XAMPLER: Learning to Retrieve Cross-Lingual In-Context Examples

Anonymous ACL submission

Abstract

Recent studies indicate that leveraging off-the-shelf or fine-tuned retrievers, capable of retrieving relevant in-context examples tailored to the input query, enhances few-shot in-context learning of English. However, adapting these methods to other languages, especially low-resource ones, poses challenges due to the scarcity of cross-lingual retrievers and annotated data. Thus, we introduce **XAMPLER: Cross-Lingual Example Retrieval**, a method tailored to tackle the challenge of cross-lingual in-context learning **using only annotated English data**. XAMPLER first trains a retriever based on Glot500, a multilingual small language model, using positive and negative English examples constructed from the predictions of a multilingual large language model, i.e., MaLA500. Leveraging the cross-lingual capacity of the retriever, it can directly retrieve English examples as few-shot examples for in-context learning of target languages. Experiments on the multilingual text classification benchmark SIB200 with 176 languages show that XAMPLER substantially improves the in-context learning performance across languages.

1 Introduction

Large language models (LLMs) have shown emergent abilities in in-context learning, where a few input-output examples are provided with the input query. Through in-context learning, LLMs can yield promising results without any parameter updates (Brown et al., 2020). However, the efficacy of in-context learning is highly dependent on the selection of the few-shot examples (Liu et al., 2022).

Recent studies (Luo et al., 2024) have uncovered a more strategic approach to example retrieval. Rather than relying on random selection, these studies advocate for retrieving examples tailored to the input query, resulting in notable performance enhancements in in-context learning. The retrievers employed by these methods can be categorized into two main types: general off-the-shelf retrievers (Liu et al., 2022), e.g., Sentence-BERT (Reimers and Gurevych, 2019), and task-specific fine-tuned retrievers (Rubin et al., 2022), which are trained based on LLM signals using labeled data.

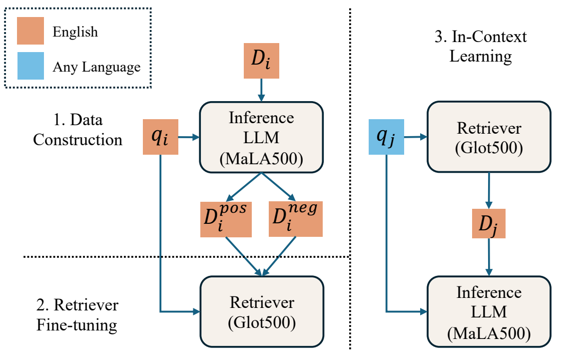

Figure 1: XAMPLER involves three steps: 1. Data Construction: given a query in English q_i, we divide the candidate examples D_i into positive examples D_i^{pos} and negative examples D_i^{neg} based on the prediction of MaLA500 (Lin et al., 2024b); 2. Retriever Fine-tuning: we fine-tune the retriever based on Glot500 (Imani et al., 2023) using the constructed data; 3. In-Context Learning: given a query in any language q_j, we use the fine-tuned retriever to retrieve relevant English examples as few-shots for in-context learning. **During training, XAMPLER is English-only, but for evaluation via in-context learning, it extends to any of the 500+ languages supported in MaLA500 and Glot500.**

Utilizing off-the-shelf retrievers has been further validated as an effective approach in multilingual settings (Nie et al., 2022; Winata et al., 2023; Tanwar et al., 2023). However, this method encounters limitations when applied to low-resource languages. Existing multilingual retrievers, e.g., SBERT (Reimers and Gurevych, 2020), cover a limited number of languages (i.e., 50+), and language-model-based retrievers (Hu et al., 2020) struggle to effectively align distant languages (Cao et al., 2020; Liu et al., 2023). Additionally, relying on off-the-shelf retrievers might lead to sub-optimal performance. Conversely, adopting task-specific fine-tuned retrievers has been demonstrated as a more effective approach (Rubin et al., 2022). Nonetheless, the availability of data for fine-tuning task-specific retrievers in low-resource languages is limited.

To tackle these challenges, we propose a simple yet effective method that relies solely on annotated English data, termed XAMPLER (Cross-Lingual Example Retrieval). As shown in Fig. 1, given an English query q_i and an English example from the candidate pool D_i, we employ in-context learning with MaLA500 (Lin et al., 2024b), a 10B multilingual LLM covering 534 languages, to predict the label of the query. Based on the correctness of the prediction, we classify the candidate example as either positive or negative, i.e., D_i^{pos} and D_i^{neg}. Then, leveraging the curated dataset, we train a retriever based on Glot500 (Imani et al., 2023), a multilingual small language model covering 534 languages, aiming to minimize the contrastive loss (Rubin et al., 2022; Cheng et al., 2023; Luo et al., 2023). Finally, the trained retriever is directly applied to retrieve valuable few-shot examples in English for the given query in the target language. The retrieved English few-shot examples, along with the input query, are then fed into MaLA500 for in-context learning. Experiments across 176 languages on SIB200 show that XAMPLER effectively retrieves cross-lingual examples, thereby enhancing in-context learning across languages.

2 Approach

2.1 Problem Definition

Given an input query q_i in any language, our objective is to enhance in-context learning for predicting the label of q_i by retrieving tailored few-shot examples from the pool of candidate examples D. Due to the scarcity of annotated data in low-resource languages, we introduce XAMPLER, namely, Cross-Lingual Example Retrieval. On one hand, we leverage in-domain English examples as the pool of candidate examples D, from which we retrieve cross-lingual examples in English for q_i in any target language. On the other hand, we only consider q_i sourced from English training data to train the task-specific retriever, which is then directly applied for evaluation across languages.

2.2 Data Construction

To train the task-specific retriever aimed at retrieving informative examples for the given query q_i, we consider contrastive learning, which requires both positive and negative examples for each query q_i. We define examples as positive when the LLM accurately predicts the ground truth of q_i while utilizing the example as a one-shot example appended to q_i for in-context learning. Conversely, examples are categorized as negative if the LLM's prediction deviates from the ground truth.

Scoring all pairs of training examples presents a quadratic complexity in $|D|$, making it resource-intensive. Inspired by Rubin et al. (2022), we mitigate this by selecting the top k similar examples as candidates. We utilize Sentence-BERT (SBERT) (Reimers and Gurevych, 2020)[1] for candidate selection. Based on our experiments detailed in Section C, we set $k = 10$. The top k candidates for q_i are denoted as $D_i = \{d_{i,1}, \cdots, d_{i,k}\}$, where each candidate $d_{i,j}$ is represented as $(x_{i,j}, y_{i,j})$, with $x_{i,j}$ being the input and $y_{i,j}$ the corresponding label.

After obtaining the candidate-query pairs $\{(q_i, d_{i,1}), \cdots, (q_i, d_{i,k})\}$, we conduct 1-shot in-context learning with MaLA500 (Lin et al., 2024b) to predict the class of the q_i given the candidate $d_{i,j}$, resulting in a predicted label $\hat{y}_{i,j}$. If MaLA500 correctly predicts the label of q_i (i.e., $\hat{y}_{i,j} = y_i$), we consider the candidate $d_{i,j}$ as a positive example ($d^+_{i,j}$); otherwise a negative example ($d^-_{i,j}$). Finally, we divide D_i into sets of positive and negative examples, denoted as D_i^{pos} and D_i^{neg}, respectively.

2.3 Retriever Fine-tuning

We utilize the contrastive loss (Rubin et al., 2022; Cheng et al., 2023; Luo et al., 2023) to train the task-specific retriever, aiming to maximize the similarity between q_i and $x_{i,j}$ if $x_{i,j}$ is a positive example while minimizing the similarity if $x_{i,j}$ is a negative example. We opt for Glot500 (Imani et al., 2023) with a model size of 395M as the base model for training the retriever, considering the significant cost of fine-tuning an LLM. We train for 50 epochs with a learning rate of 2e-5. Due to the multilingual nature of Glot500, the fine-tuned retriever can be effectively transferred to retrieve in-context examples for other languages.

2.4 In-Context Learning

At test time, when employing in-context learning across languages, where q_i can be in any language, we use the fine-tuned task-specific retriever to retrieve a few cross-lingual examples in English tai-

[1]We use version distiluse-base-multilingual-cased-v1.

lored to q_i. The retrieved examples are appended to q_i as input for MaLA500 (Lin et al., 2024b) to predict the label of q_i through in-context learning.

3 Experiment

3.1 Setup

Benchmark We evaluate XAMPLER on a massively multilingual text classification benchmark, SIB200 (Adelani et al., 2023). SIB200 involves seven classes: science/technology, travel, politics, sports, health, entertainment, and geography. Our evaluation spans a diverse set of 176 languages, obtained by intersecting the language sets of SIB200 and MaLA500 (see §A). The English training set contains 701 samples, and each language has 204 samples for evaluation.

Our evaluation framework follows the prompt template used in Lin et al. (2024b): 'The topic of the news [sent] is [label]', where [sentence] represents the text for classification and [label] is the ground truth. [label] is included when the sample serves as a few-shot example but is omitted when predicting the sample. We opt for English prompt templates over in-language ones due to the labor-intensive nature of crafting templates for non-English languages, especially those with limited resources. MaLA500 takes the concatenation of few-shot examples and q_i as input, then proceeds to estimate the probability distribution across the label set. We measure the performance with accuracy.

Baselines *Random Sampling.* We randomly select examples from the English candidate pool D.

Off-the-shelf Retriever. We utilize SBERT (Reimers and Gurevych, 2019), a model covering 50+ languages trained with parallel corpora based on mBERT (Devlin et al., 2019). Additionally, we employ two massively multilingual language models, namely Glot500 and MaLA500, as retrievers, denoted as Glot500 RET and MaLA500 RET, respectively. Tailored examples are retrieved based on the cosine similarity between the sentence representations of the candidate and the query. For Glot500, we utilize mean pooling over hidden states of the selected layer. For MaLA500, we adopt a position-weighted mean pooling method on the selected layer, assigning higher weights to later tokens (Muennighoff, 2022). We use K-Nearest Neighbors (KNN) to select the layer that performs best across layers (see §B). The selected layers for Glot500 and MaLA500 are 11 and 21, respectively.

Cross-lingual Transfer. Cross-lingual transfer is another baseline exploiting English data. In this approach, the multilingual language model is fine-tuned with English data and then deployed for evaluation across target languages. Both Glot500 and MaLA500 are included, with their corresponding cross-lingual transfer baselines denoted as Glot500 XLT and MaLA500 XLT. For Glot500, we opt for full-parameter fine-tuning. For MaLA500, which is trained by incorporating LoRA (Hu et al., 2022) into LLaMA 2-7B (Touvron et al., 2023), we only update the LoRA parameters with prompt tuning.

K-Nearest Neighbors. We consider K-Nearest Neighbors (KNN) with the fine-tuned task-specific retriever of XAMPLER and the baselines with retrievers for comparison. Specifically, we adopt majority voting based on the labels of the examples retrieved by the given retriever.

	KNN			ICL		
	Latin	Non-Latin	Avg	Latin	Non-Latin	Avg
Random	-	-	-	56.38	58.66	57.14
SBERT	44.90	37.13	42.29	63.90	62.93	63.57
Glot500 RET	51.16	60.26	54.21	65.21	70.09	66.85
MaLA500 RET	31.90	32.63	32.15	61.02	63.58	61.88
Glot500 XLT	-	-	-	67.09	74.30	69.51
MaLA500 XLT	-	-	-	69.15	71.39	69.90
XAMPLER	**67.05**	**75.97**	**70.04**	**73.59**	**79.80**	**75.67**

Table 1: Average macro-accuracy across all 176 languages, 117 languages in Latin scripts and 59 in non-Latin scripts on SIB200 using XAMPLER and the baselines. KNN results are only provided for retriever-based methods. Results are based on the 3-shot setting.

3.2 Main Results

The comparison between the baselines and XAMPLER is illustrated in Table 1. Our analysis reveals several insights based on the performance with In-Context Learning (ICL) across different methods. Notably, the random baseline exhibits the worst performance among the baselines using ICL, emphasizing the critical role of example selection for effective in-context learning. Leveraging an off-the-shelf retriever notably boosts performance. For instance, Glot500 RET outperforms the random baseline by 9.71%, and it also outperforms MaLA500 RET, showcasing the validity of selecting Glot500 as the base retriever. Leveraging English data further enhances performance, with both Glot500 XLT and MaLA500 XLT surpassing their corresponding RET baselines by 2.66% and 8.02%.

Among all the methods, XAMPLER achieves the highest performance, surpassing the second-best method, MaLA500 XLT, by 5.77%. This highlights the effectiveness of training a task-specific retriever solely with English data. Moreover, XAMPLER outperforms MaLA500 XLT by 4.44% on languages written in Latin scripts and by 8.41% on languages written in non-Latin scripts. It shows that XAMPLER can provide greater benefits for languages with non-Latin scripts, which are relatively isolated from English written in Latin.

3.3 Effect of Task-Specific Retriever

The results from various retrievers utilizing KNN, as shown in Table 1, indicate that XAMPLER's fine-tuned retriever excels at retrieving more examples within the same classes as the query in the target language. Notably, XAMPLER with KNN outperforms the second-best retriever, Glot500 RET, by a notable margin, i.e., 15.83%. This superiority enables XAMPLER to leverage majority label bias in in-context learning (Zhao et al., 2021), thereby enhancing overall performance.

3.4 Effect of In-Context Learning

In Table 1, we compare XAMPLER with KNN against that with ICL. Notably, XAMPLER with ICL surpasses XAMPLER with KNN by 5.63%. Moreover, XAMPLER with ICL demonstrates an average improvement of 6.54% for languages in Latin and 3.83% for languages in non-Latin. This disparity may be attributed to the ability of in-context learning to effectively model queries written in the same script as the provided examples, while facing challenges in handling queries written in different scripts from the few-shot examples.

We further compare XAMPLER's performance with KNN and ICL using varying numbers of retrieved examples, as illustrated in Figure 2. Interestingly, XAMPLER with ICL exhibits inconsistent superiority over KNN, with performance variances ranging from 3% to 10%. Specifically, XAMPLER with KNN achieves its peak performance with 5 examples, whereas ICL achieves impressive results with only 2 examples. Notably, in comparison to KNN's optimal performance, recorded at 73.26% with 5 shots, XAMPLER with ICL demonstrates a notable improvement of 2.58%. These findings underscore the efficacy of applying in-context learning in effectively leveraging the retrieved examples.

4 Related Work

Early studies (Gao et al., 2021; Liu et al., 2022; Rubin et al., 2022) on retrieving informative examples for few-shot in-context learning often rely on off-the-shelf retrievers to gather semantically similar examples to the query.

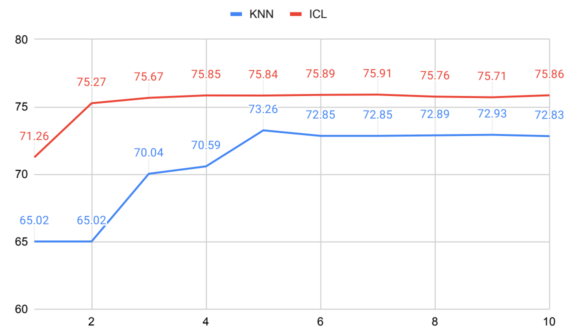

Figure 2: KNN (K-Nearest Neighbors) vs. ICL (In-Context Learning) with different number of shots. X-axis: number of shots. Y-axis: Macro-average accuracy.

While off-the-shelf retrievers have shown promise, the examples they retrieve may not always represent optimal solutions for the given task, potentially resulting in sub-optimal performance. Hence, Rubin et al. (2022) delve into learning-based approaches: if an LLM finds an example useful, the retriever should be encouraged to retrieve it. This approach enables direct training of the retriever using signals derived from query and example pairs in the task of interest.

Several works (Shi et al., 2022; Nie et al., 2022; Winata et al., 2023; Tanwar et al., 2023) extend these methods to non-English languages. A study closely related to ours is Shi et al. (2022), which trains a cross-lingual example retriever via distilling the LLM's scoring function and evaluates it on four languages for the Text-to-SQL Semantic Parsing task. However, our contribution lies in addressing the more challenging low-resource scenario, thereby extending the applicability and robustness of the approach proposed by Shi et al. (2022).

5 Conclusion

In this paper, we introduce XAMPLER, a novel approach designed for cross-lingual example retrieval to facilitate in-context learning in any language. Relying solely on English data, XAMPLER trains a task-specific retriever capable of retrieving cross-lingual English examples tailored to any language query, thereby facilitating few-shot in-context learning for any language. Our experiments on SIB200 across 176 languages show that XAMPLER can outperform previous methods by a notable margin.

Limitations

We did not consider other models and benchmarks due to the absence / unavailability of massively multilingual ones. Additionally, while it is acknowledged that English may not universally serve as the optimal source language for cross-lingual transfer across all target languages (Lin et al., 2019; Wang et al., 2023; Lin et al., 2024a), our study does not explore the selection of different source languages due to the predominant availability of training data in English for many tasks.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2023. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *CoRR*, abs/2309.07445.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12318–12337. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André F. T. Martins, and Hinrich Schütze. 2024a. mplm-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 276–310. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024b. Mala-500: Massive language adaptation of large language models. *CoRR*, abs/2401.13303.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3125–3135. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schütze. 2023. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. *CoRR*, abs/2311.08849.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Seyed Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2023. Dr.icl: Demonstration-retrieved in-context learning. *CoRR*, abs/2305.14128.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *CoRR*, abs/2401.11624.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *CoRR*, abs/2202.08904.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2022. Cross-lingual retrieval augmented prompt for low-resource languages. *CoRR*, abs/2212.09651.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022. XRICL: cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5248–5259. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6292–6307. Association for Computational Linguistics.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *CoRR*, abs/2305.00090.

Genta Indra Winata, Liang-Kang Huang, Soumya Vadlamannati, and Yash Chandarana. 2023. Multilingual few-shot learning via language model retrieval. *CoRR*, abs/2306.10964.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Detailed Results

The language list of SIB200 and the results of XAMPLER and the compared baselines are shown in Table 2 and Table 3.

B KNN Performance Across Layers

We show the 10-shot KNN results across layers with Glot500 and MaLA500 as retrievers in Figure 3 and 4. As shown, layer 21 of MaLA500 and layer 11 of Glot500 achieve the best performance across layers. Therefore, the retrieved results based on these two layers are used in the baselines.

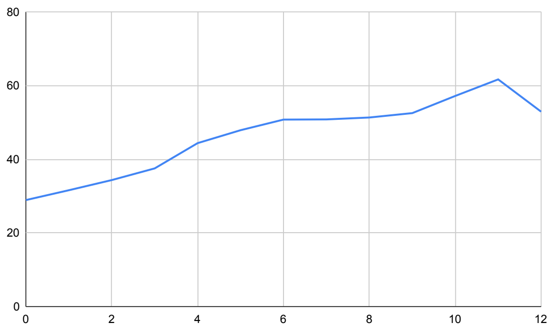

Figure 3: Results of 10-shot KNN (K-Nearest Neighbors) with Glot500 as retriever across layers.

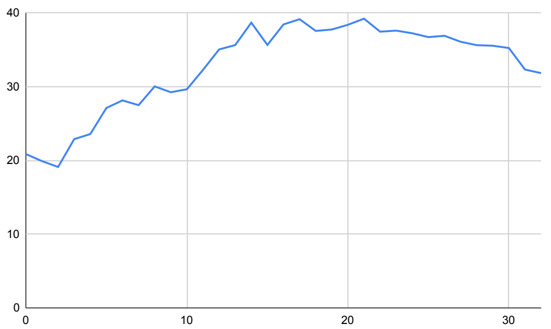

Figure 4: Results of 10-shot KNN (K-Nearest Neighbors) with MaLA500 as retriever across layers.

C Effect of k

We conduct additional experiments to analyze the impact of the parameter k, with the results presented in Figure 5. Our findings indicate that XAMPLER performs optimally when $k = 10$. However, as k exceeds 10, there is a slight decrease in performance. This trend may be attributed to the possibility that increasing k leads to fewer hard negatives for training the retriever.

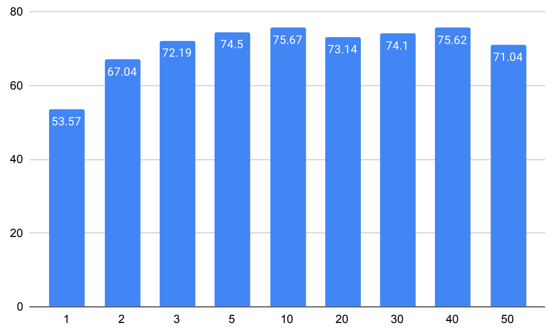

Figure 5: In-context learning with XAMPLER with different k.

	Random	SBERT (KNN)	SBERT	Glot500 RET (KNN)	Glot500 RET	MaLA500 RET (KNN)	MaLA500 RET	Glot500 XLT	MaLA500 XLT	XAMPLER (KNN)	XAMPLER
ace_Latn	61.76	40.20	64.71	50.00	69.12	28.92	69.12	65.12	66.18	75.49	75.98
acm_Arab	55.39	79.90	79.41	64.22	73.04	35.29	62.75	76.81	74.51	76.47	79.90
afr_Latn	65.20	75.00	83.33	60.78	78.92	43.63	72.06	76.10	78.92	82.84	86.27
ajp_Arab	58.82	76.47	79.41	62.25	71.57	32.84	63.24	75.44	72.55	82.84	83.33
als_Latn	64.71	46.57	75.49	63.73	73.53	36.76	67.65	76.76	81.86	65.20	83.33
amh_Ethi	53.92	29.41	58.33	46.57	55.39	22.55	57.35	68.46	61.27	71.57	73.53
apc_Arab	59.80	79.41	79.41	62.25	74.51	31.37	62.25	77.52	75.00	82.84	84.80
arb_Arab	58.33	83.33	81.37	61.76	70.10	36.76	67.16	77.40	78.43	78.43	85.78
ary_Arab	56.37	79.41	78.92	55.39	67.65	25.98	56.37	74.90	67.16	80.39	80.88
arz_Arab	55.39	74.51	74.02	62.25	69.61	33.33	63.73	77.08	71.08	78.92	80.88
asm_Beng	61.27	22.55	61.76	59.80	77.45	22.55	66.67	74.93	76.96	78.92	84.80
ast_Latn	70.59	73.04	81.86	63.73	79.90	49.51	81.37	80.81	84.31	75.49	89.71
ayr_Latn	37.25	24.51	40.69	32.84	41.67	19.12	42.65	49.93	48.53	38.24	50.98
azb_Arab	44.61	26.47	47.55	41.18	59.31	23.04	47.55	65.56	58.82	72.55	73.04
azj_Latn	66.18	58.33	78.92	72.55	76.47	20.10	68.63	78.73	87.25	75.49	85.78
bak_Cyrl	58.33	50.98	74.02	59.80	72.55	28.92	63.73	77.87	75.00	72.55	80.88
bam_Latn	36.76	31.37	42.65	31.86	43.14	19.61	41.18	49.71	44.12	46.08	46.08
ban_Latn	66.18	44.61	70.59	53.92	70.59	31.86	72.06	73.01	72.06	75.98	82.35
bel_Cyrl	64.22	38.24	72.06	59.31	76.96	43.63	73.04	77.16	79.90	69.12	84.31
bem_Latn	42.16	28.43	43.63	32.35	50.00	18.14	39.71	56.89	58.33	60.78	61.76
ben_Beng	63.73	20.10	61.27	60.29	75.00	32.35	69.12	73.95	80.88	79.90	82.35
bjn_Latn	63.73	31.37	62.75	59.31	70.10	33.33	70.59	69.80	69.61	76.96	81.86
bod_Tibt	44.61	09.31	33.33	38.24	48.53	19.61	47.55	62.79	47.06	54.90	59.31
bos_Latn	70.10	55.39	77.45	72.55	79.90	46.08	75.00	81.99	83.33	71.57	87.75
bul_Cyrl	65.69	75.00	82.84	67.16	80.39	52.45	78.92	78.87	81.86	79.90	85.29
cat_Latn	69.61	74.51	84.31	64.22	74.51	61.27	80.88	77.97	86.76	71.57	88.73
ceb_Latn	66.18	45.59	73.53	63.73	80.88	31.37	72.55	78.80	81.86	82.35	85.29
ces_Latn	69.61	53.92	76.96	63.24	76.47	55.88	78.92	77.25	86.27	76.96	88.24
cjk_Latn	37.75	28.43	42.16	26.96	45.10	21.57	44.61	48.41	53.43	44.61	47.06
ckb_Arab	57.84	17.65	61.27	54.90	72.06	23.53	61.76	74.46	75.00	79.90	81.37
cmn_Hani	68.63	85.29	83.82	75.98	80.88	62.75	78.92	78.55	82.84	87.25	88.73
crh_Latn	58.82	67.65	75.00	54.41	70.59	14.71	58.33	67.84	71.08	64.71	75.98
cym_Latn	63.24	24.02	65.69	49.02	73.04	27.94	70.10	72.50	77.45	72.55	81.37
dan_Latn	68.14	47.06	78.92	67.65	79.90	56.37	78.43	79.68	84.80	86.27	89.71
deu_Latn	70.59	83.82	86.27	70.10	77.45	57.84	80.88	78.48	87.25	80.88	87.75
dyu_Latn	41.67	28.92	45.59	27.94	50.49	24.51	46.57	45.17	46.57	43.14	48.53
dzo_Tibt	31.37	09.31	20.59	47.55	41.18	16.67	32.84	62.21	33.33	68.63	52.94
ell_Grek	67.65	31.37	76.47	59.80	72.55	40.20	72.06	74.71	79.90	75.98	83.82
eng_Latn	70.59	84.31	87.25	79.41	86.27	69.61	81.86	83.65	85.29	89.22	91.18
epo_Latn	65.69	57.84	75.98	61.27	76.96	33.33	68.63	75.20	77.94	78.92	82.84
est_Latn	61.76	44.61	66.67	65.69	75.49	30.88	66.18	74.07	76.47	67.65	81.37
eus_Latn	57.35	47.55	71.08	66.67	76.47	29.90	68.14	76.08	70.10	66.67	82.84
ewe_Latn	38.24	26.47	39.22	27.94	39.71	20.59	38.73	47.65	46.57	50.98	53.92
fao_Latn	56.86	31.86	58.33	45.10	63.24	29.90	60.78	76.15	69.61	82.35	81.86
fij_Latn	41.18	32.35	49.02	31.86	46.08	21.08	42.65	55.74	51.47	50.00	55.39
fin_Latn	65.20	36.76	71.57	62.75	74.02	48.53	74.51	75.98	80.88	76.96	83.82
fon_Latn	39.22	20.10	41.67	33.33	45.59	20.10	42.65	44.41	45.59	41.67	48.53
fra_Latn	69.61	86.27	84.31	60.78	79.41	63.24	79.90	81.20	85.29	85.78	89.22
ful_Latn	37.75	35.29	48.04	23.53	48.04	21.57	45.59	45.54	45.10	47.06	50.49
fur_Latn	60.29	63.73	72.55	54.90	67.16	36.76	70.59	67.75	73.53	75.49	79.90
gla_Latn	54.90	17.16	52.45	39.22	59.80	25.00	61.76	59.00	60.29	59.31	63.73
gle_Latn	57.84	23.53	57.84	49.02	65.20	31.86	62.75	62.94	69.12	70.10	74.02
glg_Latn	72.06	76.96	81.86	67.65	80.88	53.43	79.90	79.36	84.31	81.37	87.75
grn_Latn	53.92	62.75	69.61	43.14	61.76	27.45	60.78	69.85	64.71	76.47	75.00
guj_Gujr	59.80	13.73	53.43	68.14	73.04	32.84	65.20	77.87	73.53	81.86	86.27
hat_Latn	61.76	50.49	72.06	59.31	74.02	25.49	68.14	73.55	81.86	79.41	82.84
hau_Latn	56.86	23.04	55.88	47.06	60.78	23.53	57.35	62.33	70.10	58.82	70.10
heb_Hebr	45.59	26.96	47.06	58.33	58.33	29.90	43.63	72.60	56.86	75.00	75.00
hin_Deva	60.29	12.75	56.86	67.65	71.08	34.80	65.20	76.84	76.47	84.31	85.78
hne_Deva	56.86	15.69	54.41	55.88	70.10	28.43	66.67	70.07	75.49	79.41	80.88
hrv_Latn	70.10	52.94	78.43	73.53	82.35	51.47	74.51	81.86	85.78	71.57	88.24
hun_Latn	64.71	38.24	66.18	69.61	81.86	50.49	72.06	81.20	84.80	83.82	85.78
hye_Armn	67.16	16.18	67.16	60.78	70.59	31.37	68.14	76.62	77.45	80.88	83.33
ibo_Latn	57.35	37.25	63.24	49.02	70.10	23.53	56.86	69.09	72.06	74.02	79.41
ilo_Latn	61.76	50.49	69.12	47.55	68.14	32.35	65.69	70.81	77.45	75.00	79.90
ind_Latn	69.61	64.22	82.84	75.98	81.37	50.98	75.98	80.34	83.82	88.24	89.71
isl_Latn	60.78	31.86	67.65	55.39	69.61	24.51	62.75	73.04	75.49	77.94	78.92
ita_Latn	71.08	82.84	88.24	68.14	81.37	64.22	77.94	78.80	86.76	79.90	88.24
jav_Latn	64.22	40.69	71.08	56.37	75.00	30.88	70.10	73.19	75.00	79.90	82.35
jpn_Jpan	71.08	76.96	87.25	68.14	76.96	65.69	82.84	78.33	80.88	83.33	87.75
kab_Latn	28.43	20.10	27.94	27.45	36.27	22.06	29.41	38.70	27.94	31.37	37.75
kac_Latn	35.29	28.92	41.18	40.20	42.65	20.10	36.76	54.09	45.10	40.69	46.57
kam_Latn	38.73	30.88	47.55	34.31	46.08	17.65	34.31	46.99	50.49	46.08	50.00
kan_Knda	58.82	15.69	56.37	64.22	71.57	24.02	59.80	74.73	73.04	75.00	77.94
kat_Geor	65.20	15.69	58.82	63.73	75.98	27.94	70.10	77.79	75.98	72.55	83.33
kaz_Cyrl	61.76	43.14	68.63	69.12	74.51	31.37	64.71	75.47	78.43	67.16	79.41
kbp_Latn	37.25	25.00	43.14	34.31	44.61	17.65	44.12	49.83	47.06	41.18	48.04
kea_Latn	64.71	61.27	78.92	48.53	75.49	33.82	75.98	66.81	80.39	74.51	79.41
khm_Khmr	69.12	34.31	69.61	58.82	76.47	38.24	74.51	76.42	79.41	82.84	86.76
kik_Latn	46.57	33.33	51.47	43.63	50.98	19.12	44.12	55.47	52.94	58.82	59.31
kin_Latn	45.59	26.96	50.00	39.22	51.47	19.61	48.53	57.01	67.65	64.71	64.22
kir_Cyrl	58.82	42.65	67.65	68.63	75.98	24.51	55.39	75.93	75.00	63.73	77.45
kmb_Latn	36.76	24.51	41.67	28.43	42.65	24.02	42.65	44.80	48.04	46.57	50.98
kmr_Latn	51.96	27.45	58.82	50.00	69.12	20.10	56.37	63.55	66.18	52.94	70.59
kon_Latn	49.02	50.49	63.24	44.61	63.73	20.10	54.41	58.85	61.27	57.84	64.22
kor_Hang	69.61	81.37	85.29	62.25	78.43	48.53	78.43	77.33	79.90	82.35	87.25
lao_Laoo	64.71	33.33	69.12	58.33	74.02	30.39	70.59	77.33	75.00	79.90	84.31
lij_Latn	64.22	61.76	77.45	50.49	69.61	35.29	66.67	70.07	76.96	70.59	80.39
lim_Latn	61.76	68.63	79.41	55.88	72.55	38.24	72.55	69.49	65.20	71.08	75.98
lin_Latn	48.04	46.57	61.27	48.53	61.27	20.59	54.90	64.95	58.33	63.24	68.63

Table 2: 3-shot accuracy with KNN or ICL (Part I). Methods with '(KNN)' denote that KNN is used with retrieved samples, otherwise ICL is used.

	Random	SBERT (KNN)	SBERT	Glot500 RET (KNN)	Glot500 RET	MaLA500 RET (KNN)	MaLA500 RET	Glot500 XLT	MaLA500 XLT	XAMPLER (KNN)	XAMPLER
lit_Latn	61.27	41.18	64.71	62.25	72.55	29.41	63.24	78.04	79.41	69.61	85.29
lmo_Latn	63.24	57.35	73.04	48.53	66.67	33.33	66.18	66.25	73.53	72.55	77.45
ltz_Latn	62.75	60.78	76.96	56.37	72.55	35.29	70.10	69.26	80.39	72.06	79.41
lua_Latn	38.24	42.65	50.49	36.27	45.10	24.51	47.06	55.51	52.45	47.55	53.43
lug_Latn	38.24	29.41	41.67	33.82	50.00	20.10	40.69	53.28	51.47	52.45	58.33
luo_Latn	37.75	26.47	40.69	31.86	43.63	18.63	44.12	51.03	50.00	55.39	59.31
lus_Latn	48.53	41.18	57.84	39.71	50.98	30.39	51.96	60.83	61.76	65.69	68.14
lvs_Latn	64.22	39.22	68.63	62.25	73.53	31.86	64.22	78.87	79.90	71.08	82.84
mai_Deva	57.35	17.16	50.49	62.25	69.61	28.92	59.31	76.91	68.63	83.82	78.92
mal_Mlym	57.84	15.69	52.45	58.33	67.16	26.47	59.80	73.33	68.63	77.94	78.92
mar_Deva	59.31	15.69	55.88	60.78	75.00	27.94	63.73	75.96	77.94	73.04	79.41
min_Latn	67.65	44.61	66.67	58.82	72.55	32.84	74.02	70.42	78.43	78.43	79.90
mkd_Cyrl	68.63	57.84	77.94	67.16	75.98	49.02	75.49	76.50	80.88	72.55	83.82
mlt_Latn	68.14	50.00	72.55	60.29	77.45	31.86	67.65	77.87	81.37	78.92	84.31
mon_Cyrl	58.82	21.08	59.31	63.73	73.53	25.00	62.25	75.12	71.57	79.41	82.84
mos_Latn	37.75	39.22	46.08	23.04	39.22	21.08	42.65	47.06	45.10	44.61	50.49
mri_Latn	52.94	22.55	54.41	18.63	46.08	17.16	46.08	54.49	60.78	59.80	68.63
mya_Mymr	53.92	16.67	49.02	53.92	66.18	23.04	50.98	73.68	54.90	74.02	80.39
nld_Latn	68.63	82.84	85.29	75.00	81.37	56.37	81.37	78.24	86.27	87.25	88.73
nno_Latn	68.14	46.57	73.53	64.22	75.49	42.65	74.51	78.63	83.82	86.27	89.22
npi_Deva	63.24	15.20	59.31	67.16	74.51	29.41	70.10	77.99	75.49	82.35	87.75
nso_Latn	41.67	30.39	52.45	39.22	50.49	21.57	43.63	60.71	55.39	54.41	58.82
nya_Latn	45.10	35.29	52.94	50.49	61.76	23.04	49.02	68.36	58.82	66.67	68.14
oci_Latn	67.65	69.61	81.86	58.33	70.10	44.12	77.45	76.50	84.31	74.02	85.78
orm_Latn	34.80	13.73	34.80	40.69	42.16	15.69	35.29	50.59	46.08	42.65	48.53
ory_Orya	51.47	29.41	47.06	59.31	67.65	23.53	57.35	73.55	62.25	80.88	78.92
pag_Latn	55.88	63.24	72.06	59.31	71.57	26.47	62.25	73.31	75.98	80.39	80.88
pan_Guru	57.35	14.22	53.92	51.96	67.16	26.47	60.78	70.44	73.04	71.08	75.98
pap_Latn	65.20	63.73	75.49	58.82	76.96	34.31	70.10	72.23	78.43	77.94	80.39
pes_Arab	66.67	37.25	67.65	69.61	76.96	34.31	71.08	77.72	79.41	87.75	86.76
plt_Latn	52.94	29.41	57.35	42.65	60.29	21.08	50.49	64.22	69.61	48.04	67.16
pol_Latn	68.14	78.43	87.25	67.16	78.92	57.84	78.43	77.97	84.80	75.00	85.29
por_Latn	72.06	81.86	87.75	60.78	81.37	59.31	83.33	79.71	85.78	85.78	90.69
prs_Arab	65.20	32.84	64.22	62.75	76.96	31.86	71.57	78.70	78.92	86.76	86.27
pus_Arab	49.51	25.98	52.94	53.43	59.80	23.53	50.98	69.34	62.25	60.78	71.08
quy_Latn	49.02	29.41	52.45	42.65	52.94	19.61	53.92	59.31	57.84	56.37	62.25
ron_Latn	67.65	58.33	77.94	66.18	75.98	55.39	76.96	78.14	83.33	80.88	86.76
run_Latn	43.14	22.06	44.61	41.18	49.51	18.14	45.59	51.86	60.78	62.75	66.18
rus_Cyrl	69.61	83.82	86.27	72.06	79.90	56.86	77.45	79.53	81.37	86.27	89.71
sag_Latn	43.14	39.71	49.02	39.71	54.41	20.59	45.59	52.50	49.02	56.37	59.31
san_Deva	51.96	15.20	52.94	48.04	59.80	23.53	59.31	70.34	60.29	66.67	68.63
scn_Latn	68.14	55.39	77.45	49.51	71.57	34.31	73.53	67.50	74.51	72.06	80.39
sin_Sinh	61.27	22.06	62.25	66.18	73.53	26.47	65.20	74.29	72.55	79.90	82.84
slk_Latn	66.67	49.51	70.59	63.73	74.02	45.10	75.98	77.99	81.86	74.02	84.31
slv_Latn	63.73	49.02	74.51	64.22	74.02	51.96	75.00	75.51	78.92	66.67	81.86
smo_Latn	54.41	32.84	63.24	36.76	62.75	18.63	52.94	69.14	63.73	66.18	73.04
sna_Latn	45.10	24.51	47.06	42.16	51.47	21.57	40.20	59.31	60.29	55.39	62.75
snd_Arab	44.12	29.90	49.51	50.49	52.45	20.10	44.61	65.83	47.55	59.80	67.16
som_Latn	42.65	20.59	42.65	49.02	59.80	17.16	42.16	56.52	63.24	50.00	58.82
sot_Latn	45.10	24.51	49.02	43.63	58.33	21.57	48.04	63.63	57.84	64.71	67.65
spa_Latn	71.57	84.80	85.78	67.16	78.92	59.80	80.39	79.14	85.29	82.84	90.20
srd_Latn	63.73	64.22	75.98	49.51	70.59	34.80	71.08	66.10	66.67	61.76	78.43
srp_Cyrl	68.63	55.88	79.41	65.69	82.84	47.55	79.41	78.53	82.35	67.65	80.88
ssw_Latn	45.10	24.51	44.12	40.69	56.86	26.96	51.96	63.19	59.80	59.80	63.24
sun_Latn	68.14	43.14	70.10	63.24	78.43	29.90	73.53	75.93	81.37	82.84	85.29
swe_Latn	66.18	46.08	72.55	71.57	77.94	59.31	77.94	79.73	85.29	84.80	87.75
swh_Latn	61.27	28.92	60.29	56.86	66.67	12.75	54.90	72.94	76.96	64.22	75.98
szl_Latn	62.75	67.65	75.98	54.41	70.10	40.69	71.57	68.65	75.49	63.73	77.45
tam_Taml	54.90	15.20	50.98	64.22	71.08	27.45	62.25	75.37	75.98	80.88	78.92
tat_Cyrl	61.76	46.57	71.57	63.73	75.00	24.51	68.63	79.17	78.92	70.59	82.35
tel_Telu	57.84	18.63	50.98	65.69	70.59	25.98	60.29	74.53	74.51	84.80	81.37
tgk_Cyrl	60.29	26.96	61.27	63.73	73.53	35.29	63.73	77.08	80.88	82.35	80.39
tgl_Latn	68.14	38.73	69.61	63.24	79.41	30.88	70.10	78.21	83.82	83.33	86.76
tha_Thai	65.20	25.98	56.86	65.20	75.49	41.18	67.65	77.43	70.10	79.90	86.76
tir_Ethi	44.12	29.41	51.47	42.16	48.53	18.63	47.06	55.91	50.00	55.39	57.35
tpi_Latn	65.20	53.92	75.98	57.84	70.59	27.94	64.71	76.94	74.51	81.37	84.80
tsn_Latn	49.02	31.37	49.02	35.78	50.49	20.10	42.16	56.76	56.86	54.90	59.80
tso_Latn	45.59	25.98	44.12	38.73	51.47	22.06	50.98	56.23	55.39	55.88	62.25
tuk_Latn	53.43	50.00	68.63	51.47	69.61	14.22	50.98	73.33	66.67	61.27	77.94
tum_Latn	42.16	26.96	47.06	32.35	56.86	24.02	43.63	62.55	52.94	65.20	64.71
tur_Latn	68.14	82.84	86.27	68.63	76.96	15.69	56.86	78.06	81.37	73.53	84.80
uig_Arab	45.10	12.75	40.20	55.39	53.43	18.63	37.25	72.18	49.51	72.55	69.12
ukr_Cyrl	66.18	59.31	77.45	67.16	79.90	56.86	79.41	78.63	81.86	76.47	86.27
umb_Latn	33.82	29.90	37.75	33.82	46.57	22.06	39.71	48.97	49.02	47.06	49.51
urd_Arab	55.88	29.90	54.90	59.80	69.12	36.27	66.67	74.29	69.12	64.22	75.98
uzb_Latn	57.35	38.24	62.25	59.80	73.04	24.02	62.75	73.14	76.96	60.78	77.45
vec_Latn	67.65	68.14	82.84	57.84	73.53	36.76	77.45	72.72	79.41	75.00	81.86
vie_Latn	72.06	28.92	65.69	67.65	78.92	56.86	81.86	79.63	77.45	79.90	87.25
war_Latn	64.22	50.00	76.96	58.82	75.98	28.92	71.57	75.76	81.86	83.82	84.80
wol_Latn	45.59	42.16	55.88	29.90	50.98	22.06	48.53	51.03	50.00	47.55	57.84
xho_Latn	49.51	27.45	52.94	43.14	61.27	23.04	50.98	62.79	63.73	54.90	67.16
yid_Hebr	40.20	20.59	44.12	39.22	50.49	24.02	46.57	55.91	51.96	59.31	62.75
yor_Latn	45.10	31.86	46.57	29.41	45.10	23.04	44.12	49.22	49.02	57.35	58.82
yue_Hani	69.61	82.35	84.80	73.53	83.33	60.78	81.86	79.66	82.35	85.78	85.29
zsm_Latn	66.18	58.82	82.35	75.49	78.43	48.04	76.96	80.74	82.35	85.78	89.71
zul_Latn	55.88	24.51	49.51	50.98	66.18	28.43	51.96	69.02	65.69	69.61	74.02
Avg	57.14	42.29	63.57	54.21	66.85	32.15	61.88	69.51	69.90	70.04	75.67

Table 3: 3-shot accuracy with KNN or ICL (Part II). Methods with '(KNN)' denote that KNN is used with retrieved samples, otherwise ICL is used.