Measuring Linguistic Competence of LLMs on Indigenous Languages of the Americas

Justin Vasselli Arturo Martínez Peguero
Frederikus Hudi Haruki Sakajo Taro Watanabe
Nara Institute of Science and Technology
vasselli.justin_ray.vk4@is.naist.jp, martinez_peguero.arturo.ma3@is.naist.jp,
frederikus.hudi.fe7@is.naist.jp, sakajo.haruki.sd9@naist.ac.jp, taro@is.naist.jp

Abstract

This paper presents an evaluation framework for probing large language models' linguistic knowledge of Indigenous languages of the Americas using zero- and few-shot prompting. The framework consists of three tasks: (1) language identification, (2) cloze completion of Spanish sentences supported by Indigenous-language translations, and (3) grammatical feature classification. We evaluate models from five major families (GPT, Gemini, DeepSeek, Qwen, and LLaMA) on 13 Indigenous languages, including Bribri, Guarani, and Nahuatl. The results show substantial variation across both languages and model families. While a small number of model-language combinations demonstrate consistently stronger performance across tasks, many others perform near chance, highlighting persistent gaps in current models' abilities with Indigenous languages.

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 [1, 2]. These factors make Indigenous languages a valuable test case for evaluating model generalization.

Multilingual NLP benchmarks such as XTREME [3] and XGLUE [4] concentrate primarily on high- and medium-resource languages with substantial digital presence. More inclusive initiatives like FLORES [5] and the Americas-NLP shared tasks [2] have introduced datasets for machine translation involving Indigenous languages, enabling evaluation in specific translation contexts. Still, there remains a gap in benchmarks that assess general language understanding.

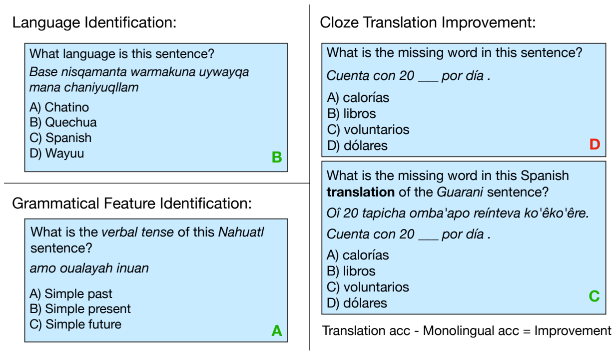

Figure 1 Overview of the tasks included in this benchmark

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 is asked to predict the missing word. In the grammatical feature identification task, the model is shown a sentence in an Indigenous language and is 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 [6]. Recent extensions include CLiMP for Chinese [7], JBliMP for Japanese [8], RuBLiMP for Russian [9], and MultiBLIMP [10], 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 [11], X-FACTR [12], and Multilingual LAMA [13] evaluate factual recall with cloze prompts. LM-PUB-QUIZ [14] converts these into multiple-choice form. WDLMPro [15] 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 [16] enables the evaluation of semantic inference in 10 Indigenous languages through translations of the XNLI corpus. MasakhaNER [17] and MasakhaPOS [18] benchmarks named entity recognition and part-of-speech tagging for African languages and explore cross-lingual transfer using multilingual and region-specific models. XTREME-UP [19] introduces a broad benchmark spanning 88 underrepresented languages across multiple user-facing tasks, enabling large-scale evaluation under low-resource constraints.

3 Methodology

Linguistic Corpora We use development data from the AmericasNLP 2025 Shared Task [20], 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.

Pre-trained Language Models We evaluate ten large language models spanning a range of architectures and sizes. These include GPT-4.1 [21], Gemini 2.0 Flash [22], and DeepSeek-V3-0324 [23], all of which we access through the API. We also test open-weight instruction-tuned models: LLaMA-3.1-8B, LLaMA-3.2-3B [24], and Qwen-3B, 7B, and 14B [25]. 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 A.

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, such as 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.

Cloze Translation Completion To test whether LLMs can understand 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 `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**, where only the Spanish cloze sentence is shown, and **Bilingual**, where both the Spanish cloze and Indigenous translation are shown.

We measure accuracy by comparing results between the two settings. The option order is fixed between settings to avoid position bias. Accuracy gains in the bilingual condition suggest some understanding of the Indigenous input. The final dataset contains 5,529 problems spanning 13 languages. Most languages contribute over 400 examples.

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.

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 (94%–100%), 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 2 shows per-language accuracy under this condition. Larger models performed more consistently, but even the smallest models correctly identified Quechua and Nahuatl, suggesting that these languages are relatively well represented in the pretraining data. Spanish was included as a high resource language, with Llama-3.1-8B performing the worst at 98%.

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

Figure 3 shows GPT-4.1' s confusion matrix in the full multiple-choice setting. Confusions are notably asymmetric: Rarámuri was rarely predicted and was often confused with Bribri; Bribri sentences were frequently mislabeled as Otomí. This suggests that 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

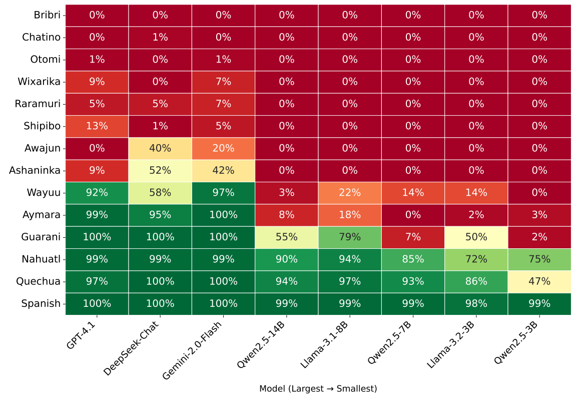

Figure 2 Indigenous Language Identification Accuracy on Open setting. Sorted by average performance.

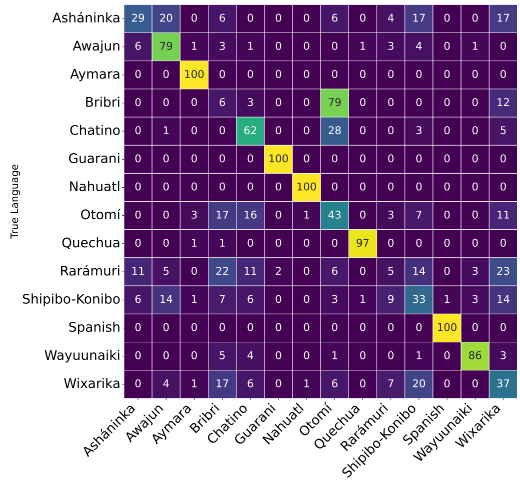

Figure 3 Confusion Matrix (%) for GPT-4.1 in Zero-shot setting. Vertical axis represents reference languages while horizontal axis represents the predictions.

models showed little improvement from additional shots.

We also examined the few-shot effects per language. Some languages showed substantial gains. For example, Bribri accuracy increased from 8.8% to 72%, and Otomí from 43.4% to 71.1%. Shipibo-Konibo and Wixarika also showed steady improvement with more examples.

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.

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 1, these three models show statistically significant improvements for Guarani, Quechua, and Nahuatl, with Gemini outper-

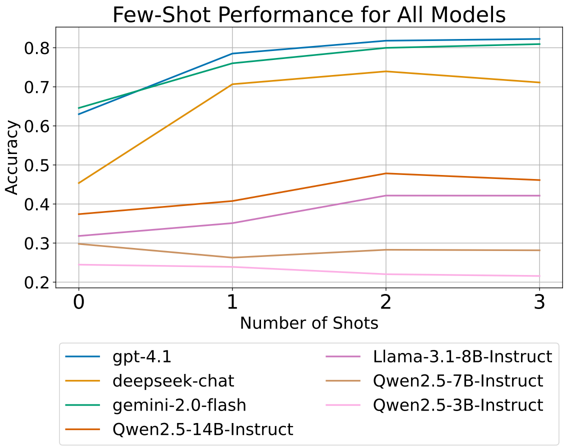

Figure 4 Few-shot accuracy comparison of various large language models across 0–3 shot settings.

Table 1 Accuracy improvement (ΔAcc) for each language on the cloze task when given the aligned Indigenous sentence. Significance levels from McNemar's test: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Language	Gemini-2.0	GPT-4.1	DeepSeek
Asháninka	1	4	3
Aymara	31***	22***	3
Awajún	12***	10***	11***
Bribri	4	7**	6*
Chatino	3	8	-1
Guarani	42***	40***	23***
Nahuatl	39***	26***	29***
Otomí	10**	7*	1
Quechua	45***	34***	36***
Rarámuri	-1	0	3
Shipibo-Konibo	11***	8*	11***
Wayuunaiki	9**	0	0
Wixarika	7*	5	6*

forming 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 2, 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,

Table 2 Accuracy (%) for each model on grammatical feature identification.

Model	Bribri	Nahuatl
GPT-4.1	27	55
DeepSeek-Chat	29	60
Gemini-2.0-flash	25	66
Qwen2.5-14B-Instruct	27	31
Llama-3.1-8B-Instruct	27	31
Qwen2.5-7B-Instruct	19	25
Qwen2.5-3B-Instruct	24	28
Random	22	27

with scores barely above chance.

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 some 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.

References

[1] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics.

[2] Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors, **Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas**, pp. 202–217, Online, June 2021. Association for Computational Linguistics.

[3] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. **CoRR**, Vol. abs/2003.11080, , 2020.

[4] Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation. **arXiv**, Vol. abs/2004.01401, , 2020.

[5] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics.

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

[7] Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 2784–2790, Online, April 2021. Association for Computational Linguistics.

[8] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In Andreas Vlachos and Isabelle Augenstein, editors, **Findings of the Association for Computational Linguistics: EACL 2023**, pp. 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[9] Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. RuBLiMP: Russian benchmark of linguistic minimal pairs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 9268–9299, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs, 2025.

[11] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[12] Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 5943–5959, Online, November 2020. Association for Computational Linguistics.

[13] Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 3250–3258, Online, April 2021. Association for Computational Linguistics.

[14] Max Ploner, Jacek Wiland, Sebastian Pohl, and Alan Akbik. LM-pub-quiz: A comprehensive framework for zero-shot evaluation of relational knowledge in language models. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)**, pp. 29–39, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[15] Lutfi Kerem Senel and Hinrich Schütze. Does she wink or does she nod? a challenging benchmark for evaluating word understanding of language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 532–538, Online, April 2021. Association for Computational Linguistics.

[16] 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. AmericasNLI: Machine Translation and Natural Language Inference Systems for Indigenous Languages of the Americas. **Frontiers in Artificial Intelligence**, Vol. 5, p. 995667, 2022.

[17] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. MasakhaNER: Named entity recognition for African languages. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1116–1131, 2021.

[18] 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, et al. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10883–10900, Toronto, Canada, July 2023. Association for Computational Linguistics.

[19] 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, et al. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 1856–1884, Singapore, December 2023. Association for Computational Linguistics.

[20] Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, et al. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In Manuel Mager, Abteen Ebrahimi, Robert Pugh, Shruti Rijhwani, Katharina Von Der Wense, Luis Chiruzzo, Rolando Coto-Solano, and Arturo Oncevay, editors, **Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)**, pp. 134–152, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics.

[21] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report, 2024.

[22] Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models, 2025.

[23] DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, et al. Deepseek llm: Scaling open-source language models with longtermism, 2024.

[24] 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. Llama: Open and efficient foundation language models, 2023.

[25] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report, 2023.

A Prompts

Below are the prompts used for each of the tasks. The italics represent the parts that change for each instance.[1)]

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:

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 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.

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.
2) Easy uses the same prompt but with incorrect options: English, German, and French. Full includes the complete list of languages as options