Morphological Segmentation of Low-Resource Languages with Large Language Models

Anonymous ACL submission

Abstract

In this work, we show how large language models (LLMs) perform on morphological segmentation and annotation tasks, relative to baseline transformers when trained with or without full sentence context. We find that with full sentence context LLMs exceed the performance of fine-tuned transformer models for some low-resource languages, while requiring far less training data. We also observe that the digital resource status of a language is not a clear indicator of segmentation and annotation performance. For example, Claude-Haiku can achieve an F1 score of 0.635 for morphological segmentation on English (exceeding a baseline F1 score of 0.442), while for Natugu (a very low-resource language) Claude-Haiku can achieve an F1 score of 0.624 (exceeding a baseline F1 score of 0.413). These results indicate that LLMs can lower the barrier of entry for linguistic researchers or speakers of a language who wish to develop literacy materials or language technology from annotated texts, without requiring computationally expensive training or custom user interfaces.

1 Introduction

Morphological analysis benefits both downstream NLP (Park et al., 2021) and the scientific study of under-documented languages (Moeller, 2021). For linguists, studying text with annotated morphologies is an important part of understanding the workings of a target language. However, manually annotating morphology is a slow and labor-intensive process that requires linguistic training and expertise in the target language. Previous works demonstrated fine-tuned transformer models could facilitate the morpheme segmentation and annotation tasks, both for widely-spoken languages and low-resource languages (Wu et al., 2021).

Morphology comprises word-building properties in human languages as well as their accompanying (morpho-)syntactic phenomena. Several tasks comprise the morphological analysis process, two of which are segmentation and lemmatization. Segmentation identifies word internal boundaries between minimal meaningful units (morphemes or (surface) morphs), for example `word` → `word + -s`. Lemmatization identifies or replaces the word with its lemmas (citation form), for example `was` → `be`.

This work compares the performance of LLM in-context learning with transformer models at four tasks: morphological segmentation, lemmatization, and two additional approaches to morphological annotation at different levels of detail. We assess model accuracy across three major languages and six low-resource language datasets, for which LLMs have little or no exposure. We also quantify the performance gains from in-context learning on LLMs by varying both the number and selection of few-shot samples in an LLM prompt.

2 Background

ALthough LLMs presumably have very little expesure to low-resource languages, they have shown some success with them (Guo et al., 2024; Su et al., 2024; Tanzer et al., 2024). This success is dependent on prompt strategies, and this study expands on work to understand the effects prompting strategies for in-context learning with low-resource languages (Cahyawijaya et al., 2024; Nguyen et al., 2024).

Although Pranjić et al. (2024) leveraged default LLM subword tokenization to train a segmenter that shows improvement on several low-resource languages, no work has tested LLM performance on morphological segmentation in under-documented languages. However, it is known that LLMs can generalize "knowledge" of morphology in major and low-resource languages (Dang et al., 2024) although they underperform task-specific

systems (Weissweiler et al., 2023).

Previous NLP models specific to morphological segmentation of low-resource languages (Nicolai et al., 2024; Moeng et al., 2022; Wu et al., 2021; Liu et al., 2021; Kann et al., 2016) are usually trained on individual words, ignoring any clues in the context, and is often combined with morpheme glossing (Barriga Martínez et al., 2021). Our work performs segmentation without glossing and includes the full sentence context, comparing the results to word-level context.

3 Languages and Data

We performed morphological segmentation on a diverse set of languages (Table 1). The data for the three major languages English [eng], Czech [cse], and Mongolian [mon] is taken from the 2022 SIGMORPHON Shared Task 2 (Batsuren et al., 2022). This data combined derivational and inflectional morphology but did not indicate the types of morphemes. Therefore, it only allows us to perform simple segmentation (Task 1). Table 1 lists additional information on the specific languages, language families, and total number of annotated sentences. The data for the other languages is from mostly unpublished collections compiled for linguistic studies. These include:

Bonggi [bdg] is spoken by about 1,500 people in Malaysia (Eberhard et al., 2024) and has non-concatenative morphology. Some of the data is published (Boutin, 1994).

Alas-Kluet [btz] is spoken by 200,000 people on Sumatra, Indonesia (Eberhard et al., 2024). Its morphology features reduplication, infixation, and circumfixation.

Choctaw [cho] is spoken by about 10,000 people in Mississippi and Oklahoma (USA) (Broadwell, 2006). Its morphology is polysynthetic but the orthography in the data separates complex words to be easier to read.

Lezgi [lez] has about half a million speakers in Russia and Azerbaijan. Its morphology is highly agglutinative (Haspelmath, 1993).

Natügu [ntu] is spoken by about 4,000 people in the Temotu Province of the Solomon Islands. It has primarily agglutinative morphology (Åshild Næss and Boerger, 2008). Some of the data is published.[1]

Upper Tanana [tau] is an official language of Alaska but is critically endangered, with barely 50 adult speakers in Alaska (USA) and the Yukon Territory (Canada) (Lovick, 2020). Its complex morphology features non-continuous lexical, derivational, and inflectional prefixes on verbs. Some of the original data is available at the Alaska Native Language Archive.[2]

This collection of low-resource languages covers diverse geographic, genealogical, syntactic, and morphological categories. It is subject to little or no LLM dataset "contamination" that could otherwise taint performance benchmarks. We can say this with confidence for three reasons: the data is mostly unpublished and could not be harvested for LLM training, the amount of data in the languages is quite limited (see Table 1), and morphological segmentation is not a common task in NLP, making relevant data relatively scarce even for major languages.

4 Tasks

To fully evaluate the segmentation capabilities and to compare across different types of languages, we lay out four separate morphological annotation tasks. For all four tasks, we select from each language dataset 100 random test sentences. The rest of the dataset is used for few-shot prompting or training. These four tasks are as follows:

1. Given an input, separate words by their morphemes. The symbol "@@" denotes a boundary between morphemes. For example: `John's dogs are jumping.` → `John@@'s dog@@s are jump@@ing.`

2. Given an input, lemmatize each word. This task requires accurate morphological segmentation and recognition of morpheme types: root morphemes vs affixes. For example: `John's dogs are jumping.` → `John dog be jump`

3. Given an input, separate words into composite morphemes and further distinguish between roots and affixes (denoted "@=" for suffixes/enclitics and "=@" prefixes/proclitics). For example: `John's dogs are jumping.` → `John@='s dog@=s are jump@=ing.`

4. For a given sentence, separate words by their morphemes and further distinguish between

[1] https://www.langlxmelanesia.com/tilp

[2] https://www.uaf.edu/anla

Language	Language Family	ISO 639 Code	Total Sentences	Digital Support
English	Indo-European	[eng]	1783	Thriving
Czech	Indo-European	[ces]	500	Vital
Mongolian	Mongolic	[mon]	500	Vital
Alas-Kluet	Austronesian	[btz]	302	Emerging
Bonggi	Austronesian	[bdg]	3955	Still
Choctaw	Muskogean	[cho]	1441	Ascending
Lezgi	Nakh-Daghestanian	[lez]	1531	Ascending
Natügu	Austronesian	[ntu]	1606	Ascending
Upper Tanana	Dene-Yeniseian	[tau]	2398	Emerging

Table 1: Languages used across tasks. Data for [eng], [mon], and [ces] are sourced from the 2016 SIGMORPHON shared task. Digital support taken from Eberhard et al. (2024).

suffixes (denoted "@="), prefixes (denoted "=@"), enclitics (denoted "@=="), and proclitics (denoted "==@"). This is the same as the previous task, except for the added distinction between suffixes vs. enclitics, and prefixes vs. proclitics. For example: `John's dogs are jumping.` → `John==@'s dog=@s are jump=@ing.`

We assess the performance of LLMs versus current state-of-the-art transformer models on these tasks using two levels of input context. One level provides the full sentence as an input to the model. The second level inputs one word at a time without its context of use. We evaluate the accuracy of a model-inferred output using an F1 metric assessed across each morpheme in a test input and across each word in the sentence-level context.

Since the SIGMORPHON benchmark only separates morphemes between words without indicating the type of separation, we only consider English [eng], Czech [ces], and Mongolian [mon] for the first task. The remaining languages, which are all classified as low-resource, are considered for the first three tasks based on the availability of manually annotated lemmatized words and clear distinctions on affixes and clitics. We only consider Bonggi [bdg], Alas-Kluet [btz], and Upper Tanana [tau] for Task 4 since the Choctaw [cho], Lezgi [lez], and Natugu [ntu] datasets have very few to no instances of clitic morphemes. Tasks 1, 3, and 4 are common tasks in linguistics but not in NLP, whereas lemmatization (Task 2) is common in NLP but not in linguistics (although it is subsumed under the other tasks).

5 In-Context Learning

In addition to our four tasks, which are primarily motivated by use cases in linguistics, we also perform some of the tasks with three tests that aim to assess the LLM model capabilities under varying circumstances. We opt for assessing these tests on simple morphological segmentation (Task 1) with `claude`, as it has been a popular and effective choice among LLMs for low-resource languages (Kholodna et al., 2024; Jayakody and Dias, 2024) and held the overall best performance on our tasks as well.

Amount of Context in Input. First, we instruct each model to run simple morphological segmentation (Task 1) across each word individually and then on whole sentences. We compare the results to assess how much context informs overall LLM performance. We compared each context level to the `baseline` performance.

Optimal In-Context Prompt Size. Second, we take advantage of the 200K-token length context window on `claude` to measure the effect of greater or fewer few-shot example sentences on morphological segmentation (Task 1). We assess six "shots", starting with no examples (zero-shot), 10 examples, 20, 100, and up to 500 segmented sentences in the few-shot prompt.

Effect of Strategic Data Selection on LLM. In general, running LLM inference with in-context learning is far easier than fine-tuning a custom-trained baseline transformer, and requires a much smaller initial set of annotated examples. Even if the `baseline` can outperform the LLM, the lower demand on data may allow linguists to bring a more diverse, and presumably more informative, selec-

tion of available data in the target language. To explore the importance of data diversity with the advantage of few-shot examples, we select 20 sentences in the training set that are the most similar to the sentences in the test set. We measure similarity by a term frequency-inverse document frequency (TF-IDF) metric (Ramos et al., 2003).

6 Models and Methods

In this work, we consider several options for LLMs (open-source and commercial) and benchmark these against a state-of-the-art transformer implementation. The setup, training cost, and manual annotation required to effectively use this benchmark implementation was significantly greater than performing in-context learning on an LLM.

We contrast LLM performance against a fairseq implementation (Ott et al., 2019) of a transformer model (`baseline`), using settings (Wu et al., 2021) which previously demonstrated state-of-the-art performance (Pimentel et al., 2021) with computational morphology. For each language, the `baseline` model is trained on all annotated sentences except for 100 sentences kept in the test set. Additionally, the training dataset for the `baseline` is split 9/1 into train and development data. The performance of the `baseline` model is then assessed based on its accuracy when annotating the test set. On average, training one baseline model on a single task for one low-resource language takes roughly 5 minutes on an NVIDIA A100 GPU.

We evaluate four LLMs on each task and context: Llama2 with 7 billion tunable parameters (`llama2`) (Touvron et al., 2023), Llama3 with 7 billion tunable parameters (`llama3`) (AI@Meta, 2024), Claude-Haiku from Anthropic (`claude`), and GPT-4o-mini from OpenAI (`openai`). The `llama2` and `llama3` models were queried using Ollama (Ollama, 2024), while `claude` and `openai` were queried using API calls.

We found that a suitable combination of LLM prompt size and performance uses a few-shot prompt containing 20 randomly selected examples from the training set, and use 20 examples unless otherwise specified. The prompt presents each sentence both before and after morpheme annotation or lemmatization. The prompt is constructed as follows, with minor adjustments between tasks: "`You are a morpheme annotator. You separate morphemes by '@@', and words by ' '. Respond with the output, excluding the word "Output". For example: {fewshot_examples}.`" Figure 1 shows results on an example sentence from the test set of [btz] with `claude` for morphological segmentation, lemmatization, and annotation for all tasks.

After constructing a few-shot prompt on 20 randomly selected examples, we iteratively pass each sentence from the test set to be annotated by the LLM. We repeat this experiment across ten trials with different randomly selected few-shot examples, so that we can determine the degree of performance variability caused by the composition of the randomly selected few-shot examples. Assuming statistical independence between trials, we assess whether F1 scores are similarly distributed across the ten trials using a Kolmogorov-Smirnov (K-S) test (Massey Jr, 1951). For each trial $i, j \in \{1, ..., 10\}$ $(i < j)$, we compute a P-value $p_{i,j}$ using K-S. We combine these p-value statistics using Fisher's method, written as $X^2_{2k} = -2\sum_{i,j} ln(p_{i,j})$ (Fisher, 1970). Larger values of X^2_{2k} indicate greater variability in performance based on the randomly selected few-shot examples.

7 Results

Across all four morphological tasks, LLMs exceed the `baseline` model in some instances, while falling short on others as seen in Figures 2 thru 4. For some low-resource languages such as [bdg] and [cho], the `baseline` model performed significantly better than the LLMs across all tasks. Among the LLMs we assessed, `openai` and `claude` consistently outperformed `llama2` and `llama3` (with perhaps a slight overall advantage to `claude`). Claude tends to outperform the `baseline` on full sentence context (Figure 5), but the `baseline` is better when asked to segment individual words. The performative advantage of `claude` comes with the trade-offs of increased cost, decreased transparency, and a required network connection compared to open-source LLMs such as `llama2` and `llama3`.

We observe no clear correlation between the resource status of a language and the LLM performance across the four tasks. For more information on tabular scoring across all tasks, see Appendix A. For simple morphological segmentation (Task 1), we find that `claude` and `openai` perform best for [eng]. This is unsurprising, given the presumed abundance of English in the training sets for both models and their stated benchmark performance at

Source`	Sentence
Raw	sang khaje te bekhune engo balik me kute ne ni kekhan khut wajah mesekhi~sekhi te siap menjabut saat~saat si nitimai soh.
Task 1 (Gold)	sang khaje te bhekhu@@ne enggou bhalik bhe kute@@ne ni kekhan khut wajah me@@sekhi te siap n@@jabut saat@@saat si ni@@timai soh.
Task 1 (Model)	sang khaje te bhekhu@@ne enggou bhalik bhe kute@@ne ni kekhan khut wajah me@@sekhi te siap n@@jabut saat~sa@@at si ni@@timai soh.
Task 2 (Gold)	sang khaje te bhekhu enggou bhalik bhe kute ni kekhan khut wajah sekhi te siap jabut saat si timai soh
Task 2 (Model)	sang khaje te bhekhu enggou bhalik bhe kute ni kekhan khut wajah sekhi te siap jabut saat si nitime soh
Task 3 (Gold)	sang khaje te bhekhu@=ne enggou bhalik bhe kute@=ne ni kekhan khut wajah me=@sekhi te siap n=@jabut saat=@saat si ni=@timai soh.
Task 3 (Model)	sang khaje te bekhun@=e enggou bhalik bhe kute@=ne ni kekhan khut wajah me=@sekhi te siap men=@jabut saat=@saat si nitimai soh.
Task 4 (Gold)	sang khaje te bhekhu@==ne enggou bhalik bhe kute@==ne ni kekhan khut wajah me=@sekhi te siap n=@jabut saat=@saat si ni=@timai soh.
Task 4 (Model)	sang khaje te=@bekhun enggou bhalik bhe kute@==ne ni=@kekhan khut wajah =@sekhi te silap n=@jabut =@saat saat si ni=@timai soh.

Figure 1: Example [btz] sentence across all four tasks using `claude`, with F1 scores. Strings in red were incorrectly segmented by the model; green strings are correct.

multi-task language understanding (MMLU) tasks (Hendrycks et al., 2020). What is more surprising is the high F1 scores on low-resource languages, in particular [lez], [btz], and [ntu] for which there is likely little or no content in the training sets (for reference, none of these languages are present in either the FLORES-200 (Costa-jussà et al., 2022) or the MADLAD-400 databases (Kudugunta et al., 2024)). We also find that morphological segmentation is especially challenging for the two other major languages [ces] and [mon], although why both `baseline` and `llama2` models struggle on these languages in particular is not clear. One possibility is that, since these SIGMORPHON datasets contain both predictable and more unpredictable derivational morphological segmentation, performing accurate segmentation is more challenging and improved performance would require a larger training set. However, other datasets such as [ntu] and [bdg] also contain derivational segmentation. Another explanation might be that both languages use the Cyrillic alphabet; however, [lez] uses the same alphabet.

Lemmatization (Task 2) requires segmenting morphemes (e.g., "running" becomes "runn@@ing"), identifying which morpheme is the root ("runn"), and, when necessary, transforming to the base form of that root ("run"). The increased linguistic information necessary for this task is reflected in the decreased F1 scores between Task 1 and Task 2 on some languages, with F1 scores for Task 2 presented in Figure 3. This decrease in performance is best exhibited by [cho], where LLM performance drops precipitously from Task 1 to Task 2. This may be caused by the orthographic representation of Choctaw, which leaves some tokens in texts without any root at all. It may also be due to the language's polysynthetic morphology type, where root morphemes can completely change form as multiple morphemes are combined. However, Upper Tanana [tau] is also polysynthetic and does not see such a large difference between tasks. Decreased model

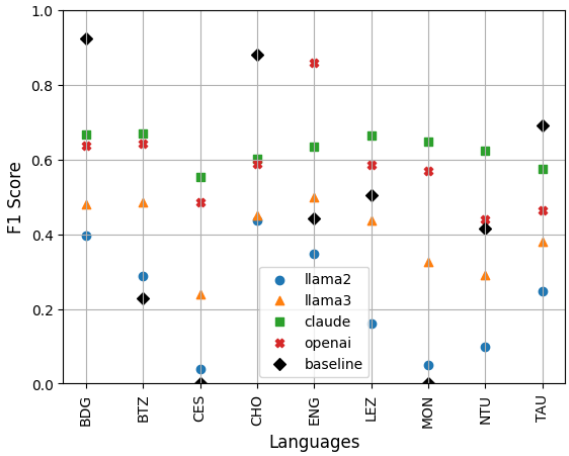

Figure 2: F1 scores on morphological segmentation of sentences (Task 1) across models and languages. The `baseline` model was given the entire training set, while LLMs were given 20 random few-shot samples.

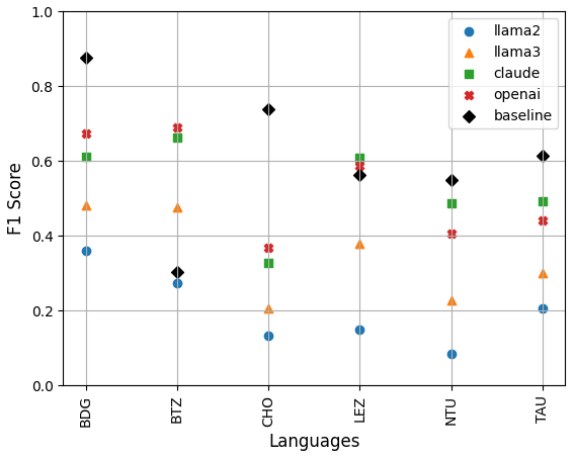

Figure 3: F1 scores on lemmatization of sentences (Task 2) across models and languages.

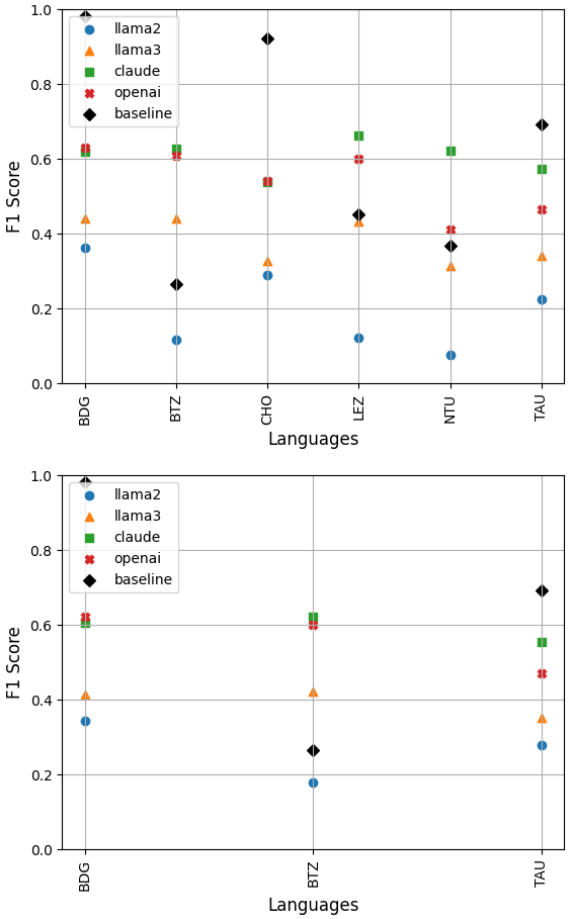

Figure 4: Top: F1 scores on morphological segmentation of sentences with an indication of suffix/enclitic or prefix/procilitic (Task 3) across models and languages. Bottom: F1 scores on morphological segmentation of sentences with an indication of suffix, enclitic, prefix, procilitic (Task 4) across models and languages.

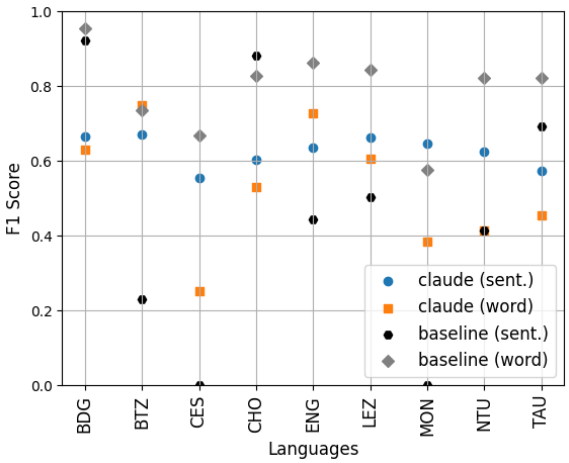

Figure 5: F1 scores of sentence-level morphological segmentation (Task 1) versus word-level morphological segmentation across languages, comparing the `baseline` model with `claude`.

performance across other languages from Task 1 to Task 2 is less apparent, suggesting that the greatest challenge to lemmatization is in fact the initial segmentation step.

The F1 scores for Tasks 3 and 4 are shown in Figure 4. Distinguishing between suffixes/enclitics and prefixes/proclitics (Task 3) is similar to the segmentation task, but requires correctly identifying the root (as with Task 2) as well as all morpheme boundaries. Any morphemes after the root (to its right in a left-to-right orthography) are suffixes or enclitics, and any morphemes before the root are prefixes or proclitics. We see very little decrease in F1 scores between the of Task 1 and Task 3, suggesting that the adding boundary markers identifying the different types of morphemes to the simple segmentation does not add to the complexity of the task. Whereas in Task 2, accurately isolating the root of a word is challenging. This is most clearly conveyed by the LLM performance on [cho], which is similar between Task 1 and Task 3 but is far worse for Task 2. Task 4 is very similar to Task 3 but makes one final distinction of morpheme types. Each model must distinguish between prefixes and proclitics as well as between suffixes and enclitics. Clitics are far less common than prefixes or suffixes and do not occur many languages. They appear in [bdg], [btz], and [tau] (one clitic appears 3 times in [ntu]). We only considered these languages for Task 4 since clitics are rare or not present in other language datasets. Task 4 shows only a minor decrease in performance across models relative to Task 3. All models tend to err towards misidentifying clitics as prefixes or suffixes.

In addition to segmentation on full sentences, we also consider segmentation on individual words in Task 1 using both the `baseline` model and `claude`. As shown in Table 5, we find that, in nearly every case, the `baseline` model performs better when segmenting individual words in a sentence, rather than an entire sentence at once. We find the opposite is true for `claude`: in nearly every case, sentence-level segmentation outperforms word-level segmentation.

Increasing the number of training or example sentences increases computational inference time and cost, and requires obtaining a larger manually annotated dataset. Whereas the `baseline` requires all available training data to achieve its performance, for the LLMs a moderate number of few-shot sentences (20 sentences, for our tasks) is sufficient to achieve relatively high performance.

This means LLMs provide an opportunity to find an efficient balance between dataset size and performance. Additional few-shot examples beyond 20 notes diminishing returns. This trend is shown in Figure 6. Alternatively, if a larger manually annotated dataset is available, actively selecting the annotated sentences most similar to the target sentences (according to TF-IDF) yields a significant increase in F1 scores. Figure 6 shows this performance increase both for an increasing number of random few-shot samples as well as for 20 few-shot sentences selected using TF-IDF.

Since `claude` is the least sensitive to few-shot sample selection, we predict that other LLMs might experience a greater increase in performance with strategic selection of annotated examples. Considering Fisher statistics (X^2_{2k}) across all tasks, it appears `llama2` is far more sensitive how the 20 few-shot examples are selected than the other models. We find that the sensitivity to the few-shot prompt decreases as the size of the LLM increases. Furthermore, `claude` exhibits the least sensitivity across all tasks. Figure 7 shows Fischer test statistics for Task 1. In addition to the variability between 20 few-shot sample selections, we also find that increasing the number of few-shot samples yields better overall performance with `claude` for [lez], [ntu], and [tau].

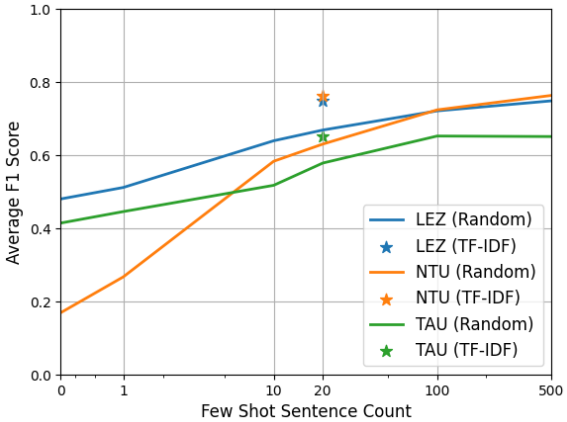

Figure 6: F1 scores of sentence-level morphological segmentation (Task 1) with a varying number of randomly selected few-shot training samples, along with few-shot training samples selected using TF-IDF. These tests were performed on [lez], [ntu], and [tau] with `claude`

8 Conclusion and Future Work

We run four morphological segmentation tasks on three major languages and six low-resource languages and compare the results across four LLMs

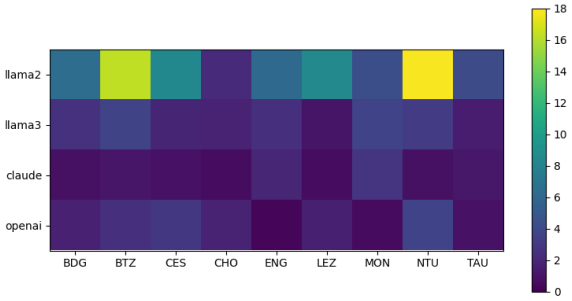

Figure 7: Fisher X^2_{2k} statistics for few-shot trials (20 annotated sentences) on Task 1 across models and languages. Higher values indicate a greater degree of sensitivity to few-shot training samples and overall segmentation performance.

and baseline transformer model. For low-resource languages such as Nutugu, Alas-Kluet, and Lezgi, we find that commercial models from both Anthropic and OpenAI far exceed a baseline transformer model. We conclude that although LLMs cannot reliably outperform transformers, they offer a competitive opportunity for performing linguistic annotation tasks on low-resource languages because they achieve optimal performance with very few examples. The optimal LLM performance is achieved with 20 examples.

Additionally, we show that selecting sentences similar to the test set for few-shot inference can improve annotation performance. This constitute a far smaller fraction of the training set supplied to the `baseline` model. This gives linguists the luxury to select from their manually annotated data those sentences that most similar to those in their unannotated data. LLMs easily outperformed the baseline when asked to segment full sentences. We hypothesize that this difference arises from LLMs making better use of contextual clues within a sentence than the `baseline` model or because its large training data allows it to generalize knowledge of long sequences across languages. Overall, our work indicates that LLMs, even though if they cannot yet achieve state-of-the-art performance, significantly lower the barrier of entry for linguistic research on low-resource languages, and are especially useful when there is an absence of manually annotated data.

In future work, we hope to explore opportunities for applying tools that combine in-context learning with active learning to accelerate the manual annotation for under-documented languages. Furthermore, given how our morpheme annotation tasks

can build off one another with increasing specificity and linguistic complexity, approaches such as chain-of-thought reasoning (Wei et al., 2022) may also increase annotation performance. Finally, given recent investments in lightweight hardware capable of running quick LLM inference, we envision that such strategies might be leveraged with open-source models that can be used locally. This would allow linguists or community speakers to use LLM-based annotation tools without requiring network connectivity or expensive hardware.

Limitations

Our work highlights how LLMs can perform morphological annotation and lemmatization tasks on low-resource language data. We are fairly confident that the LLMs we assess have no exposure, or extremely limited exposure, to the low-resource languages we are testing with. However the degree of possible data contamination is unclear, particularly if related languages in the same families are considered. We do not assess whether similarities in morphological annotation on languages within the same family might be responsible for increased performance in some cases. The degree to which LLMs are relying on in-context information, versus out-of-context training, is not readily measurable from our experiments. It is also unclear whether the results we supply for the six low-resource languages could be generalized to other low-resource languages with different morphological, semantic, or syntactic structure. Although we tried to understand the nature of the field data annotations, we cannot describe with certainty the completeness, consistency, or type of morphological segmentation (e.g. canonical or surface, inflectional only or mixed with derivational).

Ethics Considerations

This work is part of a growing body of literature on the under-representation of non-English languages within LLMs, and the biases (statistical, linguistic, and cultural) that these models perpetuate. Although we have express permission to share most of the low-resource language data used in this paper, we choose to do so only upon request because the minority communities represented by the data are more vulnerable to such biases and to commericial exploitation of their limited language resources.

Code and Data Availability

The code and tests associated with morphological segmentation and in-context learning tests are made available for the SIGMORPHON benchmark languages ([eng], [ces], and [mon]). The code for other morphological annotation tasks is also made available. These are all provided at [GH LINK]. Annotated datasets for the six low-resource languages may be available upon request.

Conflict of Interest

The authors have no conflicts to declare.

References

AI@Meta. 2024. Llama 3 model card.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic Interlinear Glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinović, Kyle Gorman, Zdeněk Žabokrtskỳ, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, et al. 2022. The sigmorphon 2022 shared task on morpheme segmentation. *arXiv preprint arXiv:2206.07615*.

Michael E. Boutin. 1994. *Aspect in Bonggi*. Ph.D. thesis, University of Florida.

George Aaron Broadwell. 2006. *A Choctaw Reference Grammar - Nebraska Press*. Studies in the Anthropology of North American Indians. University of Nebraska.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Anh Dang, Limor Raviv, and Lukas Galke. 2024. Morphology Matters: Probing the Cross-linguistic Morphological Generalization Abilities of Large Language Models through a Wug Test. In *Proceedings*

of the Workshop on Cognitive Modeling and Computational Linguistics, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seven edition. SIL International, Dallas, Texas.

Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Martin Haspelmath. 1993. *A grammar of Lezgian*. Mouton de Gruyter, Berlin; New York.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Ravindu Jayakody and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. In *2024 International Conference on Asian Language Processing (IALP)*, pages 162–167. IEEE.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas. Association for Computational Linguistics.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 397–412. Springer.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological Segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Olga Lovick. 2020. *A Grammar of Upper Tanana, Volume 1: Phonology, Lexical Classes, Morphology*. U of Nebraska Press, Lincoln. Google-Books-ID: jdrKDwAAQBAJ.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Sarah Moeller. 2021. Computational Morphology for Language Documentation and Description. *Colorado Research in Linguistics*, 25.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2022. Canonical and Surface Morphological Segmentation for Nguni Languages. In *Artificial Intelligence Research*, pages 125–139. Springer International Publishing, Cham.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin. 2024. *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Mexico City, Mexico.

Ollama. 2024. Ollama: Get up and running with large language models.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold

Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level Morphological Segmentation Using Large Language Model. In *LREC-COLING 2024*, pages 10665–10674. ELRA Language Resource Associatio.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Jim Su, Justin Ho, George Broadwell, Sarah Moeller, and Bonnie Dorr. 2024. A comparison of fine-tuning and in-context learning for clause-level morphosyntactic alternation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 179–187.

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. Vienna Austria.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the Bugs in ChatGPT's Wugs: A Multilingual Investigation into the Morphological Capabilities of a Large Language Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907. Association for Computational Linguistics. Cited in dissertation.

Åshild Næss and Brenda H. Boerger. 2008. Reefs–santa Cruz as Oceanic: Evidence from the verb complex. *Oceanic Linguistics*, 47:185–212.

A Tabular Scores Across All Tasks

Below are tables containing F1 and Fisher statistics for all four tasks. Scores are given for all four LLMs as well as the fine-tuned baseline model. Table 2 corresponds to Task 1, Table 3 corresponds to Task 2, Table 4 corresponds to Task 3, and Table 5 corresponds to Task 4.

B Few-Shot Performance Similarity

While we present the Fisher statistics for Task 1 (Figure 8) in the main body of the paper, in this appendix we present the remaining maps for Fisher statistics for Task 2 (Figure 9), Task 3 (Figure 10), and Task 4 (Figure 11).

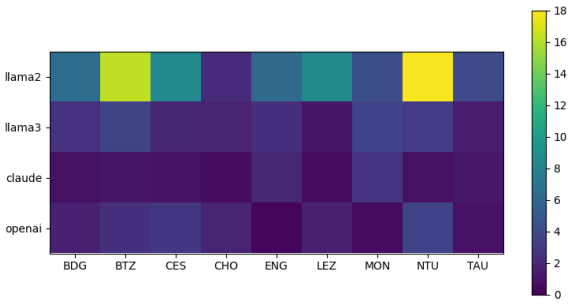

Figure 8: Fisher X^2_{2k} statistics for few-shot trials on Task 1 across models and languages. Higher values indicate a greater degree of dependence on few-shot training sample selection and overall segmentation performance.

Lang. Code	Baseline F1	`llama2` F1 (X^2_{2k})	`llama3` F1 (X^2_{2k})	`claude` F1 (X^2_{2k})	`openai` F1 (X^2_{2k})
ENG	0.442	0.347 (6.124)	0.499 (2.395)	0.635 (1.911)	**0.858 (0.219)**
CES	<0.001	0.039 (8.388)	0.238 (1.845)	**0.554 (0.908)**	0.485 (2.894)
MON	<0.001	0.051 (4.292)	0.325 (3.649)	**0.647 (2.752)**	0.569 (0.515)
BTZ	0.229	0.287 (16.260)	0.485 (3.724)	**0.669 (1.107)**	0.643 (2.393)
LEZ	0.504	0.160 (8.491)	0.438 (1.022)	**0.664 (0.590)**	0.586 (1.574)
NTU	0.414	0.099 (17.857)	0.291 (3.170)	**0.624 (0.781)**	0.439 (3.611)
TAU	**0.692**	0.247 (4.198)	0.380 (1.377)	0.574 (1.151)	0.464 (0.858)
CHO	**0.882**	0.437 (2.191)	0.449 (1.792)	0.603 (0.591)	0.590 (1.803)
BDG	**0.923**	0.397 (6.414)	0.481 (2.538)	0.666 (0.803)	0.636 (1.645)

Table 2: Morpheme segmentation (Task 1) F1 scores and Fisher statistics across all models and languages.

Lang. Code	Baseline F1	`llama2` F1 (X^2_{2k})	`llama3` F1 (X^2_{2k})	`claude` F1 (X^2_{2k})	`openai` F1 (X^2_{2k})
BTZ	0.303	0.273 (10.902)	0.475 (2.497)	0.662 (2.031)	**0.690 (0.907)**
LEZ	0.560	0.149 (2.338)	0.377 (1.734)	**0.607 (2.781)**	0.587 (1.533)
NTU	**0.548**	0.082 (7.725)	0.225 (1.729)	0.486 (3.208)	0.404 (3.496)
TAU	**0.612**	0.204 (3.474)	0.300 (2.268)	0.492 (1.017)	0.441 (0.315)
CHO	**0.738**	0.132 (3.128)	0.205 (0.974)	0.328 (1.109)	0.367 (0.697)
BDG	**0.874**	0.359 (3.114)	0.482 (3.651)	0.611 (1.924)	0.674 (1.096)

Table 3: Lemmatization (Task 2) F1 scores and Fisher statistics across all models and languages.

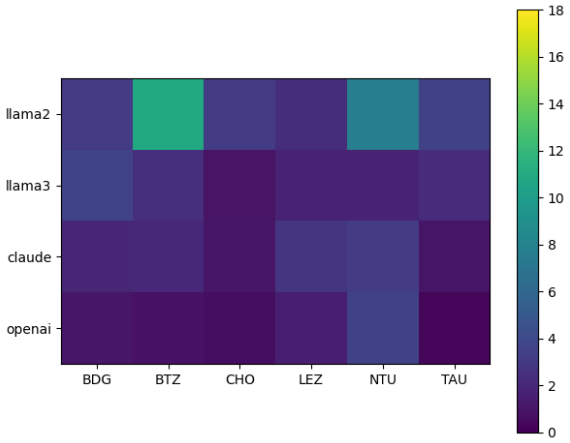

Figure 9: Fisher X^2_{2k} statistics for few-shot trials on Task 2 across models and languages. Higher values indicate a greater degree of dependence on few-shot training sample selection and overall segmentation performance.

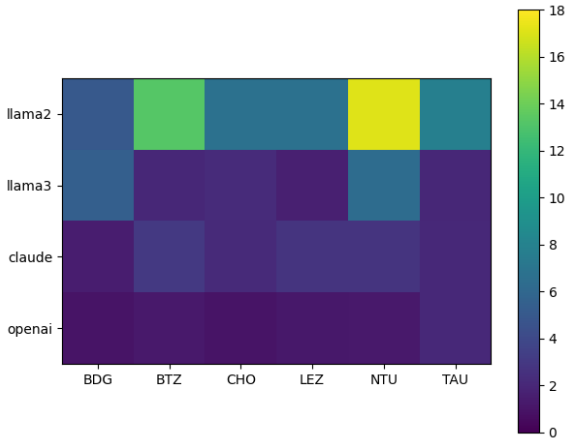

Figure 10: Fisher X^2_{2k} statistics for few-shot trials on Task 3 across models and languages. Higher values indicate a greater degree of dependence on few-shot training sample selection and overall segmentation performance.

Table 4: Morpheme segmentation and annotation by suffix/enclitic and prefix/proclitic (Task 3) F1 scores and Fisher statistics across all models and languages.

Lang. Code	Baseline F1	`llama2` F1 (X^2_{2k})	`llama3` F1 (X^2_{2k})	`claude` F1 (X^2_{2k})	`openai` F1 (X^2_{2k})
BTZ	0.264	0.116 (13.221)	0.440 (1.992)	**0.627 (2.967)**	0.608 (1.213)
LEZ	0.450	0.122 (6.732)	0.431 (1.603)	**0.663 (2.688)**	0.599 (1.127)
NTU	0.368	0.074 (17.130)	0.314 (6.329)	**0.620 (2.674)**	0.411 (1.211)
TAU	**0.691**	0.224 (7.795)	0.340 (2.033)	0.573 (2.042)	0.464 (2.043)
CHO	**0.923**	0.287 (6.693)	0.327 (2.231)	0.536 (2.170)	0.539 (0.916)
BDG	**0.982**	0.362 (5.049)	0.439 (5.483)	0.618 (1.418)	0.629 (0.917)

Table 5: Morpheme segmentation and annotation by suffix, enclitic, prefix, and proclitic (Task 4) F1 scores and Fisher statistics across all models and languages.

Lang. Code	Baseline F1	`llama2` F1 (X^2_{2k})	`llama3` F1 (X^2_{2k})	`claude` F1 (X^2_{2k})	`openai` F1 (X^2_{2k})
BTZ	0.264	0.177 (15.598)	0.422 (1.840)	**0.622 (1.238)**	0.599 (0.805)
TAU	**0.691**	0.277 (2.008)	0.349 (0.875)	0.553 (1.104)	0.471 (2.639)
BDG	**0.982**	0.341 (8.103)	0.412 (3.152)	0.604 (1.247)	0.621 (0.620)

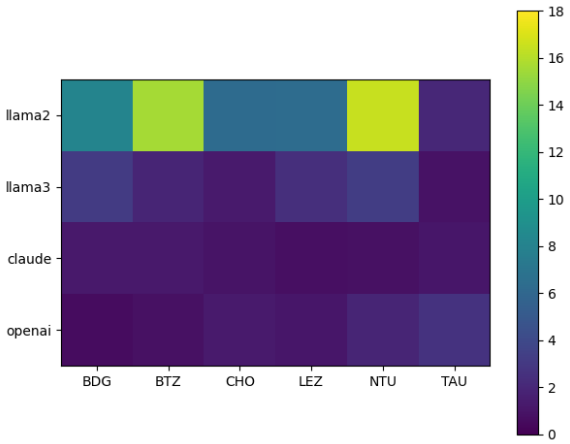

Figure 11: Fisher X^2_{2k} statistics for few-shot trials on Task 4 across models and languages. Higher values indicate a greater degree of dependence on few-shot training sample selection and overall segmentation performance.