at https://github.com/victorshi119/Cultural_Image_Captions_2026

designed to reorganize and condense the original materials into more model-friendly representations. In a second, separate session, we supplied Claude with the raw text resources together with the generated prompt in order to produce the final distilled resources.

Dynamically retrieved pools. For each test image a text description is generated and used as the BM25 query (Robertson and Zaragoza, 2009). Four resource pools are indexed and queried per image:

1. **DAMPY**: 111 culturally grounded entries from DAMPY, Paraguay's official audiovisual dictionary (42 food, 47 fauna, and 22 flora entries), each with bilingual Spanish/Guaraní labels. We note that the DAMPY dataset we scraped from the internet does not provide captions. We therefore generated bilingual captions in two steps: first, a VLM produced baseline Guaraní captions for each image; then, Claude-Opus-4.7 in Adaptive Thinking mode refined those captions for cultural accuracy and idiomatic Guaraní usage.

2. **FLORES**: English-Guaraní parallel sentence pairs from FLORES-200 (NLLB Team et al., 2022), a publicly available multilingual benchmark dataset providing in-domain Guaraní sentence structure across a broad topical distribution (1,012 sentence pairs used).

3. **DEV**: the 50 gold Guaraní dev-set captions released by the shared task organizers, retrieved by BM25 text matching per image to provide in-domain captioning style. Since the competition only provides Guaraní labels, we augmented the dataset with parallel English captions generated by Claude, which serve as the BM25 query surface for retrieval.

4. **CUL**: a curated Guaraní cultural knowledge base consisting of 22 thematic sections written in Spanish, covering food, drink, dress, architecture, household objects, flora and fauna, festivals, children's games, occupations, crafts, music, religion, mythology, and idiomatic Guaraní caption style. Each section is oriented toward visually identifiable entities and includes Guaraní vocabulary, "visual cue" sub-entries, and an explicit anti-stereotype list. The base file was compiled by the authors from Paraguayan web sources, then expanded via a two-step LLM-assisted process described in Appendix A. Since CUL is small enough to fit in the context window, it is always injected in full into the system prompt; BM25 retrieval is applied additionally to surface the most image-relevant sections.

To be specific, the retrieval works as follows: we first build a BM25 index over either grammar sections for CUL, which is organized by section, or paired entries for DAMPY, FLORES, and DEV. Then, depending on the dataset, we ask a language model (Gemma 4 31B) to first describe the image in the relevant language, using English for FLORES and Spanish for DAMPY, CUL, and DEV. We then use BM25 to retrieve the corresponding relevant Guaraní examples.

For the visual few-shot run (v7), instead of injecting text-retrieved DAMPY entries, DAMPY images and their Guaraní captions are injected as multimodal few-shot examples in the prompt. In the submitted run, these are selected statically (the first 10 images from the dataset); BM25-driven visual retrieval, where exemplars would be dynamically matched per test image, was not implemented in time (see Appendix B).

Prompt structure. Each request follows a four-step process:

1. **System prompt:** Contains the captioning guidelines, fixed grammar resources, and dynamically retrieved context blocks from CUL, DAMPY, FLORES, and the dev captions.

2. **Multimodal few-shot prompting:** Optional multimodal few shot examples are injected depending on the retrieval setting. For visual few shot runs, the retrieved DAMPY image caption pairs are inserted here as multimodal demonstrations.

3. **Test image:** The query image is provided to the model.

4. **Generation:** The model generates a caption in Guaraní.

Configurations. Table 1 summarizes the nine submitted system configurations evaluated in the shared task. Runs v0-v5 and v8 use text+RAG mode, where examples from all four pools are dynamically selected using retrieval-augmented generation (RAG) based on similarity scores. Run v6 instead uses text+static mode, where a fixed set

Ver	Mode	DMP	FLR	DEV	CUL
v0	text+RAG	5	100	5	5
v1	text+RAG	10	100	10	5
v2	text+RAG	15	100	10	5
v3	text+RAG	20	100	10	6
v4	text+RAG	15	120	10	8
v5	text+RAG	15	150	10	10
v6	text+static	3	30	0	0
v7	visual+static	10	30	0	0
v8	text+RAG	10	100	10	5

Table 1: Nine submission configurations. **DMP**: DAMPY shots. **FLR**: FLORES shots. **DEV**: dev-set shots. **CUL**: top-k CUL sections additionally surfaced by BM25 (full CUL is always injected). In **RAG** mode, shots are selected by retrieval score; in **static** mode, the first n examples from the dataset are used. In **visual** mode, DAMPY shots are injected as multimodal image–caption pairs; in **text** mode, they are injected as text only.

Ver	Mode	chrF++
v0	text+RAG	22.40
v1	text+RAG	22.02
v2	text+RAG	21.59
v3	text+RAG	21.90
v4	text+RAG	21.44
v5	text+RAG	20.43
v6	text+static	22.49
v7	visual+static	24.17
v8	text+RAG	22.02
baseline	—	20.82

Table 2: Nine submission configurations and their dev-set chrF++ scores, alongside the official baseline for reference.

of examples is selected statically from DAMPY and FLORES without retrieval. Run v7 uses visual+static mode, replacing DAMPY text examples with visual shots and using the first 10 images from the dataset as fixed demonstrations. Across all runs, the grammar book (157 lines) and interlinear glossed text (144 blocks) are always included as fixed context.

3 Results

3.1 Dev-set Results

Table 2 shows the results for each configuration on the dev set. The dev set contains only 50 image-caption pairs, which creates a tension in evaluation: using dev-set captions as few-shot examples (as in the full pipeline) would contaminate the evaluation, so DEV shots are excluded when scoring on the dev set.

Team	Ver	chrF++	CIDEr
IUHoosiers	v4	**24.67**	0.149
IUHoosiers	v5	24.42	**0.167**
IUHoosiers	v1	24.41	0.135
IUHoosiers	v8	24.41	0.135
IUHoosiers	v2	24.41	0.143
IUHoosiers	v0	24.39	0.124
IUHoosiers	v3	24.16	0.128
IUHoosiers	v6	24.04	0.112
IUHoosiers	v7	22.43	0.074
gators	v0	23.10	0.124
baseline	v0	20.14	0.005
Mila	v1	19.77	0.031
usp	v0	19.73	0.024
NAIST	v0	19.41	0.046

Table 3: Official test-set leaderboard (selected entries). Eight of nine IUHoosiers submissions outperformed all other teams. v5 achieved the highest CIDEr despite not having the best chrF++.

3.2 Test-Set Results

Table 3 shows official test-set chrF++ and CIDEr for all IUHoosiers submissions alongside competing teams. The best submission (v4) achieves 24.67 chrF++ and 0.149 CIDEr, a 4.53-point gain over the official baseline (20.14) and a 1.57-point margin over the next-best team (gators, 23.10). IUHoosiers placed first for Guaraní in human evaluation with a mean annotator rating of 3.45/5, versus 3.30 for the second-ranked team; eight of our nine submissions outscore every other team's best result.

4 Analysis

4.1 Qualitative Analysis

Submission selection. We selected these nine configurations from a larger candidate pool, with final choices informed by chrF++ scores alongside our personal judgement and our team's working knowledge in Guaraní and Jopara. While chrF++ was a useful signal for tracking progress during the pilot phase, final selection was not driven by it alone, as the metric measures character n-gram overlap against a single gold reference and is therefore blind to semantic similarity and cultural correctness. A caption that correctly identifies a cultural item but uses different surface forms than the reference can score no better than one that gets the cultural content wrong while sharing incidental n-grams. Instead, selection applied two criteria. First, *output style*: captions should match the concise, visually specific register of the dev-set gold labels and avoid formulaic boilerplate appended regardless of image content. Second, *cultural accuracy*:

captions should correctly identify culturally specific items, especially when the discriminating cue is visual and subtle. The BM25 retrieval pipeline supports both: FLORES parallel context shapes linguistic register and fluency, while DAMPY retrieval anchor culturally grounded vocabulary.

Figure 2: Test image grn_052. v4 (text+RAG) misidentifies the drink as *tereré* and offers a generic cultural gloss, while v7 (visual few-shot) correctly identifies *mate* and provides richer cultural context as this version situates it as a morning drink shared in the cool breeze and reflecting its role in Paraguayan relational life.

[wrong] v4 (text+RAG)

Tereré ekípo oĩva poyvi isañyju ári: guámpa yvyrágui, vombílla ha término. Ko tembi'u ha y'u ha'e hína ñande reko imbaretevéva ha omojoaju ñande reko Paraguáipe.

[correct] v7 (visual few-shot)

Mate ha guampakuéra oĩva peteĩ pano pytã ha hovy ári. Ko tembi'u ha imombe'upy oñembopyre hína ñane reko Paraguáipe, oje'u pyhareve yvytu ro'ysã jave.

Case study: mate vs. tereré. One notably culturally specific image among the first ten test examples is **grn_052**, depicting a *mate* setup (gourd, bombilla, *pava* kettle). Most text+RAG configurations label it *tereré*; only v2 primarily identifies it as *mate*, while v0 mentions both drinks. The text+static run (v6) and the visual few-shot run (v7) both correctly identify *mate*, but differ in descriptive accuracy. v6 lists the objects present and appends a generic cultural phrase (*"Mokõi guampa ha peteĩ pava oĩ peteĩ ao póipe. Ko'ãva ojeipuru mate pyahúpe, ñane rembi'u ha y'u rekojera. Ohechauka ñane ñemoirũ ha ñane tekopytã paraguáigua."* — "Two guampa and a pava are in a pocket; these are used in new mates, our food and drink recipes; it shows our solidarity and our Paraguayan culture"), while v7 visually grounds the scene (noting the red and blue cloth), provides the correct ideal weather setting (a morning drink taken in the cool breeze), and articulates its social significance (reflecting relationships and the surrounding environment). That is to say, v7 shows more depth in understanding than v6.

Text runs also tend to add repeated generic closing phrases (e.g., *"Ko tembi'u ha'e peteĩ rembiapokue tee Paraguáigui"*, "this dish is a true cultural product of Paraguay") regardless of image content, while gold captions use more varied, image-specific language. The visual run produces more image-specific descriptions, which explains their higher qualitative character despite lower chrF++ on the test set. Overall, the human evaluation's preference for IUHoosiers over competitors reflects the generally higher Guaraní output quality compared to translation-based baselines.

5 Conclusion

IUHoosiers achieved first place for Guaraní at AmericasNLP 2026 using Gemma 4 31B with BM25 augmented retrieval from four Guaraní knowledge sources. Beyond the result itself, the most practically useful contribution is the pipeline design: language specific knowledge injection at inference time without fine tuning rather than any particular hyperparameter choice. The broader takeaway is that carefully curated grammar and cultural resources, injected at inference time, can remain competitive without parameter updates. That is the durable result worth carrying forward. Visual few shot retrieval remains a promising but underexplored direction: dev set calibration (Appendix B) showed visual shots outperforming text injection by 1.41 chrF++ points, and BM25 driven visual retrieval, where exemplars are dynamically matched to each test image, was not implemented in time for submission and warrants future investigation.

Acknowledgments

We thank Indiana University Research Technologies for REALLMS API access and the BigRed 200 HPC cluster, and the AmericasNLP 2026 organizers for the dataset and evaluation infrastructure.

References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs really learn to translate a low-resource language from one grammar book? arXiv preprint arXiv:2409.19151.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, and

15 others. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Luis Chiruzzo, Santiago Castro, Mariella Cardenas, Gustavo Gimenez González, Yliana Gimenez, and Dina Wonsever. 2020. Development of a Guaraní-Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 893–898, Marseille, France. European Language Resources Association.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guaraní*. UCL Press.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google DeepMind. 2025. Gemma 4 technical report. https://blog.google/technology/google-deepmind/gemma-4/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yuchen Huang, Zhiyuan Fan, Zhitao He, Sandeep Polisetty, Wenyan Li, and Yi R. Fung. 2025. CultureCLIP: Empowering CLIP with cultural awareness through synthetic images and contextualized captions. *Preprint*, arXiv:2507.06210.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025. It's all about in-context learning! teaching extremely low-resource languages to LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29544–29559, Suzhou, China. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. arXiv preprint arXiv:2309.16575.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3977–3997, Vienna, Austria. Association for Computational Linguistics.

A Construction of the CUL Cultural Knowledge Base

The CUL resource was built in three LLM-assisted steps using Claude Opus 4.

Step 1 — Seed file. The authors compiled a 16-section seed knowledge file in Spanish, drawing on their own working knowledge of Guaraní culture, Wikipedia, and the *Every Culture* encyclopedia entry on Guaranís. The seed covered cosmovision, food, drink, traditional medicine, flora and fauna, music, religion, and mythology, but lacked visual grounding, ethnic distinctions between mestizo Paraguayan and Indigenous Guaraní communities, and Guaraní vocabulary for common visual entities.

Step 2 — Prompt generation. We first prompted Claude Opus 4 to design a specialized research prompt for expanding the seed file, targeting gaps most relevant to image captioning: traditional dress, architecture, festivals, children's games, household utensils, and a Guaraní visual vocabulary. The generated prompt instructed the model to (1) audit the existing file for visual grounding and accuracy, (2) search authoritative Paraguayan sources to identify gaps, (3) draft additions in the same Spanish-language bullet-point style as the seed, and (4) consolidate everything into a single deliverable.

Step 3 — Expansion. In a separate Claude Opus 4 session, the generated prompt was supplied together with the seed file. The model queried sources including ABC Color, Portal Guarani, Última Hora, IWGIA, and Tierraviva, and produced the consolidated CUL file containing 22 thematic sections with explicit visual-grounding cues throughout. An anti-stereotype section warns the model against projecting folkloric cues onto ordinary modern Paraguayan scenes. The final document is included in the project code repository.

B Visual Few-Shot: Dev-Set Calibration and Future Directions

Table 4 shows chrF++ from pre-submission calibration experiments on the 50-image dev set using *static injection* (fixed FLORES sentence counts and DAMPY shot counts). These experiments explored a wider parameter range than the final v0–v8 submissions.

Mode	Config (FLR / DMP)	chrF++
visual	15 shots / FLR=50	**24.43**
visual	10 shots / FLR=30	24.17
visual	5 shots / FLR=30	23.49
text	FLR=100, DMP=5	**23.02**
text	FLR= 30, DMP=3	22.49
text	FLR= 80, DMP=3	22.21
text	FLR= 50, DMP=5	22.09
text	FLR= 20, DMP=2	21.80
text	FLR= 50, DMP=3	21.43

Table 4: Dev-set calibration results (static injection, 50 images). **FLR**: FLORES sentences injected. **DMP**: DAMPY shot count. For visual runs, DAMPY images are injected as few-shot context. The 10-shot/FLR=30 visual configuration corresponds to submitted run v7.

Visual few-shot consistently outperformed text injection across all shot counts on the dev set. We see that the best visual configuration (15 shots, FLR=50) achieved 24.43 chrF++, a 1.41-point lead over the best text configuration (FLR=100, DMP=5: 23.02). These findings motivated including visual few-shot among the nine official submissions.

Two stronger visual directions were not pursued due to time constraints. First, the 15-shot/FLR=50 configuration (24.43 on dev) was not submitted; the submitted visual run (v7, 10 shots/FLR=30) achieved 24.17 on dev and 22.43 on the test set. Second, BM25-driven visual retrieval, which dynamically selects the most image-relevant DAMPY exemplars per test image rather than using static selection, was not implemented.

The submitted v7 trailed text BM25 (v4) by 2.24 chrF++ on the test set (22.43 vs. 24.67). We hypothesize three contributing factors. First, the 50-image dev set is small, so the best static shot count may overfit to it. Second, the 111-entry DAMPY pool covers only food, flora, and fauna; when a test image falls outside this distribution, static visual selection may return misleading exemplars. Third, broad FLORES context (k=120/150) generalizes across diverse test topics in a way that static visual selection cannot when no relevant DAMPY entry exists. BM25-driven visual retrieval would address all three by dynamically matching exemplars to each image, and is a natural direction for future work.

6fanle Submission to the AmericasNLP 2026 Shared Task on Wixarika Image Captioning

Ji Wang and Hanqi Yang
Uppsala University

Abstract

This paper describes the 6fanle system for the Wixarika track of the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages. We report the data, preprocessing, model components, development experiments, and final results. Our system uses Spanish as a pivot language: Qwen3-VL generates Spanish caption candidates, CLIP retrieval supplies visually related examples, the official Sheffield-compatible machine translation model translates candidates into Wixarika, and a character n-gram language model reranks the translated outputs. The selected configuration achieved 19.1468 chrF++ in local 5-fold validation. In official Wixarika automatic evaluation, our v0 submission obtained 19.1569 chrF++ and 0.02145 CIDEr. In the final overall ranking, the team placed third.

1 Introduction

The AmericasNLP 2026 Cultural Image Captioning shared task asks systems to produce captions for images in Indigenous languages of the Americas (Bui et al., 2026). We participated in the Wixarika track, where only 70 official image-caption examples were available before test inference: 20 pilot examples and 50 development examples. The test set contains images and metadata, and the required output is a JSONL file with a `predicted_caption` field.

Because the available paired image-caption data are very small, we did not train an end-to-end image-to-Wixarika model. Instead, we used a modular pipeline that separates visual description from low-resource text generation. A vision-language model produces Spanish descriptions, and the official machine translation component converts them to Wixarika. Retrieval and Wixarika-side reranking are used to adapt the outputs to the limited in-domain data. We release the code and configuration files needed to reproduce the pipeline at `https://github.com/ousyu66-pixel/americasnlp2026-wixarika-captioning`.

2 Related Work

Our system follows the general pattern of previous AmericasNLP system descriptions, which document the resources used, preprocessing decisions, model variants, validation protocols, and final submissions. The official baseline and the Sheffield-compatible translation component motivate our use of Spanish-Wixarika MT as the low-resource generation stage (Gow-Smith and Haddow, 2023).

We also considered ideas from low-resource multimodal data construction and weakly supervised visual data generation (Xie et al., 2024; Xiao et al., 2025). However, the final system does not synthesize new images and does not train a new multimodal model. Instead, we use a lightweight version of this idea: additional Spanish descriptions are rewritten into short visual captions and used only as retrieval support.

3 Data

Table 1 summarizes the data sources used by the submitted system and by a rejected ablation involving lexical filtering. We distinguish final-system resources from resources used only in rejected experiments to avoid overstating what entered the submitted run.

3.1 Preprocessing and Post-processing

The preprocessing stage normalizes input JSONL records, resolves image paths, and prepares three kinds of retrieval examples: pilot examples with Spanish captions, development examples back-translated from Wixarika into Spanish, and augmented examples rewritten from longer Spanish descriptions into short caption-style Spanish. Held-out validation examples are removed from the retrieval bank and from the Wixarika language-model

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 243–247

Source	Count / size	Use in this work
Official Wixarika pilot images	20	Used for validation, retrieval support, and Wixarika language-model training. Pilot Spanish captions are used when available.
Official Wixarika development images	50	Used for validation and Wixarika language-model training. Spanish retrieval text is produced by Wixarika-to-Spanish backtranslation.
Official Wixarika test images	201	Used only for final inference. No test labels were used.
Wixarika Research Center augmented images	65	Spanish descriptions were rewritten into short caption style and used as auxiliary retrieval support (Wixarika Research Center, 2026).
AmericasNLP 2023 Spanish-Wixarika parallel text	9,960 pairs	Used as Wixarika-side text for the character language model.
Native Languages Huichol word list	small word list	Used only in the rejected filtering experiment, not in the final submitted configuration (Native Languages of the Americas, 2026).

Table 1: Data sources and their use.

training text during cross-validation.

The final post-processing stage is audit-based candidate cleanup. The pipeline stores all Spanish and Wixarika candidates in `predictions.audit.jsonl`. If the selected Wixarika output shows clear degeneration, such as repeated plus-marked fragments or repeated tokens, the cleanup script chooses a better already generated candidate for the same image. It does not manually write captions and does not use test references. The cleanup step was automatic and candidate-constrained. It never created a new caption from scratch: for each affected test image, it selected one of the Wixarika candidates that had already been generated by the same pipeline for that image. No test references or manually written test captions were used. This makes the cleanup a deterministic post-selection step rather than manual test-set annotation. On the final test run, this changed 37 captions and reduced the heuristic bad-output count from 20 to 7 out of 201 predictions.

4 Methods

The final pipeline is: image → CLIP retrieval → Qwen3-VL Spanish caption candidates → official Spanish-to-Wixarika MT → Wixarika character-LM reranking → audit cleanup → JSONL submission.

4.1 Retrieval

We embed images with `openai/clip-vit-large-patch14-336`, a CLIP model (Radford et al., 2021), and retrieve the top $k = 4$ visually similar examples. The retrieval code uses `CLIPModel.get_image_features` and normalizes image vectors before nearest-neighbor search. Retrieved Spanish captions are inserted into the Qwen3-VL prompt as style and content support.

4.2 Spanish Caption Generation

We use Qwen/Qwen3-VL-8B-Instruct (Qwen Team, 2025) to generate Spanish caption candidates. Spanish is used as a pivot because the VLM is more reliable in Spanish than in Wixarika. The selected configuration requests eight candidates per image with `max_new_tokens=72`, temperature 0.85, `top_p=0.95`, and sampling enabled. The prompt asks for concise descriptions based only on visible content, with attention to people, clothing, textiles, art objects, ritual objects, music, labor, architecture, and actions.

4.3 Machine Translation

The official Sheffield-compatible translation model is used in both directions. Wixarika-to-Spanish translation provides retrieval text for development examples, while Spanish-to-Wixarika translation converts generated Spanish candidates into target-language candidates. The MT beam size is 5. To avoid memory and device-mixing issues on a single A100 40GB instance, Qwen3-VL and CLIP run on GPU, while the official MT subprocess can run on CPU when needed. We observed that the official model produced more useful outputs in the Spanish-to-Wixarika direction than in the Wixarika-to-Spanish direction. Therefore, Wixarika-to-Spanish translation was used only as an auxiliary step for building retrieval text, while final caption generation always used the Spanish-to-Wixarika direction.

System	Protocol	chrF++	Decision
Organizer baseline	Baseline report	∼17.7	Reference
Selected pipeline	Local 5-fold CV	19.1468	Final
Output filter + word list	Local 5-fold CV	18.9689	Rejected
Restrictive prompt	Local 5-fold CV	16.9403	Rejected
Decoding sweep	Local 2-fold screen	18.4174	Rejected

Table 2: Development validation and ablation results.

4.4 Reranking

Each image receives multiple Wixarika candidates. A character 5-gram language model scores target-language fluency, and the final score combines fluency, a length preference, candidate order, and a penalty for very short outputs. For final test inference, the language model is trained on all available non-test Wixarika text.

5 Experiments

We used deterministic 5-fold cross-validation over the 70 official pilot and development examples. Each fold contained 14 validation examples. Held-out examples were excluded from both the retrieval bank and the character language model. We used chrF++ for local validation, matching the automatic evaluation emphasis of the shared task.

5.1 Development Results

Table 2 reports the selected configuration and the main rejected variants. We mention the organizer baseline only as a reference point; the main comparison is among our own validated variants.

The decoding sweep was an early 2-fold screening run used to discard an unpromising setting under limited compute; it is reported for transparency but is not directly comparable to the 5-fold validation runs.

The selected pipeline achieved fold scores of 18.1857, 19.7800, 19.4415, 19.5702, and 18.7564. The filtering patch improved two folds but lowered the overall mean, as shown in Table 3. The selected pipeline was also more stable across folds, with a sample standard deviation of 0.6601 compared with 0.9309 for the filtering patch. The restrictive prompt reduced chrF++ substantially, suggesting that over-constraining the Spanish caption removed useful in-domain wording.

Fold	Selected	Filter	Delta
1	18.1857	19.1539	+0.9682
2	19.7800	19.0545	-0.7255
3	19.4415	20.1932	+0.7517
4	19.5702	18.8574	-0.7128
5	18.7564	17.5856	-1.1708
Mean	19.1468	18.9689	-0.1779

Table 3: Fold-level comparison between the selected pipeline and filtering patch.

5.2 Test Results

The final inference run produced 201 predictions for 201 Wixarika test images. In the official automatic Wixarika evaluation, our submitted v0 system obtained 19.1569 chrF++. In the final human evaluation, the team ranked third overall with a mean rating of 2.48 over 201 test images. This suggests that, despite competitive automatic chrF++, fluency, naturalness, and image-specific adequacy remained important limitations.

5.3 Error Analysis

We analyzed the final 201-line submission file and the saved audit logs. The submission itself was complete: all 201 test images received a non-empty prediction, all IDs were unique, and all required JSONL fields were present. The remaining errors were therefore generation-quality errors.

Table 4 summarizes the main automatic error categories observed in the final output. The most frequent issue was residual Spanish or untranslated lexical material. We found 35 predictions containing likely Spanish words or named expressions, such as words for streets, buildings, materials, murals, and other concrete objects. This suggests that the Spanish-to-Wixarika MT component often preserved source-language nouns when the relevant Wixarika lexical item was unavailable or uncertain.

A second issue was repetition and translation degeneration. Although the final cleanup stage replaced 37 suspicious predictions by selecting alternative candidates from the audit file, 17 final predictions still showed repeated words, repeated character sequences, or abnormal forms. These cases likely reduced human-perceived fluency. We also found three groups of exactly repeated captions, affecting five additional rows, which indicates that some outputs were too generic and insufficiently image-specific.

For example, `hch_111` preserved Spanish words such as *camioneta*, *calle*, *aceras*, and *baldosas*.

Error type	Count
Complete predictions	201 / 201
Empty predictions	0
Likely Spanish residue	35
Repetition / degeneration suspects	17
Over-long predictions	10
Exact duplicate caption groups	3
Extra rows affected by exact duplicates	5
Cleanup replacements before submission	37

Table 4: Automatic error analysis of the final Wixarika test submission.

Similarly, hch_234 retained words such as *mural*, *girasoles*, and *bordados*. Repetition also remained in some outputs, such as hch_246, where *katixex-exeiyat+* contains an abnormal repeated character sequence.

The audit logs also showed that some Spanish caption candidates were visually underspecified. When the Spanish description only mentioned generic people, rural scenes, or objects, the translated Wixarika caption could be fluent but not sufficiently tied to the image. This helps explain why the system was competitive under character-level chrF++ but weaker in image-specific adequacy and fluency. Overall, the main bottleneck was the interaction between generic visual descriptions, untranslated Spanish lexical items, and MT degeneration in the final Wixarika output.

6 Conclusions

We presented a translation-centered system for Wixarika image captioning. The selected system combines CLIP retrieval, Qwen3-VL Spanish caption generation, official Spanish-Wixarika translation, and Wixarika character-LM reranking. It achieved competitive automatic results and produced a complete 201-image submission, but the final ranking shows that Wixarika fluency and human-perceived naturalness remain the main areas for improvement.

Limitations

The system relies on Spanish as a pivot language, so visual details lost during Spanish captioning cannot be recovered by translation. The Wixarika language model is a surface-level character model and cannot guarantee semantic correctness. The cleanup step removes obvious degeneration but may still select semantically imperfect candidates.

Ethics Statement

This system is an assistive research prototype, not an authoritative Wixarika captioning tool. Automatic captions may be inaccurate or culturally inappropriate. We did not manually create test captions, and any real deployment should involve Wixarika speakers and community stakeholders.

References

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Edward Gow-Smith and Barry Haddow. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Native Languages of the Americas. 2026. Vocabulary in Native American Languages: Huichol Words. https://www.native-languages.org/huichol_words.htm. Accessed May 2026.

Qwen Team. 2025. Qwen3-VL-8B-Instruct. https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct. Accessed May 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Wixarika Research Center. 2026. Wixarika Research Center Online Archive. https://www.wixarika.org/. Accessed May 2026.

Bushi Xiao, Qian Shen, and Daisy Zhe Wang. 2025. From text to multi-modal: Advancing low-resource-language translation through synthetic data generation and cross-modal alignments. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages*, pages 24–35.

Yangchen Xie, Xinyuan Chen, Hongjian Zhan, Palaiahnakote Shivakumara, Bing Yin, Cong Liu, and Yue Lu. 2024. Weakly supervised scene text generation for low-resource languages. *Expert Systems with Applications*, 237:121622.

Culturally Grounded Image Captioning in Indigenous Languages with Vision-Language Models: Cascaded and Single-Stage Approaches

Mirelle Bueno
Motorola Mobility
mirellec@motorola.com

Sushil Garg
Motorola Mobility
sushilgarg@motorola.com

Abstract

Culturally grounded image captioning for under-resourced Indigenous languages is challenging due to severe data scarcity and the need to describe culturally specific visual content. This paper describes our submission to the AmericasNLP 2026 shared task, where we evaluate two architectural paradigms for caption generation across Bribri, Guaraní, Yucatec Maya, Wixárika, and Orizaba Nahuatl. First, we implement a cascaded system that combines a large vision-language model with a machine translation pipeline, showing that culturally contextualized, persona-based prompting improves over the official baseline in most comparable settings. Second, we develop a direct, end-to-end Single-stage approach by adapting PaliGemma 2 using LoRA fine-tuning, continued pre-training, and multilingual joint training. Our single-stage experiments show that, despite severe domain mismatch and reliance on synthetic training data, multilingual training and continued pre-training improve automatic chrF++ relative to single-language LoRA fine-tuning in some settings. Overall, cascaded pipelines remain the strongest among the evaluated approaches under current data constraints, while single-stage models remain a promising but currently data-limited path toward direct Indigenous-language image captioning.

1 Introduction

Recent progress in vision-language modeling has led to strong performance on image captioning benchmarks, especially in high-resource languages and domains with large-scale annotated data. However, culturally grounded image captioning for Indigenous languages remains substantially more difficult.

In this paper, we describe our system for the AmericasNLP 2026 shared task (Bui et al., 2026) on culturally grounded image captioning for under-resourced Indigenous languages. The task requires participants to generate natural-language image captions for five target languages: Bribri, Guaraní, Yucatec Maya (hereafter Maya), Wixárika, and Orizaba Nahuatl (hereafter Nahuatl). We approach the task through two paradigms. The first is a cascaded architecture in which a general-purpose vision-language model generates a Spanish caption, which is then translated into the target Indigenous language using a machine translation system derived from the AmericasNLP 2023 shared task submission (Gow-Smith and Sánchez Villegas, 2023) designed to translate from Spanish into eleven Indigenous languages. The second is a single-stage architecture that adapts a vision-language model to generate captions directly in the target language, avoiding an explicit intermediate translation step. Furthermore, to the best of our knowledge, this is among the first efforts to adapt a single-stage image-captioning system specifically for Indigenous-language caption generation in this shared-task setting.

Our experiments demonstrate that the cascaded persona-based system is the strongest among the evaluated approaches overall across the development and test sets. Specifically, on the test set, it outperforms the baseline by 1.46 chrF++ for Wixárika, 0.92 for Bribri, and 4.54 for Nahuatl, while underperforming the baseline for Guaraní. These results indicate that persona-based prompting is beneficial in most settings but remains sensitive to language-specific translation quality. Furthermore, we observe that employing a powerful vision-language model for caption generation does not inherently guarantee superior final quality, as performance remains heavily contingent upon the quality of the translation model.

For the Single-stage approach, we explored diverse training methodologies, including continued pre-training, LoRA supervised fine-tuning (Hu et al., 2021), and multilingual training. Our findings show that multilingual training improves over

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 248–256

single-language LoRA fine-tuning in our development experiments. This suggests that multilingual training may provide useful transfer for image captioning, although our experiments do not isolate transfer from the effects of increased training data volume. A similar phenomenon has been widely documented in multilingual neural machine translation (Aharoni et al., 2019; Iyer et al., 2024). For Wixárika, continued pre-training provided substantial gains over alternative techniques, narrowing the gap with the official baseline to 2.99 chrF++ points on the test set. These insights may support the development of more effective vision-language models (VLMs) for Indigenous languages, reducing dependence on translation steps or intermediate modules.

Consequently, the primary contributions of this paper are twofold: first, we describe the systems submitted to the AmericasNLP 2026 shared task and evaluate the impact of prompting on culturally contextualized caption generation; second, we present an initial empirical comparison of several VLM adaptation strategies aimed at integrating extremely low-resource languages previously unseen by VLMs. The datasets and code are available at: https://github.com/MirelleB/InclusionVLM

2 Task and Data Description

The shared task challenges participants to develop multimodal systems that generate culturally grounded natural-language descriptions of images in under-resourced Indigenous languages. Formally, given an input image I, the objective is to generate a caption $S = \{w_1, w_2, ..., w_n\}$ in a target Indigenous language L. In contrast to conventional benchmarks such as MS-COCO (Lin et al., 2014), which prioritize generic object detection, this task emphasizes cultural granularity. Models must move beyond superficial labeling (e.g., "a colorful craft"), where visually appropriate, to identify culturally specific entities, such as Wixárika ceremonial artifacts, requiring a deeper integration of visual features and community-specific semantic knowledge.

For the evaluation phase, the shared task adopted a two-stage ranking methodology. Initially, all submitted systems were ranked automatically using the chrF++ metric. Subsequently, the top five systems advanced to a human evaluation stage, where they were assessed against a fixed set of standardized criteria.

2.1 Data Description

The corpus curated for this shared task is designed to reflect the sociocultural heritage and daily practices of the participating Indigenous communities. In contrast to generic image-captioning benchmarks, this dataset prioritizes domain-specific semantic density and cultural authenticity. The primary tracks encompass Bribri, Guaraní, Maya, Wixárika and Nahuatl. For each language, the development set consists of approximately 50 high-quality, manually annotated image-caption pairs. A blinded test set is utilized for the final benchmarking process. To ensure the integrity of the evaluation and prevent overfitting to the test distribution, ground-truth references are withheld until the conclusion of the competitive cycle. The annotations exhibit a high degree of descriptive specificity that poses a challenge to standard multimodal architectures. While a conventional vision-language model might generate a generic description such as "a person sitting," the shared task dataset annotations frequently employ a specialized lexicon to denote specific traditional activities. Consequently, successful systems must move beyond broad visual categorization toward a fine-grained understanding situated within the community's cultural context.

2.2 Evaluation Metric

We evaluated all developed systems using the mean sentence-level chrF++ score (Popović, 2017) via sacreBLEU (Post, 2018) implementation across the development subset, as recommended by the shared task organizers.

3 System Description

Two distinct architectural approaches were implemented and evaluated:

(i) A Cascaded framework, which utilizes a frozen Vision-Language Model (VLM) to generate culturally situated captions, followed by a frozen machine translation (MT) model to translate the output into the target language; a similar configuration also served as the baseline of the shared task.

(ii) A Single-stage approach, involving the development of a unified, end-to-end model capable of simultaneously interpreting the image and generating culturally situated captions directly in the target language.

Although the cascaded system yielded results competitive with the baseline, we hypothesize that

the single-stage architecture offers potential long-term advantages for under-resourced Indigenous-language settings. By generating captions natively, these models avoid an explicit translation step at inference time and may reduce the semantic distortion and information loss introduced by intermediate translation pipelines.

3.1 Cascaded approach

The cascaded architecture utilizes a VLM to generate an initial caption, which is subsequently processed by an MT model for translation. This framework is frequently adopted in low-resource linguistic scenarios, particularly where foundation models lack exposure to the target languages during pretraining (Geigle et al., 2025). A key architectural advantage is its modularity, allowing for the independent substitution of either the VLM or the MT component. However, the use of machine translation can lead to a loss of cultural nuance (Venuti, 2008; Szymańska, 2017).

In our cascaded pipeline, which is similar to the official shared task baseline, Gemini 2.5 Flash (Gemini Team et al., 2025) generated the Spanish image captions. These captions were subsequently translated into the target Indigenous languages utilizing the Sheffield translation system developed for the AmericasNLP 2023 Shared Task (Gow-Smith and Sánchez Villegas, 2023) for Bribri, Guaraní, Wixárika, and Nahuatl; for Maya, we utilized the separately trained Spanish-to-Maya MT model described in Section 3.7.

The Spanish captioning phase was executed with hyperparameters set to a temperature of 0.01 and top-p of 0.01. We conducted an ablation study of the following prompting strategies. The prompts were instantiated separately for each target community by replacing the language- and culture-specific descriptors; Appendix A shows the Wixárika instantiation used in the prompt ablation.

i) Baseline: The official shared task prompt, which includes comprehensive instructions regarding the cultural context of the target languages.

ii) Persona-based: A significantly more concise prompt focused on cultural perspective. This approach uses an explicit cultural-role framing to encourage concise captions that include culturally relevant details. Because such prompting may also encourage over-specific interpretation, we interpret improvements primarily in terms of automatic chrF++ rather than validated cultural fidelity.

iii) Question-Answer Chain (QA Chain): Drawing inspiration from prior work (Ibrahim et al., 2025; Bai et al., 2025), this three-stage approach begins with the LLM generating five questions based on image elements. In the second stage, the model answers these questions by leveraging its internal knowledge of the target culture. Finally, the model synthesizes these inputs to produce a contextualized caption.

While the QA Chain strategy offers a more exhaustive extraction of cultural detail, it is susceptible to error propagation; any inaccuracies in the intermediate reasoning phase are carried through to the final caption and may remain undetected by standard automatic metrics.

These strategies were initially evaluated on the Wixárika development set. To further investigate potential performance gains, we also assessed the impact of Gemini 2.5 Pro, constraining this analysis strictly to the Persona-based approach.

As demonstrated in Table 1, the Persona-based approach outperformed the baseline by a margin of 1.5 chrF++ points. Interestingly, the results also revealed a 0.53-point drop in the score when employing Gemini 2.5 Pro. These results suggest that using the larger Gemini 2.5 Pro model does not necessarily translate into higher final chrF++ scores, likely because downstream MT quality can dominate the final output. Consequently, we adopted the Gemini 2.5 Flash in conjunction with the Persona-based approach for the final evaluation across all target languages.

Method	**chrF++**
Baseline	17.77
Persona + Gemini 2.5 Flash	19.25
Persona + Gemini 2.5 Pro	18.72
QA Chain	17.07

Table 1: chrF++ comparison between prompt strategies on the Wixárika development set.

Table 2 presents the evaluation results on the development and test sets for the persona-driven prompting strategy across the target languages. These findings demonstrate that the Persona-based prompting improves over the baseline on the development set for all comparable languages and on the test set for Bribri, Wixárika, and Nahuatl. However, it underperforms the baseline on the Guaraní test set, and no Maya baseline is available for direct comparison. We hypothesize that persona-based prompting improves chrF++ in most

comparable settings, suggesting that culturally oriented prompts may help, though automatic scores alone do not establish cultural fidelity.

Language	Approach	Dev	Test
Bribri	Baseline	7.57	7.01
	Persona-based	8.49	7.93
Wixárika	Baseline	17.77	16.91
	Persona-based	19.25	18.37
Maya	Baseline	–	–
	Persona-based	9.00	16.96
Guaraní	Baseline	20.82	20.13
	Persona-based	21.75	16.48
Nahuatl	Baseline	11.53	9.52
	Persona-based	15.68	14.06

Table 2: chrF++ scores for the cascaded baseline and persona-based prompting approach across target languages. Dashes indicate scores that were unavailable.

3.2 Single-stage approach

Our primary objective is to train a single-stage VLM capable of generating culturally situated captions in the target languages without relying on intermediate modules. While many high-resource image-captioning settings are now effectively handled by proprietary and open-source models, it presents a formidable challenge for extremely low-resource languages due to the acute scarcity of both monolingual corpora and task-specific datasets.

To address this challenge, we evaluated three distinct training strategies:

(i) LoRA fine-tuning: Training directly on image-caption pairs in the target languages.

(ii) Continued Pre-training: Adapting the VLM's language component to the target language through unsupervised learning, followed by fine-tuning on image-caption pairs.

(iii) Multilingual Joint Fine-tuning: Developing a single unified model trained simultaneously on image-caption pairs across Wixárika, Bribri, Guaraní, and Maya.

Each strategy serves a specific investigative purpose: approach (i) assesses the model's linguistic abilities in languages not seen during pre-training; approach (ii) quantifies the impact of continued pre-training on the language backbone; and approach (iii) investigates whether cross-lingual transfer is effective for image captioning.

All experiments utilized the PaliGemma 2 (3B) pre-trained model (Steiner et al., 2024). This architecture was selected for its competitive performance in image captioning relative to significantly larger models and for its integration within the Gemma ecosystem, which streamlines the optimization of the language module. For strategy (ii), our experiments were focused exclusively on Wixárika.

The following sections provide a description of the training and evaluation datasets, alongside the technical specifications of our training protocols.

3.3 Data preparation

While constructing massive image-captioning datasets for high-resource languages remains a significant undertaking, doing so for extremely low-resource languages presents a formidable challenge. In response to this data scarcity, existing literature frequently leverages machine translation models to synthesize captions across multiple target languages. Adopting a similar methodology, we utilized the winning system from the AmericasNLP 2023 Shared Task to translate captions from the Polaris dataset (Wada et al., 2024). Polaris was selected due to its status as a large-scale, human-annotated corpus specifically designed for evaluating image captioning systems, encompassing approximately 131,000 human judgments. To accommodate training time constraints and ensure data quality, we filtered the dataset to include only examples with a human-annotated score exceeding 0.75. This threshold reflects high confidence in the captions' quality, accounting for critical factors such as fluency, relevance, and descriptive granularity. The filtering yielded a training dataset of 13,562 samples. It is important to note that Polaris captions differ fundamentally from the shared task dataset in that they lack specific cultural context. Our primary objective in utilizing Polaris is to facilitate cross-modal alignment, enabling the model to map visual elements to their corresponding linguistic representations in the target language. However, a potential limitation of this approach is that the use of synthetic (translated) data may introduce cascading noise during the training phase.

3.4 Monolingual data

For the continued pre-training phase, Wixárika source data were aggregated from previous AmericasNLP shared tasks; specifically, 93,247 tokens were sourced from AmericasNLI (Ebrahimi et al., 2022) and 189,362 tokens from AmericasNLP 2023 (Ebrahimi et al., 2023). These corpora provide human-curated or human-translated text rather

than model-generated synthetic captions. In total, the monolingual corpus utilized in this study comprised 282,609 tokens for Wixárika; it is important to note that this corpus is extremely small relative to the data typically used for language-model adaptation, highlighting the severe resource constraints under which the model was trained and the importance of exploring external datasets.

3.5 LoRA Fine-tuning

The training was conducted over ten epochs using the Low-Rank Adaptation (LoRA) approach with a rank (r) of 8. We used LoRA for parameter-efficient adaptation while limiting updates to the base model parameters; specifically, we aimed to preserve the model's pre-established cross-modal capabilities while simultaneously extending its linguistic proficiency to previously unseen target languages. The optimization protocol utilized a learning rate of 1×10^{-5} and a batch size of 4. For the training objective, we implemented a concise instruction format: each input was prepended with the mandatory PaliGemma <image> token, followed by the Spanish prompt *Generar pie de foto*.

3.6 Continued Pre-training and VLM Alignment

The continued pre-training framework was executed in two distinct stages. The initial phase focused on domain adaptation of the language backbone—specifically the Gemma-2 2B model. This stage aimed to integrate knowledge of a previously unseen language into the model using the causal language modeling pretext task of next-token prediction. To prevent the degradation of existing pre-trained knowledge, we employed LoRA with a rank (r) of 16. Furthermore, we incorporated a replay buffer consisting of approximately 4% English monolingual data to serve as a regularization mechanism, ensuring the retention of the model's primary language capabilities. This phase was conducted with a batch size of 4 and a learning rate of 2×10^{-4} over three epochs, totaling 1,071 optimization steps. Following the adaptation of the language model, we proceeded to the cross-modal alignment phase. During this stage, the Vision Tower and Language Model were kept frozen, with the training objective concentrated solely on optimizing the multimodal projector. This alignment phase utilized the same instructional prefix, learning rate, and hyperparameter configuration as described in the previous LoRA fine-tuning section.

3.7 Multilingual Joint Fine-tuning

This approach extends the methodology described in the previous fine-tuning section by incorporating multiple Indigenous languages into a single training objective. The primary motivation behind this experiment is to investigate the potential for cross-lingual transfer and determine whether synergistic effects between different low-resource languages can enhance performance in the target task.

The training corpus comprised data from Bribri, Maya, Guaraní, and Wixárika. In accordance with our established fine-tuning protocol, we translated the curated Polaris dataset into each of these target languages. For Bribri, Guaraní, and Wixárika, we employed the winning translation system from the AmericasNLP 2023 Shared Task. However, because Maya was unsupported by this Sheffield-developed system, we trained an independent Spanish-to-Maya machine translation model to facilitate both cascaded inference and synthetic data generation.

To manage the multi-target nature of this stage, a multilingual prompting strategy was adopted. Each input was formatted with the <image> prefix, followed by the instruction: *'Generar pie de foto en [language]'*, where the placeholder was dynamically substituted with the corresponding target language of the training sample.

4 Results

Table 3 reports the performance of the proposed methodologies. Single-stage experiments were conducted for Bribri, Guaraní, Maya, and Wixárika; Nahuatl was evaluated only in the cascaded setting. For clarity, the Low-Rank Adaptation fine-tuning is denoted as LoRA SFT, the Continued Pre-training and VLM Alignment as CPT + Alignment, and the Multilingual Joint Fine-tuning as Multilingual SFT. The baseline values cited correspond to the official scores reported by the shared task organizers.

Our results indicate that, despite the significant domain shift between our training set and the shared task dataset—specifically regarding visual stylistic diversity and the required descriptive complexity—the trained models demonstrated promising but limited performance in both Bribri and Wixárika languages. On the development set, the best single-stage models trailed the baseline by 2.32 chrF++ for Bribri and 2.64 chrF++ for Wixárika. On the test set, the corresponding gaps were 4.47 for Bribri and 2.99 for Wixárika. Per-

formance was less competitive for Guaraní, where the best single-stage model remained 10.51 chrF++ below the development baseline and 12.52 chrF++ below the test baseline. For Maya, no official baseline was available, so the single-stage results should be interpreted without direct baseline comparison.

Furthermore, the Multilingual Joint Fine-tuning approach improves over standard LoRA fine-tuning on the development set, suggesting that multilingual training and resource sharing can improve direct caption generation under low-resource conditions.

These findings suggest that the models acquired some cross-modal mapping capabilities even under suboptimal data conditions. We posit that by refining the training corpora to include more culturally grounded samples and specifically incentivizing the model to prioritize Indigenous sociocultural nuances, the performance of these single-stage systems may narrow the gap to current baselines.

Language	Approach	Dev	Test
Wixárika	Baseline	17.77	16.91
	LoRA SFT	11.79	–
	CPT + Alignment	15.13	13.92
	Multilingual SFT	12.72	10.51
Bribri	Baseline	7.57	7.01
	LoRA SFT	2.71	–
	Multilingual SFT	5.25	2.54
Maya	Baseline	–	–
	LoRA SFT	7.69	–
	Multilingual SFT	9.00	9.13
Guaraní	Baseline	20.82	20.13
	LoRA SFT	7.65	–
	Multilingual SFT	10.31	7.61

Table 3: chrF++ scores for Single-stage approaches across target languages on the development and test sets. Nahuatl was not evaluated in the single-stage setting due to resource and development timeline constraints. Dashes indicate scores that were unavailable or configurations that were not evaluated.

5 Related Work

Recent vision-language models (VLMs) remain predominantly English-centric, struggling to capture culturally embedded concepts in low-resource settings. While prior works extend architectures to multiple languages (e.g., Maya, a multilingual VLM (Alam et al., 2025), M-MiniGPT4 (Han et al., 2026)), they primarily target broad multilingual coverage rather than culturally grounded Indigenous-language captioning. Furthermore, these efforts often rely heavily on translated data. Although translation provides a scalable foundation, models adapted with native multimodal pairs (e.g., Chinese CLIP (Yang et al., 2022), DanQing (Shen et al., 2026)) demonstrate that culturally situated data is crucial for deep semantic fidelity.

In the absence of native end-to-end models, cascaded pipelines combining high-resource VLMs with machine translation offer a practical alternative (Geigle et al., 2025). However, these pipelines risk compounding errors by generating generic descriptions (Lin et al., 2014) that strip away community-specific nuances. Unlike recent retrieval-augmented cultural captioning methods (Ibrahim et al., 2025), we operate under severe resource constraints. We therefore contrast cascaded prompting with single-stage PaliGemma adaptation strategies (e.g., LoRA, continued pretraining) to assess the viability of end-to-end culturally grounded captioning for Indigenous languages.

6 Conclusion

This paper presents the development of multimodal systems for generating culturally grounded, natural-language descriptions of images in under-resourced Indigenous languages within the AmericasNLP shared-task framework. We propose and evaluate two distinct methodologies: a cascaded pipeline comprising VLM-generated captions followed by a machine translation (MT) system, and a direct, single-stage VLM. In the cascaded setup, persona-based prompting improved over the official baseline in most comparable settings, suggesting that explicitly encouraging a culturally grounded perspective can improve automatic captioning scores.

Experiments with the single-stage VLM indicate that direct caption generation can be trained, but it remains substantially less competitive than the Cascaded approach under the present data constraints. On the development set, the best single-stage models remained 2.64 chrF++ below the Wixárika baseline and 2.32 chrF++ below the Bribri baseline, indicating partial but incomplete progress toward direct caption generation due to severe data scarcity and cross-domain mismatch. Single-stage VLMs exhibit potential, provided they are trained on extensive, culturally grounded datasets that reward community-specific semantic fidelity over generic

visual descriptions.

Overall, our findings highlight a practical trade-off: while cascaded systems currently yield stronger short-term performance than the evaluated single-stage systems, single-stage systems offer a direct approach to Indigenous-language modeling, albeit requiring more robust data curation and language adaptation. Our final shared-task submission ranked 7th overall, placing in the top five for three of five evaluated languages: Maya, Nahuatl, and Wixárika.

7 Limitations

This study is subject to several limitations. First, our findings are constrained by the shared task's restricted set of target languages and small development sets, which limits both statistical power and generalizability to other Indigenous languages. Second, reliance on synthetically translated training data from the Polaris dataset inevitably introduces translation artifacts and omits crucial community-specific cultural grounding, potentially restricting models to surface-level cross-modal alignment. Third, our evaluation relies predominantly on automatic chrF++ scores; while effective for measuring character-level similarity, this metric cannot adequately capture cultural fidelity, factual grounding, or terminological accuracy. Fourth, our prompting experiments explore a limited set of inference strategies, omitting advanced paradigms such as chain-of-thought or retrieval-augmented generation. Fifth, our persona-based and QA-chain prompts rely on explicit cultural-role framing. While this framing improved chrF++ in several settings, it may also encourage over-specific cultural interpretations or hallucinated cultural associations, and we did not conduct community or speaker validation of the generated captions. Finally, our architectural comparisons are confined to the cascaded Gemini pipeline and single-stage PaliGemma 2 models; consequently, these results may not generalize to broader vision-language architectures.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *Preprint*, arXiv:1903.00089.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S. Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S. M. Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth.S, Snehanshu Mukherjee, and Alham Fikri Aji. 2025. Behind Maya: Building a multilingual vision language model. *Preprint*, arXiv:2505.08910. Accepted at VLMs4ALL CVPR 2025 Workshop.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, Silvia Fernandez Sabido, Luis Samuel Santiago Melchor, Sotero Silverio, Robert Pugh, Raúl Vázquez, John E. Ortega, Arturo Oncevay, Rubén Manrique, Luis Chiruzzo, Rolando Coto-Solano, Elisabeth Mager, Shruti Rijhwani, David Ifeoluwa Adelani, Manuel Mager, and Katharina von der Wense. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-Solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2831–2881, Vienna, Austria. Association for Computational Linguistics.

Gemini Team et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Seung Hun Eddie Han, Youssef Mohamed, and Mohamed Elhoseiny. 2026. M-MiniGPT4: Multilingual VLLM alignment via translated data. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 11–16, Rabat, Morocco. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

George Ibrahim, Rita Ramos, and Yova Kementchedjhieva. 2025. CONCAP: Seeing beyond English with concepts retrieval-augmented captioning. *Preprint*, arXiv:2507.20411. Published as a conference paper at COLM 2025.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Quality or quantity? on data scale and diversity in adapting large language models for low-resource translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Hengyu Shen, Tiancheng Gu, Bin Qin, Lan Wu, Yuling Wu, Shuo Tan, Zelong Sun, Jun Wang, Nan Wu, Xiang An, Weidong Cai, Ziyong Feng, and Kaicheng Yang. 2026. DanQing: An up-to-date large-scale Chinese vision-language pre-training dataset. *Preprint*, arXiv:2601.10305.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. PaliGemma 2: A family of versatile VLMs for transfer. *Preprint*, arXiv:2412.03555.

Izabela Szymańska. 2017. The treatment of geographical dialect in literary translation from the perspective of relevance theory. *Research in Language*, 15:61–77.

Lawrence Venuti. 2008. *The Translator's Invisibility: A History of Translation*, 2nd edition. Routledge, London; New York.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. *Preprint*, arXiv:2402.18091.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *Preprint*, arXiv:2211.01335.

A Cascaded Approach: Prompts

The following prompts are reported verbatim for reproducibility. They reflect the prompts used during the official submission and were not revised after evaluation; therefore, the results should be interpreted as measuring the behavior of these specific prompts rather than an endorsement of their cultural assumptions.

(A) PERSONA-BASED PROMPT

Eres wixárika (huichol), conoces y practicas la cultura.
Basándote en tu cultura, usa todo tu conocimiento para generar un pie de foto conciso, respetuoso y culturalmente preciso.
Ejemplo corto (PREFERIDO):
"Cuadro de estambre wixárika con patrones que representan visiones chamánicas. Los colores brillantes y diseños simbólicos son característicos del arte ceremonial contemporáneo."
Genera pies de foto concisos siguiendo este formato.

(B) QUESTION-ANSWER CHAIN
[Question generator prompt]

Considerando la imagen, crea 5 preguntas únicas relacionadas con la cultura.
Es importante mencionar que las imágenes pertenecen a la cultura wixárika (huichol), una comunidad indígena de la Sierra Madre Occidental, México.
Formato: Pregunta 1: , Pregunta 2: ,...

[Prompt that answers the questions]
Eres un experto en la cultura Wixárika (Huichol), una comunidad indígena de la Sierra Madre Occidental en México y debes responder correctamente a las preguntas; utiliza la imagen para contextualizar tu respuesta. Formato: pregunta 1: respuesta 1 / pregunta 2: respuesta 2

[Captioning prompt]
Teniendo en cuenta la imagen y las preguntas y respuestas relacionadas con la cultura Wixárika (Huichol), una comunidad indígena de la Sierra Madre Occidental en México, utiliza esta información para generar un pie de foto conciso de la imagen en una o dos frases.

Figure 1: Prompts used to evaluate the cascaded approach: (A) persona-based prompting and (B) chained question-answer prompting.

Schema-Constrained Image Captioning for Five Low-Resource Indigenous Languages

Diego Cuadros[1], **Nicholas Leeds**[1], **Amanda Avalos**[1],
Azul Alipzar-Velazquez[1], Jared Coleman[1]
Faezeh Dehghan Tarzjani[2], **Bhaskar Krishnamachari**[2]
[1]Loyola Marymount University, [2]University of Southern California
dcuadros@lion.lmu.edu, nleeds@lion.lmu.edu, aavalo12@lion.lmu.edu,
aalpiza1@lion.lmu.edu, jared.coleman@lmu.edu
dehghant@usc.edu, bkrishna@usc.edu

Abstract

We describe our submission to all five tracks of the AmericasNLP 2026 Shared Task on Cultural Image Captioning: Bribri, Guaraní, Yucatec Maya, Orizaba Nahuatl, and Wixárika. Our system is an LLM-assisted rule-based machine translation (LLM-RBMT) captioner. For each language, a coding agent reads the small development split and open-web linguistic references and writes a complete Pydantic grammar package with a closed vocabulary. At inference time, a vision–language model sees the image and the schema, emits a structured `SentenceList` under constrained decoding, and a deterministic Python renderer produces the surface string. The model never generates target-language tokens. The same architecture handles all five languages with no fine-tuning, no parallel corpora, and no human edits to the generated packages. On the official test set, the system placed first on human evaluation in Bribri and Orizaba Nahuatl, third on Yucatec Maya, and first on ChrF++ in Yucatec Maya. We suggest that a strength of the approach is that outputs are restricted to simple sentences that are grammatically correct by construction, modulo the correctness of the generated grammar itself.

1 Introduction

The AmericasNLP 2026 Shared Task on Cultural Image Captioning (Bui et al., 2026) asks systems to produce target-language captions for culturally grounded images in five Indigenous languages of the Americas: Bribri (`bzd`), Guaraní (`grn`), Yucatec Maya (`yua`), Orizaba Nahuatl (`nlv`), and Wixárika (`hch`). The setting is intentionally adversarial for end-to-end neural captioning: each language includes only fifty development examples, no parallel image–caption corpora exist for training, and public linguistic references are uneven in quality and coverage. Submissions are first ranked by ChrF++, and a top-five-per-language subset is forwarded to human evaluation by speakers.

A direct neural approach to this task asks a vision–language model (VLM) to caption the image in the target language. This works in proportion to how much of each target language the model has seen in pretraining, which for these five languages ranges from almost nothing to occasional Bible translations and Wikipedia stubs. The model can still produce fluent-looking strings, but it has no way to tell the user that it is guessing, and the user has no way to tell which fragments of the output are grounded.

We take a different approach. Rather than trying to make a VLM better at decoding into a language it does not know, we remove its ability to decode into the target language at all. For each language we generate a *language package*: a small Python module that encodes a closed vocabulary, a set of sentence templates, and a deterministic renderer. The VLM is then constrained to emit a structured object that conforms to the package's Pydantic schema, and the rendered string is produced by Python. Because the schema's lexical fields are typed as `Literal[...]` enums drawn from the package's vocabulary, the VLM cannot emit an out-of-vocabulary lemma; the constraint is enforced during decoding by the structured-outputs API. A narrow `proper_noun` field is the single escape hatch. The full source code for the captioner, the generator agent, and the five generated language packages is publicly available.[1]

Because none of the authors speak any of the target languages, the language packages themselves are also not written by us. Instead, they are written by a coding agent that reads the development split, web search, and reference packages, then writes a self-contained Python module that imports, validates, and renders against held-out captions. The agent is required to cite each external linguistic

[1] https://github.com/kubishi/americasnlp-2026-shared-task

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 257–263

reference inline in the package's own documentation, so that any claim made by the grammar can be traced back to the source the agent learned it from. The agent is the per-language adaptation cost; once it has produced a package, captioning is a single VLM call per image.

The resulting packages are *auditable*: a reviewer can read the vocabulary, sentence templates, and renderer source, and trace any surface caption back through the structured intermediate to the specific lexical and grammatical choices that produced it, each of which is cited in the package's inline documentation. Auditability is not the same as correctness, however. A grammar that the agent generated from public references and a fifty-row development split, with no speakers in the loop, can be fully traceable and still wrong. Any practical use of this architecture requires that the generated packages be validated by native speakers and community partners.

Contributions. This paper makes the following contributions.

1. A schema-constrained image-captioning architecture for low-resource Indigenous languages in which the VLM never generates target-language tokens. Lexical fields are typed as strict `Literal[...]` enums drawn from a closed vocabulary, a deterministic Python renderer produces the surface string, and a single narrowly prompted `proper_noun` slot carries genuine named entities through verbatim.
2. A reproducible agent-driven workflow for writing language packages from the shared task's development split and open-web references, applied uniformly to five typologically distinct languages with no per-language tuning by the authors.
3. Development-set comparisons against pipeline and direct-prompting baselines, together with the official test results, and an analysis of the dissociation between ChrF++ and human-evaluation ranks under which the system placed first on human evaluation in two of five languages and top-three in a third.

Outline. Section 2 situates the work relative to LLM-RBMT, low-resource MT, and structured decoding. Section 3 describes the language packages and the captioning pipeline. Section 4 states the experimental setup. Section 5 reports development and final test results. Section 6 discusses error patterns and the gap between ChrF++ and human evaluation. Section 7 states the system's limitations and Section 9 concludes.

2 Related Work

LLM-assisted rule-based MT. Our system is a vision-input extension of the LLM-Assisted Rule-Based Machine Translation (LLM-RBMT) paradigm introduced for endangered-language text-to-text translation (Coleman et al., 2024, 2026a). The original LLM-RBMT systems used an LLM to decompose English input into a structured intermediate that an expert-designed renderer mapped into the target language. We adopt the this approach, replacing the text encoder with a VLM, and the per-language hand engineering with an agent that writes the renderer.

Low-resource neural MT. A large literature evaluates massively multilingual neural MT models (e.g., NLLB-style systems) on low-resource languages (NLLB Team et al., 2024). These systems perform well when the target language has at least modest parallel data and degrade sharply otherwise. For the AmericasNLP languages, parallel data is either tiny or absent, and the organizer baseline (Qwen3-VL $\rightarrow$ NLLB) reflects this: it covers four of five target languages and reaches a four-language mean of 14.4 ChrF++.

Structured decoding and constrained generation. Recent work uses constrained decoding, structured outputs, and JSON schemas to force LLMs to emit valid objects (OpenAI, 2024). Our contribution is not the constraint mechanism itself but its use as the *only* interface to the target language. The schema is generated per language and encodes the grammar. The LLM is structurally unable to bypass it.

Agent-authored code and grammars. Coding agents have recently become capable enough to write self-contained Python modules that pass tests. We use this capability to push per-language engineering effort into the agent rather than the authors. Earlier LLM-RBMT systems required substantial human time per language (Coleman et al., 2026a). Ours requires only that the agent's run succeed.

3 System

The system has two parts. A *generator* runs once per language and emits a language package. A *cap-*

```
yaduha-{iso}/
  pyproject.toml
  yaduha_{iso}/
    __init__.py
    vocab.py
    prompts.py
```

Figure 1: Language package layout. `vocab.py` lists the closed vocabulary; `__init__.py` defines the Pydantic `Sentence` subclasses and the deterministic `__str__()` renderers; `prompts.py` carries any language-specific guidance the agent wanted the VLM to see.

tioner runs once per image and uses the package as its schema. Once a package exists, the captioner is a single VLM call followed by a deterministic render. Our packages follow the YADUHA[2] convention used by prior LLM-RBMT work for endangered languages (Coleman et al., 2024, 2026a).

3.1 Language Packages

Each language package is a standard Python package with the layout shown in Figure 1.

The package exports a `language` object with a small number of sentence classes (typically subject–verb, subject–verb–object, and a copular or stative variant where the language has one). Each class is a Pydantic model with morphological fields (person, number, tense or aspect, possession, and so on) and lexical fields whose types are `Literal[...]` enums drawn from `vocab.py`. The class's `__str__()` method renders the structured object into a surface string in the target language. Rendering is pure Python; no LLM runs at render time.

Strict-literal lemma typing. One important schema decision is that lexical fields are strict enums. If `vocab.py` lists thirty nouns, the noun-lemma field is typed `Literal["noun_1", ..., "noun_30"]`. The VLM's structured-outputs decoder refuses to emit any other string. The effect is that out-of-vocabulary nouns cannot leak into the output disguised as fluent target-language words.

Proper-noun escape hatch. Strict vocabularies are appropriate for common nouns. They are too strict for the named entities that cultural captions routinely contain (places, monuments, markets, civic institutions). Every `Noun` model therefore exposes a single `proper_noun: Optional[str]` field. The VLM may write a name into this field when the image contains one; otherwise it is left empty. The system prompt instructs the model not to abuse the escape hatch as a way around the strict lemma constraint.

3.2 Captioning Pipeline

The submitted pipeline is one VLM call per image:

$$\text{image} \rightarrow \text{VLM} + \texttt{schema} \rightarrow \texttt{SentenceList} \rightarrow \texttt{render()} \rightarrow \text{caption}.$$

The VLM (`gpt-5`) receives the image and the Pydantic schema generated from the language package and emits a `SentenceList` JSON object through the OpenAI structured-outputs API. Python then renders each `Sentence` in the list and concatenates the results.

3.3 Generator Agent

For each target language, the generator agent receives a small training slice of the available captions, the URLs of public linguistic references, one or more existing Yaduha packages as templates, and a validation harness. The harness exposes three tools: a package importer and schema-sanity checker, an English-to-structure smoke test, and a held-out comparison against a disjoint validation slice of captions. The agent reads, writes, and tests Python code in a loop until the harness reports a valid, self-consistent package.

The agent's bootstrap prompt forbids hardcoded shortcuts of the form "if the English input is exactly X, output Y." Every behavior must route through structured `Sentence` inputs and the rendering logic. The held-out validation captions are visible to the harness for scoring but not to the agent, and the harness reports only aggregate statistics. Absent native-speaker review, this is the best automatic check we have that the agent is generalizing rather than memorizing.

We implemented the generator using Anthropic's Claude Opus 4.7 model invoked through the Claude Code coding-agent toolkit, which exposes the file-system and shell tools needed to read references, write Python modules, and run the validation harness in a loop. The architecture does not depend on this specific agent.

4 Experimental Setup

Languages and data. We use the official AmericasNLP 2026 splits without modification. Table 1 summarizes sizes. For the generator, we partition

[2] https://github.com/kubishi/yaduha

Language	ISO	Dev	Test
Bribri	bzd	50	267
Guaraní	grn	50	110
Yucatec Maya	yua	50	212
Orizaba Nahuatl	nlv	50	200
Wixárika	hch	50	201
Total		250	990

Table 1: Dataset sizes. The agent sees only a deterministic 30-row slice of each development split. The remaining 20 rows are held out for validation.

each fifty-row development split into a deterministic thirty-row training slice (exposed to the agent as input captions) and a twenty-row validation slice that is hidden from the agent and used by the harness for held-out scoring. All development numbers reported below are computed over the full fifty-row split. Test scores come from the official test set.

Metric. Official ranking uses ChrF++ (Popović, 2015) in the first stage and human evaluation by speakers (top five systems per language) in the second stage. Per-row ChrF++ in our development tables is computed with `sacrebleu`'s default settings. The mean over a language is the unweighted mean of per-row scores. The test ChrF++ in Table 4 is the official score reported by the organizers.

Configurations. We evaluate the configurations listed in Table 2. The submitted configuration is the one-step `gpt-5` captioner with the minimal v3 prompt. The two pipeline configurations differ only in their structured-translator backend. The direct three-shot baseline asks a strong VLM (`claude-sonnet-4-5`) to produce a target-language caption given three in-context examples. The organizer baseline is the shared task's official Qwen3-VL → NLLB system. We report its published per-language ChrF++ where available.

5 Results

5.1 Development Set

Table 3 reports per-language mean ChrF++ on the fifty-row development split for every evaluated configuration. The final column is the unweighted mean over the four languages on which the organizer baseline reports a number. This is the column on which our system can be compared directly with the official baseline.

Three observations follow. First, the schema-constrained captioner (which never produces a target-language token by free generation) beats the direct-prompting baseline by 0.84 ChrF++ on average and beats the organizer baseline by 3.30 ChrF++ on the four-language mean. The improvement is largest on Yucatec Maya (uncovered by the baseline) and Orizaba Nahuatl. Second, the schema-constrained captioner trails the organizer baseline on Guaraní and Wixárika, both languages where NLLB has some training data and the package coverage is comparatively thin. Third, within our family, the simplest method wins: one-step schema-constrained captioning outperforms every two-step pipeline variant on average, including the variant that uses `gpt-5` as the translator.

Configuration	Description
`gpt-5` one-step v3 (submission)	Image + Pydantic schema → `SentenceList` → Python render. Minimal prompt.
`gpt-5` one-step v2	Same as above with the earlier, more verbose prompt.
Pipeline (sonnet + 4o)	VLM English caption → `gpt-4o` structured translator → render.
Pipeline (sonnet + 4o-mini)	As above with `gpt-4o-mini` translator.
Pipeline (sonnet + gpt-5)	As above with `gpt-5` translator.
Direct 3-shot	`claude-sonnet-4-5` caption in target language given three in-context examples.
Organizer baseline	Qwen3-VL → NLLB (no coverage for `yua`).

Table 2: Evaluated configurations.

5.2 Final Test Results

Table 4 reports the official results released by the shared-task organizers. ChrF++ ranking is computed over all submitted systems. Human evaluation considers the top-five-per-language ChrF++-ranked systems, with mean rating from speaker evaluators on a 4-point scale.

The system placed first on human evaluation in Bribri and Orizaba Nahuatl, third in Yucatec Maya, and did not qualify for human evaluation in Guaraní or Wixárika. ChrF++ ranks are mid-pack in three languages and last-but-one in two. The dissociation between ChrF++ and human rank is the central empirical finding of this paper and we discuss it in Section 6.

6 Analysis

6.1 ChrF++ Versus Human Judgment

The system's results split cleanly along the ChrF++/human-rating axis. On the three languages

Configuration	bzd	grn	yua	nlv	hch	Mean†
`gpt-5` one-step v3 (submission)	**11.93**	17.47	**27.71**	**26.64**	14.84	**17.72**
`gpt-5` one-step v2	11.73	19.17	26.02	24.79	16.20	17.97
Pipeline (sonnet + gpt-5)	10.84	15.60	17.44	17.27	16.16	14.97
Pipeline (sonnet + 4o)	10.96	15.63	17.02	16.79	14.73	14.53
Pipeline (sonnet + 4o-mini)	10.38	15.32	16.48	16.81	15.56	14.52
Direct 3-shot (sonnet 4.5)	9.43	18.37	18.86	21.56	**18.14**	16.88
Organizer baseline (Qwen3-VL → NLLB)	7.57	**20.82**	—	11.53	17.77	14.42

Table 3: Development-set per-language mean ChrF++ (N=50 per language). **Bold** marks the best score per column. †Mean is over the four organizer-comparable languages (`bzd`, `grn`, `nlv`, `hch`); `yua` is excluded because the organizer baseline does not cover it.

Language	RAN	ChrF++ score	ChrF++ rank	Human rating	Human rank
Bribri	3 / 2 / es-3	10.03	5 / 7	2.895	**1 / 5**
Yucatec Maya	5 / 4 / es-4 / en-4	**23.41**	**1 / 6**	2.892	3 / 5
Orizaba Nahuatl	6 / 4 / es-4 / en-3	21.00	2 / 7	3.465	**1 / 5**
Guaraní	6 / 1 / en-6 / es-2	16.90	7 / 8	—	DNQ
Wixárika	4 / 2 / es-4 / en-4	15.61	7 / 8	—	DNQ

Table 4: Official test results for team `yaduha`. RAN (Resource Abundance Notation) (Coleman et al., 2026b) reports the order-of-magnitude count of speakers / monolingual / bilingual partners for each language. **Bold** marks first-place finishes. "DNQ" indicates that the system's ChrF++ rank did not place it in the top-5-per-language subset that was forwarded to human evaluation.

where it qualified for human evaluation, it took two firsts and a third. On the two where it did not qualify, its ChrF++ rank was near the bottom of the field. This pattern is consistent with the architecture's design intent: outputs are grammatical by construction (insofar as the package's grammar is correct), but the package's vocabulary is bounded, so character overlap with diverse human references is capped.

ChrF++ rewards character-level overlap with a single reference caption. A system that uses a closed vocabulary and a small set of sentence templates can be near-perfect on the constructions it covers, but it incurs a fixed cost whenever the reference uses a word or construction the package lacks. A system that fluently generates target-language strings, by contrast, will often score higher on ChrF++ even when the strings are grammatically or semantically broken, because the references they are scored against are themselves character n-grams. Human evaluators see through this in a way the metric does not. On the other hand, because the schema admits only a small set of sentence templates and a closed vocabulary, the captions the system produces are necessarily shorter and more uniform than what a free-form generator might write. The stylistic range a fluent speaker would draw on is unavailable to the system, and this is visible to human evaluators as well.

6.2 Where Coverage Hurts

The two languages on which the system failed to qualify, Guaraní and Wixárika, illustrate the cost of closed vocabularies, though we are limited in what we can say about why. None of the authors are linguists or speakers of any of the target languages, so the explanations below are not first-hand diagnoses but a post-hoc analysis produced by the same coding agent that wrote the packages, asked to read the failing predictions against the references and conjecture about the gap. For Guaraní, the agent attributes the failure to productive morphology (possessive prefixes and person-marked verbs) that its package modeled only partially, leaving surface strings as reasonable base forms but missing the inflectional patterns the references use. For Wixárika, the agent points to a richer property-predication system than its package captured. After we removed a degenerate `CopularSentence` type that had been producing tautologies of the form "X is X," the package lost its only path to express adjective-like content. If these explanations are correct, both gaps would respond to better packages rather than to larger models, but verifying them and identifying the gaps we are missing would require a post-mortem with linguists and speakers of the

affected languages, which we view as a valuable direction for follow-up work.

6.3 Where Constraint Helps

The Bribri result is the most striking case in this direction. The system's ChrF++ score placed it 5th of 7 submissions (just making the top-five-per-language cutoff for human evaluation) and yet speaker evaluation ranked it first. This is consistent with the picture sketched above: the captions were not the most fluent in the field, but the constrained-generation mechanism kept them grammatical and prevented the model from producing fluent-looking but ungrammatical/incorrect strings that the evaluators would have rejected, and that appears to have weighed more heavily in speaker judgment than character overlap weighed in the metric. Orizaba Nahuatl tells a milder version of the same story (2nd on ChrF++, 1st on human evaluation). The natural counterfactual question concerns Guaraní and Wixárika, where ChrF++ ranks of 7th out of 8 placed the system below the cutoff for human evaluation. We cannot say how those captions would have fared under speaker review, but the Bribri result is at least suggestive that the bottom of the ChrF++ table is not necessarily a reliable predictor of the bottom of the human-evaluation table.

6.4 One-Step Versus Two-Step

In the pipeline configurations, a VLM first produces a free-form English caption of the image, and a text model then maps that English caption into the structured `SentenceList` that the renderer consumes. The problem with this approach is that information is lost at each handoff: the VLM commits to an English description before knowing which of the package's lemmas and sentence types are available to express it, and the downstream translator has to recover the structured intermediate from text that may no longer carry the necessary cues. The one-step approach avoids this by letting the VLM select the structured fields directly, conditioned on the image. Table 3 shows that on every language except Wixárika, the one-step pipeline beats every two-step variant, including the variant that uses the same `gpt-5` model as the structured translator.

7 Limitations

We note four limitations of the system.

Package coverage is a hard ceiling. A schema-constrained captioner cannot produce a concept the package does not express. For languages where the package's vocabulary or morphology is thin, the system can produce a grammatical and fully auditable output that nevertheless misses what the reference caption says. The fix is better packages, which ultimately requires more time with the agent or (preferrably) with speakers.

ChrF++ underestimates the architecture. The dissociation between ChrF++ rank and human rank in Table 4 is suggestive, not conclusive: it could reflect a general property of schema-constrained systems, or it could reflect specifics of this particular shared task. Either way, the system is at a structural disadvantage on the metric used for the first ranking stage.

No native-speaker review of packages. We did not have native-speaker validation of the agent-generated packages during the shared-task window. The official human evaluation is the strongest external check we have, and it is positive on three of five languages, but a serious deployment of this architecture would require collaboration with speakers.

Frontier-model dependence. The submitted configuration uses `gpt-5` as the VLM. We evaluated open-weight VLMs (`qwen2.5-vl:32b`) during development and found that they satisfy strict schemas less reliably without fine-tuning, especially when vocabularies grow large.

8 Ethics Statement

This work concerns Indigenous languages and culturally grounded images. The system is designed so that errors surface in the output rather than being hidden: when the package cannot express a concept, the constrained-output mechanism leaves a visible gap rather than producing a fluent-looking guess. Aditability is not the same as correctness, though, and we did not have native-speaker validation of the agent-generated packages during the shared task. The captions produced by this system should be treated as experimental.

Any deployment of this architecture beyond an development setting should involve close collaborations with speakers and community members. The agent's web search consumes public linguistic references whose authorship includes both community-authored materials and secondary

summaries. Those resources should be cited and weighted accordingly in any published package.

9 Conclusion

We presented a schema-constrained image-captioning system for five low-resource Indigenous languages in which a vision–language model never generates target-language tokens. Surface strings are rendered deterministically from a structured intermediate whose lexical fields are typed as strict enums drawn from an agent-authored language package. The same architecture handles all five languages with no fine-tuning, no parallel data, and no human edits to the packages.

On the official shared task, the system placed first on human evaluation in Bribri and Orizaba Nahuatl, third in Yucatec Maya, and first on ChrF++ in Yucatec Maya. It did not qualify for human evaluation in Guaraní or Wixárika, where ChrF++ ranks were weak. We read this dissociation as evidence that schema-constrained systems are penalized by character-overlap metrics and rewarded by speaker evaluation, and we think the architecture's value is most visible in the latter.

References

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on image captioning in Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Jared Coleman, Ruben Rosales, Kira Toal, Diego Cuadros, Nicholas Leeds, Bhaskar Krishnamachari, and Khalil Iskarous. 2026a. Comparing LLM-based translation approaches for extremely low-resource languages. In *Proceedings of the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT)*. Association for Computational Linguistics.

Jared R. Coleman, Tainã G.D. Coleman, and Bhaskar Krishnamachari. 2026b. RAN: Resource abundance notation for languages in NLP. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OpenAI. 2024. Introducing Structured Outputs in the API. https://openai.com/index/introducing-structured-outputs-in-the-api/.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

USP at AmericasNLP 2026 Shared Task: Culturally-Aware Image Captioning for Indigenous Languages via Vision-Language Models and Fine-Tuned Neural Machine Translation

Rafael M. Fernandes
University of São Paulo (USP)
São Paulo, Brazil
rafael.macario@usp.br

Abstract

We describe the USP system for the AmericasNLP 2026 Shared Task on Culturally Relevant Image Captioning for Indigenous Languages, covering Guaraní (grn), Maya Yucateco (yua), Nahuatl (nah), Wixárika (hch), and Bribri (bzd). We propose a two-stage cascade: Qwen3-VL-8B-Instruct (Bai et al., 2025) generates Spanish captions via language-specific cultural prompts; language-specific fine-tuned NLLB-200-distilled-600M (NLLB Team et al., 2022) models then translate them into each target language. We train on AmericasNLP 2023 data (Ebrahimi et al., 2023) augmented with public parallel corpora. Our system achieves competitive results, including **3rd place in Guaraní human evaluation** (2.41/5.0) and 5th in Bribri (1.09/5.0) among 8 teams. We also report that NLLB-200-distilled-600M silently lacks vocabulary entries for Bribri and Maya Yucateco, producing English output without error.

1 Introduction

Culturally relevant image captioning for Indigenous languages requires understanding not just what is depicted, but what it *means* within a specific community (Yun and Kim, 2024; Gao et al., 2025). Standard captioning benchmarks are predominantly Western-centric (Lin et al., 2014), and even state-of-the-art VLMs exhibit substantial cultural blind spots for non-Western communities (Burda-Lassen et al., 2025; Romero et al., 2024; Lupascu et al., 2025). A picture of a ceramic vessel might be described as "a clay pot" when the culturally meaningful description is *guampa para tereré*—a vessel central to Guaraní identity.

The AmericasNLP 2026 Shared Task on Culturally Relevant Image Captioning (Bui et al., 2026) provides images from five Indigenous communities: Guaraní (Paraguay, Brazil), Maya Yucateco (Mexico), Nahuatl (Mexico), Wixárika (Mexico), and Bribri (Costa Rica). These languages have been central to AmericasNLP MT tasks since 2021 (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025) and educational materials creation (Lupicki et al., 2025; Vasselli et al., 2025), providing established corpora on which our system builds.

We present a two-stage cascade (Figure 1): a VLM generates culturally-prompted Spanish captions, which are then translated per language by fine-tuned NLLB-200. Using Spanish as a pivot (Utiyama and Isahara, 2007) is natural for our setting: it is the dominant contact language for all five communities, all major parallel corpora are Spanish-indexed, and recent low-resource captioning systems adopt the same strategy (Oduwole et al., 2026; Jain et al., 2021).

Our contributions are:

1. A two-stage cascade covering all five shared task languages, with language-specific cultural prompting requiring no VLM fine-tuning.
2. Documentation of a previously unreported failure mode: NLLB-200 silently generates English for Bribri and Maya Yucateco due to missing vocabulary tokens.
3. Empirical evidence that domain mismatch between general-domain MT corpora and image captions can outweigh gains from language-specific fine-tuning.

2 Related Work

Cultural captioning. Yun and Kim (2024) propose CIC, a VQA+LLM framework that elicits cultural elements (traditional clothing, ritual objects) to enrich captions. Karamolegkou et al. (2024) show that cultural prompting improves *human-judged* quality even when ChrF++ declines—suggesting automatic metrics undervalue culturally enriched outputs. Buettner et al. (2025) use multimodal recaptioning with native-speaker examples to correct English-centric perceptual bias. Gao et al. (2025) construct MELLA, a dataset of cultur-

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 264–271

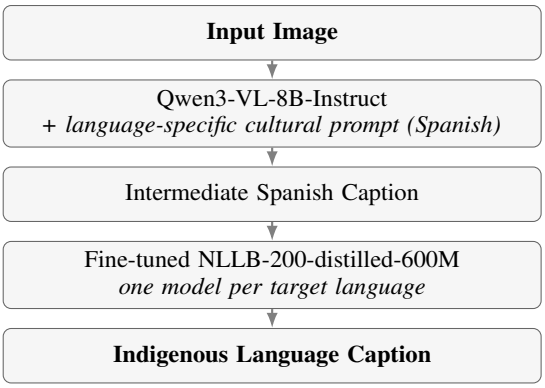

Figure 1: Two-stage cascade. Stage 1 generates a culturally-prompted Spanish caption; Stage 2 translates it into the target Indigenous language.

ally grounded "thick descriptions" for low-resource MLLMs.

Cascade MT for low-resource captioning. Jain et al. (2021) show that combining image-text with bitext training (MURAL) compensates for sparse caption data in low-resource languages. Oduwole et al. (2026) apply a nearly identical NLLB-based pivot cascade to African languages, finding that while the architecture is effective, domain mismatch between MT training data and image captions is the primary limitation—a finding we corroborate.

Cultural bias in VLMs. VLMs trained on Western-centric data fail substantially on culturally specific visual content (Burda-Lassen et al., 2025; Cao et al., 2024; Lupascu et al., 2025). The CVQA benchmark (Romero et al., 2024) quantifies this gap across 30 countries and 31 languages. Standard metrics like ChrF++ do not capture cultural correctness (Yun and Kim, 2024; Burda-Lassen et al., 2025); fine-tuned semantic embeddings (LaBSE) correlate better with human judgments for Guaraní and Bribri (Krasner et al., 2025).

AmericasNLP community. NLLB-200 fine-tuning has been the dominant approach in recent AmericasNLP MT tasks for our target languages (DeGenaro and Lupicki, 2024; García Gilabert et al., 2024; de Gibert et al., 2025). LLM prompting with community-specific context is competitive for Guaraní, Maya, Bribri, and Nahuatl (Lupicki et al., 2025; Vasselli et al., 2025). Bribri morphology presents particular challenges for sequence-to-sequence models (Anderson et al., 2025).

Language	Key cultural categories in prompt
Guaraní	*Tereré*, *ñandutí*, *tatakua*, traditional foods (chipa, sopa paraguaya), mythology (Pombero, Jasy Jatere), Jesuit missions
Maya	Henequén, *huipil*, *cenotes*, Hanal Pixán, Cochinita Pibil, Kukulkán, Chaac, archaeological sites
Wixárika	Peyote (*hikuri*), *nierika* yarn painting, beadwork (*chaquira*), *mara'akame*, Wirikuta pilgrimage
Nahuatl	Milpa, Quetzalcóatl, Day of the Dead, mole, Voladores de Papantla, copal, temazcal
Bribri	Cacao, *Sibö*, conical housing, *sukia* healer, chicha, Talamanca territory

Table 1: Cultural vocabulary categories targeted by each language-specific prompt. Full prompts in Appendix A.

3 System Description

3.1 Stage 1: Culturally-Prompted Visual Captioning

We use **Qwen3-VL-8B-Instruct** (Bai et al., 2025) to generate a Spanish description for each image (resized to 384×384 px). We design a *language-specific cultural prompt* for each of the five languages, following the approach of Yun and Kim (2024). Each prompt is written in Spanish and instructs the model to generate a concise (2–4 sentence) culturally-aware caption while foregrounding community-specific objects, practices, and symbols.

Table 1 shows the cultural vocabulary categories included in each prompt. The complete prompts are provided in Appendix A.

For Guaraní, we iterated over two prompt versions on the development set: an initial version (V1) with categorical vocabulary lists, and a refined version (V2) that additionally includes three few-shot examples of culturally correct captions. V2 was used for the final test submission; all other languages used a single prompt version with no few-shot examples.

Inference parameters: `do_sample=False`, `max_new_tokens=128`, `repetition_penalty=1.3`. All inference runs on NVIDIA T4 GPUs (Kaggle environment). The VLM is unmodified; no fine-tuning is performed.

Language	ANLP23	Total
Guaraní[†]	∼14k	∼46k
Nahuatl[‡]	∼16k	∼36k
Wixárika[§]	∼9k	∼16k
Bribri[¶]	∼7.5k	∼9.2k
Maya[‖]	—	∼11.8k

Table 2: Training data. ANLP23 = AmericasNLP 2023 (Ebrahimi et al., 2023). Extra corpora: [†]Jojajovai (Chiruzzo et al., 2022); [‡]Axolotl (Gutierrez-Vasques et al., 2016); [§]Pywirrarika (Mager et al., 2018); [¶]Feldman et al. (Feldman and Coto-Solano, 2020); [‖]Iikim Translator (Rangel and Kobayashi, 2024) (no ANLP23 for Maya).

3.2 Stage 2: Language-Specific Neural Machine Translation

We fine-tune **NLLB-200-distilled-600M** (NLLB Team et al., 2022) separately for each language, following the dominant approach in recent AmericasNLP MT tasks (DeGenaro and Lupicki, 2024; García Gilabert et al., 2024). For each language we evaluate both the fine-tuned and the base NLLB model on the development set and select the best performer.

Training data. We start from the **AmericasNLP 2023 parallel corpus** (Ebrahimi et al., 2023) and augment with additional public resources (Table 2). All data is Unicode NFC-normalized and whitespace-cleaned.

Vocabulary issues: Bribri and Maya. We discover that `bzd_Latn` (Bribri) and `yua_Latn` (Maya Yucateco) are absent from NLLB-200's vocabulary. Inspection of the pretrained tokenizer configuration confirms that neither identifier appears in the special tokens map or the sentence piece vocabulary. Querying either token returns id 3 (`<s>`), so the base model silently generates English. We add `yua_Latn` as a new special token (id 256204) for Maya, followed by `model.resize_token_embeddings()` to extend the embedding matrix; the new token's embedding is randomly initialized and trained from scratch. For Bribri, `bzd_Latn` was similarly added and the embedding matrix resized before fine-tuning. This is consistent with vocabulary gaps documented by Lupascu et al. (2025) for other under-represented languages in large multilingual models.

Training configuration. For Wixárika, Nahuatl, Bribri, and Guaraní: 3 epochs; batch 1 + 32 gradient accumulation steps (effective batch 32); lr 2×10^{-5}; 200 warmup steps; weight decay 0.01; Adafactor (Shazeer and Stern, 2018); gradient checkpointing; `fp16=False`; max sequence length 64 tokens. Hardware: NVIDIA A100 (Google Colab Pro).

Maya used different settings due to platform constraints (Kaggle T4): 5 epochs with early stopping (patience 2); batch 4 + 4 gradient accumulation steps (effective batch 16); same lr and optimizer; `fp16=True`. Training data comprised the Iikim Translator corpus augmented with the 50 AmericasNLP 2026 development captions repeated 20 times to expose the model to the image caption domain; task rules explicitly permitted using the development set for training.

Lang.	Dev ChrF++		Sub.	Test	Human (test)	
	Base	FT		ChrF++	Rating	Rank
Guaraní	19.49	17.57	FT	19.73	2.41	3rd/5
Wixárika	2.48	12.50	FT	13.68	—	—
Maya	—	9.03	FT	10.83	—	—
Nahuatl	3.29	5.23	FT	9.49	—	—
Bribri	—	2.65	FT	10.95	1.09	5th/5

Table 3: Dev and test ChrF++ for the submitted (FT) system, and test human evaluation. "—" = not evaluated or did not qualify for human evaluation (Bui et al., 2026). Human ratings on a 1–5 scale; 8 teams participated.

Inference. Fine-tuned models: beam search (k=4), `max_length=256`. For Bribri, `repetition_penalty=2.5` and `no_repeat_ngram_size=3` suppress degenerate repetition loops (Holtzman et al., 2020) caused by the combination of small training data, domain mismatch, and Bribri's complex morphology (Anderson et al., 2025).

4 Results

Table 3 reports development ChrF++ (Popović, 2017) for both base and fine-tuned NLLB models, and test-set human evaluation scores. The shared task ranked systems first by ChrF++ on the test set; the top 5 per language advanced to human annotation (1–5 scale).

For Wixárika, fine-tuning yields a large improvement (+10.02 ChrF++) over the base model, likely because the base NLLB has almost no Wixárika coverage. For Bribri, only the fine-tuned model is viable due to the vocabulary issue described above.

Qualitative analysis. Table 4 shows real system inputs and outputs from our test submissions, il-

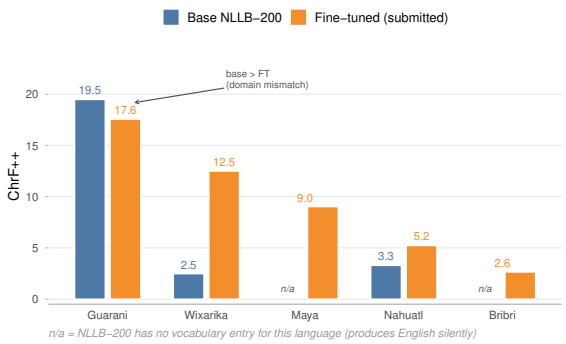

Figure 2: Dev ChrF++ for base vs. fine-tuned NLLB-200. Fine-tuning substantially improves Wixárika (+10.02) but slightly hurts Guaraní (domain mismatch). Bribri and Maya have no base score due to the vocabulary gap.

lustrating the pipeline in practice. Examples are unedited.

5 Analysis

Cultural prompting and the metric gap. Our 3rd-place Guaraní result was achieved without fine-tuning the VLM. For the other four languages, we used zero-shot cultural prompts; for Guaraní, we refined the Stage 1 prompt from a zero-shot version (used on dev) to a few-shot version with three culturally grounded caption examples (used on test), motivated by the observation that NMT fine-tuning slightly hurt dev ChrF++. The strong human evaluation result supports that culturally grounded prompting improved perceived caption quality. Notably, the Guaraní test ChrF++ (19.73) slightly exceeds the dev fine-tuned score (17.57), consistent with the few-shot V2 prompt providing richer Stage 1 captions than the zero-shot V1 used on dev. Karamolegkou et al. (2024) explain this pattern: cultural prompting improves human-judged quality even when ChrF++ declines, because culturally enriched descriptions include community-specific terms absent from generic references. Our dev ChrF++ scores likely underestimate cultural quality (Yun and Kim, 2024; Burda-Lassen et al., 2025); metrics capturing semantic and cultural fidelity (Krasner et al., 2025) would be more appropriate.

Fine-tuning does not always help. For Guaraní, the base NLLB model (19.49) outperforms the fine-tuned version (17.57) on dev. This suggests domain mismatch between our training data (general parallel text) and image captions (short, present-tense descriptions) can outweigh language-specific fine-tuning signal. Oduwole et al. (2026) and de Gibert et al. (2025) report similar findings. For Wixárika, where the base model has almost no coverage (2.48 ChrF++), fine-tuning improves substantially (+10.02).

Stage 1 — Spanish (VLM)	Stage 2 — target language
Guaraní, id grn_052 (tereré image)	
Tereré, bebida tradicional guaraní preparada con yerba mate, agua fría y una bombilla llamada "guampa". Se aprecian las calabazas nativas usadas como cántaras.	Tereré, mba'yrumýiva guaraníme ojejapóva yvyra'ípe, y ro'y ha peteî bombilla hérava guampa. Ko ta'angápe ojehecha umi ka'aguy guasúva ojeporúva cántaras.
Guaraní, id grn_053 (ñandutí image)	
Encaje ñandutí, tejido manual con hilos multicolores en forma circular, característico de Itauguá, Paraguay. Símbolo de la herencia textil guaraní.	Ñemboja'o ñandutí, ñongatupyrýva umi mba'apohára ha'éva Itauguá, Paraguái. Ha'e símbolo herencia textil guarani.
Bribri, id bzd_051 (cacao/Sibö image)	
Este es el cacao, una planta sagrada que honra a Sibö, Dios creador bribri. Sus semillas son parte de las ofrendas rituales y simbolizan vida y fertilidad.	s e'r kákoa, tsékea … Sibòa. Se' wöyök tö íyikaalaa ie'pa rö i wèna- awa- ataa. [Bribri; see submission bzd_051]
Nahuatl, id nlv_052 (failure: repetition collapse)	
Una mujer nahua descansa sobre los escalones frente a una puerta antigua, vistiendo su huipil tejido con patrones que reflejan la riqueza artesanal de sus raíces. Su postura tranquila evoca las prácticas cotidianas del campo veracricense, donde aún se cultiva milpa y se preparan tamales.	*[Repetition collapse: 119 of 200 Nahuatl outputs consist entirely of repeated Unicode variation- selector characters. No repetition penalty was applied for Nahuatl.]*

Table 4: Examples from our test submissions (unedited). Stage 1 correctly elicits culturally specific terms (*tereré*, *guampa*, *ñandutí*, *Sibö*) without VLM fine-tuning. The Nahuatl row shows a real failure: 119 of 200 Nahuatl test outputs collapsed to emoji repetition loops in Stage 2, as no repetition penalty was applied for that language.

NLLB vocabulary gaps are a silent failure. The absent `bzd_Latn` and `yua_Latn` tokens produce fluent English output with no warning. We recommend explicit verification before applying NLLB to any new language:

```
tokenizer.convert_tokens_to_ids(
```

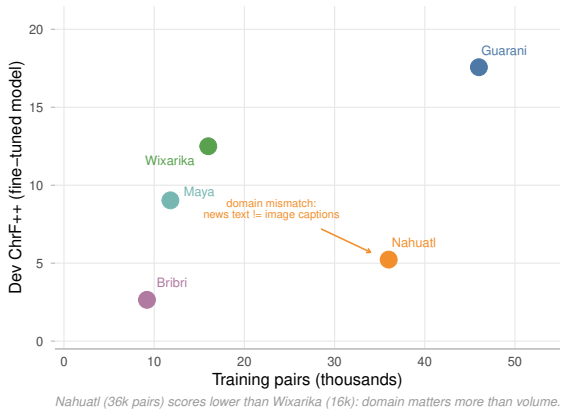

Figure 3: Training data size vs. dev ChrF++ (fine-tuned model). Nahuatl (36k pairs) underperforms Wixárika (16k pairs) due to domain mismatch between the Axolotl corpus (news text) and image captions.

```
"bzd_Latn") # -> 3 (<s>): NOT supported
```

Lupascu et al. (2025) document similar uneven coverage across large multilingual models.

Repetition collapse. Bribri's degenerate decoding (Holtzman et al., 2020) likely reflects the interaction of scarce training data (9.2k pairs), high domain mismatch, and morphological complexity (Anderson et al., 2025). A repetition penalty of 2.5 resolves the symptom; future work should address the cause via back-translation (Sennrich et al., 2016) or community-sourced captioning data (Gao et al., 2025).

6 Conclusion

We presented the USP system for the AmericasNLP 2026 image captioning shared task: a two-stage cascade combining culturally-prompted VLM captioning with per-language fine-tuned NLLB-200 NMT. Building on the AmericasNLP community's parallel corpora (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025) and cultural captioning literature (Yun and Kim, 2024; Buettner et al., 2025; Gao et al., 2025), our system achieves 3rd place in Guaraní human evaluation. We contribute two findings: (1) NLLB-200 silently lacks Bribri and Maya Yucateco vocabulary; and (2) cultural prompting may improve human-judged quality in ways ChrF++ does not reflect (Karamolegkou et al., 2024). Taken together, these findings suggest that large multilingual models fail Indigenous captioning not only due to data scarcity, but through hidden infrastructure mismatches: tokenizer coverage gaps, domain misalignment, and cultural grounding deficiencies. We release our prompts and training scripts.[1]

Ethics Statement

This work involves languages spoken by Indigenous communities whose cultural knowledge informed our prompting strategy. The cultural prompts were designed based on publicly available literature and existing NLP resources, without direct community consultation. We acknowledge that prompt-based cultural grounding risks misrepresenting or essentializing community practices, particularly if cultural elements are applied to images where they are not present. We do not claim that our system produces culturally authoritative captions, and we encourage community-led evaluation and correction of any outputs deployed in practice.

Limitations

Cultural prompts were authored without native-speaker consultation, limiting their cultural validity (Oduwole et al., 2026; Gao et al., 2025). NMT models trained on general-domain text face domain mismatch on image captions; submitted systems may not reflect the optimal configuration for all languages. Bribri and Maya models use non-standard vocabulary configurations. Human evaluation sample sizes vary by language. For Guaraní, the development and test sets used different prompt versions (V1 zero-shot vs. V2 few-shot), which introduces a confound: the observed human evaluation gain cannot be attributed solely to cultural prompting, as prompt format also changed.

Acknowledgments

The author thanks the AmericasNLP 2026 organizing committee. Compute via Google Colab Pro and Kaggle.

References

Carter Anderson, Mien Nguyen, and Rolando Coto-Solano. 2025. Unsupervised, semi-supervised and LLM-based morphological segmentation for Bribri. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 63–76, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

[1] https://github.com/rmaacario/americasnlp2026-usp

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 5 others. 2025. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

Kyle Buettner, Jacob T. Emmerson, and Adriana Kovashka. 2025. A multimodal recaptioning framework to account for perceptual diversity across languages in vision-language modeling. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1989–2006, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José Ake Pool, Reynaldo Alexander Couoh Martin, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on image captioning in Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Oliver Burda-Lassen, Aman Chadha, Sanjay Goswami, and Vikas Jain. 2025. How culturally aware are vision-language models? In *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 1–6.

Yong Cao, Wenlong Li, Jiaang Li, Yifei Yuan, and Daniel Hershcovich. 2024. Exploring visual culture awareness in GPT-4V: A comprehensive probing. *arXiv preprint arXiv:2402.06015*.

Luis Chiruzzo, Santiago Alemán, Santiago Góngora, Aldo Alvarez, Lili Ferrari, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guaraní-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2098–2107, Marseille, France. European Language Resources Association.

Ona de Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Dan DeGenaro and Tom Lupicki. 2024. Experiments in Mamba sequence modeling and NLLB-200 fine-tuning for low resource multilingual machine translation. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 188–194, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Luis Chiruzzo, Javier Garcia Gilabert, Manuel Mager, Arturo Oncevay, Robert Pugh, Shruti Rijhwani, and Katharina von der Wense. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 1–17, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 117–130, Mexico City, Mexico. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. A pipeline for spoken language documentation of Bribri. In *Proceedings of the WILDRE-5 Workshop at LREC*, Marseille, France. European Language Resources Association.

Yanbo Gao, Jianfei Fei, Nuo Chen, Ruonan Chen, Guofeng Yan, Yuanmeng Lan, and Bojun Shi. 2025. MELLA: Bridging linguistic capability and cultural groundedness for low-resource language MLLMs. *arXiv preprint arXiv:2508.05502*.

Javier García Gilabert, Aleix Sant, Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. BSC submission to the AmericasNLP 2024 shared task. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 143–149, Mexico City, Mexico. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Pompa. 2016. Axolotl: A large corpus of Spanish-Nahuatl parallel text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. European Language Resources Association.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask

representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antonia Karamolegkou, Phillip Rust, Ruixiang Cui, Yong Cao, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop (HuCLLM)*, pages 53–66, Bangkok, Thailand. Association for Computational Linguistics.

Nathaniel Krasner, Justin Vasselli, Belu Ticona, Antonios Anastasopoulos, and Chi-Kiu Lo. 2025. Machine translation metrics for indigenous languages using fine-tuned semantic embeddings. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: A survey. *arXiv preprint arXiv:2502.05568*. Accepted in Information Fusion.

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, and David Yarowsky. 2025. JHU's submission to the AmericasNLP 2025 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 105–111, Albuquerque, New Mexico. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mardiyyah Oduwole, Prince Mireku, Fatimo Adebanjo, Oluwatosin Olajide, Mahi Aminu Aliyu, and Jekaterina Novikova. 2026. AfriCaption: Establishing a new paradigm for image captioning in African languages. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 44–55, Rabat, Morocco. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Julio Rangel and Norio Kobayashi. 2024. Advancing NMT for indigenous languages: A case study on Yucatec Mayan and Chol. In *Proceedings of the 4th Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 155–164, Mexico City, Mexico. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4596–4604. PMLR.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT 2007*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Justin Vasselli, Haruki Sakajo, Arturo Martínez Peguero, Frederikus Hudi, and Taro Watanabe. 2025. Leveraging dictionaries and grammar rules for the creation of educational materials for indigenous languages. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 112–118, Albuquerque, New Mexico. Association for Computational Linguistics.

Youngsik Yun and Jihie Kim. 2024. CIC: A framework for culturally-aware image captioning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1634–1642.

A Cultural Prompts

Below are the full Spanish-language cultural prompts used in Stage 1 for each language. Each prompt instructs the model to generate a 2–4 sentence Spanish caption while foregrounding culturally relevant elements.

Guaraní (V1 — zero-shot, used on dev). *"Eres un sistema de subtitulado de imágenes diseñado para describir imágenes con relevancia cultural para el pueblo Guaraní de Paraguay. Tu tarea: Generar subtítulos concisos, respetuosos y culturalmente precisos (2–4 oraciones máximo). Contexto cultural a reconocer: tereré (bebida fría de yerba mate), chipa, sopa paraguaya, mbejú, ñandutí, ao po'i, jacarandá (tajy), mburucuyá, opy (casa ceremonial), tatakua, kambuchi, guampa, misiones jesuíticas, Jasy Jatere, Pombero, Kurupi, Luison."*

Guaraní (V2 — few-shot, used on test). *"Eres un sistema de subtitulado de imágenes para el pueblo Guaraní de Paraguay. Genera un subtítulo en ESPAÑOL, conciso y culturalmente preciso (2–4 oraciones).*

EJEMPLOS DE SUBTÍTULOS CORRECTOS:

Imagen de encaje artesanal colorido: "Ñandutí, encaje artesanal tradicional de Itauguá, Paraguay. Se teje a mano con hilos de colores formando patrones circulares como telas de araña."

Imagen de sopa con bolitas amarillas: "Vori vori, sopa tradicional paraguaya elaborada con bolitas de harina de maíz y queso. Es un plato típico de la gastronomía guaraní, especialmente consumido en invierno."

Imagen de vasija de barro con tela: "Kambuchi, vasija de barro tradicional guaraní, utilizada para transportar y conservar agua fresca."

CONTEXTO CULTURAL: tereré, chipa, sopa paraguaya, ñandutí, tatakua, kambuchi, guampa, Jasy Jatere, Pombero, Kurupi, Luison, misiones jesuíticas."

Maya Yucateco. *"Eres un sistema de subtitulado de imágenes para el pueblo Maya de México (Yucatán). Contexto cultural: henequén, huipil, milpa, cenotes, Chichen Itzá / Uxmal / Tulum, Hanal Pixán, jarana, pib/mucbipollo, sopa de Lima, cochinita pibil, Xtabay, Alux, Chaac, Kukulkán."*

Wixárika. *"Eres un sistema de subtitulado de imágenes para el pueblo Wixárika (Huichol) de México. Contexto cultural: peyote (hikuri), Wirikuta, nierika (tabletas rituales), cuadros de estambre, arte con chaquira, ojo de Dios (tsikiri), mara'akame, kuchuri, tatewarí (Dios del Fuego), Tatei Haramara, ceremonias Mitote, nawá/tejuino. Prefiere "Wixárika" sobre "Huichol"."*

Nahuatl. *"Eres un sistema de subtitulado de imágenes para el pueblo Nahua de México (Orizaba, Veracruz). Contexto cultural: mole, tamales, tlayudas, chinampas, Teotihuacán / Tenochtitlán, Quetzalcóatl, Xochitl, Día de Muertos (cempasúchil), huipil, copal, temazcal, milpa, Voladores de Papantla."*

Bribri. *"Eres un sistema de subtitulado de imágenes para el pueblo Bribri de Costa Rica. Contexto cultural: Talamanca, cacao (planta sagrada), Sibö (dios creador), clanes matrilineales, sukia (chamán), casa cónica circular, chicha de maíz/pejibaye, pejibaye, cestería, usure (ceremonia de muerte), Kéköldi, Cordillera de Talamanca."*

Nearest-Neighbor Retrieval for Indigenous Image Captioning

Justin Vasselli, Arturo Martínez Peguero, Shintaro Ozaki,
Frederikus Hudi, Haruki Sakajo, Taro Watanabe
Nara Institute of Science and Technology
vasselli.justin_ray.vk4@is.naist.jp

Abstract

This paper describes the NAIST submission to the AmericasNLP 2026 Shared Task on Indigenous Language Image Captioning. We investigate two approaches for generating captions in Bribri, Guaraní, Nahuatl, Wixárika, and Yucatec Maya. The first is a nearest-neighbor retrieval system that uses CLIP image embeddings to retrieve the most similar image from the development set and directly reuse its caption. The second is a generation pipeline that combines scene analysis, dictionary-grounded lexical planning, retrieved gloss templates, and interlinear gloss representations to constrain generation in low-resource settings.

The retrieval-based approach substantially outperformed the gloss-based pipeline under chrF++ evaluation and was competitive across all submitted systems, achieving first-place automated system rankings for Bribri and Wixárika and third place for Nahuatl. The gloss-based pipeline produced weaker automatic evaluation results and exposed problems with dictionary coverage, orthographic mismatches between resources, and unstable grammatical generation. Our results suggest that retrieval-based methods provide a strong baseline for low-resource captioning tasks when high-quality examples are available.[1]

1 Introduction

The AmericasNLP 2026 Shared Task on Indigenous Language Image Captioning focuses on generating captions for images in low-resource Indigenous languages. The task is challenging for current multimodal systems due to the limited amount of training data and the varying quality of generation available for these languages in current translation systems and large language models (Bui et al., 2026).

Our submission explores two approaches based on the retrieval of development set captions. The first uses nearest-neighbor retrieval. Given a query image, we retrieve the most similar image from the development set using CLIP embeddings (Radford et al., 2021) and return its caption directly. The second approach uses GPT-5.4 (Singh et al., 2026) to generate an interlinear gloss using retrieved captions and a dictionary-grounded lexical planning stage, then converts the gloss into the target language using language-specific conversion rules.

Our experiments showed that nearest-neighbor retrieval substantially outperformed the gloss generation pipeline on automatic evaluation metrics across all the languages we tested. Retrieval-based captioning produced fluent captions, while the pipeline often produced short or constrained outputs. We also observed differences across languages, suggesting that the difficulty of generating fluent captions varies depending on language support and dataset characteristics.

These results show that retrieval remains effective for low-resource captioning tasks, particularly under surface-overlap metrics such as chrF++ (Popović, 2015). However, the human evaluation results were more mixed across languages.

2 Nearest-Neighbor Retrieval

2.1 System Overview

We implement a nearest-neighbor retrieval approach that retrieves the caption associated with the most similar development image.

We evaluate four similarity strategies on the development set:

- CLIP similarity
- DINOv2 similarity
- English caption similarity
- Spanish caption similarity

[1] Code available at https://github.com/JVasselli/americasnlp2026-naist

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 272–278

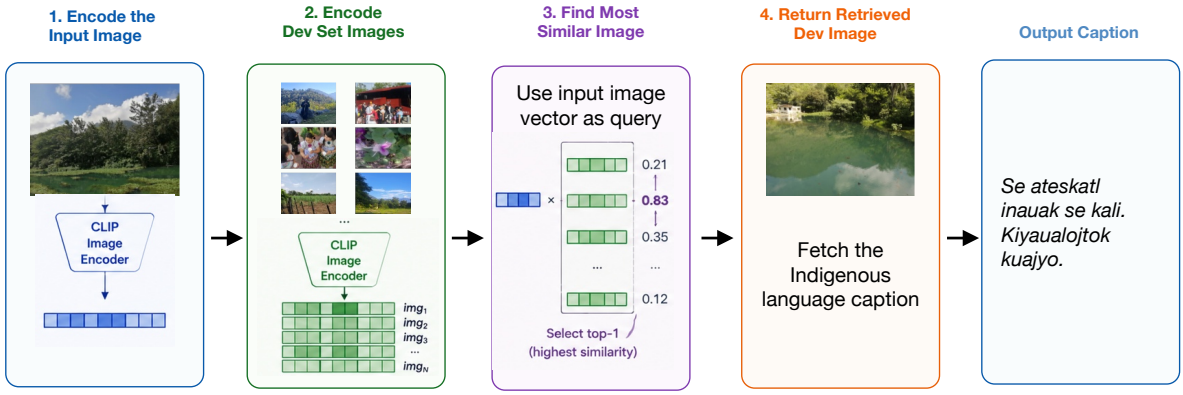

Figure 1: Overview of the nearest-neighbor retrieval system

Language	CLIP	DINOv2	Caption EN	Caption ES	Baseline
Bribri	**20.52**	19.91	18.86	18.61	7.57
Maya	**21.94**	21.65	18.68	19.66	—
Guaraní	**22.97**	22.66	22.33	21.16	20.82
Nahuatl	**25.08**	24.02	23.62	23.52	11.53
Wixárika	19.71	**19.83**	18.69	19.01	17.77

Table 1: Development set chrF++ scores for nearest-neighbor retrieval using different similarity strategies. English and Spanish denote caption similarity for each language. The official baseline did not include Maya results.

For image-based retrieval, we compute embeddings for all images and select the nearest neighbor using cosine similarity.

For CLIP-based similarity, we use OpenCLIP with a ViT-B/32 architecture pretrained on the LAION-2B dataset (Cherti et al., 2023). The resulting image embeddings have a dimension of 512, and images are processed using the default OpenCLIP preprocessing pipeline.

For DINOv2-based similarity, we use the DINOv2 large model (ViT-L/14) (Oquab et al., 2024). This model is trained with self-supervision. The resulting embeddings have a dimension of 1024, and images are processed using the default DINOv2 preprocessing pipeline (resize, center crop, and normalization).

We include caption-based retrieval to test whether the generated English and Spanish descriptions capture scene similarities missed by image embeddings. We first generate English and Spanish captions for each image in both the development and test sets using `gpt-4o-mini-2024-07-18`. These captions are then encoded using LaBSE (Feng et al., 2022), a multilingual sentence embedding model designed for cross-lingual retrieval. We compute cosine similarity between embeddings to identify the nearest neighbor.

For each query image, we select the caption associated with the most similar image in the development set. In the development set experiments, the query image is excluded from the candidate pool.

2.2 Results

As shown in Table 1, CLIP-based retrieval achieved the strongest development set performance for four of the five languages. DINOv2 retrieval remained competitive, particularly for Wixárika, where it slightly outperformed CLIP. Caption-based retrieval using generated English and Spanish descriptions consistently underperformed image-based retrieval, though the gap was smaller for Guaraní and Nahuatl. All retrieval approaches outperformed the official baseline on the languages where baseline scores were available. Based on development set performance, we selected the CLIP-based retrieval system for submission.

3 Gloss-based Pipeline Approach

3.1 System Overview

In addition to nearest-neighbor retrieval, we explored a gloss-based generation pipeline intended to constrain generation through dictionary grounding and intermediate gloss representations. Instead

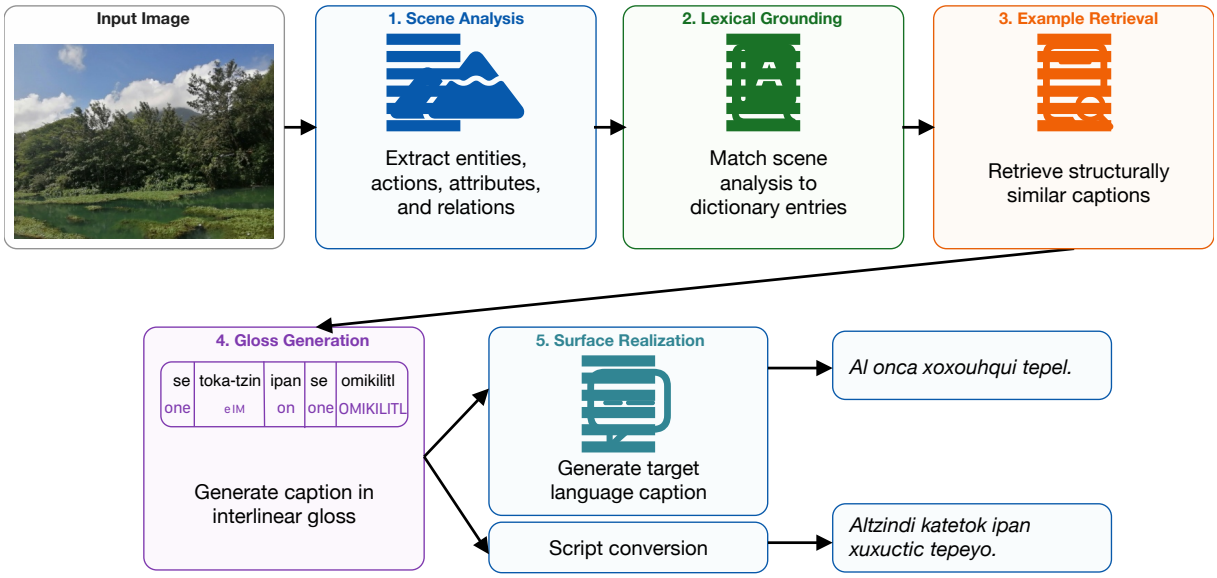

Figure 2: Overview of the gloss-based generation pipeline.

of generating captions directly from the image, the system first extracts structured scene information and maps concepts to attested lexical items before generating a caption gloss. This approach was designed to reduce hallucinations in low-resource Indigenous languages. The pipeline consists of five stages:

1. Scene analysis
2. Lexical grounding
3. Example retrieval
4. Interlinear Gloss generation
5. Surface Realization

Resources. To support the gloss-based pipeline, we constructed several synthetic linguistic resources for each language. First, we manually created a glossed caption resource from the provided target captions. Each entry contains a morphological segmentation, interlinear gloss, and approximate English translation. We created these glosses using publicly available dictionaries and grammar references, including grammar references for Guaraní (Estigarribia, 2020), Bribri (Jara, 2018), Yucatec Maya (Bolles and Bolles, 2001), and Nahuatl (Tuggy, 2004); they were intended for experimental generation support rather than authoritative linguistic annotation. Below is a sample entry:

```
"Guarani": "Chemandu'áta nderehe.",
"morphology": "che-mandu'a-ta nde=rehe",
"gloss": "1SG.INACT-remember-FUT 2SG=on",
"Spanish": "Me acordaré de ti.",
"English": "I will remember you."
```

From these glossed examples, we automatically extracted gloss-to-morpheme alignment tables that record how gloss units were morphologically realized in the examples.

```
...
3SG.front tu_táan 1
3SG.on y-óok'ol 1
ADD tak_xan 1
CL.tree kúul 2
DET le 26
...
```

For Guaraní and Nahuatl, which contain richer agglutinative morphology, we additionally created subword-level alignment resources.

```
...
3SG.OBJ ki 14
3SG.POSS i 5
3SG.POSS tlan 2
ABS tli 12
APPL li 1
APPL lia 2
CAUS lti 1
COND skia 1
...
```

We also collected bilingual dictionaries for each language and used them during lexical planning and gloss conversion, including the Bribri dictionary released by Vasselli et al. (2024), the Guaraní dictionary *Avañe'ẽ del Taragui* (Ministerio de Educación de la Provincia de Corrientes, 2022), the Yucatec Maya dictionary *Diccionario de Uso del Maya Yucateco* (Yoshida and Ucan Dzul, 2025),

Language	Systems	Nearest Neighbor	Gloss Generated	Gloss Converted	Baseline
Bribri	14	**19.37 (1)**	4.21 (13)	5.29 (11)	7.01 (10)
Maya	9	**15.80 (5)**	12.27 (6)	10.20 (8)	-
Guaraní	24	19.41 (18)	9.71 (22)	8.33 (23)	**20.14 (12)**
Nahuatl	12	**20.93 (3)**	10.09 (10)	15.34 (8)	9.52 (11)
Wixárika	14	**19.84 (1)**	-	-	16.91 (8)

Table 2: Test set results on automated metric. Number in the parentheses is the system ranking of all submitted systems.

Language	CLIP NN	Generated	Converted
Bribri	20.52	4.66	5.38
Maya	21.94	12.52	11.79
Guaraní	22.97	7.47	6.97
Nahuatl	25.08	10.13	14.23

Table 3: Development set chrF++ scores for the gloss-based pipeline variants (Generated, Converted) and CLIP-based nearest-neighbor (NN) retrieval.

the Wixárika dictionary by SIL International (SIL International, 2012), and the Nahuatl dictionary by SIL International (SIL International, 2002).

Step 1: Scene analysis. We first prompt GPT (`gpt-4o-mini-2024-07-18`) to analyze an input image and produce a structured scene representation. The representation contains entities, actions, attributes, and spatial relations, together with corresponding Spanish lexical items. An example scene representation is shown below:

```
{
  "attributes": ["green", "cloudy"],
  "main_entity": "water",
  "predicate_type": "existential",
  "secondary_entities":
    ["trees", "mountain", "sky"]
}
```

The system also produces Spanish lexical mappings, such as *green* → *verde* and *water* → *agua*.

Step 2: Lexical grounding. The extracted concepts are matched against digitized dictionaries. Concepts without dictionary matches are pruned from the scene representation. This stage produces a lexical plan containing only attested vocabulary. For example, the scene representation above may be reduced to concepts such as *green*, *water*, and *mountain* if no dictionary matches are found for the remaining items.

Step 3: Example retrieval. We retrieve candidate captions from the development set using manually assigned structure labels such as `ENTITY + STATE + LOCATION`. Retrieved examples provide interlinear glosses and morphological patterns for generation.

Step 4: Interlinear gloss generation. GPT-5.4 (`gpt-5.4-2026-03-05`) is then prompted to generate an interlinear gloss caption, together with a target-language caption. The prompt includes the image, the lexical plan, and retrieved gloss examples. The model is instructed to use only lexical items appearing in the dictionaries or retrieved examples.

For example, the system may generate the gloss:

```
water be-PROG on green mountain
```

Step 5: Surface realization. Finally, we convert the gloss into Indigenous language captions in two ways: LLM generation and rule-based script. The LLM is prompted to generate a caption immediately after the gloss and has all of the information from the captioning stage (image, scene plan, examples, dictionary entries). In the above example GPT-5.4 generated:

```
Al onca xoxouhqui tepel.
```

We additionally apply a deterministic gloss conversion to produce a second version of the caption. This stage combines dictionary matches, rule-based morphology, and morpheme alignments extracted from the synthetic gloss resources. For the previous example, the conversion stage produced:

```
Altzindi katetok ipan xuxuctic tepeyo.
```

3.2 Results

Table 3 shows the development set results for the gloss-based pipeline. Across all languages, both variants scored well below nearest-neighbor retrieval. Performance also varied across languages. The deterministic conversion stage improved results for Bribri and Nahuatl but slightly reduced performance for Maya and Guaraní. Due to incomplete language-specific resources and conversion rules, the gloss-based pipeline was not finalized for Wixárika and was therefore not included in these experiments.

Language	Reference	Pipeline
Bribri	14.92	7.40
Maya	11.28	6.02
Guaraní	21.64	6.52
Nahuatl	8.34	5.26

Table 4: Average number of tokens per caption for the reference captions and the gloss-based pipeline outputs.

Lang.	Avg rating	Sys. Ranking
Bribri	2.219	3
Maya	1.934	4
Guaraní	1.978	4
Nahuatl	1.220	4
Wixárika	3.790	1

Table 5: Test set results of nearest neighbor retrieval after human evaluation.

Lang.	Improved	Avg Δ	Min Δ	Max Δ
Bribri	32/50	0.72	-3.48	10.32
Maya	22/50	-0.85	-16.43	10.38
Guaraní	27/50	-0.24	-6.63	4.77
Nahuatl	43/50	4.88	-7.52	17.90

Table 6: Change in chrF++ score after deterministic gloss conversion relative to direct caption generation on the development set. "Improved" indicates the number of examples for which gloss conversion increased the chrF++ score.

4 Results and Discussion

4.1 Automatic Evaluation

Table 2 shows the official test set chrF++ results for our submitted systems. The nearest-neighbor retrieval system achieved the strongest performance of our submissions across all languages. The system ranked first overall for Bribri and Wixárika, third for Nahuatl, fifth for Maya, and eighteenth for Guaraní.

The gloss-based pipeline consistently produced lower chrF++ scores than nearest-neighbor retrieval. One reason is that the pipeline produced much shorter captions than the references, as shown in Table 4. The lexical planning stage aggressively pruned concepts without reliable dictionary matches, which reduced hallucinated vocabulary but also removed many scene details. As a result, generated captions often contained only a small subset of the entities and actions present in the reference captions.

We also observed several additional issues during gloss generation and lexical planning. The system relied heavily on dictionary matches and often ignored development set vocabulary. In some cases, valid dictionary matches were missed due to inflectional or orthographic variation, such as plural forms or differences in accent marking between the dictionaries and the shared task data. The gloss generation stage also occasionally produced grammatical constructions that did not appear in the retrieved examples or alignment resources.

Despite its simplicity, nearest-neighbor retrieval consistently produces fluent captions. Under surface-overlap metrics such as chrF++, this translates to stronger results than the constrained gloss generation pipeline.

4.2 Human Evaluation

Table 5 shows that the nearest-neighbor retrieval system achieved strong rankings on the automatic evaluation metric. The system ranked first for Wixárika, third for Bribri, and fourth for Maya, Guaraní, and Nahuatl. However, the corresponding human evaluation ratings were more mixed across languages.

This difference suggests that strong chrF++ performance does not always correspond to strong human evaluation performance. Retrieved captions are fluent and grammatical because they originate from attested examples in the development set, but the retrieved image may still differ from the query image in important ways. Surface-overlap metrics, such as chrF++, do not strongly penalize these semantic mismatches.

4.3 Gloss Conversion versus Direct Generation

Table 6 compares the direct caption generation against the rule-based gloss conversion script output on the development set. The effect of gloss conversion varied by language.

Gloss conversion improved the majority of examples for Bribri, Nahuatl, and Guaraní. However, only Nahuatl showed a substantial increase in the average chrF++ score after conversion. Guaraní and Maya both showed slight decreases on average despite improvements on many individual examples.

These results suggest that deterministic gloss conversion can improve lexical and morphological consistency when the generated gloss aligns well with the target language resources. However, errors introduced during lexical planning or gloss gen-

eration often propagate into the conversion stage. Since the conversion system relied heavily on dictionary lookups and synthetic alignment resources, that meant that mismatches in orthography, morphology, or gloss structure could lead to degraded outputs.

The relatively strong improvement for Nahuatl may also reflect differences in language support within current language models. Vasselli et al. (2026) found that large language models demonstrate comparatively stronger understanding of Guaraní and Nahuatl than several other Indigenous languages. However, the shared task uses Orizaba Nahuatl (`nlv`), while many multilingual resources and evaluations focus on broader Nahuatl varieties such as `nah`. These differences may have affected both gloss generation and gloss conversion quality.

5 Related Work

Early image captioning systems explored retrieval-based approaches that reused captions from visually similar images rather than generating captions directly. Farhadi et al. (2010) mapped images and captions into a shared semantic representation for caption retrieval, while Ordonez et al. (2011) demonstrated that nearest-neighbor image retrieval combined with direct caption transfer could produce competitive image descriptions using large captioned image collections.

More recent work has explored linguistic resources and explicit grammatical structures for low-resource language generation. Ginn et al. (2024) investigates gloss generation for endangered languages using large language models, demonstrating that intermediate linguistic representations can support generation for low-resource languages. Taguchi and Sproat (2026) shows that large language models can generalize grammatical patterns from linguistic descriptions and curated examples, motivating our use of retrieved gloss templates and structured prompting. Vasselli et al. (2024) applies dictionary-guided prompting and retrieved linguistic examples for educational material generation in Indigenous languages, demonstrating how lexical grounding and example retrieval can help constrain generation in low-resource settings. Our gloss-based pipeline combines lexical grounding, retrieved caption structures, and intermediate gloss representations to constrain generation.

6 Conclusion

We presented two approaches for the AmericasNLP 2026 Indigenous language image captioning shared task: a nearest-neighbor retrieval system based on CLIP image similarity and a gloss-based generation pipeline using lexical grounding and intermediate interlinear gloss representations.

Our experiments showed that nearest-neighbor retrieval substantially outperformed the gloss-based pipeline across most languages under chrF++ evaluation. Despite its simplicity, retrieval produced fluent captions drawn directly from attested examples in the development set and achieved strong rankings on the shared task leaderboard. In contrast, the gloss-based pipeline often produced short and overly constrained captions due to aggressive lexical pruning and limitations in the available linguistic resources.

The gloss-based pipeline nevertheless highlighted several challenges for low-resource Indigenous language generation, including orthographic variation across resources, sparse dictionary coverage, and difficulty in constraining grammatical generation. The mixed results from deterministic gloss conversion further suggest that improvements in lexical alignment and morphological normalization may be necessary before rule-based conversion can reliably improve generated outputs.

Overall, our results suggest that retrieval-based approaches remain competitive for low-resource captioning tasks, particularly when high-quality attested examples are available. However, the gap between automatic metrics and human evaluation shows that fluent retrieved captions are not always accurate. Future work may benefit from combining the fluency advantages of retrieval with stronger semantic grounding and more robust linguistic resource integration.

References

David Bolles and Alejandra Bolles. 2001. *A Grammar of the Yucatecan Mayan Language*. Foundation for the Advancement of Mesoamerican Studies, Crystal River, FL.

Minh Duc Bui, David Guzmán, Abteen Ebrahimi, Franklin Morales, Marvin Agüero-Torales, Raquel Insfrán, Cecilia González, Ramón Araujo, Luca Cernuzzi, Carlos Raul Noh Chi, Carlos Eduardo Tec Cahun, Sindi Estrella Poot Cohuo, Daniel Ricardo Benítez Chi, Santos Natanael Palomo Arévalo, Jessica Elizabeth Canul Canche, Deysi Aracely Poot Poot, Wendy Marleny Dzib Dzib, Eduardo José

Ake Pool, Reynaldo Alexander Couoh Martin, and 15 others. 2026. Findings of the AmericasNLP 2026 shared task on cultural image captioning for Indigenous languages. In *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, San Diego, California. Association for Computational Linguistics.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. Grammars of World and Minority Languages. UCL Press.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Carla Victoria Jara. 2018. *Gramática de la lengua bribri*. éditeur non identifié.

Ministerio de Educación de la Provincia de Corrientes. 2022. *Avañe'ẽ del Taragui: Diccionario guaraní-español, español-guaraní*. Ministerio de Educación de la Provincia de Corrientes, Corrientes, Argentina. Coordinación de Educación Intercultural Bilingüe.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

SIL International. 2002. Diccionario náhuatl de la sierra norte de puebla.

SIL International. 2012. Diccionario huichol–español.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, and 1 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Chihiro Taguchi and Richard Sproat. 2026. Creating conlangs to probe the metalinguistic grammatical knowledge of llms. *Preprint*, arXiv:2510.07591.

David H. Tuggy. 2004. Náhuatl: Lecciones para principiantes. Originally published in 1991. Electronic edition copyright 2004.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

Justin Vasselli, Arturo Mp, Frederikus Hudi, Haruki Sakajo, and Taro Watanabe. 2026. Measuring linguistic competence of LLMs on indigenous languages of the Americas. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 287–296, Rabat, Morocco. Association for Computational Linguistics.

Shigeto Yoshida and Angel Abraham Ucan Dzul. 2025. *Diccionario de Uso del Maya Yucateco*, primera edición corregida edition. Publicación independiente, México. Open Educational Resource (REA).

Findings of the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages

Minh Duc Bui[1] **David Guzmán**[2] **Abteen Ebrahimi**[3] **Franklin Morales**[4,5] **Marvin Agüero-Torales**[6,7,8] **Raquel Insfrán**[6,9] **Cecilia González**[6,9] **Ramón Araujo**[6,9] **Luca Cernuzzi**[6,9] **Carlos Raul Noh Chi**[24] **Carlos Eduardo Tec Cahun**[24] **Sindi Estrella Poot Cohuo**[24] **Daniel Ricardo Benítez Chi**[24] **Santos Natanael Palomo Arévalo**[24] **Jessica Elizabeth Canul Canche**[24] **Deysi Aracely Poot Poot**[24] **Wendy Marleny Dzib Dzib**[24] **Eduardo José Ake Pool**[24] **Reynaldo Alexander Couoh Martin**[24] **Silvia Fernandez Sabido**[25] **Luis Samuel Santiago Melchor**[11] **Sotero Silverio**[4] **Robert Pugh**[12,13] **Raúl Vázquez**[14] **John E. Ortega**[15] **Arturo Oncevay**[16] **Rubén Manrique**[17] **Luis Chiruzzo**[18] **Rolando Coto-Solano**[19] **Elisabeth Mager**[20] **Shruti Rijhwani**[21] **David Ifeoluwa Adelani**[2,22] **Manuel Mager**[23] **Katharina von der Wense**[3,1]

[1]Johannes Gutenberg University Mainz [2]Mila-Quebec AI Institute, McGill University [3]University of Colorado Boulder [4]Independent Researcher [5]Mẽkichawak [6]Centro Tecnológico en Ingeniería (CIDIT), Paraguay [7]Universidad de Granada, Spain [8]Fujitsu, Spain [9]Universidad Católica Nuestra Señora de la Asunción, Paraguay [11]Ximomachti [12]Indiana University [13]Mozilla Data Collective [14]University of Helsinki [15]Northeastern University [16]Pontificia Universidad Católica del Perú [17]Universidad de Los Andes [18]Universidad de la República, Uruguay [19]Dartmouth College [20]Universidad Nacional Autónoma de México [21]Google DeepMind [22]Canada CIFAR AI Chair [23]Universidad Iberoamericana, Mexico [24]Universidad Intercultural Maya de Quintana Roo, Mexico [25]CentroGeo, Mexico

Abstract

Indigenous languages of the Americas face severe endangerment, and the scarcity of culturally grounded resources remains a critical barrier to revitalization efforts. We present the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages, the first shared task dedicated to generating captions for images depicting Indigenous cultures of the Americas, written in the Indigenous languages themselves. To support this, we introduce and publicly release a newly constructed dataset spanning five cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika. Evaluation follows a two-stage process, combining automatic evaluation using ChrF++ with human evaluation of the top-performing systems for each language. Eight teams participate, submitting 27 systems in total. Results indicate that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short on descriptive detail and, critically, cultural grounding.

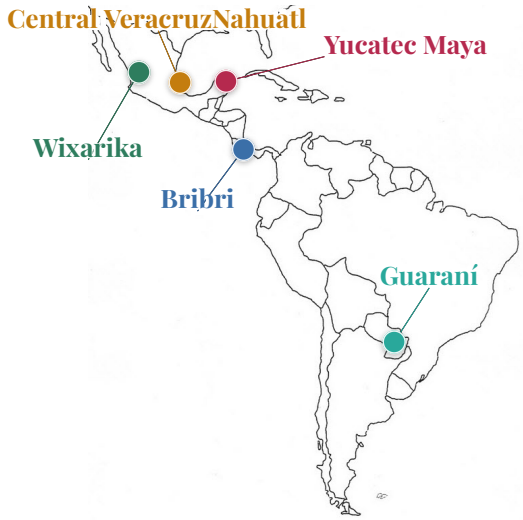

Figure 1: Approximate geographic distribution of the five Indigenous cultures covered in this work

1 Introduction

Many Indigenous languages of the Americas are endangered, spoken by small communities and at high risk of extinction. Revitalization depends on teaching materials that are costly and slow to produce (Chiruzzo et al., 2024). Image captioning is unusually well-suited to address this gap: pairing a culturally specific image with a caption in the target language teaches language and culture at once, since the learner takes in not only the words but the practices, objects, and settings they name. Yet producing accurate captions is hard on two fronts. First, even in the text-only setting, the most novel

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 279–295

Bribri
Ù tsg̱bita g̱ tg̱ idir yirîrî. E' yawèke ùkö wa, tsi̱ne̱t a̱ñì ǿmi̱k.
Palm segments are woven together at the top of the conical house. The suita palm leaves are woven really tight together.

Guaraní
Peteĩ kuimba'e oguata hína tatapỹi ári; "Tata ári jehasa" niko peteĩ ritúal [...].
A man walking barefoot on burning embers; "Tata ári jehasa" is a ritual [...].

Yucatec Maya
Junkúul puut jach ya'ab u yich beyxan yaan tu paachil ya'abach u le' ja'as.
A papaya tree with many fruits and many banana leaves behind it.

Central Veracruz Nahuatl
Se siuatsintli ika itlaken xochijyo itlanpa se aueuetl uejkapantik.
A girl in an embroidered dress under a huge ahuehuete tree.

Wixárika
Ukáa tú mati tatsiu h+xi in+net+ hipame yu iwama wa há mat+a.
The woman is harvesting green tomatillos with several colleagues.

Figure 2: Representative examples from each language. Each image is shown with the sentence in the target language (italicized) and its English translation. Note that the English translation is not released.

NLP techniques still struggle with low-resource languages (Mager et al., 2024; Weerasinghe et al., 2025; Hettiarachchi et al., 2025); adding the visual, multimodal dimension only raises the bar. Second, accurate captions demand not just linguistic competence but cultural knowledge—models frequently exhibit substantial limitations in visual cultural understanding, defaulting to Western-centric depictions and interpretations (Nayak et al., 2024; Winata et al., 2025; Liu et al., 2025; Bui et al., 2025). Both challenges are thus barriers to building systems that serve Indigenous communities.

We introduce the AmericasNLP Shared Task on Cultural Image Captioning for Indigenous Languages, the first effort to develop systems that generate accurate, culturally grounded captions for culturally relevant images, written in the Indigenous languages themselves. To make this possible, we contribute a new dataset of images and captions from five cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika; see Figure 1. Evaluation goes beyond automatic metrics: we complement them with a human evaluation to ensure that the cultural and linguistic quality of generated captions is assessed rigorously.

Eight teams participate, submitting 27 systems. Our results show that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short on detail and, critically, cultural grounding. These findings highlight the need for continued investment in culturally aware, multimodal NLP for Indigenous languages.

The remainder of the paper covers the dataset creation (Section 2), our evaluation process (Section 3.1), the submitted systems (Section 3.2), as well as results and additional insights (Section 4 and Section 5).

2 Dataset Creation

We describe the construction of our dataset, covering the cultures and languages represented, the image collection and caption annotation procedures, data collector recruitment, and dataset statistics. Representative examples are shown in Figure 2. We publicly release the development data in our Github repository[1] under the Creative Commons NonCommercial license (CC NC-BY).[2]

2.1 Cultures and Languages

The shared task features 5 cultures and their dominant languages: Bribri, Guaraní, Yucatec Maya, Central Veracruz Nahuatl, and Wixárika.

Bribri The Bribri are an Indigenous people of southern Costa Rica. They speak Bribri (BZD), a Chibchan language used by an estimated 7,000 people (INEC, 2011). It is a vulnerable language (Sánchez Avendaño, 2013), in that increasingly fewer children speak the language. The language has been documented through dictionaries (Margery, 2005; Krohn, 2020), grammars (Jara Murillo, 2018a), collections of oral literature (Jara Murillo, 2018b; García Segura, 2016; Constenla, 1996, 2006), digital corpora

[1] https://github.com/AmericasNLP/americasnlp2026

[2] https://creativecommons.org/licenses/by-nc/4.0/

Split	Metric	BZD	GRN	YUA	NLV	HCH
Dev (Train)	Instances	50	50	50	50	70
	Avg. Words	$14.9_{\pm 5.0}$	$21.6_{\pm 7.7}$	$11.3_{\pm 5.5}$	$8.3_{\pm 4.4}$	$16.0_{\pm 5.5}$
	Avg. Characters	$93.3_{\pm 30.7}$	$154.6_{\pm 51.1}$	$71.1_{\pm 40.9}$	$61.0_{\pm 31.3}$	$92.7_{\pm 28.1}$
Test	Instances	267	101	212	200	201
	Avg. Words	$13.4_{\pm 4.1}$	$21.9_{\pm 7.3}$	$15.2_{\pm 9.1}$	$5.9_{\pm 3.1}$	$12.0_{\pm 4.3}$
	Avg. Characters	$83.1_{\pm 22.3}$	$146.1_{\pm 48.9}$	$95.4_{\pm 59.4}$	$43.2_{\pm 22.5}$	$69.8_{\pm 22.7}$

Table 1: Dataset statistics for captions across our five Indigenous languages. Note that the dev split was originally intended solely for evaluation, but subsequently released for training as the task proved challenging without any in-language resources.

(Flores Solórzano, 2017), and learning materials for adults (Constenla et al., 2004; Jara Murillo and García Segura, 2013) as well as children (Sánchez Avendaño, 2020).

Guaraní The Guaraní are an Indigenous people of South America, primarily associated with Paraguay but also present in parts of Bolivia, Argentina, and Brazil. They speak Guaraní (GRN), a Tupi–Guaraní language with approximately 6.5 million speakers. Guaraní is a co-official language of Paraguay and is spoken by the vast majority of the population throughout the country, not being confined only to certain regions or social groups. This has resulted in many varieties of the language with different levels of code-switching and borrowing from Spanish, Portuguese, and other European languages.

Yucatec Maya The Yucatec Maya are an Indigenous people of the Yucatán Peninsula of Mexico, northern Belize, and parts of Guatemala. They speak Yucatec Maya (YUA), a Mayan language with approximately 800,000 speakers (INEGI, 2020). The Maya civilization of the Yucatán Peninsula represents a living ancestral legacy whose worldview deeply integrates nature, mathematical knowledge, and a cyclical conception of time. Far from being reduced to an archaeological past, contemporary Maya culture remains fully active within the daily practices and community dynamics of the region, where the language acts as the backbone of cultural identity and collective memory, as well as the social fabric. Essential architectural and environmental elements, such as the traditional Maya house, valued as a symbolic and sacred space, along with detailed linguistic expressions used to describe the surroundings and emotional states, reflect a unique way of inhabiting and understanding the world (Universidad Autónoma de Yucatán, n.d.).

Thus, far from being a static system, Yucatec Maya (*maayat'aan*) operates as a dynamic vehicle through which communities conceptualize their current environment, traditional agricultural knowledge, and everyday experiences. Nevertheless, the transition of this language toward modernity demands going beyond mere patrimonialization or discursive recognition; it requires a critical and effective exercise of refunctionalization that grants it shared practical and technological spaces, thereby guaranteeing its vitality in the face of contemporary challenges (Briceño Chel, 2021).

Central Veracruz Nahuatl Central Veracruz Nahuatl (alternatively Orizaba or Zongolica Nahuatl, *Náhuatl central de Veracruz*, ISO 639-3 NLV) is one of approximately 30 formally recognized varieties of Nahuatl, a Uto-Aztecan language (Valiñas Coalla, 2020). Many aspects of this language, spoken in several municipalities in the state of Veracruz, Mexico, in and around Orizaba, have been discussed in linguistics research (Goller et al., 1974; Tuggy, 1992, 1998). The Nahuas, the Indigenous ethnic group associated with the Nahuatl language, have diverse cultural practices that reflect their wide geographic distribution throughout Mexico. In Veracruz, many Nahuas continue to dress in traditional clothing and cultivate *milpa* in the traditional Mesoamerican manner, in both rocky/mountainous and marshy terrain.

Wixárika The Wixáritari (or Huichol) are an Indigenous people of Mexico. They live in the mountainous regions between the states of Jalisco, Nayarit, Durango, and Zacatecas (Gómez, 1999). They speak Wixárika (HCH), a polysynthetic language belonging to the Uto-Aztecan language family, spoken by approximately 50,000 people (INEGI, 2020). This ethnic group possesses a strong cultural identity, which is expressed through their clothing, speech, and ceremonies, such as the drum,

deer, and toasted corn ceremonies, among many others (Anguiano, 1978; Neurath, 2002). Their spiritual strength resides in the trinity of the peyote (*hik+ri*), deer (*kayumari*), and maize (esquite). The Wixaritari "hunt" the *hik+ri* during their pilgrimage to Wirikuta (Lumholtz, 1902), near the town of Real de Catorce in the state of San Luis Potosí (Martínez, 2006). From this plant, the *marakate* receive the spiritual strength to guide the Wixárika people and perform healings. Their primary economic activities are agriculture—involving seasonal migration to agricultural fields in Mexico's coastal regions—and the creation of handicrafts (Zingg, 1982).

2.2 Images

This section describes the image collection process, including general guidelines shared across all languages and culture-specific details for each community.

General Instruction For each language and culture pair, we recruit members of the respective Indigenous communities for data creation and annotation (see Section 2.4). We inform them about the goals, methodology, and reasons why the data collection is done and ask them to photograph everyday life in their communities, spanning a broad range of domains—food, work, ceremony, and nature. This community-driven approach ensures that the images themselves are culturally authentic, reflecting the lived experience of each community. We ask for pictures where humans are absent or not recognizable, to maintain the privacy of the community members. In the cases where people are shown, we either anonymize the images removing faces, or get permissions from the individuals to be part of the dataset.

Culture-Specific Details For Bribri, photographs are taken by cellphones or sourced from publicly available collections (see Appendix B.1). Images focus on important elements of Bribri culture (e.g., agriculture, architecture, crafts and material culture, and Talamancan landscapes) while also depicting aspects of Indigenous life elsewhere in Costa Rica and archaeological objects.

For Guaraní, images are drawn from the daily life of community members, supplemented by open-source material from the web. As the language is spoken throughout all society in Paraguay, the images present a mix of indigenous and criollo cultures, including food, activities, flora, fauna, clothing, and some archaeological items.

For Yucatec Maya, community members photograph everyday elements and activities, spanning local food, domestic and *milpa* agriculture, transportation, regional buildings, endemic plants, and campus life. Data collectors show a shared interest in documenting traditional and identity-defining aspects of their communities, such as hammock use, flowers along pathways, domestic animals, and local cuisine.

For Central Veracruz Nahuatl, a community member travels to multiple neighboring Nahuatl-speaking communities and collaborates with local contributors to photograph daily life. Images are captured using a Samsung Galaxy A53 smartphone.

For Wixárika, the collection reflects day labor, life and nature in the sierra, and elements of everyday community life.

2.3 Captions

For all images, we ask the same members of the respective Indigenous communities to provide captions in their Indigenous languages. We additionally collect Spanish translations to make the information about the images easier accessible. However, we do not pass the Spanish captions on to the shared task participants. Annotators are provided with annotation guidelines and an illustrative example (see Appendix B.2 for the full guidelines). The guidelines encourage annotators to produce *culturally enriched captions* where appropriate: rather than only describing the most salient object, captions should elaborate on the function, purpose, or significance of objects, clothing, gestures, or settings—but only where such elaborations are grounded in what is visually present in the image.

2.4 Recruitment

The dataset is constructed with active involvement of Indigenous community members, who contribute both images and captions. We briefly describe each group below.

Bribri The Bribri annotations come from an L1 speaker of the language, who has worked as a school teacher in the community.

Guaraní The annotators of the Guarani dataset are four fluent native speakers who also participate in other NLP projects. They are two women and two men, and their ages range between 24 and 39.

Yucatec Maya For the construction and evaluation of the Yucatec Maya corpus, nine native-speaking annotators are recruited from various towns across the Maya region of Quintana Roo. The team comprised both current students and alumni from the Language and Culture, as well as the Information and Communication Technologies programs at the Universidad Intercultural Maya de Quintana Roo (UIMQROO), with the assistance of relatives and neighbors.

Central Veracruz Nahuatl The Nahuatl dataset collection is organized and carried out largely by a Nahuatl teacher and translator in the area of Rafael Delgado, who has worked extensively on pedagogical material, cultural communication, and linguistic resources. He travels to multiple neighboring Nahuatl-speaking communities and works with other local collaborators.

Wixárika The dataset is created with the help of the Zoquipan community members. All pictures are either collected by the authors or by the recruited community members.

2.5 Dataset Statistics

We report the dataset statistics in Table 1. Each language has 50 instances in the development split—originally intended solely for evaluation, but subsequently released for training and inference as the task proved challenging without any in-language resources—except for HCH, which has 70 as it served as our pilot language. Test sets are considerably larger, ranging from 101 instances for GRN to 267 for BZD. Sentence length varies across languages: GRN exhibits the longest sequences on average (21.6 words in training, 21.9 in test), while NLV is the most concise (8.3 and 5.9 words, respectively).

3 Experimental Setup

3.1 Evaluation Process

We conduct a two-step evaluation process: (1) automatic evaluation and (2) human judgment of the top-5 systems per language.

Automatic Evaluation We use chrF++ (Popović, 2017) to automatically evaluate generated captions against our ground truth references. The top-5 systems for each language then proceed to the second stage, the human evaluation.

Score	Description
5	Fluent, natural, and culturally accurate. No significant errors.
4	Well-written with minor flaws in detail or cultural vocabulary.
3	Understandable, but inaccurate, incomplete, or too vague.
2	Serious grammatical errors; description mostly inaccurate.
1	Wrong language, incomprehensible, or unrelated to the image.

Table 2: Human evaluation scoring rubric (shortened, see full description in Appendix C.1). We capture two dimensions with the score: (1) *language quality* and (2) *fidelity to the image and correct use of cultural terminology*.

LANGUAGE	NUMBER OF RATINGS	IMAGES
BZD	320	267
GRN	228	101
YUA	212	212
NLV	200	200
HCH	201	201

Table 3: Number of ratings and images per language used in human evaluation.

Human Evaluation Automatic metrics such as chrF++ cannot capture the multifaceted ways in which an image can be described. Moreover, since our annotations contain only a single reference caption per image, human judgment provides a more robust and nuanced assessment of system outputs.

To keep the annotation process lightweight, we design a 1–5 rating scale intended to capture two dimensions: (1) *language quality* and (2) *fidelity to the image and correct use of cultural terminology*. Our scoring rubric is presented in Table 2. Prior to annotation, each annotator is shown a calibration example containing five captions representative of each score level. When multiple annotators rate the same example, their scores are averaged per example before computing the overall mean. All test-set samples receive at least one annotation. We report the number of ratings per language in Table 3. We further report the full guidelines with details about the annotators in C.1.

To facilitate annotation, we develop a platform which presents annotators with an image alongside five captions submitted by the top-performing teams, displayed in randomized order to avoid position bias. The reference caption is also provided to assist annotators in their judgment. Figure 7 presents a screenshot of the annotation interface

through which evaluators rate the generated descriptions.

Overall Winner Points are awarded based on each team's rank in the human evaluation for each language: 1st place receives 5 points, 2nd place 4 points, 3rd place 3 points, 4th place 2 points, 5th place 1 point, and teams not selected for human evaluation receive 0 points. A team's total score is the sum of their points earned across all five languages.

3.2 Baseline

We provide participants with a baseline that follows a two-stage *generate-then-translate* pipeline: a vision–language model (VLM) first produces a caption in Spanish, which is then translated into the target Indigenous language. We adopt Spanish as a pivot language because the machine translation resources available for Indigenous languages of the Americas are predominantly paired with Spanish; generating first in a high-resource language lets the baseline leverage existing translation systems.

Stage 1: Captioning in Spanish We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) to generate a Spanish caption for each image. The model is conditioned on a culturally-informed system prompt—drawn from publicly available encyclopedic sources—that is specific to each culture. The prompt instructs the model to first describe what is visually present, add only essential and visually grounded cultural context, include target-language terms where possible, and keep the caption concise.

Stage 2: Translation into the Target Language The Spanish caption is translated into the target language using the winning system (Gow-Smith and Sánchez Villegas, 2023) from the Americas-NLP 2023 Shared Task on Machine Translation into Indigenous Languages (Ebrahimi et al., 2023). As this system does not cover Yucatec Maya, we report no baseline for that language.

3.3 Submitted Systems

We summarize each participating team's approach below. All results are shown in Tables 4, 5 and 6.

Gators (Dhawan et al., 2026) uses a two-stage retrieval-augmented translation pipeline. A VLM, either Qwen2.5-VL-72B (Bai et al., 2025b) or Qwen3-VL-8B (Bai et al., 2025a), generates a Spanish caption, which is then translated by Gemini 2.5 Flash (Comanici et al., 2025) using retrieval-augmented many-shot in-context prompting: BM25 retrieves similar Spanish–target pairs from per-language parallel banks and supplies them as in-context examples alongside development examples, with the number of retrieved and development examples tuned per language.

Mila (Lara and Raval, 2026) post-trains Aya Vision 32B (Dash et al., 2025) in multiple stages: supervised fine-tuning on Spanish–Indigenous-language machine translation, optional reinforcement learning with verifiable rewards, and a final fine-tuning stage on image captioning, such that the model generates captions directly in the target language rather than via a Spanish pivot. They additionally submit a zero-shot GPT-5.5 (OpenAI, 2026) direct-captioning system.

IUHoosiers (Shi et al., 2026) submits for Guaraní only, using inference-time knowledge injection, without any fine-tuning. For each image, Gemma 4 31B (Farabet and Lacombe, 2026) produces a description that is used as a BM25 (Robertson and Zaragoza, 2009) query over four Guaraní knowledge sources. The retrieved items, together with a fixed grammar-book excerpt and interlinear-glossed examples, are injected into the prompt to generate the caption in a single pass.

6fanle (Wang and Yang, 2026) submits for Wixárika only, using a modular Spanish-pivot pipeline: CLIP (Radford et al., 2021) retrieves visually similar images to provide grounding examples, Qwen3-VL-8B-Instruct (Bai et al., 2025a) generates Spanish caption candidates, the Sheffield 2023 MT system (Gow-Smith and Sánchez Villegas, 2023) translates these candidates into Wixárika, and a character 5-gram language model reranks the translations to select the final caption.

InclusionVLM (Bueno and Garg, 2026) compares two approaches. Their cascaded system pairs a VLM—Gemini 2.5 Flash (Comanici et al., 2025)—, using concise persona-based cultural prompting, with the Sheffield 2023 MT system (Gow-Smith and Sánchez Villegas, 2023) as well as a separate Spanish–Maya model for Yucatec Maya. Their single-stage system adapts PaliGemma 2 3B (Steiner et al., 2024) end-to-end via LoRA fine-tuning (Hu et al., 2021), continued pretraining, and multilingual joint training.

Yaduha (Cuadros et al., 2026) uses a schema-constrained, LLM-assisted rule-based approach in

Team \ Lang.	BZD	GRN	YUA	NLV	HCH
baseline	7.01	20.14	–	9.52	16.91
6fanle	–	–	–	–	19.16†
IUHoosiers	–	**24.67**†	–	–	–
InclusionVLM	7.94	16.48	16.97†	14.06†	18.37†
Mila	11.73†	19.77†	15.99†	20.66†	19.01†
NAIST	**19.37**†	19.41†	15.80†	20.93†	**19.84**†
gators	17.90†	23.10†	21.11†	**25.42**†	17.58†
usp	10.95†	19.73†	10.83	9.49	13.68
yaduha	10.03†	16.90	**23.41**†	21.00†	15.61

Table 4: Automatic evaluation results for all participating teams. We report the best ChrF++ score per team across all submitted systems. † marks the top-5 teams per language, as those are selected for human evaluation.

which the VLM never emits target-language text directly. For each language, a coding agent—Claude Opus 4.7 (Anthropic, 2026)—authors a *language package*—a Python module with a closed vocabulary, Pydantic sentence schemas, and a deterministic renderer—based on the development split and public linguistic references. At inference, a VLM—GPT-5 (Singh et al., 2026)—sees the image and schema and emits a structured representation under constrained decoding, which the renderer converts into the surface caption.

USP (Fernandes, 2026) uses a two-stage cascade pipeline in which Qwen3-VL-8B-Instruct (Bai et al., 2025a) generates a culturally-prompted Spanish caption and a fine-tuned NLLB-200-distilled-600M (NLLB et al., 2022) model, one per language, translates it into the target language, trained on AmericasNLP 2023 data augmented with public parallel corpora. The team documents a failure mode in which NLLB-200 lacks vocabulary entries for Bribri and Maya and silently produces English output.

NAIST (Vasselli et al., 2026) explores two strategies. Their primary system performs nearest-neighbor retrieval: it embeds the test image with CLIP (Radford et al., 2021), finds the most similar development image and returns its caption directly. Their second system is a generation pipeline that analyzes the scene, grounds the identified concepts in dictionary entries, and retrieves gloss templates alongside interlinear gloss representations to constrain generation in low-resource settings.

Team \ Lang.	BZD	GRN	YUA	NLV	HCH
6fanle	–	–	–	–	2.48
gators	2.758	3.390	3.175	3.375	2.90
IUHoosiers	–	**3.448**	–	–	–
InclusionVLM	–	–	1.108	1.185	2.33
Mila	1.994	1.764	**3.203**	1.560	2.21
NAIST	2.219	1.978	1.934	1.220	**3.79**
usp	1.086	2.410	–	–	–
yaduha	**2.895**	–	2.892	**3.465**	–

Table 5: Aggregated human evaluation scores. Each entry reports the mean rating on a 1–5 scale, where higher scores indicate better language quality and greater image fidelity. Bold indicates the highest score per language. Underline indicates the second highest score per language.

Team \ Lang.	BZD	GRN	YUA	NLV	HCH	Total
gators	4	4	4	4	4	**20**
NAIST	3	2	2	2	**5**	14
yaduha	**5**	0	3	**5**	0	13
Mila	2	1	**5**	3	1	12
IUHoosiers	0	**5**	0	0	0	5
InclusionVLM	0	0	1	1	2	4
usp	1	3	0	0	0	4
6fanle	0	0	0	0	3	3

Table 6: Points per language and total. Bold indicates the highest score per column; underline indicates the second highest.

4 Results

We first report the automatic evaluation, and then human evaluation.

Automatic Evaluation We report the per-language ChrF++ scores for the best systems per team and language in Table 4. Complete results can be found in Appendix D.

Overall, most participating teams surpass the baseline for at least one language. Notably, three teams—Mila, NAIST, and gators—rank in the top-5 across all five languages, reflecting strong and consistent performance. Among these, NAIST achieves the highest scores for BZD (19.37) and HCH (19.84), while Gators leads for NLV (25.42). Outside this group, Yaduha achieves the best score for YUA (23.41), and IUHoosiers comes first for GRN (24.67).

Human Evaluation Table 5 presents the human evaluation results for the top-5 systems for each of the five languages, and Table 6 the derived points used to determine the overall winner.

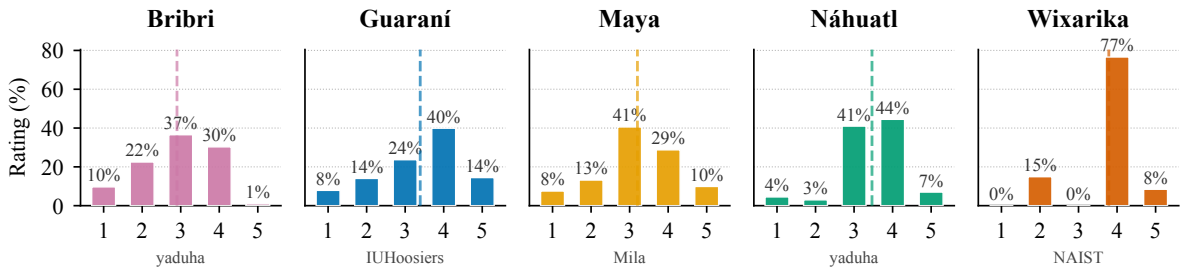

Figure 3: Rating distributions (1–5) of the best-performing system per language based on human evaluation. The dashed vertical line indicates the mean rating for each system.

The gators team demonstrates the most consistent performance, finishing second for every language, which translates into the highest total point score of 20. Per-language winners vary considerably: Yaduha achieves the highest scores for BZD (2.895) and NLV (3.465), Mila leads for YUA (3.203), IUHoosiers for GRN (3.448), and NAIST for HCH (3.79), where the margin over second place is the largest, with 0.89 points. With 14 points, NAIST finishes as runner-up overall.

Discussion Human evaluation scores suggest that the task remains largely unsolved across all five languages. Even the strongest per-language systems score between 2.895 (BZD) and 3.79 (HCH), indicating that outputs are at best understandable but lack the detail and cultural grounding that accurate captioning demands. The gators team, our overall winner, reflects this pattern well: despite finishing second across all five languages, their ratings range only from 2.76 (BZD) to 3.38 (NLV), reflecting consistent but imperfect outputs. Performance varies considerably across languages, though lower scores may reflect more challenging images or greater difficulty of generation in that language for the models, rather than being a property of the language itself: Wixárika (HCH) sees the most reliable system, with NAIST achieving a notably strong score of 3.79, while Bribri (BZD) and Yucatec Maya (YUA) prove most challenging, with lower means.

Taken together, these results indicate that no system achieves robust, culturally grounded captioning across all five languages. The gap between current outputs and fluent, culturally accurate captions underscores both the difficulty of the task and the need for continued investment in culturally aware, multimodal NLP for Indigenous languages.

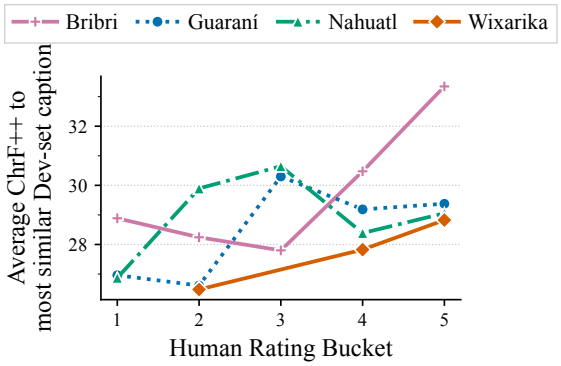

Figure 4: Average similarity between test captions and their most similar development-set captions across human rating buckets. Similarity is measured using ChrF++.

5 Analysis

5.1 Quantitative Analysis

We further provide a quantitative analysis of the outputs of the best performing system per language according to our human evaluation (Table 5).

Rating Distribution Figure 3 shows the score distributions of the best-performing system per language. Wixárika (NAIST) stands out with a strongly peaked distribution at rating 4 (77%) and near-zero low-quality outputs. Note that NAIST performed better than the second team by a substantial amount (3.79 vs. 2.90). Systems for Bribri (Yaduha) and Maya (Mila) exhibit broader distributions centered around rating 3, reflecting more variable output quality. Guaraní (IUHoosiers) and Náhuatl (Yaduha) fall in between, with moderate concentrations at ratings 3–4. Overall, the distributions suggest that caption quality is lower for Bribri and Maya while Wixárika benefits from the most reliable system.

Development–Test Caption Similarity vs. Human Ratings We examine whether human evaluation scores on the test set correlate with the similarity between test captions and captions in the development set. To measure this, we compute ChrF++ between each test caption and the most similar caption from the development set. We exclude Maya (Mila) from this analysis, since their system did not make use of the development data.

Figure 4 reports the average ChrF++ score between most similar captions within each human rating bucket. Overall, we observe a weak but consistent positive trend: captions receiving higher human ratings tend to be slightly more similar to a development-set caption. This effect is more pronounced for Wixárika and Bribri, where the increase in similarity across rating buckets is clearer, whereas other languages show a flatter, noisier pattern. Notably, for Wixárika, the top-ranked system (NAIST) directly returns development-set captions, making this relationship explicit: its high ratings are by construction tied to the development set.

Taken together, these results suggest that test-set examples that are more similar to the development split elicit higher-rated predicted captions.

5.2 Qualitative Analysis

We further present a qualitative overview based on annotator comments collected following the human evaluation.

Bribri The higher scoring descriptions for the Bribri content focus on producing oversimplified but grammatically plausible descriptions. These are not capable of capturing the cultural intricacies of the images, but they are at least able to provide a readable description of some part of the picture (e.g., a pot with a traditional stew of pork and yucca in it, held by a woman, is described as *Chkà tso' ù a. Pë' tö chkà alòk* "The food is in the house. People prepare food"). The submitted systems also produce numerous hallucinations. For example, some systems report seeing canoes on a river, where the picture merely shows a river with ripples on the water. In a picture of the important tradition of the *Ák kuk* "pulling the stone" (Brenes Mora, 2024), numerous men are seen wading in a river and carrying a large stone on a mesh of trunks similar to a palanquin. Here the same hallucination shows up again: one of the systems describes this as "*Pë' dàmi taîë kanò kï*" meaning "Lots of people are coming by canoe."

Guaraní Systems in general had a good grasp of what they needed to generate and were mostly comprehensible, but often lost important details that could be easily seen by humans.

Yucatec Maya Our annotators express surprise at how accurately some models are able to describe the images in their native language. They also note instances where certain systems produce comical descriptions of the visual content.

Central Veracruz Nahuatl Many of the captions are impressive in their specificity and naturalness. However, there are also many descriptions that veer off-topic, are editorialized, and exaggerated aspects of a seemingly idealized Nahua culture. In some cases, captions are provided in Spanish instead of Nahuatl.

Wixárika Overall, evaluation identifies a considerable number of duplicated phrases. The annotator reports that evaluation is hard as most descriptions are hard to read and confusing. Nevertheless, the systems also generate a set of good or excellent descriptions. Finally, an important number of generations contain code-switching, or consist mostly of Spanish words.

6 Conclusion

We present the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages, the first shared task dedicated to generating captions for images depicting Indigenous cultures of the Americas, written in the Indigenous languages themselves. To support this effort, we create a novel dataset for the task. Eight teams participate, submitting 27 systems in total. Results indicate that the task remains largely unsolved: while the strongest systems produce understandable captions, they fall short in descriptive detail and, critically, in cultural grounding. These findings highlight both the inherent difficulty of the task and the pressing need for continued investment in multimodal, culturally aware models for Indigenous languages.

Limitations

We acknowledge several limitations of our data collection and evaluation process. First, we collect only a single caption per image, which limits the reliability of the reference captions. Similarly, human evaluation annotations are sparse, with most languages covered by only one or two annotators

per item, which may introduce noise into the ratings. Second, while we provide annotation guidelines, the asynchronous nature of the collection process leads to inconsistencies across languages. This is particularly evident in the development set, which reflects the challenges of an initial annotation round. Future work could address these quality issues and refine the dataset.

Ethics Statement

This work contains pictures and descriptions taken in the context of Indigenous communities of the Americas. These groups have been historically subject to discrimination, oppression, and colonialism. This work is done with the aim of closing the technological gap between Indigenous languages and the majority language in NLP. That being said, we recognize the risks of working in this setting. Therefore, we take the following measures: all annotators and participants are informed about this work, the goals, and where to download and read the results; all annotators are community members and are recognized for their work with an hourly salary equivalent to that of a high school teacher in their respective region. We additionally avoid using pictures that contain human faces. In the case of the inclusion of human faces, we either anonymize the images or obtain explicit approval. Most of the dataset is publicly available—the exception is the held-out test set, which we retain to avoid data contamination (when scraped by LLMs without our permission). The data has been released under the CC-BY-NC license, upon agreement with the annotators. All annotators retain ownership of the dataset. All decisions while creating the data collections are taken based on the standards defined by our field (Bird, 2020; Mager et al., 2023).

Acknowledgments

We would like to thank all teams for participating in the shared task, all data contributors, and community members that participated in this effort.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956.

Mariana Anguiano. 1978. *La endoculturación entre los huicholes*. México INI, 1978.

Anthropic. 2026. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-29.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Samantha Brenes Mora. 2024. Pueblos bribri y cabécar celebran su cultura con la ancestral "jala de piedra".

Fidencio Briceño Chel. 2021. ¿ hacia dónde va la lengua maya de la península de yucatán? entre institucionalización y patrimonialización. *Maya America: Journal of Essays, Commentary, and Analysis*, 3(1):12.

Mirelle Bueno and Sushil Garg. 2026. Culturally grounded image captioning in indigenous languages with vision-language models: Cascaded and single-stage approaches. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous*

Languages of the Americas (AmericasNLP 2024), pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Adolfo Constenla. 1996. *Poesía tradicional indígena costarricense*. Editorial Universidad de Costa Rica.

Adolfo Constenla. 2006. *Poesía bribri de lo cotidiano: 37 cantos de afecto, devoción, trabajo y entretenimiento*. Editorial Universidad de Costa Rica.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Diego Cuadros, Nicholas Leeds, Amanda Avalos, Azul Alipzar-Velazquez, Jared Coleman, Faezeh Dehghan Tarzjani, and Bhaskar Krishnamachari. 2026. Schema-constrained image captioning for five low-resource indigenous languages. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *Preprint*, arXiv:2505.08751.

Aashish Dhawan, Christopher Driggers-Ellis, Dzmitry Kasinets, Christan Grant, and Daisy Wang. 2026. Retrieval-augmented long-context translation for cultural image captioning: Gators submission for AmericasNLP 2026 shared task. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *Preprint*, arXiv:2405.19092.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Clement Farabet and Olivier Lacombe. 2026. Gemma 4: Byte for byte, the most capable open models. `https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/`. Accessed: 2026-05-28.

Rafael M. Fernandes. 2026. USP at AmericasNLP 2026 shared task: Culturally-aware image captioning for indigenous languages via vision-language models and fine-tuned neural machine translation. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

Alí García Segura. 2016. *Ditsö Rukuö Identity of the Seeds: Learning from Nature*. IUCN.

Theodore R Goller, Patricia L Goller, and Viola G Waterhouse. 1974. The phonemes of Orizaba Nahuatl. *International Journal of American Linguistics*, 40(2):126–131.

Paula Gómez. 1999. El huichol de san andrés cohamiata. *Jalisco, Archivo de lenguas indígenas de México 22, México: El Colegio de México,*.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage, editors. 2025. *Proceedings of the First Workshop on Language Models for Low-Resource Languages*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

INEGI. 2020. Censo de población y vivienda 2020. Accedido: 2026-05-28.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*, second edition. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Haakon S. Krohn. 2020. Diccionario digital bilingüe bribri.

Luis Lara and Param Raval. 2026. From machine translation to image captioning: Training vision-language models for indigenous languages of the americas. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *Preprint*, arXiv:2501.01282.

Carl Lumholtz. 1902. El méxico desconocido [2 vols.]. *México, Editora Nacional*.

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors. 2024. *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Association for Computational Linguistics, Mexico City, Mexico.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*, second edition. Editorial de la Universidad de Costa Rica.

Isabel Martínez. 2006. Gutiérrez del angel, arturo. la peregrinación a wirikuta: El gran rito de paso de los huicholes. méxico: Etnografía de los pueblos indígenas de méxico, instituto nacional de antropología e historia, universidad de guadalajara, 2002, 310 p.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Johannes Neurath. 2002. Las fiestas de la casa grande. *Universidad de Guadalajara, Instituto Nacional de Antropología e Historia, Guadalajara, Mexico*.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-28.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 11479–11505. Curran Associates, Inc.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Wenchen Shi, Phakphum Artkaew, and Luke Gessler. 2026. Culturally-aware image captioning for Guaraní with multimodal prompting: IUHoosiers at AmericasNLP 2026. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Carlos Sánchez Avendaño. 2020. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. DIPALICORI.

David Tuggy. 1992. The affix-stem distinction: A cognitive grammar analysis of data from orizaba nahuatl. *Cognitive Linguistics*, 3(3):237–300.

David Tuggy. 1998. *Giving in Nawatl*, page 35–66. John Benjamins Publishing Company.

Universidad Autónoma de Yucatán. n.d. Portal de la Cultura Maya. `https://www.mayas.uady.mx/`. Accessed: 2026-05-30.

Leopoldo Valiñas Coalla. 2020. *Lenguas originarias y pueblos indígenas de México. Familias y lenguas aisladas*. Academica Mexicana de la Lengua, México.

Justin Vasselli, Arturo Martínez Peguero, Shintaro Ozaki, Frederikus Hudi, Haruki Sakajo, and Taro Watanabe. 2026. Nearest-neighbor retrieval for indigenous image captioning. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji Wang and Hanqi Yang. 2026. 6fanle submission to the AmericasNLP 2026 shared task on Wixarika image captioning. In *Proceedings of the 6th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2026)*, San Diego, California. Association for Computational Linguistics.

Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors. 2025. *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*. Association for Computational Linguistics, Abu Dhabi.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Robert Mowry Zingg. 1982. *Los huicholes una tribu de artistas*. México INI, 1982.

A Related Work

Image Captioning Benchmarks. Early captioning benchmarks such as MS-COCO (Lin et al., 2015) and Flickr30k (Young et al., 2014) established standard evaluation protocols for image description, but focus primarily on everyday Western scenes. Nocaps (Agrawal et al., 2019) extended this to novel object categories drawn from Open Images. Dong et al. (2024) proposed a benchmark and evaluation metric for *detailed* image captioning, addressing the shortcomings of short-caption benchmarks.

Our work extends these efforts by encouraging culturally enriched captions.

Cultural Visual Understanding. Several benchmarks assess the cultural awareness of vision-language models (VLMs). CulturalVQA (Nayak et al., 2024) probes VLM understanding across clothing, food, drinks, rituals, and traditions from 11 countries, but is restricted to English. CVQA (Romero et al., 2024) presents a multilingual VQA setting spanning 26 languages and 28 countries with human-written questions. WorldCuisines (Winata et al., 2025) provides a large-scale food-centric VQA benchmark across 30 languages sourced from Wikipedia, and ALM-bench (Vayani et al., 2025) scales further to 100 languages, targeting low-resource languages and diverse cultural aspects.

However, none of these works addresses Indigenous communities and languages, a gap our dataset fills.

B Detailed Dataset Creation

B.1 Bribri Image Creation Sources

For Bribri, we additionally source images from three Costa Rican cultural institutions: the Sistema de Información Cultural de Costa Rica (Sicultura),[3] the Dirección de Gestión Sociocultural of the Ministerio de Cultura y Juventud,[4] and the Universidad de Costa Rica.[5].

B.2 Caption Guidelines

The full annotation guidelines are provided in Figure 5. We further provide one correct and one incorrect example caption. The correct caption reads: "A wooden house, the so-called *carretón*, built specifically to store food such as corn, also serves as living quarters for people." The incorrect caption reads: "A Wixárika woman shelling corn to make nixtamal"—chosen because the image does not clearly convey that the corn is being prepared for nixtamal.

C Detailed Experimental Setup

C.1 Human Evaluation Detail

For Yucatec Maya, the human evaluation phase involved two computer science alumni who are native Maya speakers from the Yucatán Peninsula, one from the Universidad Autónoma de Yucatán (UADY) and the other from UIMQROO. For the remaining languages, the same annotators who contributed to data collection also conducted the human evaluation.

We report the full annotation guideline in Figure 6. Note that the guideline is in Spanish originally. Furthermore, we show a screenshot of our annotation tool in Figure 7.

D All Systems Results

We report the performance of all submitted systems in Table 7. Alongside chrF++, we also report CIDEr (Vedantam et al., 2015), a metric originally proposed for image captioning and widely adopted in vision-language benchmarks. To our knowledge, CIDEr has not previously been applied to Indigenous language evaluation; we therefore adopt chrF++ as our primary metric, given its stronger track record in low-resource and morphologically rich language settings (Ebrahimi et al., 2024).

[3] https://si.cultura.cr
[4] https://www.dircultura.go.cr
[5] https://www.ucr.ac.cr

Team	**Ver.**	Bribri		Guaraní		Maya		Nahuatl		Wixárika	
		chrF	CIDEr	chrF	CIDEr	chrF	CIDEr	chrF	CIDEr	chrF	CIDEr
6fanle	v0	—	—	—	—	—	—	—	—	19.16	0.0214
IUHoosiers	v0	—	—	24.39	0.1238	—	—	—	—	—	—
IUHoosiers	v1	—	—	24.41	0.1351	—	—	—	—	—	—
IUHoosiers	v2	—	—	24.41	0.1428	—	—	—	—	—	—
IUHoosiers	v3	—	—	24.16	0.1284	—	—	—	—	—	—
IUHoosiers	v4	—	—	**24.67**	0.1489	—	—	—	—	—	—
IUHoosiers	v5	—	—	24.42	**0.1668**	—	—	—	—	—	—
IUHoosiers	v6	—	—	24.04	0.1122	—	—	—	—	—	—
IUHoosiers	v7	—	—	22.43	0.0738	—	—	—	—	—	—
IUHoosiers	v8	—	—	24.41	0.1351	—	—	—	—	—	—
InclusionVLM	v0	7.94	0.0059	16.48	0.0465	16.97	0.0100	14.06	0.0031	18.37	0.0084
InclusionVLM	v1	2.54	0.0002	7.61	0.0055	9.14	0.0059	—	—	10.52	0.0014
InclusionVLM	v2	—	—	—	—	—	—	—	—	12.34	0.0078
InclusionVLM	v3	—	—	—	—	—	—	—	—	13.93	0.0079
Mila	v0	11.73	0.0152	19.63	0.0314	—	—	19.42	0.0332	18.81	0.0246
Mila	v1	10.89	0.0086	19.77	0.0310	—	—	19.85	0.0372	18.85	0.0249
Mila	v2	11.31	0.0140	19.42	0.0520	—	—	20.66	0.0415	19.01	0.0299
Mila	v3	4.56	0.0058	12.80	0.0510	15.99	0.0950	19.72	0.0615	10.98	0.0096
NAIST	v0	**19.37**	**0.1167**	19.41	0.0458	15.80	0.0424	20.93	0.0711	**19.84**	**0.0422**
NAIST	v1	4.21	0.0007	9.71	0.0350	12.27	0.0918	10.09	0.0123	—	—
NAIST	v2	5.29	0.0018	8.33	0.0270	10.20	0.0555	15.34	0.0597	—	—
baseline	v0	7.01	0.0005	20.14	0.0050	—	—	9.52	0.0015	16.91	0.0066
gators	v0	10.64	0.0212	23.10	0.1243	21.11	**0.1765**	**25.42**	**0.1795**	17.58	0.0326
gators	v1	17.90	0.1081	21.63	0.1271	—	—	—	—	—	—
usp	v0	10.95	0.0007	19.73	0.0236	10.83	0.0100	9.49	0.0001	13.68	0.0027
yaduha	v0	10.03	0.0084	16.90	0.0379	**23.41**	0.1116	21.00	0.0284	15.61	0.0266

Table 7: Automatic evaluation results per team and submission version. Best result per language in **bold**.

Annotation Guidelines

Please follow these rules when writing image descriptions:

1. Start with only describing what is visually present.
- Focus on people, objects, animals, settings, and clearly visible actions.
- Focus on the most salient visual element(s).

2. IMPORTANT: Add cultural explanations
- Cultural enrichment: You must elaborate on/explain the function, purpose, or typical use of objects, clothing, gestures, or settings, but only when these explanations are grounded in what is visually observable.
- Visually observable: The function, purpose, or typical use must be immediately clear, without ambiguity, to any member of the community.
- Do not infer identity, background, profession, or cultural group unless it is explicitly indicated by visible, unambiguous cues. Better describe it with words, e.g. a person wearing traditional ⟨Culture X⟩ clothes vs. a ⟨Culture X⟩ person.

3. Be objective and neutral.
- Use factual, descriptive language.
- Avoid opinions or value-laden terms such as "beautiful," "cute," "poor," "aggressive", or "angry."
- Avoid emotional interpretations ("sad situation", "exciting moment", . . .) unless the emotion is visually undeniable (e.g., a person visibly crying).

4. Description Format
- Sentences should neither be too short nor too long. Try to be concise.
- Each sentence must contain a verb.
- Please pay attention to grammar and spelling.
- The description should be 1–2 sentences long.
- Additional comments are welcome.

Figure 5: Annotation guidelines provided to annotators for writing enriched image captions. Guidelines emphasize visual grounding, cultural context, objectivity, and concise formatting.

Human Evaluation Guidelines

How to Rate

For each system description, assign a single overall score from 1 to 5. Evaluate the description along two dimensions:

1. Language Quality: Is it written in the target language? Is it grammatically correct, fluent, and natural? If the description is poorly written or illegible, it cannot be considered a good description.

2. Image Fidelity and Correct Use of Cultural Terms: If the language quality is good—does the description reflect what is actually seen in the image? Does it use the correct cultural terms? Would it seem respectful and accurate to a member of the community?

Rating Scale

5 — Excellent: Fluent and natural language with an accurate, culturally grounded description of the image. Uses correct cultural and technical terms. Nothing significant to correct.

4 — Good: Clear, well-written language with a correct description, but with minor flaws, e.g., minor language errors, missing details in the description, or imprecise cultural vocabulary.

3 — Mixed: Language is mostly understandable, but the description is mistaken about the image, omits important content, or is too vague to be clearly useful.

2 — Poor: Language has serious problems (incorrect grammar, frequent errors, hard to follow) and the description is largely inaccurate. Still recognizable as an attempt at a description.

1 — Unusable: Not in the target language, not understandable as language, or bears no relation to the image.

Figure 6: Rating guidelines provided to annotators for evaluating enriched image captions (originally in Spanish).

Imagen 69

Descripción de referencia *(sólo como guía)*
Ojehecha kesu apu'a oñemboaty peteĩ mba'e ári. Ko tembi'u ha'e kesu paraguái, peteĩ kesu pyahu Paraguáipegua.

*Para la **Imagen 69** de arriba, otorgue a cada descripción **una sola puntuación general del 1 al 5**. Evalúe primero *la calidad del idioma* y, si esta es buena, *la fidelidad a la imagen y el uso correcto de términos culturales*.

	1 - Inutilizable	2 - Deficiente	3 - Mixto	4 - Bueno	5 - Excelente
Tres ñandutí (queso fresco paraguayo) oñevendéva tendápe, ojejapóva tuju renondépe, orekóva mba'erechaukaha hũ pukukue.	○	○	○	○	○
Peteĩ ajaka henyhe "chipa argolla"guí, tembi'u ojehayhuvéva ñane retã Paraguáipe, oĩnte va'erã jehomombyrýpe ha aty guasu rogayguákuéra ndive, akói hyakuãvurei hese kesu ha aramirõ.	○	○	○	○	○

Figure 7: Partial screenshot of the human evaluation annotation tool used in our study.

Author Index