The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure?

Coleman Haley
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
coleman.haley@ed.ac.uk

Abstract

This paper describes the submission of Team "Giving it a Shot" to the AmericasNLP 2024 Shared Task on Creation of Educational Materials for Indigenous Languages. We use a simple few-shot prompting approach with several state of the art large language models, achieving competitive performance on the shared task, with our best system placing third overall. We perform a preliminary analysis to determine to what degree the performance of our model is due to prior exposure to the task languages, finding that generally our performance is better explained as being derived from in-context learning capabilities.

1 Introduction

This paper describes the submission of Team "Giving it a Shot" to the AmericasNLP 2024 Shared Task on Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024). This task covers three indigenous languages of the Americas: Yucatec Maya (yua), Guaraní (grn), and Bribri (bzd). The task is similar to the clause-level reinflection task described by Goldman and Tsarfaty (2022) and explored in a 2022 MRL shared task (Goldman et al., 2022). However, it is more challenging in a number of ways. The first is structural: the present shared task provides an input sentence, and what values of features should be changed, while the previous task provided all feature values present in the input and what they should be changed to. As such, the features describing the source must be learned latently. Other challenges come from differences in the languages covered: all three languages in this shared task are relatively low-resource, and correspondingly the training data in the shared task is also very limited (595 training examples at most).

However, as a morphological/morphosyntactic task[1], the input-output functions are relatively simple compared to many common tasks in NLP, being in all likelihood context-free or even regular (Karttunen and Beesley, 2005; Pullum and Gazdar, 1982; Roark and Sproat, 2001). Increasingly in NLP, even computationally complex tasks such as sentiment analysis are being framed as few-shot tasks for large language models (LLMs), with impressive results being obtained by presenting a few examples to a language model and allowing it to perform next-token prediction (Wang et al., 2024; Wei et al., 2022; Brown et al., 2020). The ability of such paradigms to improve performance over raw language model probabilities has been termed in-context learning; however, this term has been the subject of controversy, as it is not learning in the traditional machine learning sense, nor is it clear exactly how much information is being extracted from the context. For example, in the "in-context learning" of sentiment analysis, much of the relation between a sentence and a sentiment label is presumably latent in the pre-trained weights, and the examples serve moreso to "extract" that information from the model, enabling better generalization than could be expected from the information in the provided in-context examples alone.

This setting therefore represents an interesting case: if few-shot prompting works well here, will it be due to prior language exposure, or an ability to generalize simple functions from limited data? To explore this question, we create three simple few-shot prompting-based systems, based on two closed-source LLMs (GPT-3.5 and GPT-4) and one openly available model (Command R+), finding they perform competitively on the shared task. We permute the characters in the dataset to preserve the problem stucture while ablating language information, finding some evidence that the models primarily generalize in-context data, rather than using prior language exposure.

[1] Note that, as in prior work on clause-level morphology, the functions involved sometimes operate at the clause level

June 21, 2024 ©2024 Association for Computational Linguistics

```
Here’s some examples.
Source,Change,Target
Táan a bin koonol tu k’íiwikil koonol,TYPE:NEG,Ma’ táan a bin koonol tu k’íiwikil koonoli’
Táan u bin koonol tu k’íiwikil koonol,TYPE:NEG,Leti’e’ ma’ táan u bin koonol tu k’íiwikil koonoli’
Jach k’a’abéet in bin tu k’íiwikil koonol,TYPE:NEG,Ma’ jach k’a’abéet in bin tu k’íiwikil koonoli’
Táan a bine’ex ich kool,TYPE:NEG,Ma’ táan a bine’ex ich kooli’
Teche’ ka bin xíimbal tu yotoch,TYPE:NEG,Teche’ ma’ ta bin xíimbal tu yotochi’
...
Now fill in the third column:
Te’exe’ táan a bine’ex koonol tu k’íiwikil koonol,TYPE:NEG,Te’exe’ ma’ táan a bine’ex koonol tu k’íiwikil
koonoli’
```

Figure 1: Sample prompt (examples abbreviated). The real output of GPT-4 is shown in **bold**. The same prompt format is used for all systems and languages.

2 Method

We treat the task as a simple few-shot prompting problem, using no external data. We consider three models: two closed-source (`gpt-4-0125` and `gpt-3.5-turbo-0125` from OpenAI) and one open-source (Command R+ from Cohere). Our prompt is minimal, de-emphasising problem specific factors. It simply presents relevant examples from the training data in a CSV format, then asks the model to complete the third column of a test item CSV row. We use no additional data besides the provided training set, and perform no fine-tuning. A sample prompt is shown in Figure 1. In contrast to prior work showing the utility of expert prompting (Xu et al., 2023), describing the task domain (Zhang et al., 2024), and tipping (Salinas and Morstatter, 2024), preliminary evidence showed limited effects of any of these techniques when augmenting our prompt format. Indeed, treating the problem as a simple CSV completion task seems to have triggered interesting behavior in all 3 models: almost without exception, the first line contained either just the predicted target, or all three completed columns separated by commas. Indeed, even on our worst-performing model and language pair, Bribri using the Cohere model, only 4/480 examples in the test set are miss-parsed by these heuristics (i.e., yielding something other than the model’s prediction), in contrast to prior works where large language models typically place their completions unpredictably in unstructured text, causing parsing errors.

While the format of the prompt is simple, some heuristics are required to best make use of the provided training data and compute costs. Typically, only examples of the requested change are shown. In cases where more than 10 examples of a particular change occur in the training data, the training data exhibiting this change is sorted according to the sum of BLEU (Papineni et al., 2002) and chrF (Popović, 2015) with the source of the test item as a reference, and the 10 examples with the highest score used in the prompt. Further, when the specific change occurs fewer than 3 times in the training data (as is often the case in Bribri), we back off to similar changes: we break the queried change into the component feature changes, and take up to 3 instances of each component change. If a feature change does not occur on its own in the training data, we add one example containing the feature change which maxmizes the sum of BLEU and chrF between its source and the target source. Finally, we add up to 8 examples which contain some of the component changes, again chosen by their source BLEU+chrF similarity.

We use temperature 0.1 for the OpenAI models and temperature 0.3 for Cohere models. Preliminary evidence suggested that lower temperatures aided consistency.

3 Results

Our results on the test set are shown in full in Table 1. All of the models improve over the provided baseline for at least one of the languages. Command R+ struggles the most with the task, scoring below the baseline for Bribri and Guaraní (though improving substantially in terms of BLEU and chrF for the former). All systems improve dramatically over the baseline for Maya, which had the most provided training examples (595), with few requiring backoff. The systems performed competitively in the shared task, with

(e.g. the addition of a particle to express negation)

	Bribri			Maya			Guaraní		
System	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
`command-r-plus`	*7.08*	31.68	62.45	49.03	73.09	88.54	*9.34*	*22.64*	*73.40*
`gpt-3.5-turbo-0125`	11.67	33.80	65.51	50.97	75.09	89.76	18.13	31.94	79.36
`gpt-4-0125`	**17.71**	**39.48**	**69.28**	**53.87**	**78.54**	**91.66**	**25.00**	**40.55**	**81.71**
Baseline (edit trees)	8.75	22.11	52.73	25.81	53.69	80.23	14.84	25.03	76.10

Table 1: Performance of our three submissions for each language compared to the provided baseline. Our simple GPT-4-based system is our best across all metrics (shown in **bold**), placing third overall in the shared task. Scores below the baseline are shown in *italics*.

Command R+ placing 8th, `gpt-3.5-turbo-0125` placing 6th, and `gpt-4-0125`, placing third overall and coming in first for Maya (tying the second place system in accuracy but out-performing in terms of the secondary metrics). Overall, these results indicate that even very simple approaches using large language models can be useful for low-resource morpho-syntactic tasks, when training data is limited. However, choice of model remains important–despite the fact that Command R+ both out-performs `gpt-3.5-turbo` on MMLU and ranks higher on the LMSys Chatbot Arena[2], it substantially under-performs on both Bribri and Guaraní.

4 In-context generalization or prior exposure?

While our results suggest that large language models *can* solve complex clause-level reinflection tasks for some indigenous languages, it is unclear *what drives* this behavior. One hypothesis is that it is driven largely by prior exposure to these languages. The Glot500 dataset, which attempts to collate large amounts of data for low-resource languages for language model pretraining, contains 610,052 Maya sentences; 87,568 Guaraní sentences, and none for Bribri (ImaniGooghari et al., 2023). Attempts to develop a large corpus for Bribri have so far maxed out at just around 100,000 tokens, even with manual gathering of data from books not on the internet (Coto-Solano, 2022). This lines up relatively neatly with our results, with Maya > Guaraní > Bribri.

Another possibility is that the model is primarily generalizing the patterns of in-context examples it is provided. Support for this account is provided from the observation that the same pattern of Maya > Guaraní > Bribri is evident in the baseline, which has no prior language exposure; suggesting that the inherent difficulty of the task may vary between the languages/their datasets. As such, our primary results alone are ambiguous between these two hypotheses.

To differentiate these hypotheses, we develop a simple test involving *permuting* the alphabet for each language, such that most characters are mapped to other characters. This should provide a problem of an equivalent difficulty to the original, but which has a very different distribution over tokens, which should limit the degree to which the model uses information from prior exposure to the languages. To ensure the difficulty characteristics of the problem are preserved, letters with diacritics are permuted analogously to their counterparts without diacritics. This is due to the observation that a positive quality of our systems is their tendency to generalize patterns that apply to one set of diacritics on a letter to different diacritics on that letter. As an example, here is a real Maya sentence from the dataset followed by its permuted counterpart:

(1) Teche’ ka bin xíimbal tu najil *Original*

(2) Kitsi’ pe dun cúumder ko neyur *Permuted*

As the structure of the shared task prevents us from evaluating on a permuted test set, we present results on the development set for this experiment, shown in Table 2. We note that the results should be interpreted in light of the fact that there is substantial variability (on the order of ≈ 5 percentage points) from run-to-run. Ideally, we would run this experiment repeatedly to compute confidence intervals, but resource constraints prevent this.

Overall, our results suggest that our model performance is mostly a result of generaliz-

[2] https://leaderboard.lmsys.org

		Bribri			Maya			Guaraní		
System	**Data**	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
`command-r-plus`	orig.	**10.85**	**40.61**	**55.93**	**44.96**	**72.96**	**88.55**	**22.78**	35.42	**76.09**
	perm.	5.66	34.47	55.22	43.62	72.77	88.14	21.52	**41.95**	75.22
`gpt-3.5-turbo-0125`	orig.	4.72	37.23	57.52	**51.68**	76.20	**90.16**	**34.18**	44.95	**82.07**
	perm.	**9.91**	**37.72**	**58.22**	**51.68**	**76.62**	89.97	32.91	**49.17**	81.83
`gpt-4-0125`	orig.	15.57	41.35	62.88	53.02	75.32	**91.05**	**39.24**	**52.68**	83.06
	perm.	**18.87**	**42.15**	**63.94**	**55.03**	**76.98**	90.86	31.65	46.90	**83.10**

Table 2: We isolate the role of in-context generalization for our models using a permuted version of the development set. For each model, we show in **bold** whether performance is better on the permuted variant of the development set (lower), or the original development set (upper). Generally, systems perform similarly on both datasets, suggesting performance is primarily derived from the in-context examples.

ing in-context information, rather than applying language-level knowledge. For Maya, all 3 models retain their level of performance on the permuted test set. For Bribri, we see a moderate decrease in performance for Command R+, but an *increase* in performance for GPT-3.5 and GPT-4. This suggests an effect on the (e.g. distributional) properties of subword tokens on in-context generalization behaviours. On the other hand, Guaraní performance clearly degrades for GPT-4, suggesting either a subword issue as in the case of Bribri, or some amount of prior knowledge of the language from pre-training or instruction tuning being recruited. Taken together, though, these results suggest that in-context learning in these models is able to generalize a small set of examples in a linguistically plausible way, even in the absence of prior exposure to the language of the stimuli.

5 Conclusion

We present a simple few-shot learning setup for the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages, applied to three state-of-the-art large language models. We find even simple few-shot prompting techniques are able to beat the baseline, with our best system (GPT-4) placing third in the shared task. We investigate the extent to which the performance of our approach is due to a model's prior exposure to the language, by using a character-permuted version of the development set to maintain the problem structure while ablating the language information. We find from this preliminary evidence that the performance of these systems is driven more by in-context learning capabilities than prior exposure to these low-resource indigenous languages. We also find preliminary evidence of performance sensitivity to subwords, as we find that sometimes the model performs *better* on the permuted language than the original language.

One question not addressed here is the cause of the relative performance of the models on each of the three languages. The differences in performance mirror the performance of the baseline, suggesting that in some sense perhaps e.g. the Maya data is simpler or the training data is more informative than for some of the other languages. However, future work could characterize this further, investigating what kind of data sparsity these systems can generalize over and what kinds of functions they are better or worse at generalizing. For example, anecdotally for Bribri we found the systems struggled to generalize morphophonological stem changes (e.g., sú + ök should be sawök, but the model produces súök).

Acknowledgments

This work was in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology Language Sciences. The author would like to thank Paul J.W. Schauenburg, Verna Dankers, Sharon Goldwater, and Edoardo Ponti for their comments on this work. Any errors are my own.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. 2022. The MRL 2022 shared task on multilingual clause-level morphology. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning (MRL)*, pages 134–146, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*. CSLI Publications, Stanford, CA, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, 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.

Geoffrey K. Pullum and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504.

Brian Roark and Richard Sproat. 2001. The Formal Characterization of Morphological Operations. In *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *Preprint*, arXiv:2401.03729.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is ChatGPT a good sentiment analyzer? A preliminary study. *Preprint*, arXiv:2304.04339.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *Preprint*, arXiv:2305.14688.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *Preprint*, arXiv:2402.18025.