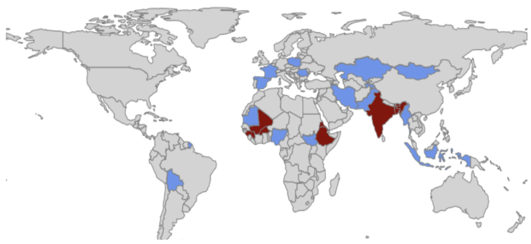
It's All About In-Context Learning!
Teaching Extremely Low-Resource Languages to LLMs

Yue Li, Zhixue Zhao and Carolina Scarton
Department of Computer Science, University of Sheffield, UK
{yli381,zhixue.zhao,c.scarton}@sheffield.ac.uk

Abstract

Extremely low-resource languages, especially those written in rare scripts, as shown in Figure 1, remain largely unsupported by large language models (LLMs). This is due in part to compounding factors such as the lack of training data. This paper delivers the first comprehensive analysis of whether LLMs can acquire such languages purely via in-context learning (ICL), with or without auxiliary alignment signals, and how these methods compare to parameter-efficient fine-tuning (PEFT). We systematically evaluate 20 under-represented languages across three state-of-the-art multilingual LLMs. Our findings highlight the limitation of PEFT when both language and its script are extremely under-represented by the LLM. In contrast, zero-shot ICL with language alignment is impressively effective on extremely low-resource languages, while few-shot ICL or PEFT is more beneficial for languages relatively better represented by LLMs. For LLM practitioners working on extremely low-resource languages, we summarise guidelines grounded by our results on adapting LLMs to low-resource languages, e.g., avoiding fine-tuning a multilingual model on languages of unseen scripts.



Language	Script	Translator	Accuracy
Dzongkha (dzo)	རིགས་འགྱུར་འདི་གིས་ (Tibetan)	Y	0.128
Santali (sat)	ᱥᱟᱱᱛᱟᱲᱤ (Ol Chiki)	N	0.172
Nko (nqo)	ߒߞߏ ߞߊ߲ߜߍ (NKo)	N	0.137
Tamasheq (taq)	ⵏⴱⴰⵜⴰ (Tifinagh)	N	0.118
Tigrinya (tir)	ምዩቴሽን (Ge'ez)	Y	0.186

Figure 1: Regional distribution of the languages studied in this paper. Red denotes the five languages with rare scripts, and blue represents the other 15 languages. *Y* (yes) and *N* (no) denote whether it's supported by Google Translate[1]. *Accuracy* represents performance on topic classification (SIB-200) with DeepSeek (7b) in zero-shot ICL (majority voting = 0.25, English = 0.83).

1 Introduction

Current large language models (LLMs) are typically pre-trained with data in more than 50 languages, offering robust support for high-resource languages, such as German and French (Le Scao et al., 2023; Grattafiori et al., 2024; Team et al., 2023). However, their coverage of low-resource languages remains limited. Since these languages are often spoken in developing regions, insufficient LLM support risks reinforcing socio-economic disparities and further isolating affected communities (Shen et al., 2024; Jadhav et al., 2024).

Extending LLMs to support extremely low-resource languages via continued pre-training (Yong et al., 2023a) is possible but often impractical, due to the need for large-scale monolingual corpora and substantial computational resources (Joshi et al., 2020). Although LLM support can be achieved for downstream tasks via resource-efficient training, such as parameter-efficient fine-tuning (PEFT), it still requires a non-trivial amount of labeled data. Therefore, with recent advances in in-context learning (ICL), we ask **whether LLMs can *learn* new languages purely through ICL** (Yong et al., 2023b; Zhang et al., 2024; Cahyawijaya et al., 2024). Specifically: (1) Is ICL alone sufficient to enable LLMs learn extremely low-resource or entirely unseen languages? (2) Can auxiliary signals in the prompt be useful enabling or improving ICL? (3) ICL or PEFT, which one is

[1] https://cloud.google.com/translate/docs/languages

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29533–29548

a better choice for learning a new language?

In this study, we consider 20 low-resource languages, including five extremely low-resource ones (Figure 1) and 15 written in Latin, Arabic, or Cyrillic scripts (referred to as *target* languages[2]). ICL with auxiliary signals (i.e., class category, language alignment) is explored. Our setup (Table 1) includes few-shot ICL, sentence-level alignment of unlabelled or labelled examples in zero-shot or few-shot ICL, and word-level alignment for the target language in zero-shot ICL.

To our knowledge, this is the first study to systematically analyse whether ICL can enable LLMs to *learn* extremely low-resource languages[3]. Our main findings are:

- In contrast to prior work (Razumovskaia et al., 2024), small-scaled fine-tuning is generally ineffective when a language and its script are highly under-represented or entirely absent from both the tokenizer and pre-training data (e.g., `sat`, `nqo` and `taq` on DeepSeek).
- In such cases, zero-shot ICL with language alignment yields substantial gains, potentially surpassing vocabulary extension on multilingual pre-trained language models (PLMs) through continue pre-training.
- Zero-shot ICL with language alignment is especially effective for languages with minimal LLM support, often exceeding few-shot ICL and performing comparably to, or better than, fine-tuning.
- Few-shot ICL and especially PEFT perform best for low-resource languages for which LLMs exhibit a certain level of support.

2 Related Work

Language Adaptation in Pretraining Continued pretraining LLMs on a monolingual corpus in a target language is a common strategy to extend support to languages not (well) represented in the original pretraining data, also enhancing ICL performance in the target language (Yong et al., 2023a). Various methods have been explored for training efficiency (Zhang et al., 2021; Yong et al., 2023a; Cui et al., 2023), vocabulary and tokeniser adaptation (Yamaguchi et al., 2024a; Balachandran, 2023; Cui et al., 2023; Larcher et al., 2023) and data efficiency (Yamaguchi et al., 2024b; Shaham et al., 2024; Kurz et al., 2024). However, the effectiveness of these pretraining-based methods often depends on the availability of large-scale training data, an assumption that does not hold for extremely low-resource languages in real-world scenarios.

Adapting LLMs to Low-Resource Languages for Downstream Tasks ICL with different strategies has been explored to improve LLMs' adaptation to low-resource languages, including techniques such as code-switching (Yong et al., 2023b; Schlicht et al., 2025), demonstration example selection (Winata et al., 2022; Zhang et al., 2024; Tanwar et al., 2023), prompt format optimization (Zhang et al., 2023; Cahyawijaya et al., 2024), machine translation (Bandarkar et al., 2024), and dictionary-based prompting (Lu et al., 2024; Zhang et al., 2024). Another promising direction is PEFT, which has demonstrated superior performance with computational costs comparable to few-shot ICL (Liu et al., 2022). However, most existing studies focus on languages that are: (1) relatively high-resource (e.g., German); (2) low-resource but written in widely supported scripts (e.g., Zhuang in Latin script (Zhang et al., 2024)); and (3) written in rare scripts but already included in model pre-training (Razumovskaia et al., 2024). Consequently, how to effectively adapt LLMs to extremely low-resource languages such as `nqo` (Figure 1) is still unclear.

3 Learning Extremely Low-Resource Languages

Table 1 summarises the experimented approaches and assumed available data resources. Standard cross-lingual transfer that aims to improve and then transfer task knowledge from English is not considered in this study (i.e, fine-tuning with English data or few-shot ICL with English examples), as the LLMs have already demonstrated high accuracy on English in zero-shot ICL (Table 4)[4].

Baseline The vanilla zero-shot ICL when the LLMs are prompted only with task description and the target language input tgt. We use English task

[2] represented by language code ISO 639-3 (in Table 2)

[3] The term "extremely low-resource languages" refers to the five languages with rare scripts in this study

[4] Machine translating target-languages into English is not considered, since our aim is to teach low-resource languages to LLMs. Reliable machine translators are also not available for 3 out of 5 rare-script languages. In practice, developing a high-quality machine translator is significantly more expensive than creating the data resources we consider here.

Method		Data EN	Data Target	Data Other	Training	Annotator
Zero-shot ICL	baseline	✗	✗	✗	✗	✗
	sentence-level alignment	✓(k)	✓(k)	✗	✗	*translate*
	word-level alignment	✗	✗	*dict.*	✗	✗
	word-level translation	✗	✗	*dict.*	✗	✗
Few-shot ICL	demonstration in target language	✗	✓(k)	✗	✗	*label*
	demonstration with alignment	✓(k)	✓(k)	✗	✗	*label+translate*
Parameter efficient fine-tuning		✗	✓(N)	✗	✓	*label*

Table 1: List of methods used in this paper and resources they may rely on: (1) Data: in-domain data in English (EN) or target language (Target), or a dictionary (dict.) for word translation; k denotes k-shot examples, and N denotes the full training data ($k \ll N$); (2) Training: whether model parameters are updated; (3) Annotator: whether native speakers are needed to *label* the data, or *translate* data to English to enable alignment.

description as it has been widely adopted showing improvements in ICL performance (Zhang et al., 2023; Razumovskaia et al., 2024). The prompt format is: [<task description> + <input^{tgt}>].

Zero-Shot ICL with alignment we experiment with adding word- or sentence-level alignment between English and a target language in the prompt, without providing labelled examples.

- *Word-level alignment*: We provide a translation for each word in the target-language input using a dictionary, inspired by prior work on machine translation (Zhang et al., 2024; Lu et al., 2024). The prompt format for an input^{tgt} with N words {w$_1^{tgt}$, w$_2^{tgt}$, ..., w$_N^{tgt}$} is: [<task description> + <input^{tgt}> + <w$_1^{tgt}$ means w$_1^{eng}$ in English; ...; w$_N^{tgt}$ means w$_N^{eng}$ in English>]. We use the NLLB translator[5] (Costa-jussà et al., 2022) to create the dictionaries following Lu et al. (2024). For languages not supported by NLLB (`nqo`, `sat`[6], and `min`), we train the word alignment tool *fast_align* (Dyer et al., 2013) to simulate a high-quality dictionary (See Appendix B).

- *Word-level translation*: We directly concatenate the English word translations in their orders in target languages as the "English" translation (i.e., input$^{eng'}$ = concat (w$_1^{eng}$, ..., w$_N^{eng}$)), and prompt LLMs with: [<task description> + input$^{eng'}$].

- *Sentence-level alignment*: Assuming there is a limited k number of parallel in-domain unlabelled sentences in English {s$_1^{eng}$,...,s$_k^{eng}$} and target language {s$_1^{tgt}$,...,s$_k^{tgt}$}. The prompt format is: [<target language: s$_1^{tgt}$; English: s$_1^{eng}$; ...; target language: s$_k^{tgt}$; English: s$_k^{eng}$> + <task description> + <input^{tgt}>]. We select the target-language example sentences from the training data through random sampling or BM25 (Robertson et al., 2009).

Few-Shot ICL Assuming there is a limited number of labelled in-domain data samples in the target language or English, demonstration examples from the training data are retrieved by BM25, inspired by Zhang et al. (2024).

- *Demonstration in target language*: We prompt LLMs with k-shot demonstration examples in the target language D$_1^{tgt}$, ..., D$_k^{tgt}$. The prompt format is [<task description> + <D$_1^{tgt}$, ..., D$_k^{tgt}$> + <input^{tgt}>].

- *Demonstration with alignment*: LLMs are prompted with parallel demonstration examples in both English and target languages: [<task description> + <D$_1^{tgt}$ means D$_1^{eng}$, ..., D$_k^{tgt}$ means D$_k^{eng}$> + <input^{tgt}>].

PEFT We preliminarily experiment with competitive methods such as LoRA (Hu et al., 2022), DoRA (Liu et al., 2024) and IA3 (Liu et al., 2022). Same as Yong et al. (2023a), we found that IA3 is the most effective and efficient approach. Therefore, due to computational constraints, we only experiment with IA3 as a representative of the PEFT methods. We also discuss the comparison with fully fine-tuned multilingual PLMs in Section 4.4.

3.1 Experimental Setups

Target Languages We mainly ground our research on the *SIB-200* seven-way topic classification dataset (Adelani et al., 2024), as it offers parallel training and evaluation data with the broadest multilingual coverage in natural language understanding (NLU) tasks. We also analyse the generalisability of our findings on reading comprehension

[5] `https://huggingface.co/facebook/nllb-200-3.3B`

[6] NLLB repeats the same word without stopping

(i.e., BELEBELE (Bandarkar et al., 2024)), which is a more challenging task than topic classification. Since most prevalent LLMs do not disclose comprehensive lists of languages present in their pretraining data, we select the low-resource languages for which LLMs exhibit significantly limited capability. We measure the LLMs' capability on each language with *Information Parity* (IP) (Tsvetkov and Kipnis, 2024) on the SIB training data. Given a text in the target language and its English translation, IP is defined as the ratio between the negative log-likelihood of the target-language text and that of its English counterpart under the same model. A very low IP score indicates that the LLM struggles to represent information in the target language, likely due to limited or no exposure during pretraining. Based on this criterion, we select the languages with the average lowest IP scores across the LLMs we study on. This includes five languages written in relatively rare and distinct scripts (Figure 1) plus 15 languages using more commonly supported scripts (i.e., seven in Latin, four in Arabic, and four in Cyrillic). For the latter, we select languages that do not have the same linguistic roots as English (Latin script), Modern Standard Arabic (Arabic script) and Russian (Cyrillic script), which are commonly represented in LLMs' training data. The full list of the languages is shown in Table 2.

Models We experiment with three recent open-source instruction-tuned LLMs with multilingual ability: *DeepSeek*[7], *LlaMA-3.2*[8] and *Gemma-2*[9]. Due to computational constraints, their medium-sized variants are considered.

Setups We adopt accuracy as the evaluation metric following Adelani et al. (2024). We use greedy search in decoding for the purpose of reproducibility. The prompt template for baseline zero-shot ICL is also used in PEFT for a fair comparison. SIB-200 dataset's official train/dev/test set split is used. The examples included in zero-shot and few-shot ICL are retrieved from the training data. As preprocessing for BM25, only white-space splitting is applied. The hyper-parameter tuning and training details for IA3 are included in Appendix B.

[7]https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat

[8]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[9]https://huggingface.co/google/gemma-2-2b-it

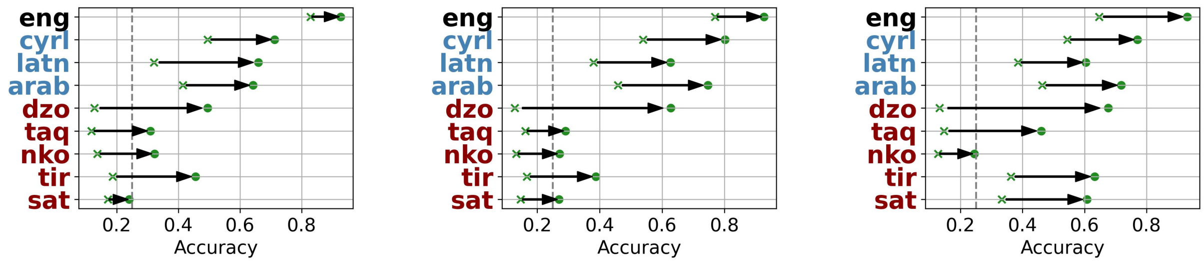
4 Results

4.1 Limitation of Fine-Tuning

Fine-tuning Improvement Disparity Figure 2 illustrates the performance improvement after fine-tuning with the training data in the target language. In most cases, fine-tuning leads to enhanced performance, although the degree of improvement varies notably. For low-resource languages using common scripts, accuracy scores can rise to more than 0.6 on average, resulting in an acceptable performance (full results in Table 4). In contrast, results on the five languages written in rare scripts are inconsistent. For instance, while DeepSeek performs worse than majority voting on all of these five languages in the baseline zero-shot ICL setting, PEFT raises the accuracy scores of `dzo` and `tir` to above 0.45. In contrast, gains for the remaining three languages are rather modest, particularly for `sat`, which still stays slightly below majority voting. We observe that this discrepancy is due to overfitting, which appears to occur at a very early stage in the fine-tuning for languages showing limited improvement.

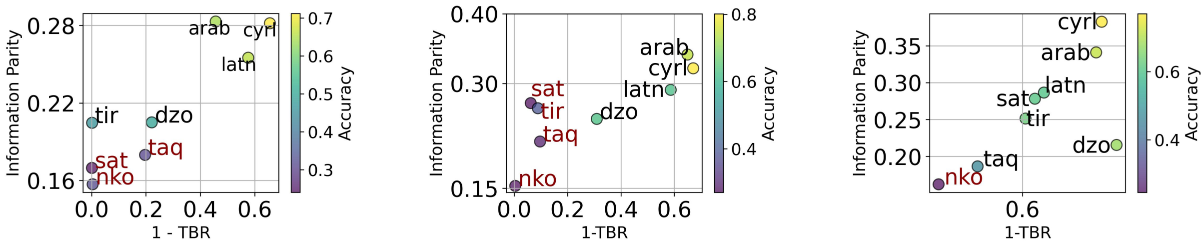
Risk of Overfitting and Impact Factors To gain more insights on why certain languages suffer more severe overfitting, we analyse:

- *Tokenization efficiency:* Most LLMs' tokenisers, including those used by the three models in our study, adopt byte-level Byte Pair Encoding (BPE) (Wang et al., 2020) or SentencePiece (Kudo and Richardson, 2018). When encountering texts in rare scripts not seen during tokeniser training, characters are often segmented into raw bytes, resulting in a vocabulary with drastically reduced effectiveness. For instance, BPE tokenisers based on UTF-8 encoding may end up representing an entire rare-script language using only 256 raw-byte token values (Wang et al., 2020), limiting the model's ability to learn generalisable linguistic patterns with small training data (Zhao and Aletras, 2024). Tokenisation efficiency for a given text i is measured using *Token-to-Byte Ratio* (TBR) $= \frac{num^i_{\text{tokens}}}{num^i_{\text{bytes}}}$, where num^i_{bytes} is the number of bytes required to represent the text with the same encoding system used in the tokeniser, and num^i_{tokens} is the number of tokens produced by the LLM's tokeniser. A TBR score close to 1 indicates that the tokeniser is operating nearly at the raw byte level, signalling extremely



(a) DeepSeek (b) LLaMA-3.2 (c) Gemma-2

Figure 2: Accuracy improvement from baseline zero-shot ICL (denoted as ×) to PEFT (denoted as •). The performance over languages in Latin (*latn*), Arabic (*arab*), and Cyrillic (*cyrl*) scripts is averaged. The performances of English (eng) and the majority voting baseline (*vertical dashed line, accuracy = 0.25*) are for reference.



(a) DeepSeek (b) LLaMA-3.2 (c) Gemma-2

Figure 3: Correlation between information parity (IP), token-to-byte ratio (TBR) and accuracy score after fine-tuning. For improved visualisation, the x-axis represents ($1-$TBR). The language bubbles close to the left corner denote languages that are under-represented by both tokeniser and model. Bubbles with darker colour denote lower performance after fine-tuning. Names of the languages with PEFT performances lower than 0.4 are marked in red.

poor tokenisation. For example, the average TBR for `sat`'s training data with DeepSeek's tokeniser is 0.99, suggesting that nearly every character is segmented into raw bytes and that DeepSeek significantly lacks meaningful representations for `sat`.

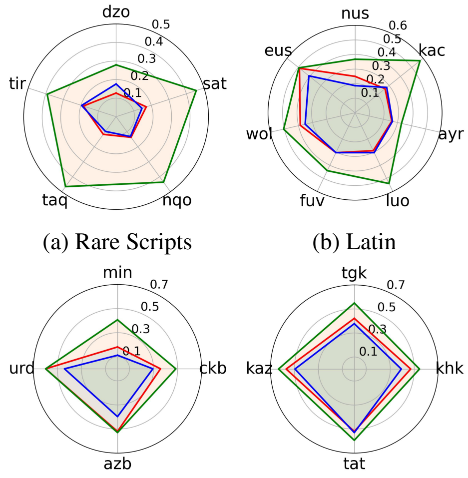
- *Multilingual capability:* Fine-tuning is usually more effective when the LLM has already acquired some linguistic competence in the target language during pre-training. IP is used again to estimate the LLMs' capabilities for each target language prior to fine-tuning. As discussed in Section 3.1, the higher the IP value the more efficiently the LLM represents information provided in the target language. Conversely, a low IP score suggests under-representation of a language.

Figure 3 shows the average TBR and IP scores for each target language in the training data, also presenting their correlations with fine-tuning performance. Languages with high TBR and low IP scores are the ones where fine-tuning tends to encounter more severely overfitting, resulting in very limited generalisation. This suggests that small-scaled fine-tuning on downstream tasks is unlikely to be beneficial when a language and its script are highly under-represented or even unseen in both tokeniser training and model pre-training. This finding also highlights the importance of improving representation of low-resource languages and scripts during pre-training. Even modest improvements in representation, either at tokeniser or model pre-training level, can lead to notably more effective find-tuning adaptation. For instance, although `tir`, `sat`, and `nqo` are nearly entirely tokenised as raw bytes by DeepSeek's tokeniser (TBR > 0.99), DeepSeek exhibits stronger pre-trained capabilities on `tir` (higher IP) compared to `sat` and `nqo`, translating into more substantial gains from fine-tuning. Similarly, while LLaMA-3.2 shows comparable IP scores for both `dzo` and `tir`, `dzo` benefits from better tokenisation (lower TBR, i.e. higher $1-$TBR), potentially leading to larger performance improvement after fine-tuning.

4.2 Alignment in Zero-Shot ICL

4.2.1 Sentence-Level Alignment

Effectiveness of Semantic Similar Examples Figure 4 shows DeepSeek's performance when prompted with one unlabeled example in the target



(a) Rare Scripts (b) Latin

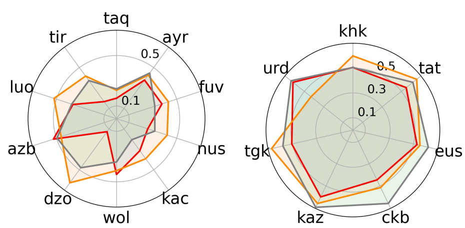
(c) Arabic (d) Cyrillic

Figure 4: Accuracy scores for DeepSeek in zero-shot ICL with sentence-level alignment when one unlabelled sentence is BM25-based (green) or randomly sampled (blue). Red denotes baseline zero-shot ICL.

language alongside its English translation. Similar trends are observed for LLaMA-3.2 and Gemma-2, with detailed results in the Appendix C. Although the topic label of the example is not included in the prompt, incorporating semantically similar texts in both target language and English significantly enhances performance for low-resource languages with low baseline zero-shot ICL performances, especially those using rarer scripts. However, this benefit diminishes as the baseline zero-shot ICL performance improves. For languages with baseline accuracy scores below 0.3, all three LLMs show an average improvement exceeding 0.22, with a peak gain of 0.36 for `sat` on LLaMA-3.2. In contrast, for languages with baseline scores above 0.3, the average improvement falls below 0.07. In some cases, LLaMA-3.2 and Gemma-2 exhibit minor performance declines, with the largest drops being 0.1 for `kaz` on both LLaMA-3.2 and Gemma-2 (baseline = 0.59 and 0.61 respectively). Furthermore, the models' performance often degrade when examples are randomly sampled from the training set, highlighting that the effectiveness of the alignment hinges on the semantic similarity between the input text and the example in target language. Moreover, improving semantic search for low-resource languages or even enhancing text pre-processing approaches (e.g., lemmatisation and stemming), beyond the simple whitespace-based tokenisation used here, has the potential to increase performance in this sentence-alignment setting.

Impact of the Number of Examples We further analyse the performance when varying the number of unlabelled parallel examples provided in the prompt, from two to five examples. We find that increasing the number of randomly sampled parallel examples in the prompt could slightly improve the performance, although the gap between BM25 sampled examples is still notable. However, providing more semantic related examples does not consistently improve performance across all languages. We hypothesize that it may be influenced by the relative length of the target-language sequences compared to English. To test this, we compute the average *tokenizer parity* (TP) scores (Petrov et al., 2023) for each language in the training set (in Table 2). Given a text in its target language and English translation, TP is defined as the ratio of the number of tokens in English to the number of tokens in the target language. A lower TP score indicates that the target language is tokenized into relatively longer sequences compared to English. We define a binary variable indicating whether adding more examples is beneficial: it is set to true if at least 3 out of the 4 multi-shot settings (2, 3, 4, and 5 examples) outperform the 1-shot setting. We then calculate the point-biserial correlation coefficient (Lev, 1949) between the TP score and this binary indicator. The results show a statistically significant correlation, suggesting that languages with lower TP scores are less likely to benefit from additional parallel unlabelled examples.

4.2.2 Word-Level Alignment



(a) chrf++ ≤ 0.5 (b) chrf++ > 0.5

Figure 5: Accuracy scores for LLaMA-3.2 over low-resource languages in zero-shot ICL with word-level alignment (gray) or word-level translation (orange) settings. Red denotes baseline zero-shot ICL.

The zero-shot ICL performance with word-level alignment or translation on LLaMA-3.2 is shown in Figure 5. The reported chrf++ score[10] of each language on NLLB is used as an indicator of the qual-

[10] https://tinyurl.com/nllb200dense3bmetrics

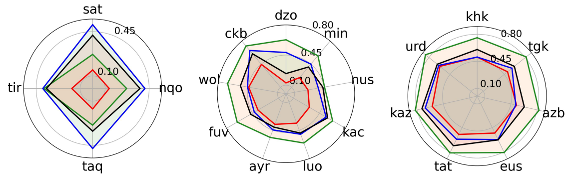
ity of the dictionary. For DeepSeek and Gemma-2 (full results in Appendix C), adopting word-level alignment or translation always improve the performance over the baseline. However, LLaMA-3.2's performance is highly influenced by the dictionary's quality. When the chrf++ score is lower than 0.5 (Figure 5a), including the low-quality word-level alignment (gray lines) can harm the performance (e.g., fuv, kac, wol, and azb). In contrast, when chrf++ score is higher than 0.5 (Figure 5b), word-level alignment is always beneficial. For nqo, sat, and min, we train *fast_align* on the SIB-200 dataset to simulate a high-quality dictionary (see Section 3). LLMs' performance for these languages is always improved with word-level alignment (around 0.6 of accuracy improvement), highlighting the importance of the dictionary quality. We provide discussion on potential impact of *fast_align* quality in the Appendix C.

ICL with word-level translation significantly reduces the inference time than using word-level alignment by reducing input length. However, its performance exhibits variable superiority across languages and LLMs. Specifically, it is consistently better than world-level alignment on LLaMA-3.2, while it is always worse than world-level or even baseline on DeepSeek and Gemma-2.

4.3 Alignment in Few-Shot ICL

For languages with baseline accuracy lower than majority voting, including English translations in few-shot ICL often improves results across all the LLMs, when using more than one demonstration example. However, in 1-shot ICL, removing the English translation tends to yield better performance on DeepSeek and LLaMA-3.2, while Gemma-2 continues to benefit from them. For languages with strong zero-shot ICL performance, both DeepSeek and Gemma-2 benefit from the alignment, whereas LLaMA-3.2 performs best without them. Overall, unlike the consistent trends across LLMs in zero-shot ICL with alignment, we observe more variations how LLMs respond to aligned prompts in few-shot scenarios (full results in Appendix C).

Although not examined by prior work (Cahyawijaya et al., 2024), we find that model performance can be highly sensitive to the order of the task description and demonstration examples in the prompt, for certain languages, especially on DeepSeek. For example, when prompting Deepseek with 1-shot labelled nqo example with English translation, the accuracy jumps from 0.30



(a) extremely low (b) low (acc<0.45) (c) low (acc>0.45)

Figure 6: Accuracy comparison among baseline (red), PEFT (green), best zero-shot ICL (blue) and best few-shot ICL (**black**) on LLaMA-3.2. Languages are categorised into: (a) Both language and script are severely under-represented (names in red in Figure 3b, baseline accuracy < 0.2): **zero-shot** > **few-shot** > **PEFT**; (b) Better represented, but baseline < 0.45: **PEFT** > **zero-shot** ≥ **few-shot**; (c) Baseline > 0.45: **PEFT** > **few-shot** > **zero-shot**.

to 0.42 if the task description is moved to after the demonstration example. In most of the cases, prompting with demonstrations at the beginning lead to better performance for DeepSeek and Gemma-2, while LLaMA-3.2 slightly prefers task description at the beginning. Results presented for few-shot ICL are based on the optimal task description position selected on the validation set.

4.4 Comparison across PEFT, Zero- and Few-Shot ICL

Based on our analysis, zero-shot ICL with word-[11] or sentence-level alignment and few-shot ICL with or without alignment can all achieve promising improvement over the baseline across languages and LLMs. Next, we discuss which approach is more effective from different aspects. We present the results in Figure 6 for LLaMA-3.2. DeepSeek and Gemma-2 show same trend (see Appendix C).

Fine-Tuning vs. ICL When the low-resource language is extremely under-represented in both tokeniser and model (Figure 6a), fine-tuning LLMs or even PLMs[12] yields minimal improvement, while zero-shot ICL with either sentence- or word-level alignment offers significant improvements (Figure 2a vs. 4a). Adelani et al. (2024) extended the vocabulary of XLM-R for nqo with continue-pretraining, leading to fine-tuning accuracy on SIB-200 test set rising 0.17 points. However, it is still lower than the performances of LLMs in zero-shot ICL with alignment, which are above 0.41.

[11] Results based on *fast_align* are excluded, as it potentially leaks gold standard English translations into the prompt.

[12] Based on results for XLM-R (large) from Adelani et al. (2024), see Table 4.

For the remaining low-resource languages (Figure 6b and 6c), fine-tuning normally has better performance than ICL. Overall, the average difference between PEFT and the best ICL approach on DeepSeek, LLaMA-3.2 and Gemma-2 is 0.13, 0.13, and 0.08, respectively.

Zero-Shot vs. Few-Shot We observe that for languages with low baseline zero-shot ICL performance (i.e., accuracy < 0.45 on all LLMs), in most cases (at least more than 50%), zero-shot ICL with word/sentence-level alignment leads to better performance than few-shot ICL regardless if the alignment is provided or not (Figure 6b, comparing black and blue lines). When baseline performance is higher than 0.45 (Figure 6c), few-shot ICL always provides best results. In our study, these languages are `khk`, `tgk`, `azb`, `eus`, `tat`, `kaz`, and `urd`, consistent across all the LLMs. Overall, if the LLM significantly lacks capability on the target language, providing label in ICL might be useless.

5 Discussion

5.1 NLP Tasks beyond Topic Classification

Reading Comprehension We test our findings on BELEBELE, a reading comprehension parallel multilingual dataset, covering 11 out of the 20 languages studied here. It contains questions with four multiple-choice answers linked to a passage. `tir` is available among the five languages with rare scripts. We split the data into training, validation and test, following SIB-200, to enable a consistent comparison (See Appendix C). We conduct experiments only on LLaMA-3.2, due to its long context length and computational constraints. We retrieve the passage from the training data with BM25 as example and provide its English translation as passage alignment. We adopt accuracy for evaluation (Bandarkar et al., 2024).

Most results align with our observations on SIB-200 dataset. Specifically, PEFT still shows no improvement for `tir`, while being more effective for other languages. Zero-shot ICL with passage alignment could still improve over the baseline zero-shot ICL in most cases, especially for the languages with lower baseline performance (e.g., accuracy < 0.35). However, as the task is more challenging, the level of improvement is not as notable as on topic classification. The model also potentially requires more unlabelled parallel data for consistent improvements across all languages. Similar to topic classification, in most cases, zero-shot ICL with passage alignment surpasses few-shot ICL when baseline performance is lower than 0.5.

Conversely, word-level alignment is not effective on reading comprehension, which may be explained by the quality of the dictionary created. Unlike topic classification, whose prediction can be made based on one or two topic-related words, reading comprehension relies less on such cue words, requiring a higher quality of word translations.

Machine Translation We further experiment with the FLORES-200 machine translation dataset (Costa-Jussà et al., 2022). Among the ICL settings we explore in this study (see Table 1), only the few-shot ICL (demonstration with alignment) setting is relevant to machine translation, which is equivalent to standard few-shot ICL for translation[13]. Similar with BELEBELE, we evaluate on the official *dev-test* set with LLaMA-3.2. We retrieve one demonstration example from the official *dev* set with BM-25, and adopt the *chrf++* score for evaluation. We present the full results in Table 5 (see Appendix C).

The results still align with our observations on classification tasks. Only one-shot demonstration example yields improvements in translation quality across both translation directions, except for `urd` where LLaMA-3.2 achieves significantly higher performance than for other languages in zero-shot ICL. Also, the gains are generally larger when translating from English into target languages than in the reverse direction. However, overall machine translation quality remains poor. For example, the *chrf++* score increases substantially from 0.4 to nearly 11 when translating English to `sat`, but such translation quality is still far from adequate for practical human use.

5.2 Suggestions to Practitioners

In practice, performance is not the only consideration. Investment in data and computational resources needs to be carefully considered, especially for low-resource languages. Aiming to adapt an LLM to a low-resource language for a downstream task, which approach should be prioritized, and what types of data should be created?

[13] Zero-shot ICL with sentence alignment corresponds to few-shot ICL, as demonstration examples of translations are provided in the prompt. As for zero-shot ICL with word alignment, we cannot rely on *fast_align* to simulate the dictionary for the three languages that are not supported by NLLB as we did for classification, because it would directly provide the gold-standard English translations to the LLMs.

For low-resource languages that are extremely under-represented in both tokeniser and model (e.g., nqo) fine-tuning is not effective and zero-shot ICL with alignment shows promising improvements. We suggest *prioritizing investment in human translation to create a small-scale in-domain parallel data for zero-shot ICL with alignment*.

For low-resource languages where LLMs demonstrate limited capability few-shot ICL might lead to better performance than zero-shot ICL with alignment. However, in most cases, these gains are modest and may even come at the risk of performance degradation. With fine-tuning LLM/PLM being effective and with acceptable zero-shot ICL performance for these languages, *decisions should be made by comparing the financial costs between human translation (for zero-shot ICL) and human annotation (for fine-tuning)*.

For low-resource languages where LLMs demonstrate a certain level of capability few-shot leads to better performance than zero-shot ICL with alignment. Human annotation tends to be required for a notable improvement for these languages. Practitioners should *consider the trade-off between the amount of data to annotate (effective fine-tuning may require more data than few-shot ICL) and the computational costs (LLM inference is more expensive than fine-tuning PLMs)*.

Since low-resource and extremely low resource languages are tokenised into long sequences, including extra word and sentence alignment in the prompt would further increase the computational costs. We estimate the computational costs for all the ICL settings in this study in Table 7 (see Appendix D).

6 Conclusion

This work provides a systematic analysis on whether ICL can enable LLMs to effectively support extremely low-resource languages on downstream tasks. As some of the key findings that contrast to prior work, we reveal the limitation of fine-tuning when languages and their scripts are both highly under-represented. In such cases, zero-shot ICL augmented with word- or sentence-level alignment yields promising results. Meanwhile, few-shot ICL or PEFT tends to perform better for languages relatively better represented during pre-training. Our study highlights the importance of language and script coverage in LLMs, and the strong potential of ICL for language adaptation.

Limitations

Although we conducted more than 450 experiments, our study did not include other popular LLMs, such as Mistral and Qwen. Due to our computational constraints and consideration of fair comparison with PEFT on same LLM size, we did not experiment with LLMs with large sizes, such as Gemma-2 (9b) [14] or LLaMA-3.3 (70b) [15]. Future work could explore whether a larger LLM could enable even more improvement in ICL with word- or sentence level language alignment.

Due to very limited datasets with parallel data available for these extremely low-resource languages, we covered topic classification and reading comprehension in this study. On reading comprehension, we were only able to experiment with 11 of the 20 target languages. Both SIB-200 and BELEBELE are constructed based on Flores-200 dataset (Costa-Jussà et al., 2022). With the increase of language coverage for NLP tasks in the future, our findings could be tested on other tasks (e.g., common-sense reasoning or summarization.) and other domains (e.g., medical, social media).

As a lack of native speakers and reliable gold-standard word translations, we were not able to accurately access the quality of the dictionary that we created using NLLB translator or fastalign. Our study does not show promising results when using the created dictionary to assist reading comprehension in ICL. However, as discussed in the main content, the results might be improved with a better dictionary. But the effectiveness on SIB-200 and ineffectiveness on BELEBELE imply that for challenging tasks with long input length, word translation quality is more important than for tasks with shorter input. Also, we directly translate words into English to simulate a dictionary following prior work. However, in practice, using a real-world dictionary raises additional challenges such as handling lexical ambiguity and polysemy, which may also impact the performance of word-level alignment in zero-shot ICL.

As very limited parallel corpus available for our target languages, we did not systematically analyse how the ICL performance would be impacted if the included unlabelled parallel text is out-of-domain

[14] https://huggingface.co/google/gemma-2-9b

[15] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

(e.g., from another dataset). However, since randomly sampled examples from the training data already poses risk of performance degradation, we hypothesise that zero-shot sentence-level alignment with out-of-domain examples might demonstrate limited benefit.

Acknowledgments

This work is supported by the UK's innovation agency (InnovateUK) grant number 10039039 (approved under the Horizon Europe Programme as VIGILANT, EU grant agreement number 101073921) (https://www.vigilantproject.eu).

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

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, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2024. On limitations of llm as annotator for low resource languages. *arXiv preprint arXiv:2411.17637*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2024. Investigating language-specific calibration for pruning multilingual large language models. *arXiv preprint arXiv:2408.14398*.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Joseph Lev. 1949. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1):125–126.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Evgeniia Razumovskaia, Ivan Vulic, and Anna Korhonen. 2024. Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet? *CoRR*, abs/2403.01929.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ipek Baris Schlicht, Zhixue Zhao, Burcu Sayin, Lucie Flek, and Paolo Rosso. 2025. Do llms provide consistent answers to health-related questions across languages? In *European Conference on Information Retrieval*, pages 314–322. Springer.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989, Miami, Florida, USA. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024a. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6760–6785, Miami, Florida, USA. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024b. How can we effectively expand the vocabulary of llms with 0.01 gb of target language text? *arXiv preprint arXiv:2406.11477*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023a. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023b. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, and 1 others. 2021. Cpm-2: Large-scale cost-effective pre-trained language models. *arXiv preprint arXiv:2106.10715*.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. 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 3226–3244, Mexico City, Mexico. Association for Computational Linguistics.

A Languages and Datasets Information

Languages The languages we experiment with are persented in Table 2, along with their IP and TP scores across LLMs. The chrf++ score from English to target language with NLLB translator, which we used to create the dictionary, is also included.

Datasets SIB-200 dataset is constructed based on the Flores-200 dataset. The data is categorised into seven topic classes: science/technology, travel, politics, sports, health, entertainment, and geography. The official training, validation and test set contain 701, 99 and 204 data points, respectively.

BELEBELE dataset is also derived from the Flores-200 dataset. It contains a passage, a question linked to the paragraph, and four choices. Following SIB-200, we split the dataset into training, validation and test set with 600, 93, and 207 data samples. No overlapping between the passages in the training/validation and test set. We preliminarily test two different random train/validation/test splits on Tigrinya and find the results are consistent.

B Implementation Details

Prompt For BELEBELE, we adopt the same prompt used by the authors for baseline zero-shot ICL[16]. As for SIB-200, we use the following prompt for baseline zero-shot ICL: "*What is the topic discussed in the following {language name} text? There are seven options: "science/technology", "travel", "politics", "sports", "health", "entertainment", and "geography". Now complete the following example without explanations. Text: {text}. Topic option is:*", as we found it performing better on the validation set than the one used by the original authors of SIB-200. Additionally, we observed that explicitly indicating the language of the input text had no impact on performance. For extremely low-resource languages such as Nko and Santali, LLaMA-3.2 and Gemma-2 refuse to perform the task if prompted with "*....complete the following example*", stating that they do not recognize the input language, no matter whether the name of the language is explicitly given in the prompt or not. However, they would produce a prediction when prompted with "*complete the following example without explanations*". For sentence-level alignment, the LLMs are instructed as "*Use the following pairs of {language name} texts and their English translations to help you understand {language name}.{alignment example}. Now based on your understanding, answer the question below without explanation.*". For word-level alignment, we instruct LLMs with "*Please use the provided English translation of each word to help you understand the {language name} text.*". Experiments are conduct on NVIDIA A100-PCIE-40GB.

Dictionary For each data sample in the test set, we extract words in target language based on white-space splitting only. Then we use NLLB-200 translator (3.3B)[17] to translate each word into English. For the three languages that are not supported by NLLB, we train the word alignment tool *fast_align* (Dyer et al., 2013) with SIB-200 training data and then align the English words and target-language words in the test set. We use the default training and alignment settings in *fast_align*[18].

[16] https://github.com/facebookresearch/belebele/blob/main/sample_zero_shot_instructions.md
[17] https://huggingface.co/facebook/nllb-200-3.3B
[18] https://github.com/clab/fast_align

Language Code	Language	Script	Family	Information Parity			Tokenizer Parity			NLLB	
				DeepSeek	LLaMA	Gemma	DeepSeek	LLaMA	Gemma	eng-X	X-eng
nqo_Nkoo	Nko	NKo	Manding	0.16	0.15	0.16	0.10	0.10	0.17	-	-
sat_Olck	Santali	Ol Chiki	Austroasiatic	0.17	0.27	0.28	0.08	0.08	0.20	28.4	39.9
taq_Tfng	Tamasheq	Tifinagh	Afro-Asiatic	0.18	0.22	0.19	0.13	0.11	0.21	18.8	26.2
tir_Ethi	Tigrinya	Ge'ez	Afro-Asiatic	0.20	0.26	0.25	0.13	0.14	0.31	24.8	49
dzo_Tibt	Dzongkha	Tibetan	Sino-Tibetan	0.20	0.25	0.22	0.08	0.09	0.26	32.6	40.1
nus_Latn	Nuer	Latin	Nilotic	0.21	0.22	0.22	0.31	0.27	0.37	28.9	38.2
min_Arab	Minangkabau	Arabic	Austronesian	0.22	0.23	0.22	0.24	0.37	0.43	-	-
tgk_Cyrl	Tajik	Cyrillic	Indo-European	0.23	0.28	0.32	0.36	0.36	0.42	49.8	59.5
ayr_Latn	Central Aymara	Latin	Aymaran	0.24	0.26	0.25	0.49	0.49	0.54	29.6	28.7
kac_Latn	Jingpho	Latin	Sino-Tibetan	0.24	0.25	0.25	0.45	0.46	0.51	38	39.3
wol_Latn	Wolof	Latin	Atlantic-Congo	0.25	0.28	0.27	0.53	0.56	0.61	28.1	39.8
azb_Arab	South Azerbaijani	Arabic	Turkic	0.25	0.29	0.31	0.28	0.55	0.58	23.8	43.6
tat_Cyrl	Tatar	Cyrillic	Turkic	0.25	0.32	0.37	0.34	0.34	0.47	48.7	56.7
luo_Latn	Luo	Latin	Nilotic	0.26	0.28	0.27	0.55	0.57	0.61	39	45.8
fuv_Latn	Nigerian Fulfulde	Latin	Atlantic-Congo	0.28	0.31	0.30	0.58	0.59	0.65	23.2	32.4
ckb_Arab	Central Kurdish	Arabic	Indo-European	0.30	0.32	0.35	0.21	0.26	0.36	45.2	58.5
khk_Cyrl	Halh Mongolian	Cyrillic	Mongolic-Khitan	0.32	0.33	0.38	0.33	0.33	0.40	42	52.6
eus_Latn	Basque	Latin	Basque	0.32	0.44	0.45	0.54	0.56	0.62	48.5	57.5
kaz_Cyrl	Kazakh	Cyrillic	Turkic	0.32	0.36	0.46	0.32	0.35	0.45	50.7	59.4
urd_Arab	Urdu	Arabic	Indo-European	0.36	0.52	0.49	0.22	0.34	0.54	48.3	61.7

Table 2: Full list of the 20 languages we experiment with in this study from the SIB-200 dataset, along with their information parity and tokenizer parity scores on DeepSeek, LLaMA-3.2 and Gemma-2. The reported chrf++ scores from two directions (eng-X: English to target language; X-eng: target language to English) with NLLB-200 translator (3.3B variant) is also included. Language code represents language (ISO 639-3)_script (ISO 15924).

IA3 The rescale vectors are learnt for key, value of the attention modules and feed-forward network in each layer. We use batch size as 4, and early stopping strategy based on validation loss, with max training epochs as 10. We use AdamW (Loshchilov and Hutter, 2019) for optimization. We perform hyper-parameter search on learning rate of {1e-3, 5e-3, 8e-3, 1e-2}. The optimal learning rate on validation set is 8e-3. For extremely low-resource languages where IA3 show limited improvement, we also search learning rate from {1e-4, 3e-3, 7e-3}. We run experiments 3 times and report the average performance. Experiments are conducted on NVIDIA GH200 480GB.

C Full Results

BELEBELE The results on LLaMA-3.2 are in Table 3.

SIB-200 The results of baseline zero-shot ICL, PEFT, along with fine-tuning multilingual PLM (from Adelani et al. (2024)) are presented in Table 4.

The results of zero-shot ICL with word-level, word translation and sentence-level alignment and few-shot ICL are presented in Table 6.

FLORES-200 The chrf++ scores for translating English into target-languages and target-languages to English are presented in Table 5.

Potential Impact of Alignment Quality by *fast_align* Since we lack gold-standard word alignment between English and nqo/sat/min, we estimate the potential impact of alignment quality on model performance with the following two sets of experiments: (1) randomly pairing the words in the target-language sentence with those in its English translation; and (2) randomly pairing the words in the target-language sentence with those in another randomly sampled English sentence.

The results show that random alignment with words in the gold-standard English translation sentence yields nearly the same ICL performance as *fast_align alignment*. However, pairing with words in randomly sampled English sentence leads to dramatically performance drop. The observation suggests that the word alignment quality between gold-standard English translation by *fast_align* is not directly related to topic prediction, unlike tasks such as machine translation, since LLMs could make predictions directly based on the presence of gold-standard English words in the prompt.

D Computational Costs Estimation

Following previous work (Liu et al., 2022; Kaplan et al., 2020), we estimate that a decoder-only LLM with N parameters uses $2N$ FLOPs per token for inference. For a parallel corpus in target language and English, we suppose the LLM tokenises the texts in target language and English into X and Y tokens on average, respectively, and ignore the length of the task description and class labels in the prompt. Therefore, the input token length for the six ICL settings we explore could be approximated as follows:

Language Code	Baseline Zero-Shot	PEFT	Zero-Shot with Align	Baseline Few-Shot	Few-Shot with Align
kac_Latn	0.261	0.353(1)	0.329(2)	0.285	0.280
wol_Latn	0.261	0.361(1)	0.319(2)	0.256	0.261
fuv_Latn	0.266(2)	0.309(1)	0.256	0.227	0.237
tir_Ethi	0.275	0.271	0.300(1)	0.227	0.280(2)
luo_Latn	0.295	0.314(2)	0.280	0.319(1)	0.275
ckb_Arab	0.333	0.440(1)	0.372(2)	0.280	0.324
tgk_Cyrl	0.357	0.391(1)	0.386(2)	0.338	0.343
kaz_Cyrl	0.362(2)	0.464(1)	0.348	0.275	0.251
khk_Cyrl	0.372(2)	0.472(1)	0.290	0.266	0.300
eus_Latn	0.415	0.623(1)	0.420(2)	0.367	0.338
urd_Arab	0.517	0.638(1)	0.459	0.546(2)	0.473

Table 3: The accuracy scores on the BELEBELE test set with LLaMA-3.2: baseline ICL (zero-shot), PEFT, zero-shot with alignment (3 parallel examples retrieved with BM25), 3-shot baseline ICL, and 3-shot ICL with alignment. Differences between baseline zero-shot ICL is statistical significant (paired chi-squared test). The number in parentheses denotes the rank of the performance on the target language.

Language Code	Baseline Zero-Shot ICL			PEFT			XLM-R
	DeepSeek	LLaMA	Gemma	DeepSeek	LLaMA	Gemma	
taq_Tfng	0.118	0.162	0.147	0.309	0.290	0.461	0.269
dzo_Tibt	0.128	0.127	0.132	0.495	0.627	0.676	0.242
nqo_Nkoo	0.137	0.132	0.127	0.323	0.271	0.245	0.232
sat_Olck	0.172	0.147	0.333	0.240	0.270	0.608	0.245
tir_Ethi	0.186	0.167	0.363	0.456	0.387	0.632	0.677
min_Arab	0.181	0.245	0.260	0.583	0.578	0.520	0.381
nus_Latn	0.250	0.260	0.255	0.569	0.456	0.485	0.439
ayr_Latn	0.260	0.377	0.333	0.637	0.539	0.559	0.525
kac_Latn	0.265	0.314	0.319	0.672	0.627	0.574	0.627
luo_Latn	0.289	0.363	0.382	0.652	0.608	0.623	0.600
fuv_Latn	0.304	0.378	0.382	0.681	0.657	0.554	0.630
ckb_Arab	0.358	0.446	0.446	0.603	0.725	0.716	0.501
wol_Latn	0.387	0.441	0.436	0.657	0.691	0.632	0.601
tgk_Cyrl	0.422	0.505	0.485	0.696	0.814	0.716	0.598
khk_Cyrl	0.471	0.505	0.490	0.681	0.755	0.691	0.885
eus_Latn	0.490	0.529	0.588	0.750	0.809	0.804	0.892
azb_Arab	0.520	0.525	0.539	0.721	0.824	0.789	0.829
tat_Cyrl	0.520	0.549	0.583	0.706	0.814	0.799	0.819
kaz_Cyrl	0.569	0.598	0.618	0.765	0.824	0.877	0.914
urd_Arab	0.598	0.618	0.608	0.662	0.858	0.848	0.876
eng_Latn	0.828	0.770	0.647	0.926	0.926	0.931	0.921

Table 4: Baseline zero-shot ICL and PEFT performance over SIB-200 on DeepSeek, LLaMA-3.2 and Gemma-2. Differences between baseline zero-shot ICL is statistical significant (paired chi-squared test). Performance of fine-tuning XLM-R(*large*) is adopted from Adelani et al. (2024).

Language Code	en-X		X-en	
	Baseline	Few-Shot	Baseline	Few-Shot
dzo_Tibt	4.04	12.36	13.07	14.40
taq_Tfng	0.74	8.56	12.33	15.92
sat_Olck	0.40	10.95	16.79	17.86
nqo_Nkoo	0.47	9.91	9.74	16.11
tir_Ethi	2.68	4.40	14.13	16.29
min_Arab	0.88	4.16	14.64	16.58
ckb_Arab	3.57	13.31	29.41	31.74
azb_Arab	1.14	9.53	27.50	30.52
urd_Arab	33.07	32.96	50.31	50.23
nus_Latn	4.09	11.01	16.29	19.01
kac_Latn	8.63	18.77	15.52	22.22
ayr_Latn	7.66	13.11	18.85	20.31
luo_Latn	11.69	14.86	19.17	23.18
fuv_Latn	8.23	11.74	16.11	20.65
wol_Latn	7.53	11.06	16.09	22.52
eus_Latn	27.76	28.90	41.93	42.25
tgk_Cyrl	7.53	14.97	28.51	31.75
khk_Cyrl	1.95	11.09	24.01	26.61
tat_Cyrl	7.79	12.75	33.66	35.42
kaz_Cyrl	13.46	15.50	35.18	37.06

Table 5: Baseline and few-shot ICL performances (chrf++ scores) of translating English to the target languages (en-X) and translating target languages to English (X-en) with LLaMA-3.2.

- Baseline zero-shot: X;
- Zero-shot with word alignment: $X + Y$;
- Zero-shot with word translation: Y;
- Zero-shot with k number of sentence-level alignment examples: $X + k * (X + Y)$;
- Few-shot (k demonstrations only in the target language): $X + k * X$;
- Few-shot (k demonstrations with alignment)): $X + k * (X + Y)$.

We present the total FLOPs of the above six ICL approaches for all the low-resource languages with each LLM when inference on SIB-200 in Table 7.

Model	Language Code	Zero-Shot				Few-Shot	
		sentence(BM25)	sentence(random)	word	word translation	without align	with align
DeepSeek	taq_Tfng	0.466	0.098	0.265	0.221	0.338	0.328
	dzo_Tibt	0.279	0.176	0.623	0.676	0.225	0.181
	nqo_Nkoo	0.436	0.132	0.617	0.681	0.451	0.417
	sat_Olck	0.456	0.147	0.632	0.563	0.358	0.284
	min_Arab	0.407	0.113	0.621	0.627	0.377	0.368
	tir_Ethi	0.392	0.196	0.368	0.397	0.387	0.343
	nus_Latn	0.368	0.186	0.412	0.319	0.373	0.387
	ayr_Latn	0.328	0.265	0.461	0.422	0.353	0.358
	kac_Latn	0.574	0.279	0.431	0.260	0.451	0.431
	luo_Latn	0.539	0.304	0.500	0.490	0.515	0.456
	fuv_Latn	0.441	0.304	0.431	0.328	0.441	0.475
	ckb_Arab	0.485	0.294	0.505	0.627	0.422	0.417
	wol_Latn	0.505	0.353	0.505	0.392	0.505	0.534
	tgk_Cyrl	0.549	0.377	0.578	0.588	0.529	0.608
	khk_Cyrl	0.544	0.392	0.544	0.539	0.471	0.525
	eus_Latn	0.495	0.407	0.613	0.618	0.574	0.554
	azb_Arab	0.529	0.397	0.480	0.466	0.559	0.554
	tat_Cyrl	0.593	0.529	0.637	0.564	0.642	0.613
	kaz_Cyrl	0.632	0.495	0.598	0.559	0.603	0.627
	urd_Arab	0.598	0.441	0.603	0.588	0.657	0.676
LLaMA-3.2	dzo_Tibt	0.250	0.113	0.480	0.627	0.206	0.235
	nqo_Nkoo	0.417	0.201	0.523	0.730	0.377	0.333
	sat_Olck	0.505	0.176	0.593	0.754	0.422	0.368
	taq_Tfng	0.475	0.181	0.235	0.225	0.338	0.260
	tir_Ethi	0.387	0.103	0.373	0.417	0.343	0.368
	min_Arab	0.451	0.186	0.537	0.726	0.407	0.255
	nus_Latn	0.382	0.186	0.319	0.422	0.436	0.348
	kac_Latn	0.539	0.196	0.206	0.392	0.559	0.319
	luo_Latn	0.500	0.250	0.373	0.520	0.485	0.436
	ayr_Latn	0.363	0.255	0.446	0.426	0.412	0.343
	fuv_Latn	0.426	0.221	0.319	0.431	0.475	0.387
	wol_Latn	0.456	0.275	0.343	0.412	0.539	0.422
	ckb_Arab	0.529	0.333	0.657	0.515	0.608	0.485
	khk_Cyrl	0.490	0.255	0.505	0.598	0.603	0.534
	tgk_Cyrl	0.471	0.270	0.578	0.672	0.618	0.510
	azb_Arab	0.515	0.279	0.495	0.461	0.627	0.510
	eus_Latn	0.515	0.324	0.623	0.554	0.588	0.627
	tat_Cyrl	0.593	0.309	0.618	0.657	0.711	0.632
	kaz_Cyrl	0.500	0.348	0.691	0.657	0.735	0.676
	urd_Arab	0.588	0.284	0.637	0.431	0.662	0.593
Gemma-2	nqo_Nkoo	0.417	0.137	0.696	0.671	0.255	0.402
	dzo_Tibt	0.250	0.127	0.578	0.569	0.240	0.230
	taq_Tfng	0.475	0.167	0.240	0.230	0.353	0.431
	nus_Latn	0.382	0.206	0.446	0.338	0.338	0.382
	min_Arab	0.451	0.216	0.672	0.614	0.368	0.407
	kac_Latn	0.539	0.328	0.436	0.314	0.436	0.520
	ayr_Latn	0.363	0.270	0.402	0.417	0.363	0.402
	sat_Olck	0.505	0.240	0.686	0.622	0.480	0.549
	tir_Ethi	0.387	0.314	0.466	0.314	0.446	0.500
	fuv_Latn	0.426	0.333	0.485	0.358	0.500	0.520
	luo_Latn	0.500	0.328	0.569	0.377	0.480	0.515
	wol_Latn	0.456	0.412	0.505	0.363	0.529	0.603
	ckb_Arab	0.529	0.333	0.618	0.554	0.559	0.564
	tgk_Cyrl	0.471	0.407	0.696	0.593	0.608	0.593
	khk_Cyrl	0.490	0.368	0.637	0.539	0.554	0.578
	azb_Arab	0.515	0.485	0.603	0.471	0.642	0.667
	tat_Cyrl	0.593	0.466	0.686	0.598	0.735	0.721
	eus_Latn	0.515	0.495	0.711	0.613	0.716	0.711
	urd_Arab	0.588	0.495	0.691	0.480	0.779	0.765
	kaz_Cyrl	0.500	0.515	0.667	0.657	0.740	0.740

Table 6: Zero-shot ICL with language alignments and few-shot ICL with or without alignment over SIB-200 on DeepSeek, LLaMA-3.2 and Gemma-2.

Model	Language Code	Zero-Shot				Few-Shot	
		baseline	word-alignment	word-translation	sentence-alignment	without align	with align
LLaMA-3.2	taq_Tfng	1.58E+13	1.74E+13	1.62E+12	3.31E+13	3.15E+13	3.31E+13
	dzo_Tibt	1.93E+13	2.09E+13	1.62E+12	4.02E+13	3.86E+13	4.02E+13
	nqo_Nkoo	1.74E+13	1.90E+13	1.62E+12	3.63E+13	3.47E+13	3.63E+13
	sat_Olck	1.99E+13	2.16E+13	1.62E+12	4.15E+13	3.99E+13	4.15E+13
	tir_Ethi	1.21E+13	1.38E+13	1.62E+12	2.59E+13	2.43E+13	2.59E+13
	nus_Latn	6.19E+12	7.81E+12	1.62E+12	1.40E+13	1.24E+13	1.40E+13
	ayr_Latn	3.39E+12	5.01E+12	1.62E+12	8.39E+12	6.77E+12	8.39E+12
	kac_Latn	3.61E+12	5.23E+12	1.62E+12	8.84E+12	7.22E+12	8.84E+12
	luo_Latn	2.89E+12	4.51E+12	1.62E+12	7.41E+12	5.79E+12	7.41E+12
	fuv_Latn	2.91E+12	4.53E+12	1.62E+12	7.43E+12	5.81E+12	7.43E+12
	wol_Latn	2.97E+12	4.59E+12	1.62E+12	7.55E+12	5.93E+12	7.55E+12
	eus_Latn	2.97E+12	4.59E+12	1.62E+12	7.56E+12	5.94E+12	7.56E+12
	ckb_Arab	6.31E+12	7.93E+12	1.62E+12	1.42E+13	1.26E+13	1.42E+13
	min_Arab	4.51E+12	6.13E+12	1.62E+12	1.06E+13	9.03E+12	1.06E+13
	azb_Arab	3.03E+12	4.65E+12	1.62E+12	7.68E+12	6.06E+12	7.68E+12
	urd_Arab	4.87E+12	6.49E+12	1.62E+12	1.14E+13	9.75E+12	1.14E+13
	khk_Cyrl	5.08E+12	6.70E+12	1.62E+12	1.18E+13	1.02E+13	1.18E+13
	tat_Cyrl	4.89E+12	6.51E+12	1.62E+12	1.14E+13	9.79E+12	1.14E+13
	kaz_Cyrl	4.83E+12	6.45E+12	1.62E+12	1.13E+13	9.66E+12	1.13E+13
	tgk_Cyrl	4.62E+12	6.24E+12	1.62E+12	1.09E+13	9.24E+12	1.09E+13
Gemma-2	taq_Tfng	8.37E+12	9.99E+12	1.62E+12	1.84E+13	1.67E+13	1.84E+13
	dzo_Tibt	6.47E+12	8.09E+12	1.62E+12	1.46E+13	1.29E+13	1.46E+13
	nqo_Nkoo	9.62E+12	1.12E+13	1.62E+12	2.09E+13	1.92E+13	2.09E+13
	sat_Olck	8.07E+12	9.69E+12	1.62E+12	1.78E+13	1.61E+13	1.78E+13
	tir_Ethi	5.30E+12	6.92E+12	1.62E+12	1.22E+13	1.06E+13	1.22E+13
	nus_Latn	4.54E+12	6.16E+12	1.62E+12	1.07E+13	9.08E+12	1.07E+13
	ayr_Latn	3.12E+12	4.74E+12	1.62E+12	7.87E+12	6.25E+12	7.87E+12
	kac_Latn	3.27E+12	4.89E+12	1.62E+12	8.15E+12	6.53E+12	8.15E+12
	luo_Latn	2.72E+12	4.34E+12	1.62E+12	7.06E+12	5.44E+12	7.06E+12
	fuv_Latn	2.61E+12	4.23E+12	1.62E+12	6.84E+12	5.22E+12	6.84E+12
	wol_Latn	2.73E+12	4.35E+12	1.62E+12	7.09E+12	5.47E+12	7.09E+12
	eus_Latn	2.68E+12	4.30E+12	1.62E+12	6.99E+12	5.37E+12	6.99E+12
	ckb_Arab	4.59E+12	6.21E+12	1.62E+12	1.08E+13	9.19E+12	1.08E+13
	min_Arab	3.79E+12	5.41E+12	1.62E+12	9.21E+12	7.59E+12	9.21E+12
	azb_Arab	2.84E+12	4.46E+12	1.62E+12	7.31E+12	5.69E+12	7.31E+12
	urd_Arab	3.07E+12	4.69E+12	1.62E+12	7.77E+12	6.15E+12	7.77E+12
	khk_Cyrl	4.17E+12	5.79E+12	1.62E+12	9.95E+12	8.33E+12	9.95E+12
	tat_Cyrl	3.51E+12	5.13E+12	1.62E+12	8.65E+12	7.03E+12	8.65E+12
	kaz_Cyrl	3.66E+12	5.28E+12	1.62E+12	8.95E+12	7.33E+12	8.95E+12
	tgk_Cyrl	3.92E+12	5.54E+12	1.62E+12	9.47E+12	7.85E+12	9.47E+12
DeepSeek	taq_Tfng	1.40E+13	1.57E+13	1.68E+12	2.97E+13	2.80E+13	2.97E+13
	dzo_Tibt	2.18E+13	2.34E+13	1.68E+12	4.52E+13	4.35E+13	4.52E+13
	nqo_Nkoo	1.74E+13	1.90E+13	1.68E+12	3.64E+13	3.47E+13	3.64E+13
	sat_Olck	2.12E+13	2.29E+13	1.68E+12	4.41E+13	4.25E+13	4.41E+13
	tir_Ethi	1.33E+13	1.50E+13	1.68E+12	2.83E+13	2.66E+13	2.83E+13
	nus_Latn	5.68E+12	7.36E+12	1.68E+12	1.30E+13	1.14E+13	1.30E+13
	ayr_Latn	3.52E+12	5.20E+12	1.68E+12	8.73E+12	7.05E+12	8.73E+12
	kac_Latn	3.78E+12	5.46E+12	1.68E+12	9.24E+12	7.56E+12	9.24E+12
	luo_Latn	3.07E+12	4.75E+12	1.68E+12	7.83E+12	6.15E+12	7.83E+12
	fuv_Latn	3.03E+12	4.71E+12	1.68E+12	7.74E+12	6.06E+12	7.74E+12
	wol_Latn	3.18E+12	4.86E+12	1.68E+12	8.04E+12	6.36E+12	8.04E+12
	eus_Latn	3.14E+12	4.82E+12	1.68E+12	7.96E+12	6.28E+12	7.96E+12
	ckb_Arab	8.26E+12	9.94E+12	1.68E+12	1.82E+13	1.65E+13	1.82E+13
	min_Arab	6.93E+12	8.61E+12	1.68E+12	1.55E+13	1.39E+13	1.55E+13
	azb_Arab	6.02E+12	7.70E+12	1.68E+12	1.37E+13	1.20E+13	1.37E+13
	urd_Arab	7.65E+12	9.33E+12	1.68E+12	1.70E+13	1.53E+13	1.70E+13
	khk_Cyrl	5.20E+12	6.88E+12	1.68E+12	1.21E+13	1.04E+13	1.21E+13
	tat_Cyrl	5.06E+12	6.74E+12	1.68E+12	1.18E+13	1.01E+13	1.18E+13
	kaz_Cyrl	5.33E+12	7.01E+12	1.68E+12	1.23E+13	1.07E+13	1.23E+13
	tgk_Cyrl	4.75E+12	6.43E+12	1.68E+12	1.12E+13	9.50E+12	1.12E+13
	eng_Latn	1.65E+12	3.33E+12	1.68E+12	4.99E+12	3.31E+12	4.99E+12

Table 7: Total FLOPs of each ICL approach for each language on DeepSeek, LLaMA-3.2 and Gemma-2 when inference on SIB-200.