Could We Have Had Better Multilingual LLMs If English Was Not the Central Language?

Ryandito Diandaru[♠], Lucky Susanto[♢], Zilu Tang[♣], Ayu Purwarianti[♠], Derry Wijaya[♣,♡]
Bandung Institute of Technology[♠], University of Indonesia[♢], Boston University[♣], Monash University Indonesia[♡]
13519157@std.stei.itb.ac.id, lucky.susanto@ui.ac.id, zilutang@bu.edu, ayu@itb.ac.id, derry.wijaya@monash.edu

Abstract

Large Language Models (LLMs) demonstrate strong machine translation capabilities on languages they are trained on. However, the impact of factors beyond training data size on translation performance remains a topic of debate, especially concerning languages not directly encountered during training. Our study delves into Llama2's translation capabilities. By modeling a linear relationship between linguistic feature distances and machine translation scores, we ask ourselves if there are potentially better central languages for LLMs other than English. Our experiments show that the 7B Llama2 model yields above 10 BLEU when translating into all languages it has seen, which rarely happens for languages it has not seen. Most translation improvements into unseen languages come from scaling up the model size rather than instruction tuning or increasing shot count. Furthermore, our correlation analysis reveals that syntactic similarity is not the only linguistic factor that strongly correlates with machine translation scores. Interestingly, we discovered that under specific circumstances, some languages (e.g. Swedish, Catalan), despite having significantly less training data, exhibit comparable correlation levels to English. These insights challenge the prevailing landscape of LLMs, suggesting that models centered around languages other than English could provide a more efficient foundation for multilingual applications.

Keywords: Llama2, machine translation, linguistic distances

1. Introduction

Large Language Models (LLMs) have been a popular research topic in Natural Language Processing (NLP) due to their remarkable performance on various tasks including machine translation (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b). Extensive evaluations on machine translation of the popular GPT model family (OpenAI, 2023) have suggested that they can translate high-resource languages (Robinson et al., 2023; Hendy et al., 2023). However, it is rarely the case for low-resource or underrepresented languages (Robinson et al., 2023; Hendy et al., 2023; Stap and Araabi, 2023; Kadaoui et al., 2023).

A straightforward approach for the lack of training data in low-resource translation is to collect more labeled data. However, investing in data creation is nontrivial as it comes with challenges, including the cost of such endeavors. For example, Aji et al. (2022) described the absence of Wikipedia articles on Indonesian regional languages and the challenges of labeled data collection for them, which includes the lack of speakers, the diversity of dialects, and the lack of a writing standard. In addition, training large language models on more data brings environmental consequences (Strubell et al., 2019). In the long run, more training data may require longer GPU compute hours, which will release more greenhouse gas emissions.

Aside from data creation, other techniques are often employed as an alternative. A popular approach for multilingual or low-resource NLP is to leverage other languages to benefit from cross-lingual transfer. These approaches include using them as pivot (Wijaya et al., 2017; Xia et al., 2019), transfer learning (Gu et al., 2018; Nguyen and Chiang, 2017), and joint training (Neubig and Hu, 2018; Johnson et al., 2017). Improvements from such methods indicate a strong influence of the presence of other languages in the training data. Given that including related languages alongside the low-resource language can improve performance (Xia et al., 2019; Poncelas and Effendi, 2022; Gu et al., 2018; Nguyen and Chiang, 2017; Neubig and Hu, 2018; Johnson et al., 2017), it is beneficial to include proximity measurements between these languages on evaluations, which can be done using the vectors from the URIEL database (Littell et al., 2017). The utilization of the URIEL database has made evaluating multiple languages more explainable by leveraging linguistically aware feature vectors from which linguistic distances can be computed. These vectors have been utilized by previous works in various ways including determining which language to use as transfer or pivot language (Lin et al., 2019; Nambi et al., 2023) and measuring language diversity (Ruder et al., 2021).

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 43–52

It has been established that there are benefits to using other languages in the training process. However, multilingual labeled data creation is challenging. In this paper, we aim to provide hints to narrow down future data collection strategies by evaluating an existing LLM family. A constraint in previous studies that assess the GPT model series (Hendy et al., 2023; Robinson et al., 2023) has been the fact that these models are proprietary, closed-source systems that do not disclose information regarding their training data. This presents a challenge as it remains unclear which languages are included in the training of the models. On the other hand, open-source LLMs such as Meta's Llama2 (Touvron et al., 2023b), is more transparent about its training process, including the languages that are included in its training data. This makes the model more suitable as a subject for our evaluation.

In this work, we are evaluating Llama2 (Touvron et al., 2023b) for machine translation to highlight its multilingual capability in languages it has or has not seen during training. We also model a linear relationship (through correlation scores) between the linguistic feature distances and the translation metrics and use these scores as a basis for language importance analysis. The goal of the analysis is to narrow down the data investment effort by shedding light on which language(s) may improve the translation of other languages when included in the training data. An efficient data collection strategy will result in future multilingual LLMs that can be trained and deployed more efficiently, thus promoting sustainability. In summary, our contributions are as follows:

1. We evaluate Llama2 and provide machine translation scores of this model for 41 languages, 15 of which were not seen during its training.
2. We reveal that increasing model parameters is more effective in improving translation over instruction tuning and few-shot learning.
3. Our research reveals that syntactic similarity between languages is not the only linguistic aspect that is strongly linked to machine translation performance. Surprisingly, these strong correlations between linguistic feature distances and machine translation performances extend beyond English and hold true across various languages, therefore opening up the possibility of other better central languages for multilingual LMs

2. Methodology

2.1. Machine Translation Evaluation

We experiment with languages reported in the training data of Llama2 (Touvron et al., 2023b), the list of which and their respective ISO 639-3 codes can be found in Table 1. We refer to this set of languages as **inllama**. We also pick 15 languages not reported in the training data which we will refer to as **outllama**, presented in Table 2. It is important to highlight that languages not explicitly mentioned in Llama2 might still be present in the training data, albeit at a minuscule proportion of less than 0.005% of its training data (Touvron et al., 2023b). Languages in **outllama** cover various language genera and writing systems. The machine translation evaluation is conducted using the FLORES-200 (Guzmán et al., 2019) benchmark as it is available for numerous low-resource languages. We exclude X→English translation directions to mitigate the risk of potential data leakage, given that FLORES-200 uses Wikipedia for its English sentences. We also exclude zero-shot translation as LLMs often get the language wrong in this prompting setup as reported by Robinson et al. (2023). We measure translation quality using machine translation scores. Translation quality is measured with the BLEU score (Papineni et al., 2002) and a model-based machine translation metric (COMET-22 (Rei et al., 2022)) where applicable. COMET-22 is used to compensate for the drawbacks of BLEU and vice-versa.

Language	Genus	BLEU	COMET-22
German (deu)	Germanic	33.68	0.83
Swedish (swe)	Germanic	37.71	0.87
Dutch (nld)	Germanic	27.45	0.84
Norwegian (nor)	Germanic	29.54	0.86
Danish (dan)	Germanic	36.21	0.86
French (fra)	Romance	42.4	0.84
Spanish (spa)	Romance	28.54	0.84
Italian (ita)	Romance	28.78	0.85
Portuguese (por)	Romance	43.21	0.87
Catalan (cat)	Romance	35.92	0.84
Romanian (ron)	Romance	31.58	0.84
Russian (rus)	Slavic	28.21	0.85
Polish (pol)	Slavic	22.34	0.83
Ukrainian (ukr)	Slavic	26.03	0.83
Serbian (srp)	Slavic	23.96	0.81
Czech (ces)	Slavic	24.94	0.82
Bulgarian (bul)	Slavic	29.57	0.83
Croatian (hrv)	Slavic	21.3	0.81
Slovenian (slv)	Slavic	19.51	0.77
Chinese (zho)	Chinese	19.79	0.82
Japanese (jpn)	Japanese	17.02	0.84
Vietnamese (vie)	Vietic	28.77	0.82
Korean (kor)	Korean	11.08	0.78
Indonesian (ind)	Malayo-Sumbawan	31.15	0.86
Finnish (fin)	Finnic	18.08	0.82
Hungarian (hun)	Ugric	18.4	0.78

Table 1: List of **inllama** languages along with their ISO 639-3 codes, genus, and machine translation scores obtained using one-shot Llama2-7B.

We aim to experiment with open-source LLMs

Language	Genus	Writing System
Afrikaans (afr)	Germanic	Latin
Galician (glg)	Romance	Latin
Macedonian (mkd)	Slavic	Cyrillic
Slovak (slk)	Slavic	Latin
Armenian (hye)	Armenian	Armenian
Basque (eus)	Basque	Latin
Georgian (kat)	Kartvelian	Georgian
Icelandic (isl)	Germanic	Latin
Igbo (ibo)	Igboid	Latin
Javanese (jav)	Javanese	Latin
Sinhala (sin)	Indic	Sinhala
Tagalog (tgl)	Greater Central Philippine	Latin
Tamil (tam)	Dravidian	Tamil
Telugu (tel)	Dravidian	Telugu
Welsh (cym)	Celtic	Latin

Table 2: List of **outllama** languages and their ISO 639-3 codes. We also include in this table additional language information retrieved from WALS (Dryer and Haspelmath, 2013)

that replicate proprietary models such as ChatGPT (OpenAI, 2023) in terms of usability and safety. At the time the Llama2 model was released and the experiment design for this paper was constructed, none of the open-source models are suitable substitutes for production models as they may not have been aligned to match human preferences and there may be a performance gap (Touvron et al., 2023b). On account of this, we decided to move forward only with the Llama2 model family. The machine translation evaluation begins with one-shot translations for both languages in **inllama** and **outllama** using the vanilla 7B model. From this experiment, we categorize languages that yield under 10 BLEU as **unlearned** languages[1]. For the **unlearned** languages, we experiment further with model scale, chat version, and adding the shot count to maximize the potential of in-context learning. Our choice of randomly picking 5 shots from the validation set of FLORES-200 is motivated by the experimental setup used by Hendy et al. (2023) which states that increasing beyond 5 shots does not result in meaningful improvement and shows that selected quality shots do not always improve more than 1 BLEU compared to random selections for GPT (text-davinci-003) model. For translation with chat models, we use the prompt by Robinson et al. (2023) which follows the recommendation of Gao et al. (2023) for designing prompts for translation using instruction-tuned models. The prompts used in our experiments are given in Table 3

[1]Based on "Almost useless" interpretation from https://cloud.google.com/translate/automl/docs/evaluate

2.2. Correlation Score Analysis

We consider several language subsets. For every language subset, we calculate the Pearson correlation score between the linguistic similarity scores of each language in the subset to a language in **inllama** and their respective translation scores. We assume that a certain language is important if we observe a positive correlation. For example, consider the language **A** and the language subset {**B**, **C**, **D**, **E**}. When the similarity of **A** with each language in the subset {**B**, **C**, **D**, **E**} and the respective machine translation scores for {**B**, **C**, **D**, **E**} exhibit a positive correlation, i.e. the closer they are to **A** the better their machine translation scores, **A** is deemed as a valuable language and is therefore hypothesized to be more optimal for a central language when developing multilingual language models. **A** is checked for each language in **inllama**. Similarity scores are calculated on five dimensions: GENETIC, GEOGRAPHICAL, INVENTORY, PHONOLOGY, and SYNTACTIC as per the URIEL typological database (Littell et al., 2017). We exclude the FEATURAL distances to focus on each dimension as FEATURAL distances are combinations of all the other feature distances[2]. Language subsets considered are **inllama** languages only, **outllama** languages only, both **inllama** and **outllama** languages, only **Germanic** languages, only **Romance** languages, only **Slavic** languages, and languages belonging to **Other genera**.

3. Results and Analysis

3.1. Machine Translation Evaluation Results

One-shot 7B Llama2 translation results are presented in Table 1 and Table 4. From Table 1, we observe that none of the languages included in **inllama** produce a BLEU score below 10. This suggests that we can reasonably assume that Llama2 is capable of translating into all the languages it has encountered during training. However, many languages in **outllama** yield a BLEU score under 10, this is expected as Llama2 is presumably unfamiliar with these languages. On the other hand, we hypothesize that there are two possibilities for the high-performing **outllama** languages; (1) those languages are indeed included in the training data i.e. included in the 0.005% of the training data, or (2) similar languages in **inllama** indeed boosted their performance.

We move forward with languages in **outllama** that yield a BLEU score below 10 and experiment

[2]For more detailed explanation of these distances, consult https://www.cs.cmu.edu/~dmortens/projects/7_project/

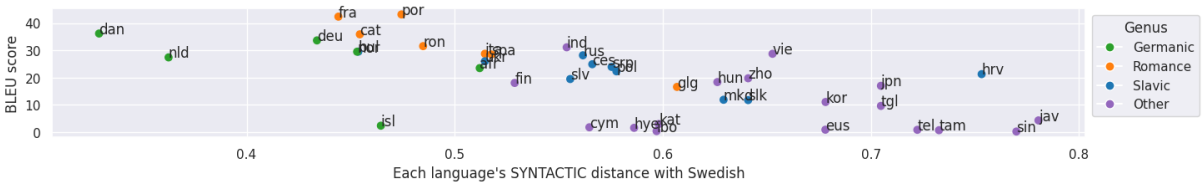


Figure 1: Scatter plot for **inllama** and **outllama** languages against the SYNTACTIC distance to **Swedish**. The correlation score is -0.67 and the p-value is 3.16×10^{-6}. The negative correlation here implies that the smaller the SYNTACTIC distance of a language to Swedish, the better is its MT performance

Model	Prompt
Non-chat	[SRC]: [src-sentence] [TGT]: [tgt-sentence] ... [SRC]: [src-sentence] [TGT]:
Chat	This is an English to [TGT] translation, please provide the [TGT] translation for these sentences: [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] ... Please provide the translation for the following sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]: [tgt-sentence] [TGT]

Table 3: Prompts used in our experiments to translate languages using the non-chat and chat versions of Llama2

Languages in outllama	BLEU	COMET-22
Afrikaans	23.52	0.74
Galician	16.62	0.76
Macedonian	11.90	0.67
Slovak	11.77	0.68
Armenian	**1.6**	0.31
Basque	**0.91**	0.33
Georgian	**2.99**	0.31
Icelandic	**2.39**	0.35
Igbo	**0.39**	0.41
Javanese	**4.33**	0.59
Sinhala	**0.25**	0.29
Tagalog	**9.65**	0.60
Tamil	**0.73**	0.30
Telugu	**0.87**	0.33
Welsh	**1.8**	0.35

Table 4: Llama2-7B one-shot translation results for languages in **outllama**. Languages with results in boldface are considered **unlearned** languages

with other variations of Llama2. We explore the effect of scale, chat version, and adding shot count and present the results in Table 5. Due to our limited compute resources we excluded the 70B and 70B-chat versions of Llama2.

Scaling up the model enhances translation ability. However, improvements from instruction–tuning and adding shot count remain inconclusive. Results presented in Table 5 demonstrate that the 13B versions of Llama2 outperform the smaller 7B versions for all **unlearned** languages. However, larger models do not seem to yield the same number of gains for every language. In best cases, 13B models increase on average as high as 2.53 BLEU with a standard deviation of 1.64. For instruction-tuning (chat) models, we observed both performance increase and decrease. The best improvements are observed in Igbo and Javanese, which improves as much as 3.16 and 2.87 respectively, and a decrease is observed in Tagalog, which performs worse on chat models with a decrease as severe as 2.64. Adding the shot count generally improves performance although it is less drastic than model scale and instruction-tuning with a mean increase of 0.47 and 0.08 for non-chat and chat Llama-13B respectively. While these model variations appear to enhance Llama2's capacity to translate into some languages greatly, there are languages where the prospects are limited. For instance, for Sinhala and Tamil, scaling up the model/adding shot count/using chat models results in less than 1 BLEU score increase.

3.2. Language Importance Analysis

We use the results from Table 1 and Table 4 for the linguistic proximity analysis. We first retrieve pre-computed distances[3] from the URIEL database and retrieve only the distances between the languages we are translating into and the languages reported in Llama2. Self or identity distances e.g. Igbo-to-Igbo distance are excluded in the Pearson correlation calculation. This correlation analysis aims to model the linear relationship between language proximity and machine translation scores to

[3] http://www.cs.cmu.edu/~aanastas/files/distances.zip

Language	7B 1S	7B 5S	7B-chat 1S	7B-chat 5S	13B 1S	13B 5S	13B-chat 1S	13B-chat 5S
Armenian	1.6	1.95	2.26	2.43	2.52	**3.03**	2.89	**3.03**
Basque	0.91	1.08	2.98	3.11	1.52	1.9	3.72	**3.88**
Georgian	2.99	3.44	4.41	4.7	5.57	**6.19**	5.97	5.88
Icelandic	2.39	3.06	3.9	3.86	4.72	5.21	**5.24**	5.04
Igbo	0.39	0.59	1.77	2.04	0.56	0.67	**3.72**	3.49
Javanese	4.33	3.71	4.94	5.06	3.15	3.76	5.92	**6.63**
Sinhala	0.25	0.38	0.57	0.52	0.48	**0.63**	**0.63**	0.62
Tagalog	9.65	10.98	10.8	10.97	16.1	**16.91**	13.82	14.27
Tamil	0.73	1.01	0.82	1.09	1.79	**2**	1.7	1.56
Telugu	0.87	1.04	1.02	0.86	2.29	**2.45**	1.77	1.68
Welsh	1.8	å 2.37	4.38	3.93	5.68	**6.8**	6.45	6.6

Table 5: BLEU scores with various Llama2 versions and shot count for languages considered **unlearned** by Llama2 (Table 4). **1S**/**5S**=one-shot/five-shot. Best result for each language is bolded.

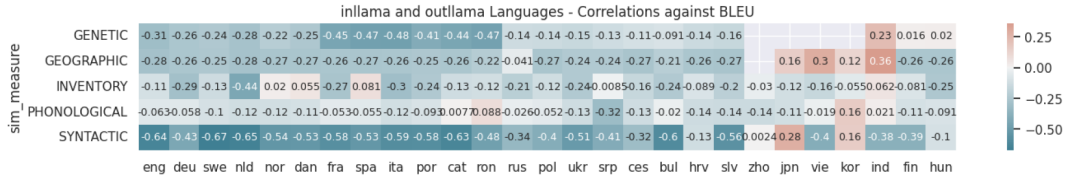
identify languages whose data may be beneficial for multilingual training.

We present our analysis as heatmaps in Figure 2 and 3 for correlations with BLEU and COMET-22 respectively. To help understand where each number came from in the heatmap, a scatter plot visualization for SYNTACTIC distance to Swedish for the combined **inllama** and **outllama** language subset against BLEU scores is presented in Figure 1 as an example. We create several different heatmaps according to the subset considered. It is important to highlight that *distance* is used as a similarity score. Therefore, a negative correlation between linguistic distance and MT scores would imply that the closer (i.e., the *smaller* the linguistic distance) a language is to this language, the higher the MT score is likely to be. In addition, since Wikipedia is a permanent fixture of LLMs' training data, we observe that there is a positive correlation between MT scores and Wikipedia article counts[4], as high as **0.64** using BLEU and **0.55** using COMET-22.

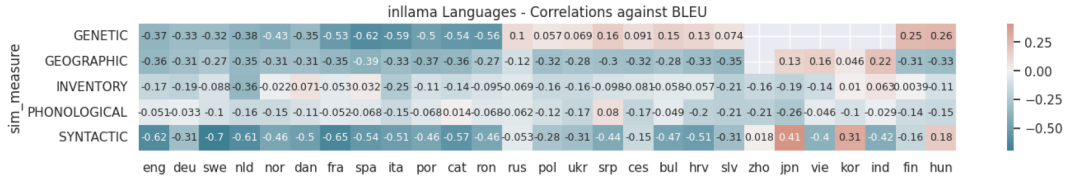
Syntactic similarity may be an important feature, but other linguistic dimensions can be too. When including every language, i.e. the **inllama** and **outllama** subset, BLEU complemented with COMET-22 scores show consistently strong correlations with syntactic features, especially with Germanic and Romance languages. This finding may not be particularly groundbreaking, as we already understand that the languages in **inllama** predominantly belong to these language genus. However, when considering only **outllama** language subset, translation performance seems to have higher correlations (either positive or negative) with GENETIC and PHONOLOGICAL distances. When considering languages in **outllama**, only SYNTACTIC similarities to certain languages e.g. Norwegian and Catalan display a strong correlation with MT performances. Correlation with features other than SYNTACTIC is also observed when considering languages in **other genera**, in which the proximity with the INVENTORY feature of Vietnamese, Dutch, German, and French are shown to correlate with COMET-22 scores.

English is not always the most syntactically important. When considering languages from **other genera** English demonstrates the most substantial syntactic correlation with MT performance, although there are other languages, such as Swedish and Vietnamese, that also display some degree of correlation. However, despite having the highest amount of training data, English is often not in the first place when considering languages by genus (e.g., **Germanic**, **Slavic**, and **Romance**). Similar to when we observe that syntactic proximity to Norwegian and Catalan have higher correlations with MT scores than syntactic proximity to English when considering only **outllama** languages, this phenomenon is accentuated when calculating correlations by genus. Among **Germanic** languages, syntactic proximity to English surprisingly shows little to no correlation with either BLEU or COMET-22 scores. Instead, **Germanic** languages' MT scores appear to correlate more with syntactic proximity to Dutch, Swedish, Catalan, and Bulgarian. This is also observed in **Slavic** languages where the MT scores generally correlate with syntactic proximity to most Germanic and Romance languages *except English*. With **Slavic** languages, syntactic proximity to English has the lowest correlation on BLEU and almost no correlation on COMET-22 scores. Finally, when focusing exclusively on **Romance** languages, it is interesting to observe that proximities to languages situated on the right side of the heatmaps i.e. **other genera**, exhibit higher correlations while they show no correlation when only considering other language subsets (Figure 2 and 3).

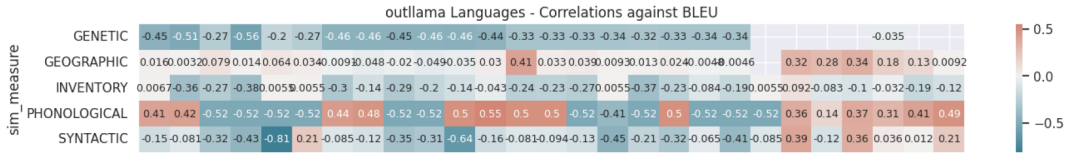
[4]Retrieved from https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group on October 2023



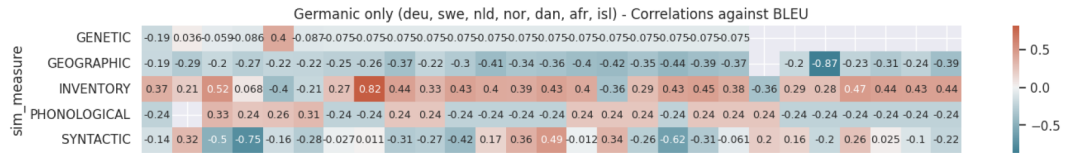
*When considering all languages (**inllama** and **outllama**), SYNTACTIC features are prominent*



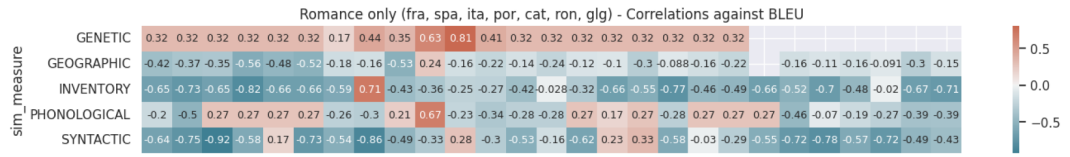
*When only **inllama** languages are considered, other linguistic features are observed to correlate stronger. Feature importance is observed to shift to GENETIC and GEOGRAPHIC features.*



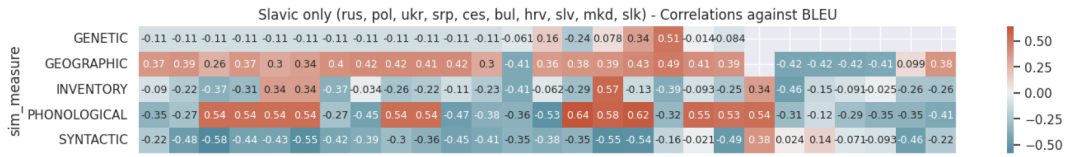
*When only **outllama** languages are considered, feature importance seem to shift away drastically from SYNTACTIC features and move toward PHONOLOGICAL and GENETIC features instead.*



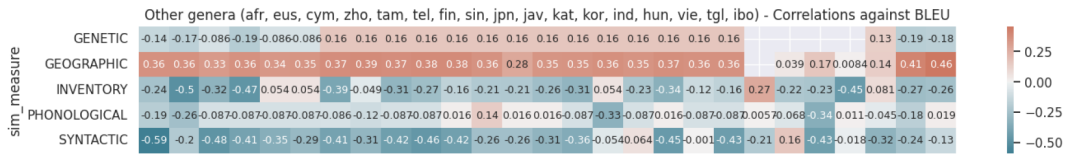
*When only **Germanic** languages are considered, the two most prominent correlation are SYNTACTIC Swedish and GEOGRAPHIC Vietnamese. Surprisingly, English shows a weaker correlation despite being a **Germanic** language.*



*When considering **Romance** languages, feature importance seem to shift into INVENTORY and SYNTACTIC features. In this language subset, languages on the right side (**Other genera**) are showing stronger correlations.*

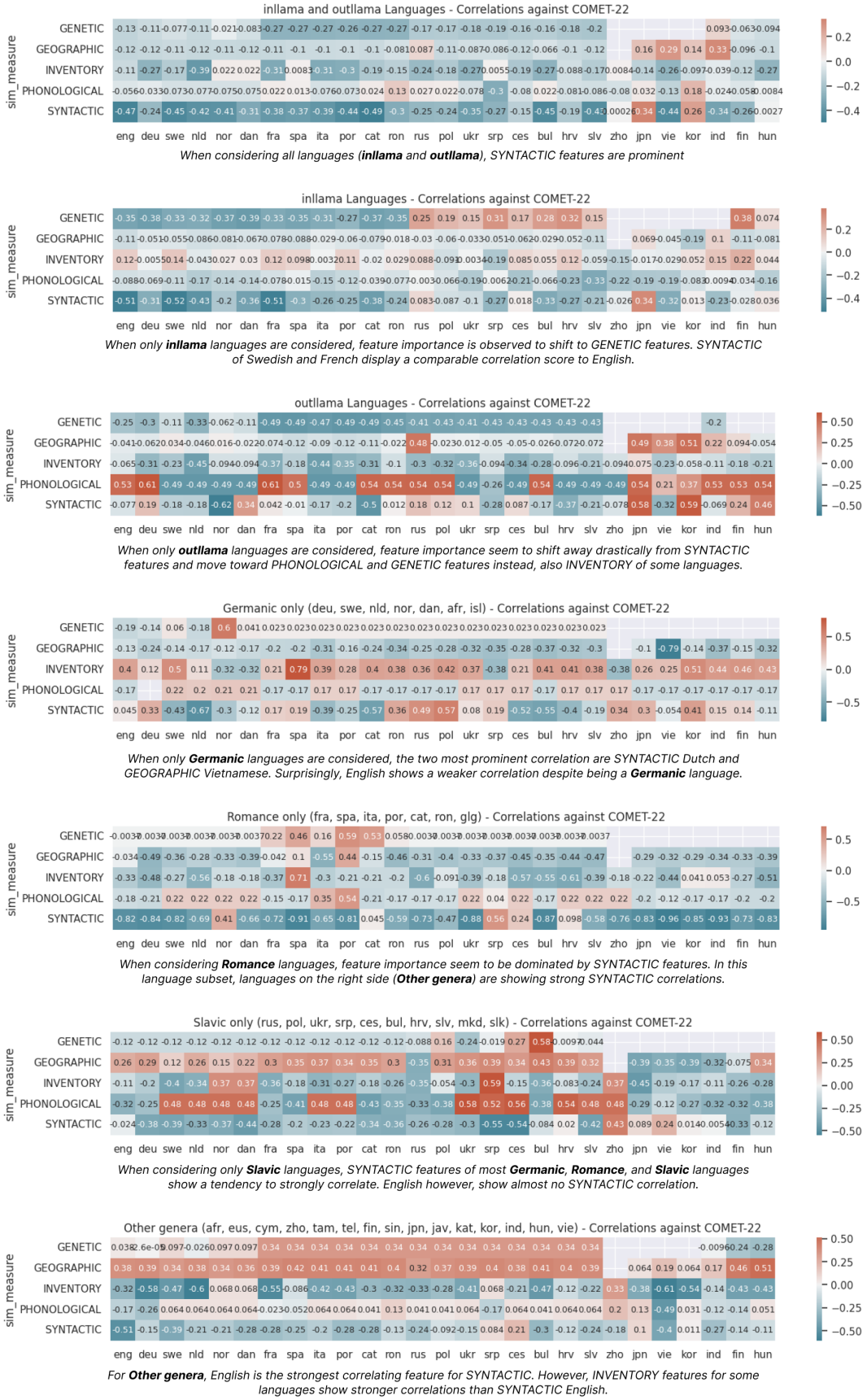


*When considering only **Slavic** languages, SYNTACTIC features of **Germanic**, **Romance**, and **Slavic** languages show a tendency to strongly correlate.*



*For **Other genera**, SYNTACTIC English is the strongest correlating feature.*

Figure 2: Heatmaps of correlations between linguistic distances with BLEU scores of the Llama2-7B one-shot prompting setup (language subset considered is written above each heatmap)



*When considering all languages (**inllama** and **outllama**), SYNTACTIC features are prominent*

*When only **inllama** languages are considered, feature importance is observed to shift to GENETIC features. SYNTACTIC of Swedish and French display a comparable correlation score to English.*

*When only **outllama** languages are considered, feature importance seem to shift away drastically from SYNTACTIC features and move toward PHONOLOGICAL and GENETIC features instead, also INVENTORY of some languages.*

*When only **Germanic** languages are considered, the two most prominent correlation are SYNTACTIC Dutch and GEOGRAPHIC Vietnamese. Surprisingly, English shows a weaker correlation despite being a **Germanic** language.*

*When considering **Romance** languages, feature importance seem to be dominated by SYNTACTIC features. In this language subset, languages on the right side (**Other genera**) are showing strong SYNTACTIC correlations.*

*When considering only **Slavic** languages, SYNTACTIC features of most **Germanic**, **Romance**, and **Slavic** languages show a tendency to strongly correlate. English however, show almost no SYNTACTIC correlation.*

*For **Other genera**, English is the strongest correlating feature for SYNTACTIC. However, INVENTORY features for some languages show stronger correlations than SYNTACTIC English.*

Figure 3: Heatmaps of correlations between linguistic distances with COMET-22 scores of the Llama2-7B one-shot prompting setup (language subset considered is written above each heatmap)

4. Related Work

Our work aligns with previous studies that assess LLMs for translation, resembling the work by Hendy et al. (2023) and Robinson et al. (2023). We aim to extend such evaluations further by investigating the influence of the languages included in the training data of the model, which was previously underexplored due to the lack of transparency of LLMs used. Our method of analysis, similar to the work of Robinson et al. (2023), investigates feature importance. Our objective is to extend that exploration by encompassing other linguistic features obtained from the URIEL typological database (Littell et al., 2017). We are interested in the phenomenon observed in the work of Lin et al. (2019) which shows that however important dataset statistics are compared to linguistic features, there are cases where using them alone to choose transfer languages results in poor performance. This phenomenon drove us to conduct a more comprehensive exploration of linguistic features.

5. Conclusion

We provide a comprehensive evaluation of machine translation in Llama2 for languages seen or unseen in its training data. In this work, we provide English→X machine translation scores of Llama2 7B for 26 languages reported to be in the training data of Llama2 models. We also evaluated 15 additional languages that are not reported to be in Llama2 training data using the 7B, 7B-chat, 13B, and 13B-chat Llama2 models. Our results show that Llama2 is capable of translating into languages it is unfamiliar with, although this phenomenon is observed only in some languages. We demonstrate that model scaling has the most substantial impact when compared to instruction tuning and adding shot count, whose improvements vary by language. We also modeled the linear relationship of linguistic distances and translation quality through correlation scores and revealed that syntactic similarity is not the only feature that displays strong correlations with machine translation scores. Furthermore, despite English having the most training data, there are other languages (e.g. Swedish, Catalan) whose linguistic distances exhibit comparable correlation scores to English albeit having much fewer training data. Our findings pose a unique perspective on the current landscape of language models, suggesting that the prevailing focus on English-centered models may not be the most optimal setup for multilingual models. We hope to open doors toward more effective and training-data-efficient multilingual systems that are shaped by languages other than English, thus promoting digital language equality and sustainability.

Limitations

Our research heavily depends on the language distances obtained from the URIEL typological database, as introduced by Littell et al. (2017). The original authors noted that many languages in the database may have missing features, which means the accuracy of our findings is constrained by the methods used to compensate for these missing features. Our evaluation with the COMET-22 metric is only done for languages supported in their models. However, the model may not be equally reliable for all languages, thus the COMET-22 correlations are only as accurate as the COMET-22 model. Furthermore, there are other ways to model the relationship between language feature distances and machine translation scores. We leave such investigations for future work. We also left out positively correlated features in our analysis as they are not readily interpretable in the context of our analysis.

In an ideal scenario, it would be advantageous to include all languages from the FLORES-200 benchmark and all available versions of Llama2 and other multilingual models to provide more evidence of the effectiveness of scaling parameter count and the overall generalizability of our findings. Unfortunately, our research is constrained by limited computational resources, preventing us from achieving this comprehensive coverage. We exclude X→English translation directions as Llama2 is likely trained on English Wikipedia. We also exclude prompting languages in **outllama** using various dictionary-based prompting techniques due to the challenging work required to collect accurate dictionary entries for low-resource languages. However, we leave this for future work.

We are also aware that the chat versions of Llama2 have been intentionally trained to prevent the generation of harmful or toxic content, and this protective design may affect the quality of translations. Moreover, the chat versions of the model generate numerous artifacts in addition to the translated sentences. We have made diligent efforts to automate the output parsing process to ensure that metrics are calculated fairly. The task of human evaluation and manual parsing of the outputs is left for future work.

Acknowledgements

Authors from Indonesian institutions are supported by the Indonesian Ministry of Education, Culture, Research, and Technology (MoECRT) ACE Open Research program. This work is also supported in part by the Indonesia-US Research Collaboration in Open Digital Technology grant funded by the Indonesian Ministry of Education, Culture, Research, and Technology. The authors would like to thank

the Indonesian government for their funding and Boston University for providing essential computing resources through the Shared Computing Cluster (SCC).

6. Bibliographical References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Akshay Nambi, Vaibhav Balloli, Mercy Ranjit, Tanuja Ganu, Kabir Ahuja, Sunayana Sitaram, and Kalika Bali. 2023. Breaking language barriers with a leap: Learning strategies for polyglot llms.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

OpenAI. 2023. Gpt-4 technical report.

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, Pennsylvania, USA. Association for Computational Linguistics.

Alberto Poncelas and Johanes Effendi. 2022. Benefiting from language similarity in the multilingual MT training: Case study of Indonesian and Malaysian. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 84–92, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463, Copenhagen, Denmark. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

7. Language Resource References

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.