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
July 3, 2026 ©2026 Association for Computational Linguistics

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}