A Comparison of Fine-Tuning and In-Context Learning for Clause-Level Morphosyntactic Alternation

Jim Su*, **Justin Ho***, **George Aaron Broadwell, Sarah Moeller, Bonnie J. Dorr**
University of Florida
{jimsu, justinho, broadwell, smoeller, bonniejdorr}@ufl.edu

Abstract

This paper presents our submission to the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages. We frame this task as one of morphological inflection generation, treating each sentence as a single word. We investigate and compare two distinct approaches: fine-tuning neural encoder-decoder models such as NLLB-200, and in-context learning with proprietary large language models (LLMs). Our findings demonstrate that for this task, no one approach is perfect. Anthropic's Claude 3 Opus, when supplied with grammatical description entries, achieves the highest performance on Bribri among the evaluated models. This outcome corroborates and extends previous research exploring the efficacy of in-context learning in low-resource settings. For Maya, fine-tuning NLLB-200-3.3B using StemCorrupt augmented data yielded the best performance.

1 Introduction

The AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024) focuses on the challenge of clause-level morphosyntactic alternation for low-resource indigenous languages of the Americas. The objective of this task is to develop a system capable of applying a set of grammatical attributes to a given source sentence, thereby generating a target sentence with the desired changes. The motivation behind this task lies in the potential for such systems to aid in the preservation and revitalization of endangered languages (Anastasopoulos and Neubig, 2019).

This task involves three indigenous languages of the Americas: Bribri, Guaraní, and Maya. Two examples are provided below:

Example 1.1. *Bribri*
Source sentence: Ye'shka' ("I walked")
Attributes: TYPE:NEG (negative polarity)
Target sentence: Ye'kë shkàne̲ ("I didn't walk")

Example 1.2. *Maya*
Source sentence: Táan in xíimbal tu jáal já' ("I'm walking on the beach")
Attributes: TYPE:NEG (negative polarity)
Target sentence: Ma' táan in xíimbal tu jáal ja'i' ("I'm not walking on the beach")

We frame this task as one of morphological inflection generation, treating each sentence as a single word. Our objective is thus twofold: to develop a system that performs sentence-level morphological inflection for low-resource indigenous languages of the Americas, and to provide insight into what techniques are effective for future practitioners who attempt this task. In pursuit of this goal, we compare the performance of two distinct approaches: fine-tuning pre-trained transformer models and leveraging LLMs through in-context learning. By evaluating these two approaches, we aim to contribute to the understanding of effective strategies for addressing the unique challenges posed by low-resource languages in tasks such as morphosyntactic alternation.

2 Background

This task is unique, as previous literature has explored morphological inflection generation on the word level rather than on the sentence level (Nicolai et al., 2023). Further, this task is challenging for two reasons:

1. Data scarcity: low-resource indigenous languages, by definition, have limited available data for training and evaluating machine learning models (Liu and Dorr, 2024). The scarcity of parallel corpora, annotated texts, and linguistic resources poses significant obstacles in developing robust morphological inflection

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 179–187
June 21, 2024 ©2024 Association for Computational Linguistics

systems (Moeller, 2021). This scarcity is compounded by the novel nature of this task as prior literature is scarce.

2. Unusual linguistic properties: Indigenous languages of the Americas exhibit a wide range of linguistic properties that diverge from those of well-studied languages like English or Spanish. These languages often feature intricate morphological, phonological, and orthographic systems (Dagostino et al., 2024). They may be polysynthetic and adhere to irregular morphological paradigms. Such linguistic properties can make it challenging to model computationally, especially in the context of limited training data.

Prior results have demonstrated the effectiveness of transformer-based models (Vaswani et al., 2023) on the word-level inflection task (Anastasopoulos and Neubig, 2019). Building upon this success, we evaluate the performance of fine-tuned transformer models on the sentence-level task, exploring their ability to capture and generate morphological inflections in context.

In an effort to extend the available data, we search for external sentence-level parallel corpora aligned with the task format. However, our search yields no suitable resources. While it may be possible to preprocess and adapt data from other formats in a separate pre-training stage, this approach is complex and may require a significant time investment for developing custom preprocessing pipelines, which is not possible in our study, given the short time-frame of this shared task.

To address the challenge of limited data resources, we opt for data augmentation using StemCorrupt (Anastasopoulos and Neubig, 2019), generating synthetic instances based on the existing data. StemCorrupt is a data augmentation technique created for generating additional instances for the word-level inflection task. The use of StemCorrupt is motivated by the availability of pre-existing code and the relative simplicity of this technique, which allow us to quickly run data augmentation and focus our efforts on other aspects of the task.

The limited supervised data challenge also prompts us to explore the use of proprietary large language models (LLMs). These models have the capability to process long context windows of arbitrary text as input and do not require fine-tuning, making them a promising alternative for low-resource settings where extensive task-specific data is unavailable.

Recent advancements have demonstrated that by scaling training compute and corpus size, LLMs may excel in tasks for they which they are not explicitly trained (Wei et al., 2022). Studies exploring the use in-context learning with LLMs on low-resource machine translation have shown promising results (Tanzer et al., 2024). More recent work in the area suggests that when paired with appropriate language resources, LLMs can even surpass human baselines in translation quality (Reid et al., 2024). These findings highlight the potential of in-context learning in LLMs for addressing the challenges posed by low-resource languages and the importance of incorporating relevant linguistic knowledge to maximize their performance.

3 Data

For fine-tuning, we use the provided training dataset[1] and an augmented dataset that we create by applying StemCorrupt to the provided training dataset. For in-context learning, we experiment with the inclusion of a grammatical description in the prompt. Previous work investigating the use of proprietary LLMs on low resource languages has shown that, when combined with grammatical descriptions, these models obtain strong performance on tasks such as machine translation (Tanzer et al., 2024). We hypothesize that using grammatical descriptions in an in-context learning setting can improve performance on this task as well.

3.1 Training Data

The training set provided by the organizers contains 1199 training instances. These instances consist of 594 Maya instances, 427 Bribri instances, and 178 Guarani instances. This dataset is somewhat imbalanced, with Guarani comprising only 14.8% of all training instances.

Each instance contains a set of *change tags*, i.e., morphosyntactic attributes that act as functors (such as TYPE:NEG to indicate negation of sentence polarity). Across all languages, there are 77 unique change tags. These follow a long-tailed distribution: some tags are shared across languages while others are unique to a particular language. Refer to the Appendix A for an exhaustive distribution of change tags.

[1] https://github.com/AmericasNLP/americasnlp2024

3.2 Data Augmentation

We perform data augmentation to generate 1000 additional instances for each language. Prior literature has demonstrated the efficacy of StemCorrupt for improving the performance of language models on word-level inflection tasks (Samir and Silfverberg, 2023). We explore the effect of StemCorrupt on this task at the sentence level.

3.3 Utilizing Grammatical Descriptions

We encounter two challenges to using published grammatical descriptions of these three languages:

1. Grammatical descriptions are difficult to find, and the orthographies used in them vary. Many of the resources that do exist were published in the 1960-80s or earlier and are only accessible online as PDF images of printed text or in digital formats that do not translate easily into correct Unicode characters.[2] We narrow our search to resources that use English as the analysis language which limit our choices further since other descriptions seem to be available in Spanish. Finally, we search for grammatical descriptions with interlinear glossed text (e.g. Umaña et al. (1998)) in order to provide information similar to the change tokens provided in the shared task data.

2. The length of data passed into an LLM is limited by its context window, establishing a hard limit on how much data (in particular, excerpts from the published resources) can be passed into the model. Even within this hard limit, particularly long input sequences can degrade performance (Li et al., 2024).

We employ the following grammatical descriptions, focusing on passages that contained interlinear glossed texts:

1. Bribri - Dickeman-Datz (1985) and Jara (1995)

2. Guarani - Estigarribia (2020)

We are unable to find a suitable grammatical description for the Yucatec Maya language that matched the orthography used in this task.

[2]For example, the scans of these typewritten Peace Corps language learning lessons: https://www.livelingua.com/project/peace-corps/guarani or this image of a Bribri grammatical description: http://journals.uvic.ca/index.php/WPLC/article/view/5054/1954

3.4 Data Processing

Since curated data is provided by the shared task organizers, minimal preprocessing is required. The Bribri data needs some additional preparation. For Maya and Guarani, no preprocessing is done.

For the Bribri language, training instances are provided in both the data/ and pilotdata/ directories. We concatenate the training sets and development sets across data/ and pilotdata/.

The Bribri data/ directory contains the straight apostrophe (') character while the pilotdata/ directory contains the right single quotation mark (’). We replace each instance of the right single quotation mark in the Bribri pilot training data with the straight apostrophe.

4 Experiments

We perform four experiments and compare the results:

1. Fine-tuning the pre-trained encoder-decoder models

2. Fine-tuning the pre-trained encoder-decoder models with data augmentation

3. In-context learning on proprietary LLMs

4. In-context learning on proprietary LLMs with a grammatical description

4.1 Experiment Setup

We apply two classes of experimental setups: fine-tuning and in-context learning. Fine-tuning adapts a pre-trained model to predict the target column one instance at a time. In-context learning includes the full training set and instances from the validation set in the prompt of an LLM, predicting multiple targets per inference run.

Both setups have strengths and weaknesses. In-context learning is constrained by a fixed context window but can work on arbitrary forms of task information such as grammatical descriptions. In contrast, fine-tuning allows the model's parameters to be updated on an arbitrarily large training dataset but requires task-specific parallel data that is challenging to find for low-resource languages.

4.1.1 Fine-Tuning

For each training instance, we concatenate the source sentence with the change tags. A separator token is used to delimit the end of the source sentence and the start of the change token. A model

Model	Bribri	Guarani	Maya
Baseline			
(Kann and Schütze, 2016)	5.66	**22.78**	26.17
BART Family			
BART-Large	7.11	2.53	44.96
MBART-50	12.89	0.00	9.39
T5-FLAN Family			
FLAN-T5-XL	1.33	0.00	2.01
NLLB-200 Family			
NLLB-200-distilled-600M	19.55	21.51	49.66
NLLB-200-distilled-600M (+ StemCorrupt)	20.00	16.45	**58.39**
NLLB-200-3.3B	24.88	16.45	53.02
NLLB-200-3.3B (+ StemCorrupt)	28.44	21.51	52.35
In-Context Learning			
Claude 3 Opus	30.53	18.99	54.36
Claude 3 Opus (+ grammatical description)	**36.73**	17.72	N/A
Gemini 1.5 Pro	8.41	N/A	44.97
Gemini 1.5 Pro (+ grammatical description)	12.21	N/A	N/A

Table 1: Dev set accuracy score for all fine-tuned models. Bold means best performing model for that language. It is worth noting that for Maya, we are not able to find grammatical descriptions that matched the orthography of the task. As for Gemini 1.5 Pro, we suspect there may be an issue with the tokenizer for Guarani as the model would generate few predictions before failing.

is trained for each language as opposed to creating a single multi-lingual inflection model, since we find the former results in better performance over the latter. We run our experiments on a single A100 GPU using a batch size of 64. We follow the same evaluation scheme proposed by the organizers using accuracy, chrF, and BLEU (Popović, 2015; Papineni et al., 2002).

4.1.2 In-Context Learning

For each in-context learning experiment, the LLM is provided the following:

1. The training set, with IDs replaced by the row number
2. The development set with changes removed
3. A relevant prompt (refer to the Appendix A for exact prompts used)

4.2 Fine-Tuning Pre-Trained Encoder-Decoder Models

We fine-tune a variety of encoder-decoder model families. Different variants of BART are used such as mBART to evaluate the effect of multi-lingual pre-training on this task (Lewis et al., 2019; Liu et al., 2020). The FLAN-T5 series of models are also evaluated as these models incorporate a unique pre-training process that is promising in terms of boosting model performance (Chung et al., 2022). The last family of models examined is the NLLB-200 family of models for their strong performance on low-resource translation (Team et al., 2022). We experiment with the 600M and 3.3B parameter version of each model. Although the NLLB-200 also includes a Mixture of Experts (MoE) model that may outperform the other versions, this model is not investigated due to its prohibitive size (54B parameters, which exceeds the memory capacity of an A100 GPU).

4.3 Fine-Tuning Pre-Trained Encoder-Decoder Models with Data Augmentation

StemCorrupt is used to generate 1000 instances for each language. Only NLLB-200 3.3B is trained using the augmented StemCorrupt data as this is the best performing model found during fine-tuning on non-augmented data.

4.4 In-Context Learning on Proprietary LLMs

We evaluate two proprietary LLMs:

1. Gemini 1.5 Pro. This model is selected for its long context window and strong performance

Model	Window Size	Strategy
Gemini 1.5 Pro	1 Million	All
Claude 3 Opus	200k	Relevant

Table 2: Context window size for each model and document strategy used.

on in-context low-resource machine translation (Reid et al., 2024)

2. Claude 3 Opus. This model is selected as the current state of the art in proprietary LLMs (Anthropic, 2024).

We briefly evaluate GPT-4 Turbo but encounter significant challenges (OpenAI et al., 2024). The model produces outputs of unacceptably low quality, rendering them effectively unusable. Additionally, GPT-4 Turbo proves unstable, consistently failing to fully process the full test dataset.

4.5 In-Context Learning on Proprietary LLMs with a Grammatical Description

Each LLM evaluated by our team is constrained by a different context window length, which affects the strategy used for passing in the grammatical description. Our team relies on two strategies: passing all grammar resources to the model and passing the most relevant grammar resources to the model. The most relevant grammar resource for each language is determined by hand, with the most frequent change tokens for each language guiding this search. The selected section for each language describes the language's morphology, verbal agreement system, and syntax of various sentence types including affirmative statements, negation, and questions.

5 Results

5.1 Dev Set Results

Table 1 shows the results of our fine-tuned models and in-context learning experiments on the dev set. For all languages except for Guarani, we are able to exceed the baseline performance significantly. For Bribri, in-context learning combined with a grammatical description is the highest performing technique with an accuracy of 36.73% over the baseline of 5.66%. For Maya, fine-tuned NLLB-200-distilled-600M with StemCorrupt augmented data is the best technique with an accuracy of 58.39% over the baseline of 26.17%.

Model	Bribri	Guarani	Maya
Baseline	8.75	**14.84**	25.81
Submission 1			
NLLB-200-3.3B	9.79	0.00	**37.42**
Submission 2			
Claude 3 Opus	**26.88**	0.00	33.23

Table 3: Test set accuracy score for all fine-tuned models. Bold means best performing model for that language.

There are unique trends that can be observed from the results of our system runs on the dev set. As anticipated, Guarani proves challenging to improve upon due to particularly limited data. Even when data augmentation techniques are applied, results of neural techniques are still below that of the statistical-backed baseline. This result reaffirms the findings of prior literature in terms of the weaknesses of neural techniques under sparse data conditions. Furthermore, this result hints at morphological or linguistic complexities in Guarani that make this task challenging.

Comparing fine-tuning and in-context learning, no technique was optimal across all languages. This result affirms two ideas: fine-tuning models is still relevant in the age of LLMs, and LLMs empowered with language resources are a viable approach for this task. For Bribri, fine-tuned models–even with data-augmentation–are not able to match the best performing in-context learning LLM.

5.2 Test Set Results

Table 3 details the test results for our submissions. Our model's performance on the test set exhibits an unexpected discrepancy compared to its performance on the dev set. Both of our best systems for the dev set underperform significantly when evaluated on the test set. Compared to the dev set, the accuracy on the test set is 20% lower for Maya and 10% lower for Bribri. This significant drop in performance warrants further investigation to identify potential causes, such as differences in domain, style, or linguistic properties between the dev and test sets.

Despite this unexpected discrepancy, it is worth noting that our team achieved the second best system submission for Bribri. Without access to the target column of the test set, the exact reason remains unclear. With such a limited number of training instances, both the in-context learning and fine-tuned model may not have enough examples to

generalize to different data distributions. Due to weak performance of our system runs applied to the dev set (and our misunderstanding that only above-baseline runs are submissible), our team has no submitted Guarani results for the test set.

6 Future Work

Much future work remains. Our search for language resources reveals a wide variety of language resources of varying types and orthographies. A future area of research is an exploration of the effect of different orthographies on the LLM performance in an in-context learning setting.

Additionally, a significant advantage of in-context learning is the reduction of restrictions on data types that can be utilized by the model. Exploring the effect of different resource types, such as dictionaries and learning worksheets, would be valuable. A historic bottleneck for the translation or inflection of low-resource languages has been data, specifically gold-standard data that adheres to a specialized format. By leveraging in-context learning, the variety of usable data is greatly increased and can offer opportunities for further exploration.

StemCorrupt has shown promise for sentence-level inflection despite initially being developed for word-level inflection. Exploring the feasibility of extending this technique to other languages is a worthwhile future endeavor.

7 Conclusion

In this paper, we present the systems submitted by our team for the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages. We find that while LLMs–through in-context learning–exhibit impressive capabilities, fine-tuning still has a role to play in the modern NLP space. Moreover, we reaffirm the results of prior literature regarding the promise of LLMs when applied to low-resource languages using in-context learning. Additional work must be done to explore the abilities of such systems, but initial results point to promising potential for the task of morphosyntactic alternation. Our work also extends prior literature on StemCorrupt and demonstrates potential applications for the technique on sentence-level inflection generation.

Limitations

The main limitation of our work is selecting only grammatical descriptions published in English. More grammatical descriptions are available in Spanish.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Carmen Dagostino, Marianne Mithun, and Keren Rice, editors. 2024. *The Languages and Linguistics of Indigenous North America*. De Gruyter Mouton, Berlin, Boston.

Margaret Dickeman-Datz. 1985. Transitivity in indefinite voice in bribri. *International journal of American linguistics*, 51(4):388–390.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. UCL Press.

Carla Victoria Jara. 1995. *Text and context of the Suwo': Bribri oral tradition*. Louisiana State University and Agricultural & Mechanical College.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

555–560, Berlin, Germany. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Zoey Liu and Bonnie J. Dorr. 2024. The Effect of Data Partitioning Strategy on Model Generalizability: A Case Study of Morphological Segmentation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Sarah Moeller. 2021. Computational morphology for language documentation and description. *Colorado Research in Linguistics*, 25.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors. 2023. *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Toronto, Canada.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evalu-

ation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, and Yuanzhong Xu. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Farhan Samir and Miikka Silfverberg. 2023. Understanding compositional data augmentation in typologically diverse morphological inflection. *Preprint*, arXiv:2305.13658.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

A.C. Umaña, F.E. Figueroa, and F.P. Mora. 1998. *Curso básico de bribri*. Editorial de la Universidad de Costa Rica.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

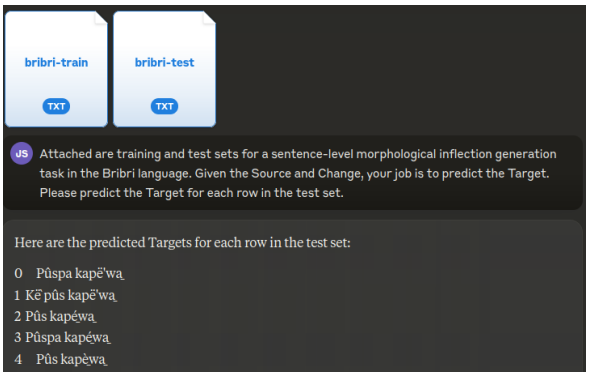
Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

A Appendix

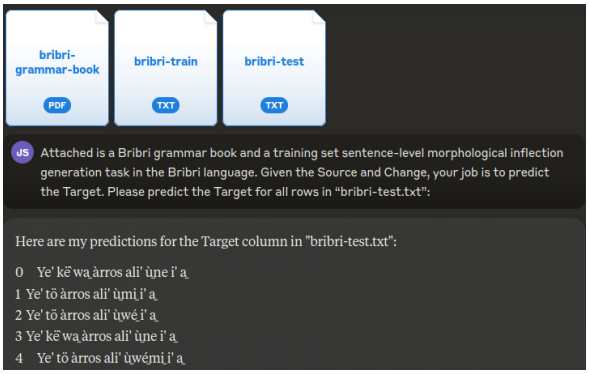
A.1 Proprietary LLM Prompts

The same prompt is used for all LLMs tested. The below screenshots are taken from Anthropic's claude.ai interface.

A.1.1 Without grammatical description



A.1.2 With grammatical description



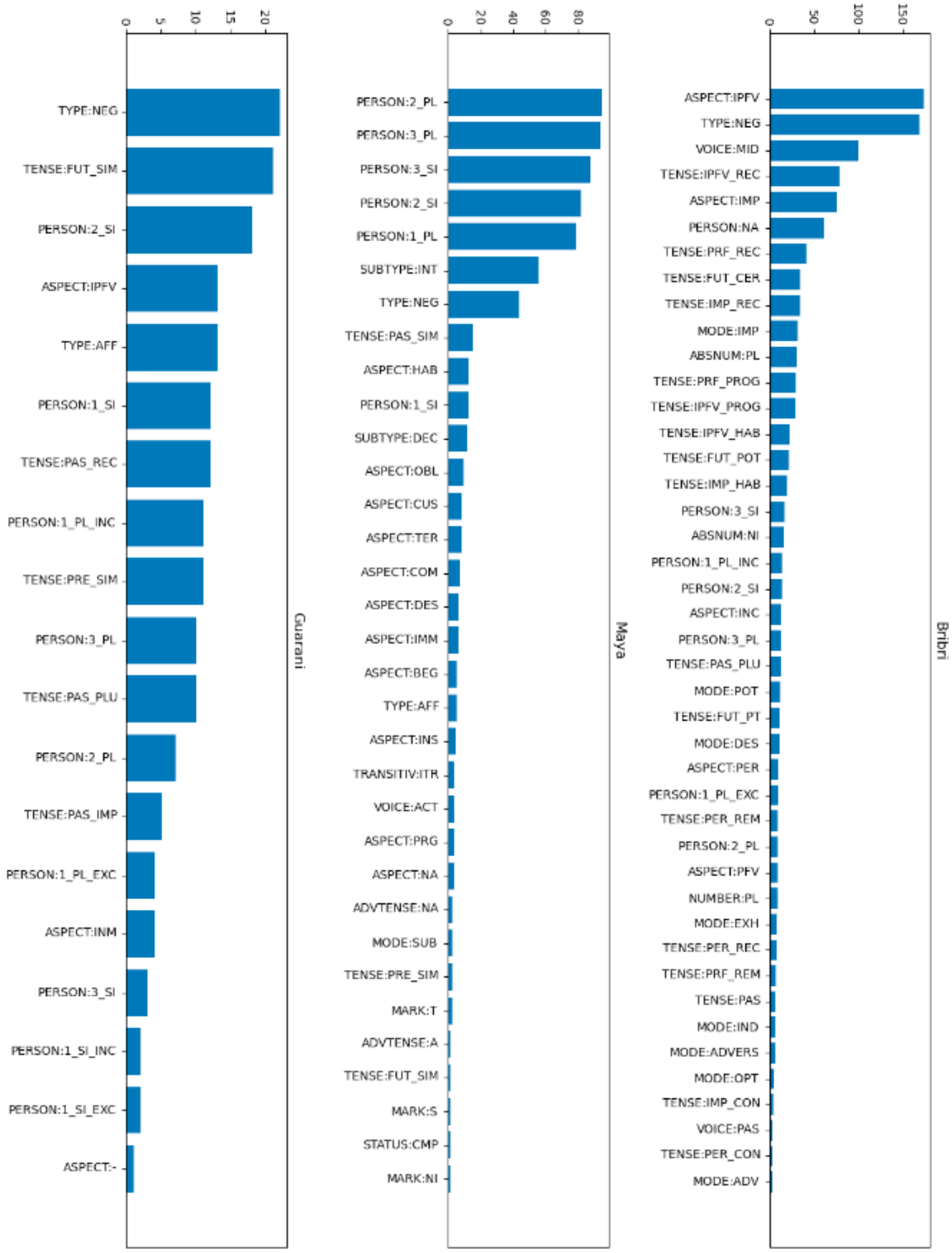


Figure 1: Distribution of change tags for each language.