Measuring Linguistic Competence of LLMs on Indigenous Languages of the Americas

Anonymous ACL submission

Abstract

This paper introduces a benchmark for evaluating Indigenous language knowledge in large language models using zero- and few-shot prompting. The benchmark includes three tasks: (1) language identification, (2) cloze completion of Spanish sentences aided by Indigenous-language translations, and (3) grammatical feature classification. We apply the benchmark to 13 Indigenous languages, including Bribri, Guarani, and Nahuatl, and evaluate models from five major families (GPT, Gemini, DeepSeek, Qwen, and LLaMA). Results reveal large differences across both languages and model families, with a small subset of model–language combinations showing consistently stronger performance across tasks, while other combinations remain close to random chance.

1 Introduction

Indigenous languages present structural properties that challenge current language models. Many are morphologically rich, with features such as polysynthesis, complex agreement, or noun incorporation. Some lack standardized orthographies, complicating tokenization and evaluation. Their typological profiles differ substantially from high-resource languages, and they are often underrepresented in pretraining data or evaluation settings (Ponti et al., 2020; Mager et al., 2021). These factors make Indigenous languages a valuable test case for evaluating model generalization.

Multilingual NLP benchmarks such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) concentrate primarily on high- and medium-resource languages with substantial digital presence.More inclusive initiatives like FLORES (Guzmán et al., 2019) and the AmericasNLP shared tasks (Mager et al., 2021) have introduced datasets for machine translation involving Indigenous languages, enabling evaluation in

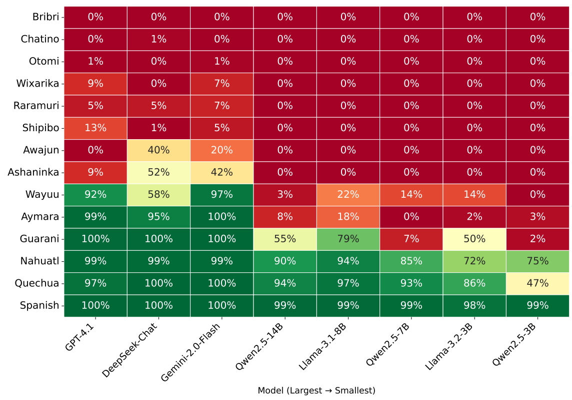

Figure 1: Indigenous Language Identification Accuracy on Open setting. Sorted by average performance.

specific translation contexts. Still, there remains a gap in benchmarks that assess general language understanding, such as lexical recognition, morphosyntactic inference, or cross-lingual reasoning, without task-specific training or fine-tuning.

To investigate how much Indigenous language knowledge large language models may encode, we introduce a probing-based benchmark designed for zero-shot evaluation. The benchmark consists of three tasks that target different aspects of linguistic understanding. In the language identification task, the model must select the correct language given a word or sentence. In the cloze completion with glosses task, the model is prompted with a Spanish sentence containing a blank and a corresponding translation or gloss in an Indigenous language, and asked to predict the missing word. In the grammatical feature identification task, the model is shown a sentence in an Indigenous language and asked to identify a specific morphosyntactic feature, such as person, number, or tense. Together, these tasks provide a targeted way to examine whether models exhibit consistent, interpretable behavior when interacting with underrepresented languages.

2 Related Work

Probing LLMs' linguistic knowledge A common approach to probing linguistic knowledge in LLMs is minimal-pair benchmarks such as BLiMP (Warstadt et al., 2020). Recent extensions include CLiMP for Chinese (Xiang et al., 2021), JBLiMP for Japanese (Someya and Oseki, 2023), RuBLiMP for Russian (Taktasheva et al., 2024), and MultiBLIMP (Jumelet et al., 2025), which covers 101 languages. These controlled formats isolate grammatical contrasts and are useful for evaluating structural generalization.

Cloze and multiple-choice formats are also used to probe model knowledge. LAMA (Petroni et al., 2019), X-FACTR (Jiang et al., 2020), and Multilingual LAMA (Kassner et al., 2021) evaluate factual recall with cloze prompts. LM-PUB-QUIZ (Ploner et al., 2025) converts these into multiple-choice form. WDLMPro (Senel and Schütze, 2021) applies this format to lexical and semantic knowledge. Our grammatical feature task similarly uses natural sentences and structured outputs to evaluate models' ability to infer morphosyntactic properties.

Low-resource language benchmarks AmericasNLI (Kann et al., 2022) enables evaluation of semantic inference in 10 Indigenous languages through translations of the XNLI corpus. MasakhaNER (Adelani et al., 2021) and MasakhaPOS (Dione et al., 2023) benchmark named entity and POS tagging for African languages and explore cross-lingual transfer using multilingual and region-specific models. XTREME-UP (Ruder et al., 2023) introduces a broad benchmark spanning 88 under-represented languages across multiple user-facing tasks, enabling large-scale evaluation under low-resource constraints.

3 Methodology

3.1 Linguistic Corpora

We use development data from the AmericasNLP 2025 Shared Task (De Gibert et al., 2025), covering 13 typologically diverse Indigenous languages of Latin America. Speaker populations range from 5,000 (Chatino) to over 7 million (Quechua), with most languages spoken in Peru or Mexico. Five have Wikipedias: Aymara, Guarani, Nahuatl, Quechua, and Wayuu. An overview of the languages is in Table 4. Full language details are in Appendix A.

3.2 Pre-trained Language Models

We evaluate ten large language models spanning a range of architectures and sizes. These include GPT-4.1 (OpenAI et al., 2024), Gemini 2.0 Flash (Gemini et al., 2025), and DeepSeek-V3-0324 (DeepSeek-AI et al., 2024), all of which we access through the API. We also test open-weight instruction-tuned models: LLaMA-3.1-8B, LLaMA-3.2-3B (Touvron et al., 2023), and Qwen-3B, 7B, and 14B (Bai et al., 2023)[1]. For API-based models, we use a temperature of 0 to ensure deterministic outputs. For open-weight models, we turn off sampling. All prompts used in evaluation are included in Appendix B.

3.3 Language Identification

We evaluate model accuracy on identifying the language of sentences, using 459 sentences per language across 13 Indigenous languages and Spanish. To ensure enough signal for identification, all sentences are at least five tokens long.

We test four prompting conditions, ranging in difficulty:

1. **Multiple Choice Easy**: Prompted to choose between four options (target + three high-resource distractors like English or French).
2. **Multiple Choice Hard**: Distractors are other Indigenous languages.
3. **Multiple Choice Full**: Prompted to choose between all 14 languages.
4. **Open**: No choices provided; model must output the language name.

Our main experiments are conducted under zero-shot prompting. For the full setting, we also test n-shot prompting, where n sentences from each of the 14 languages are included as examples.

3.4 Cloze Translation Completion

To test whether LLMs can understand the Indigenous languages, we design a cloze task based on aligned Spanish–Indigenous sentence pairs from the AmericasNLP 2025 Shared Task.

We mask one content word in each Spanish sentence, excluding proper nouns and punctuation. To avoid trivial items, we filter out examples where

[1]Inference for open-weight models was run on a single RTX A6000 GPU, totaling approximately 75 GPU hours across all tasks. API-based models were accessed via public endpoints, with total usage costing around $50.

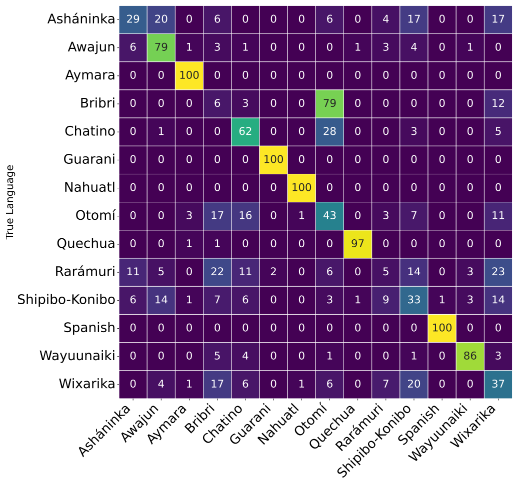

Figure 2: Confusion Matrix (%) for GPT-4.1 in Zero-shot setting. Vertical axis represents reference languages while horizontal axis represents the predictions.

Language	Example #				
	0	1	2	3	4
Asháninka	0.29	0.68	0.72	0.62	0.59
Aymara	1.00	1.00	1.00	1.00	1.00
Bribri	0.09	0.24	0.41	0.46	0.72
Nahuatl	1.00	1.00	1.00	1.00	1.00
Quechua	0.97	1.00	0.99	1.00	1.00
Shipibo-Konibo	0.35	0.48	0.52	0.64	0.78
Rarámuri	0.05	0.12	0.20	0.20	0.18
Otomí	0.43	0.70	0.79	0.74	0.71
Guarani	1.00	1.00	1.00	1.00	1.00
Spanish	1.00	1.00	1.00	1.00	1.00
Chatino	0.62	1.00	1.00	1.00	1.00
Wayuu	0.86	1.00	1.00	0.99	0.99
Wixarika	0.37	0.96	0.99	0.99	0.99
Awajún	0.79	0.82	0.83	0.89	0.86

Table 1: Zero-shot and Few-shot accuracy per language on GPT-4.1.

gpt-4o-mini correctly fills the blank. We then use the same model to generate plausible distractors, creating 4-option multiple-choice questions.

Each item includes a Spanish cloze sentence, four candidate completions, and the Indigenous translation of the Spanish. We test two settings: **Monolingual**, only the Spanish cloze sentence is shown, and **Bilingual**, the Spanish cloze and Indigenous translation are shown.

We measure accuracy, comparing results between the two settings. Option order is fixed between settings to avoid position bias.

The final dataset contains 5,529 problems spanning 13 languages. Most languages contribute over 400 examples; full counts appear in Appendix C.

3.5 Grammatical Feature Classification

To evaluate whether models can identify grammatical features from Indigenous sentences, we construct a classification task covering Bribri and Nahuatl. Each question consists of an Indigenous sentence, a target grammatical feature (e.g. person, tense, mood), the correct answer, and all possible alternative answers for that feature. Not all features are equally represented, and not all appear in every language. Data is adapted from the AmericasNLP 2025 Shared Task 2. A full list of tested features appears in Appendix D.

4 Results

4.1 Language identification

Zero-shot The easy setting acted as a sanity check, asking models to choose between the target Indigenous language and unrelated high-resource languages. All models performed well, confirming that the task and data were clear enough to distinguish Indigenous from unrelated languages.

In contrast, performance dropped sharply in the open setting. Figure 1 shows per-language accuracy under this condition. Larger models performed more consistently, but even the smallest models correctly identified Quechua and Nahuatl, suggesting these languages are relatively well represented in pretraining data.

The most reliably identified languages (Quechua, Guarani, Aymara, Nahuatl, and Wayuu) are the six in our evaluation set with their own Wikipedia editions. The Indigenous Wikipedias range in size from 24,000 articles (Quechua) to just 681 articles (Wayuu).

Figure 2 shows GPT-4.1's confusion matrix in the full multiple-choice setting. Confusions are notably asymmetric: Rarámuri was rarely predicted and often confused with Bribri; Bribri sentences were frequently mislabeled as Otomí. This suggests some languages disproportionately dominate model priors, producing directional confusion rather than mutual ambiguity.

Few-shots To evaluate the impact of few-shot prompting, we varied the number of examples per language from 1 to 4. Most models benefited from 1-shot prompting, but additional examples yielded diminishing returns. Smaller models showed little improvement from additional shots.

We also examined few-shot effects per language. Some languages showed substantial gains. For example, Bribri accuracy increased from 8.8% to

Language	Gemini-2.0	GPT-4.1	DeepSeek
Asháninka	0.01	0.04	0.03
Aymara	0.31***	0.22***	0.03
Awajún	0.12***	0.10***	0.11***
Bribri	0.04	0.07**	0.06*
Chatino	0.03	0.08	–0.01
Guarani	0.42***	0.40***	0.23***
Nahuatl	0.39***	0.26***	0.29***
Otomí	0.10**	0.07*	0.01
Quechua	0.45***	0.34***	0.36***
Rarámuri	–0.01	0.00	0.03
Shipibo-Konibo	0.11***	0.08*	0.11***
Wayuunaiki	0.09**	0.00	0.00
Wixarika	0.07*	0.05	0.06*

Table 2: Accuracy improvement (ΔAcc) for each language on the cloze task when given the aligned Indigenous sentence. Significance levels from McNemar's test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

72%, and Otomí from 43.4% to 71.1%. Shipibo-Konibo and Wixarika also showed steady improvement with more examples. Table 1 visualizes accuracy trends for different numbers of examples on GPT-4.1.

4.2 Cloze translation completions

The cloze task tests whether models can use Indigenous translations to resolve ambiguous Spanish sentences. Each example includes a masked word and several plausible completions, filtered to exclude trivial cases. Accuracy gains in the bilingual condition suggest some understanding of the Indigenous input.

All models improve when given the translation, but gains are modest in smaller models and largest in Gemini-2.0, GPT-4.1, and DeepSeek. As shown in Table 2, these three models show statistically significant improvements for Guarani, Quechua, and Nahuatl, with Gemini outperforming the rest. For most other languages, improvements were smaller and often not significant. None of the models showed evidence of understanding Asháninka, Chatino, or Rarámuri.

4.3 Grammatical Feature Identification

The grammatical feature identification task probes a model's ability to extract morphosyntactic information from a sentence, such as tense, mood, or person. This setting differs from the previous two by requiring a deeper level of linguistic knowledge.

As shown in Table 3, model performance varies widely by language. Accuracy on Nahuatl is markedly higher than on Bribri across all models, with Gemini-2.0-flash achieving 66%. In contrast, all models struggled on Bribri, with scores barely above chance.

Model	Bribri	Nahuatl
GPT-4.1	0.27	0.55
DeepSeek-Chat	0.29	0.60
Gemini-2.0-flash	0.25	0.66
Qwen2.5-14B-Instruct	0.27	0.31
Llama-3.1-8B-Instruct	0.27	0.31
Qwen2.5-7B-Instruct	0.19	0.25
Qwen2.5-3B-Instruct	0.24	0.28
Random	0.22	0.27

Table 3: Model Accuracy Summary for Grammatical Feature Task

This contrast aligns with trends observed in the previous tasks. Bribri was rarely correctly identified in Task 1 and showed only modest gains in Task 2, suggesting that models have limited usable knowledge of the language. For the larger models, Nahuatl was consistently well-identified and contributed to improved cloze performance, indicating that LLMs have degree of familiarity with the language.

5 Conclusion

We introduced a benchmark to evaluate Indigenous language knowledge in large language models through three tasks: language identification, cloze completion with bilingual context, and grammatical feature classification. These tasks target different levels of linguistic competence, from surface recognition to morphosyntactic understanding.

Our results show that strong performance is concentrated in a small subset of languages, and only the most capable models demonstrate reliable improvements across tasks. While successful language identification can be a necessary condition for deeper understanding, it is not sufficient on its own. Even among the highest-performing models, meaningful gains are limited to languages with relatively greater digital presence, such as Nahuatl.

This benchmark provides a starting point for measuring and diagnosing model behavior in low-resource and typologically diverse settings. Future work will expand coverage to additional languages and tasks to further explore the limits of current models.

Limitations

This study is limited to a subset of 13 Indigenous languages. While we include multiple language families and typological profiles, the current benchmark does not represent the full diversity of Indigenous languages of the Americas.

Second, each task relies on prompting strategies which means that performance may be affected by prompt sensitivity, and results should be interpreted in that context.

Our grammatical feature classification task is limited to just two languages and a fixed set of features derived from existing annotations, which may not generalize to other grammatical systems.

Finally, we do not perform fine-tuning or adaptation, focusing instead on zero-shot and few-shot capabilities. This design choice reflects current deployment patterns for LLMs, but it leaves open questions about how models could be improved with modest supervision in these languages.

We use data released for the AmericasNLP 2025 Shared Task, which is publicly available via GitHub. As of writing, the repository does not specify a license. We do not currently redistribute the data, but we plan to seek permission from the original organizers to provide a version of the benchmark formatted for our evaluation tasks.

Ethical Considerations

While this benchmark is designed to evaluate and promote understanding of Indigenous languages in LLMs, Indigenous languages are not public resources in the same way as high-resource languages. Though our benchmark uses publicly available data, care must be taken to respect community ownership and avoid exploiting linguistic data without engagement or consent from language communities.

We restrict our benchmark to data that is already publicly available through shared tasks and prior publications, such as the AmericasNLP 2025 shared tasks. By working only with curated and previously released datasets, we aim to respect community ownership of linguistic resources and avoid introducing new risks related to consent or provenance. Our results are intended to highlight limitations and gaps in current model performance, not to promote deployment or commercial use.

We used some generative assistance in coding and surface-level editing of the paper. All edits to code and paper were thoroughly vetted by the authors.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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.

Adolfo Constenla Umaña, Feliciano Elizondo Figueroa, and Fransisco Pereira Mora. 1998. *Curso Básico de Bribri*. Universidad de Costa Rica, San José.

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.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, and 25 others. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Gemini, 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.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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)*, pages 6098–6111, Hong Kong, China. 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.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Elisabeth Mager, Vishrav Chaudhary, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, and Ngoc Thang Vu. 2022. AmericasNLI: Machine Translation and Natural Language Inference Systems for Indigenous Languages of the Americas. *Frontiers in Artificial Intelligence*, 5:995667.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. *arXiv*, abs/2004.01401.

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, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Max Ploner, Jacek Wiland, Sebastian Pohl, and Alan Akbik. 2025. LM-pub-quiz: A comprehensive framework for zero-shot evaluation of relational knowledge in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 29–39, Albuquerque, New Mexico. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, and 8 others. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Lutfi Kerem Senel and Hinrich Schütze. 2021. Does she wink or does she nod? a challenging benchmark for evaluating word understanding of language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 532–538, Online. Association for Computational Linguistics.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

A Languages

Ash'aninka is an Arawakan language of Peru and Brazil with about 70,000 speakers. It is agglutinative and polysynthetic, featuring complex morphology including gender, realis/irrealis, and classifier systems.

Awaj'un (Aguaruna) is a Chicham language spoken by 53,000 people in northern Peru. It has rich agglutinative morphology and SOV word order, encoding spatial and aspectual distinctions.

Aymara is spoken by 1.7 million people in Bolivia, Peru, and Chile. It is agglutinative with SOV word order, evidentiality, and a unique temporal-spatial metaphor.

Bribri is a tonal Chibchan language spoken in southern Costa Rica by 7,000 people. It features morphological ergativity, SOV word order, and gendered speech registers (Constenla Umaña et al., 1998).

Chatino is a Zapotecan language group in Oaxaca, Mexico. The San Juan Quiahije variant has around 5,000 speakers. It is tonal with complex inflection and variable word order.

Guarani is a Tupi–Guarani language spoken by 4–6.5 million people, mainly in Paraguay. It is agglutinative with nasal harmony, active–stative alignment, and flexible SVO order.

Nahuatl is a Uto-Aztecan language family with 1.6 million speakers in Mexico. It is agglutinative, polysynthetic, and features pronominal affixes and flexible word order.

Otom'i is spoken by about 300,000 people in central Mexico. We focus on the Ixtenco variety. It is tonal, SVO, and morphophonologically complex.

Quechua is an agglutinative language family with over 7 million speakers across the Andes. It features SOV word order, evidentiality, and rich suffixation. We use the Ayacucho variant.

Rar'amuri (Tarahumara) is spoken in northern Mexico by about 70,000 people. It is SOV, agglutinative, and polysynthetic, with noun incorporation and postpositions.

Shipibo-Konibo is a Panoan language of 26,000 speakers in Peru. It uses SOV word order, suffixal morphology, and evidential markers.

Wayuunaiki is an Arawakan language with 420,000 speakers in Colombia and Venezuela. It is SOV, agglutinative, and actively transmitted.

Wixarika (Huichol) is a Uto-Aztecan language spoken by 35,000 people in Mexico. It is polysynthetic, agglutinative, and SOV, with noun incorporation and vowel harmony.

B Prompts

Below are the prompts used for each of the tasks. The italics represent the parts that change for each instance.

Language	Family	Approx. Speakers	Location	Wikipedia?
Asháninka	Arawakan	74,500	Peru, Brazil	
Awajun	Chicham	53,400	Northern Peru	
Aymara	Aymaran	1,700,000	Bolivia, Peru	✓
Bribri	Chibchan	7,000	Southern Costa Rica	
Chatino	Oto-Manguean	5,000	Oaxaca, Mexico	
Guarani	Tupi–Guarani	6,500,000	Paraguay, Bolivia, Argentina, Brazil	✓
Nahuatl	Uto-Aztecan	1,600,000	Mexico, Central America	✓
Otomí	Oto-Manguean	300,000	Central Mexico	
Quechua	Quechuan	7,200,000	Andean regions	✓
Rarámuri	Uto-Aztecan	70,000	Northern Mexico	
Shipibo-Konibo	Panoan	26,000	Peru	
Wayuu	Arawakan	420,000	Colombia, Venezuela	✓
Wixarika	Uto-Aztecan	35,000	Mexico	

Table 4: Overview of languages

1. Language Identification (hard) [2]

You are a language identification model. You are given a sentence and you must identify the language it is written in. You will be given a number of choices, respond with the number of the correct choice.

What language is this sentence written in? Only give the number of the correct choice.
A ni machiyé mapu ke suwiníba je'ná jípi rokóo.

1. Aymara
2. Wixarika
3. Rarámuri
4. Guarani

The correct choice is:

2. Language Identification (open)

You are a language identification model. You are given a sentence and you must identify the language it is written in. Respond with the language name.

What language is this sentence written in? (language name only)
A ni machiyé mapu ke suwiníba je'ná jípi rokóo.

Language:

[2]Easy uses the same prompt but with incorrect options English, German, French. Full includes the full list of languages as options

3. Cloze-task (monolingual)

Selecciona la mejor opción para completar esta oración:
Solo ___ una semana.

1. pasó
2. fue
3. dura
4. tiene

Solo responde con el número de la opción correcta.

4. Cloze-task (bilingual)

Oración en *Aymara*:
Mä simanakiw
Selecciona la mejor opción para completar esta traducción:
Solo ___ una semana.

1. pasó
2. fue
3. dura
4. tiene

Solo responde con el número de la opción correcta.

5. Grammatical Feature Identification

You are a language expert who can identify grammatical features of a sentence in *Bribri*. You will be given a sentence, a category of grammatical feature (e.g., tense, mood, aspect), and a list of

Language	Language Identification	Cloze Translation	Grammatical Feature
Asháninka	459	461	-
Awajun	459	445	-
Aymara	459	435	-
Bribri	459	468	1,111
Chatino	459	165	-
Guarani	459	449	-
Nahuatl	459	417	1,949
Otomí	459	452	-
Quechua	459	432	-
Rarámuri	459	444	-
Shipibo-Konibo	459	426	-
Wayuunaiki	459	492	-
Wixarika	459	443	-

Table 5: Number of instances per task per language.

options. You must select the option that best matches the grammatical feature of the sentence.

Pûs kapóulur

What is the *tense* of this sentence?

1. continuous imperfect
2. past perfect
3. continuous perfect
4. potential future
...

Only respond with the number of the correct option.

C Dataset Statistics

Table 5 reports the number of instances per language for each task. While all 13 Indigenous languages are represented in the language identification and cloze tasks, the grammatical feature classification task is currently limited to Bribri and Nahuatl. Most languages contribute over 400 examples per task.

D Grammatical Feature Identification

Table 6 lists the grammatical features included in the classification task and the number of instances per feature for each language. While both Bribri and Nahuatl share categories such as tense and person, others, such as honorific, are language specific.

Grammatical Feature	Bribri	Nahuatl
Aspect	310	193
Honorific	-	119
Mode	79	-
Mood	-	122
Number of absolutive	86	-
Person	220	937
Polarity	-	224
Tense	416	354
Total	1111	1949

Table 6: Instances of grammatical features per language