# Applying Linguistic Expertise to LLMs for Educational Material Development in Indigenous Languages

**Justin Vasselli** and **Arturo Martínez Peguero** and **Junehwan Sung** and **Taro Watanabe**
{vasselli.justin_ray.vk4, martinez_peguero.arturo.ma3, sung.junehwan.sl9, taro}@is.naist.jp

## Abstract

This paper presents our approach to the AmericasNLP 2024 Shared Task 2 as the JAJ (/dæz/) team. The task aimed at creating educational materials for indigenous languages, and we focused on Maya and Bribri. Given the unique linguistic features and challenges of these languages, and the limited size of the training datasets, we developed a hybrid methodology combining rule-based NLP methods with prompt-based techniques. This approach leverages the meta-linguistic capabilities of large language models, enabling us to blend broad, language-agnostic processing with customized solutions. Our approach lays a foundational framework that can be expanded to other indigenous languages languages in future work.

## 1 Introduction

In recent years, there has been a surge of interest in developing natural language processing (NLP) technologies for low-resource languages. This is motivated by the desire to preserve cultural heritage and foster linguistic diversity.

Among 7,000 living languages on Earth, 42% of languages are in danger of disappearing, according to Ethnologue[1]. In the Americas, there are approximately 1,200 languages (Hammarström et al., 2024), and about 70% of them are classified as endangered or already extinct by the United Nations Educational, Scientific and Cultural Organization (UNESCO) (Moseley and Alexandre, 2010). The endangered languages are losing their vitality, and intergenerational transmission is barely occurring or not occurring at all, which suggests these languages are at a high risk of becoming extinct in the near future, along with the the cultural legacy residing within. While many communities are engaged in revitalization endeavors, such as developing educational resources for language preservation, these initiatives are confronted with significant challenges in terms of cost and time investment. NLP can be one way of ameliorating the situation.

The AmericasNLP 2024 Shared Task 2 is specifically aimed at creating educational materials for indigenous languages in the Americas, such as Maya, Guarani, and Bribri (Chiruzzo et al., 2024). These languages each have their own unique linguistic features and challenges, offering a rich opportunity for the application of innovative NLP techniques.

In this shared task, we are given a dataset with base sentences and the grammatical changes expected to be applied to each base sentence. Based on this information, we will train a system that can transform the base sentences according to the indicated changes.

Base sentence: *Ye' shka'* (Bribri for "I walked")
Expected change: Polarity: Negative
Target sentence: *Ye' kë shkane* (Bribri for "I didn't walk")

We participated in this task with the goal of developing a method that effectively handles the linguistic complexities of these languages. Our hybrid methodology combines rule-based NLP methods with prompt-based techniques, making use of the meta-linguistic capabilities of large language models (LLMs). This approach enabled us to blend broad, language-agnostic processing with customized solutions that address each language's unique needs.

Maya largely does not conjugate verbs based on grammatical tense; instead, it uses aspectual marking to express time-related information. The abundance of apparently simple examples with only a single change, presented an interesting challenge. Rather than using all the available examples, we experimented with selection methods to find the most syntactically similar example sentences.

Bribri, on the other hand, had a much more com-

[1] https://www.ethnologue.com/insights/how-many-languages-endangered/



*Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 201–208
June 21, 2024 

plicated system of verb conjugation. Using more examples was not as effective as a separate system to handle the intricate verb morphology. This aspect of Bribri posed a substantial challenge, further complicated by variations in notation and spelling observed across different educational resources.

While our system addresses specific linguistic features of Maya and Bribri, we encountered time constraints that limited our ability to apply similar dedicated modifications to Guarani. Consequently, Guarani was processed using the base system, without the benefit of language-specific adaptations. Recognizing the unique linguistic structure and cultural significance of Guarani, we are committed to expanding our system in the future to include tailored strategies that cater specifically to its needs, enhancing the model's overall effectiveness.

## 2 Prior Work

Our approach is inspired by previous work on Rosetta Stone puzzles (Bozhanov and Derzhanski, 2013). These puzzles are carefully designed to encourage solvers to deduce implicit linguistic features from a low-resource language, relying solely on a limited set of bilingual sentence pairs, one of which is a low-resource language with uncommon linguistic features. Leveraging the inferred linguistic patterns, solvers are tasked with translating new sentences in both directions, which is in essence a few-shot translation task.

Şahin et al. (2020) explored the resolution of Rosetta Stone problems using a range of algorithms, from random word selection to transformer-based language models. Sung et al. (2024) conducted similar experiments, specifically focusing on assessing the metalinguistic awareness of pre-trained language models. Their findings not only underscored the limitations of various algorithms in tasks requiring linguistic inference skills but also outlined directions for advancing machines' comprehension of human language.

Recent advancements in the application of LLMs to Rosetta Stone puzzles have shown encouraging outcomes, demonstrating the potential of LLMs to extract and apply linguistic knowledge from limited data (Vamvas, 2022; Chi et al., 2024). This research supports the feasibility of our approach, suggesting that LLMs can adeptly adapt to the intricacies of low-resource languages by leveraging their inherent capabilities in pattern recognition and language comprehension.

Another relevant research area is the Machine Translation from One Book (MTOB) task (Tanzer et al., 2024). In this task, models are trained to translate between English and Kalamang using only a single Kalamang textbook. LLMs seem to be breaking into this space as well, as shown by Gemini 1.5l's state-of-the-art performance on this task. The MTOB task is particularly pertinent to our work because it demonstrates the potential of machine learning models to operate effectively even with severely limited datasets, mirroring the conditions we faced in the shared task (Pichai and Hassabis, 2024).

## 3 Method

### 3.1 Preparing the Examples

Due to the limited dataset size, ensuring the quality and consistency of data was paramount. We implemented several preprocessing steps to optimize the training data:

**Duplicate Removal and Capitalization:** We removed duplicate entries where the source and change tags were identical and standardized capitalization across the dataset to ensure uniformity. This reduced the number of example sentences in Maya from 594 to 584, but didn't affect Bribri or Guarani.

**Tag Collapsing:** Through detailed analysis of the training data, we identified and merged change tags that consistently appeared together, reducing redundancy. For example for changes that included both ASPECT and TENSE in Maya, we removed the TENSE tags as they did not appear to change the Target sentence.

**Sub-step Expansion:** We split some complex changes into simpler, sequential sub-steps. For instance, a change tagged as `TYPE:NEG, SUBTYPE:INT` was decomposed into two stages: first applying `TYPE:NEG` to reach an intermediary form, followed by `SUBTYPE:INT` to achieve the final state. This added 12 new cases to Maya and 6 new cases to Bribri.

**Backward Labeling:** When feasible, we generated additional training cases by labeling from the target back to the source, effectively doubling the data for those instances. The total number of examples used were 1199 for Maya (up from 594), 329 for Bribri (up from 309), and the 178 Guarani examples remained unchanged.

**Independence Analysis:** By consulting language-specific textbooks and analyzing the training data, we identified which changes were independent and which were interdependent, enabling more precise modeling of language rules.

### 3.2 Base System

Our base system is designed to leverage the capabilities of large language models by prompting them with relevant example cases. The system's operation during inference takes the same form, regardless of language, with some language specific modifications at key points.

Initially we simply included all available examples with the same change as the test case. However, we quickly found that not all test case changes were represented in the training data, particularly those with compound changes. To address this, we implemented a language-specific strategy for decomposing and processing the changes:

**Maya:** Changes that commonly occurred together (e.g., `ASPECT:BEG, TENSE:PAS_SIM`) were collapsed into a single step (`ASPECT:BEG`). The remaining changes were then split and processed sequentially in the order: `STATUS` → `PERSON` → `ASPECT` → `TYPE` → `SUBTYPE`.

**Bribri:** Changes that typically co-occurred (e.g., `ABSNUM:PL, PERSON:3_PL`) were combined into a single change (`PERSON:3_PL`). `PERSON` is then processed separately from the other changes.

**Guarani:** Changes are applied consecutively in the case of compound changes.

When a test case has a compound change, it is passed to the language specific function that determines which changes should be applied and if they can be applied together or sequentially. In the case of sequential changes, the system will reprompt the model with the modified output from the previous step until all specified changes are implemented. For instance, a Bribri case with the change `TENSE:PRF_REC, ASPECT:PFV, ABSNUM:PL, PERSON:3_PL` will undergo two rounds of processing, each focusing on one specific change. First the tense will be changed to recent perfect, then the result of that will be changed to 3rd person plural.

### 3.3 POS Tagging

A key component of our system is the application of custom, simplified part of speech (POS) taggers tailored to each target language. These taggers are primarily dictionary-based and are used to supplement the example sentences being passed to the LLM by explaining better the grammatical role of the words of the provided examples. The POS tagger for Maya focuses predominantly on function words, as these play a crucial role in understanding the grammatical structure of sentences. We didn't use a full dictionary for Maya, but made sure to have coverage of aspect markers such as *táan* and pronouns such as *in* or *teen* (Bolles and Bolles, 1996). Additionally our tagger is designed to recognize and handle known suffixes such as *e'ex*. For Bribri, we developed a POS tagger using a comprehensive dictionary of Bribri words from Professor Haakon S. Krohn's website[2] (Krohn, 2023).

### 3.4 The Prompt

The prompt was adapted from the one Vamvas (2022) used for the Rosetta Stone Puzzles. It integrates the retrieved examples with part of speech tags, as well as some language specific hints.

Here is a full example from Maya:

> System message:
> You are a helpful assistant with a strong background in linguistics.
>
> User message:
> This is a linguistic puzzle. Below are example sentences in a foreign language and sets of changes to apply to them.The examples are followed by the problem sentence and desired change.Your task is to look closely at the example sentences and to change the sentence correctly.
>
> Example 1:
> Sentence: Tin kaxtik a najil
> ((Tin, PRON:TA:1_S) (kaxtik, ::) (a, PRON:A:2_X) (najil, ::))
> Change(s): PERSON:3_PL, ASPECT:PRG
> Answer: Táan u kaxtiko'ob a naajil
> ((Táan, PART:ASP:) (u, PRON:A:3_X) (kaxtiko'ob, :kaxtik:3_P) (a, PRON:A:2_X) (naajil, ::))
>
> *Example 2-5*
>
> Here is the problem. Answer first, then explain your reasoning.
> Sentence: Tin wilaj u sáastal tin kaajal ((Tin, PRON:TA:1_S) (wilaj, ::) (u, PRON:A:3_X) (sáastal, ::) (tin, PRON:TA:1_S) (kaajal, ::))
> Change(s): PERSON:3_PL
> Hint:

### 3.5 Generation and Postprocessing

We tested generation on the development set using `gpt-3.5-turbo-0125`,

[2] https://www.haakonkrohn.com/bribri/index.html

`gpt-4-0125-preview` (OpenAI et al., 2024), and `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024). For the GPT models, we used temperature of 0. For Mixtral we used a greedy search.

While the generation can be stopped after "\n\n", we found that not setting the stop allowed the model to generate an explanation of the change. This was useful for interpretability, and we were able to spot some issues during development this way.

For postprocessing, the preamble (Answer:) and POS tags are stripped from the first line. Mixtral is less consistent with the answers, so we wrote some rule based processing to find potential answers in quotes.

Here is an example output from GPT-4:

> Answer: Tu yilo'ob u sáastal tu kaajalo'ob ((Tu, PRON:TA:3_X) (yilo'ob, :yil:3_P) (u, PRON:A:3_X) (sáastal, ::) (tu, PRON:TA:3_X) (kaajalo'ob, :kaajal:3_P))
>
> Explanation:
> The task involves changing the sentence from first person singular to third person plural. To do this, we follow the patterns observed in the examples.
> 1. **Pronoun Change**: In all examples, the first person singular pronoun "Tin" changes to the third person plural pronoun. In Examples 2 and 5, "Tin" changes to "Tu" for third person plural. Therefore, both instances of "Tin" in the problem sentence are changed to "Tu".
> 2. **Verb and Action Word Changes**: In the examples, verbs and action words that are in first person singular form are changed to their third person plural forms by adding an "o'ob" suffix or changing the suffix to match third person plural conventions. This is seen in Example 5 with "wilaj" changing to "yilo'ob" and "kaajali'" changing to "kaajalo'obi'". Since "kaajali'" is not present in the problem sentence, but "kaajal" is, we apply a similar transformation to "kaajal", changing it to "kaajalo'ob". The word "wilaj" similarly changes to "yilo'ob" following the pattern from Example 5.
> 3. **Maintaining Unchanged Parts**: Elements that do not pertain to person, such as "u sáastal" (indicating possession or association with "sáasta"), remain unchanged in the transformation, as seen across all examples.
> By applying these rules derived from the examples, we correctly transform the problem sentence to third person plural as requested.

### 3.6 Maya Adaptations

Initially, our system used all available examples for each change, but we refined our example selection process for Maya, where certain changes could return over 100 relevant examples.

We conducted experiments on the development set to find the optimal number of examples. These experiments varied the number of examples from one to ten per test case. The results indicated that selecting five examples struck the best balance between performance and cost-efficiency.

| System | Accuracy | BLEU | ChrF |
|---|---|---|---|
| 5 examples | 20.81 | 50.72 | 78.60 |
| + selected | 47.56 | 72.39 | 89.36 |
| + POS Tags | **56.38** | **78.26** | **91.33** |
| All matches | **56.38** | 77.63 | 91.30 |

Table 1: Different types of example retrieval on the Maya development set

The selection of these five examples was based primarily on syntactic relevance rather than semantic similarity or word form matching. By comparing the POS tagged source sentences of the examples with that of the input, we were able to identify examples that shared similar syntactic structures. This approach ensured that the examples used were not structurally analogous, enhancing the effectiveness of the model's performance.

For example if the input sentence is *Ma' ta jutaj a najili'*, an example sentence *Ta manaj a najil* overlaps significantly in word form, but the first sentence is a negative sentence, while the second is affirmative. This means they will react to the next change slightly differently. Instead, we match *Ma' ta jutaj a najili'* with *Ma' tin wa'alik u k'aaba'* which is structurally similar, despite less character-level overlap.

This method reduced computational costs by approximately 75% across the full test set, achieving comparable results to using the full example set, as shown in Table 1. Using five carefully selected examples with POS tags not only matches the performance of using all examples on the development set, but also achieves substantial cost savings, further justifying our approach.

### 3.7 Bribri Adaptations

Managing the complexity of Bribri verb conjugation, especially for irregular verbs, necessitated innovative methods to enhance translation accuracy. An initial experiment used an oracle verb conjugation hint, which leveraged correct verb forms from target sentences in the development set, to evaluate potential performance enhancements. This experiment led to a substantial increase in accuracy from 15% to 65%, signifying the crucial role of accurate verb conjugation in model performance.

Motivated by the success of the oracle hint, we

| System | Accuracy | BLEU | ChrF |
|---|---|---|---|
| Examples | 14.62 | 43.69 | 64.61 |
| + VERB form hint | 46.70 | 64.84 | 79.85 |
| + MODE hints | **47.17** | **67.01** | **80.75** |

Table 2: Ablation results for the hint categories on the Bribri development set

| Data | Accuracy | BLEU | ChrF |
|---|---|---|---|
| Maya dev | 56.38 | 78.26 | 91.33 |
| Maya test | 54.17 | 71.72 | 82.78 |
| Baseline | 25.81 | 53.69 | 80.23 |
| Bribri dev | 47.17 | 67.01 | 80.75 |
| Bribri test | 53.55 | 78.41 | 91.53 |
| Baseline | 8.75 | 22.11 | 52.73 |
| Guarani dev | 41.77 | 55.81 | 86.12 |
| Guarani test | 36.81 | 48.29 | 84.12 |
| Baseline | 14.84 | 25.03 | 76.10 |

Table 3: The results of the submitted system on the development and test sets.

developed a rule-based verb conjugator, employing a comprehensive database of verb conjugations sourced from Jara Murillo (2018)[3]. In this system, verbs identified by our POS tagger are looked up in the database to ascertain the correct conjugational form based on the grammatical context. For instance, in the sentence *Ye' tö i kötwa* with changes `TYPE:NEG, TENSE:FUT_CER, ASPECT:IPFV`, the verb *kötwa* is located by the POS tagger and looked up in the verb conjugation database. It's found to be the perfect remote form of *ujtökwa*. The conjugator transforms the verb into *ujtèpawa* for the negative certain future tense. This transformation is then included as a hint in the model's prompt:"The correct form of *kötwa* is likely *ujtèpawa*."

These enhancements, particularly the integration of verb conjugation hints, have markedly improved our system's performance, as detailed in the ablation study results in Table 2. While the rule-based conjugator does not replicate the perfect accuracy of the oracle hint, it significantly contributes to the overall effectiveness of the system in managing Bribri's complex verb conjugations.

### 3.8 Additional hints

Drawing from the Machine Translation from One Book (MTOB) task, we supplemented our approach with grammatical hints taken from textbooks. For each language, we incorporated short summaries of grammatical rules related to the changes from language textbooks. While time constraints limited the coverage of all possible changes, the preliminary results from these hints were promising and represent a straightforward avenue for further improvement.

## 4 Results

Our systems for Maya and Bribri improved on the baselines provided by the task organizers by considerable amounts. In particular, our system's Bribri accuracy was over six times higher than the edit-tree based baseline. This is likely due to the challenges of complex verb conjugation using an edit-tree approach. The complete results can be seen in Table 3.

As indicated in Tables 1, Maya became more resource efficient with example selection and POS tagging. Bribri performance saw the largest boost from verb hints and moderate improvements from mode hints, as shown in Table 2.

Our Bribri system was the best performing in the competition. Our Maya system came a close third. Due to this, and despite regrettably not submitting Guarani results, our contribution was ranked first overall. We submitted Guarani after the deadline, and report the results of all three languages on the development and test sets compared to the baseline in Table 3.

## 5 Discussion

We submitted our results using GPT-4 for system prompting due to its superior performance on the development set. Table 4 details the performance across different LLMs, noting that while Mixtral scored more competitively with GPT 3.5 for Maya, it was outperformed by both GPT models in the other languages.

### 5.1 Error Analysis

#### 5.1.1 Maya

The errors in Maya predominantly stem from inconsistencies in the example data and the complex syntactic structures that require deeper linguistic insights beyond mere pattern matching. For instance, the development set case *Te'exe' ti' kajakbale'ex tu yotoche'* (`TYPE:NEG`) changes to *Ma' kajakbale'ex tu wotochi'*, contrasting with a similar training example *Leti'obe' ti' u taalo'obi'* (`TYPE:NEG`) that

[3]https://www.lenguabribri.com/gramática-de-la-lengua-bribri

| Lang | System | Acc. | BLEU | ChrF |
|---|---|---|---|---|
| Maya | Mixtral | 44.97 | 69.19 | 83.52 |
| | GPT-3.5 | 42.28 | 67.84 | 86.04 |
| | GPT-4 | **56.38** | **78.26** | **91.33** |
| Bribri | Mixtral | 34.43 | 42.86 | 72.06 |
| | GPT-3.5 | 40.57 | 61.15 | 77.04 |
| | GPT-4 | **47.17** | **67.01** | **80.75** |
| Guarani | Mixtral | 12.66 | 20.95 | 69.84 |
| | GPT-3.5 | 36.71 | 51.38 | 83.35 |
| | GPT-4 | **41.77** | **55.81** | **86.12** |

Table 4: The results on the development set for the different LLMs.

becomes *Leti'obe' ma' ti' u taalo'obi'*, where *Leti'obe* and *ti'* are retained. These inconsistencies, including the absence of examples for changes like *tu yotoche'* to *tu wotochi'*, contribute to over 50% of the errors.

Additionally, positioning of *wáaj* in interrogative sentences varied without clear rules, leading to misplacement about 25% of the time. These idiosyncratic cases highlight the need for more robust language-specific rules in our system.

The remaining errors involved rule misapplication, such as overuse of the suffix *-o'ob*, and spelling mistakes like failing to correctly modify *tu wotoch* to *tu yotoch*.

#### 5.1.2 Bribri

For Bribri, verb conjugation continues to be a major issue, accounting for 57% of the errors. These range from minor issues like incorrect accent placement (e.g., *sùr* instead of *súr*) to significant errors such as incorrect verb forms (e.g., *kötwa* instead of *ujtèkèulur*). Another 19% of errors were due to omissions, where words present in the reference were missing in the hypothesis (e.g., *Ppö* instead of *I ppö*).

Less frequently, errors involved extraneous words in the hypothesis, making up 9% of the total errors (e.g., *Ye' wa stsa'* instead of *Ye' stsa'*). Incorrect pronoun use accounted for 8% of the errors, and the remaining 6% were due to words appearing out of order (e.g., *Kë ie' stsö* instead of *Ie' kë stsö*).

#### 5.1.3 Guarani

For Guarani, although no language-specific optimizations were implemented, the error analysis indicates that the majority of the issues are related to verb conjugation. Specifically, incorrect verb forms account for 75% of the errors observed. This suggests that developing a system similar to the rule-based verb conjugator used for Bribri, which provides hints based on accurate verb conjugation, could be highly beneficial in improving the accuracy for Guarani. Implementing such a system could significantly reduce errors and enhance the model's overall performance for this language.

## 6 Conclusion

Supplementing the capabilities of LLMs, such as GPT-4, by incorporating simple rule-based natural language processing techniques, our approach to the AmericasNLP 2024 Shared Task 2 has laid a foundational framework that can be expanded in future work to include other low-resource languages, contributing to the creation of educational materials for indigenous languages.

Particularly effective were the custom verb conjugation hints for Bribri, which markedly improved accuracy. This approach underscores the ongoing need for focused linguistic tools tailored to the specific structural complexities of each language.

Throughout the project, we encountered several challenges, including inconsistencies in the training data and the complex nature of indigenous language structures that often deviate significantly from those of more widely studied languages. These issues underscore the importance of developing tailored NLP tools that can adapt to the idiosyncrasies of any given language.

Looking ahead, we aim to extend our methodology to include Guarani more comprehensively, enhance our rule-based systems for better accuracy, and further explore the potential of LLMs in processing linguistically diverse and low-resource languages. This work not only contributes to the field of computational linguistics by providing valuable insights into the treatment of indigenous languages but also plays a crucial role in the preservation and revitalization of these vital cultural heritages.

By continuing to refine our approaches and expand our linguistic coverage, we hope to contribute to a more inclusive and equitable representation of languages in the digital age, ensuring that technology serves as a bridge rather than a barrier in the education and preservation of linguistic diversity.

## Acknowledgments

A special thank you to Professor Carla Victoria Jara Murillo and Professor Haakon S. Krohn who generously allowed us to use and repackage their

Bribri data for use in this project.

## References


David Bolles and Alejandra Bolles. 1996. *A Grammar of the Yucatecan Mayan Language*.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119, St. Julian's, Malta. Association for Computational Linguistics.

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

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. glottolog/glottolog: Glottolog database 5.0.

Carla Victoria Jara Murillo. 2018. *Gramática de la lengua bribri*, volume 1. EDigital, San José.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Haakon S Krohn. 2023. Diccionario bribri–español español–bribri.

Christopher Moseley and Nicholas Alexandre. 2010. *Atlas of the world's languages in Danger*, 3rd edition. Memory of peoples. United Nations Educational, Scientific and Cultural Organization (UNESCO).

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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav

Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sundar Pichai and Demis Hassabis. 2024. Our next-generation model: Gemini 1.5.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Junehwan Sung, Hidetaka Kamigaito, and Taro Watanabe. 2024. Exploring metalinguistic awareness in pre-trained language models through the international linguistics olympiad challenges. In *Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing*, Kobe, Japan. Association for Natural Language Processing.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Jannis Vamvas. 2022. Translation puzzles are in-context learning tasks.