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

Peiqin Lin[1,2], **André F. T. Martins**[3,4,5], **Hinrich Schütze**[1,2]
[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning
[3]Instituto Superior Técnico, Universidade de Lisboa (Lisbon ELLIS Unit)
[4]Instituto de Telecomunicações [5]Unbabel
linpq@cis.lmu.de

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 for 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 two multilingual text classification benchmarks, namely SIB200 with 176 languages and MasakhaNEWS with 16 languages, demonstrate that XAMPLER substantially improves the in-context learning performance across languages. Our code is available at https://github.com/cisnlp/XAMPLER.

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 (whether an example is helpful) using labeled data.

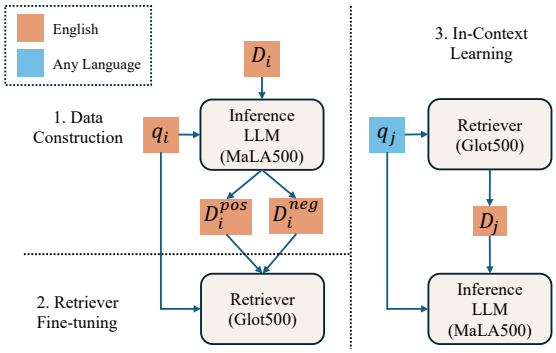

Figure 1: XAMPLER involves three steps: 1. Data Construction: given a query in English q_i, we divide the candidate English examples D_i into positive examples D_i^{pos} and negative examples D_i^{neg} based on the prediction of MaLA500 (Lin et al., 2024b); 2. Retriever Fine-tuning: we fine-tune the 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 D_j as few-shots for in-context learning. **For training, XAMPLER requires English data only. Once trained, the model can be applied to any of the 500 languages covered by MaLA500/Glot500 without any need for (often unavailable) labeled low-resource data.**

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.,

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 3968–3977

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 and 16 languages on MasakhaNEWS 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 B, 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 using the AdamW optimizer with a learning rate of 2e-5 and a batch size of 16. Due to the multilingual nature of Glot500, the fine-tuned retriever

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

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 tailored 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 two text classification benchmarks: SIB200 (Adelani et al., 2023a) and MasakhaNEWS (Adelani et al., 2023b). The partitioning for these datasets was predefined by the respective benchmarks. SIB200 is a massively multilingual text classification benchmark with seven classes. Our evaluation spans a diverse set of 176 languages, obtained by intersecting the language sets of SIB200 and MaLA500 (see §C). The English training set contains 701 samples, with 204 evaluation samples per language. MasakhaNEWS is a news classification task for 16 African languages, covering six topics. The English training set contains 3.31k samples, with 175 to 948 evaluation samples per language.

Our evaluation framework follows the prompt template used in Lin et al. (2024b): 'The topic of the news [sentence] 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 We compare XAMPLER with the following retrieving strategies:

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

Multilingual Language Models. We use two massively multilingual language models, Glot500 and MaLA500, as retrievers. 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 to select the layer that performs best across layers (see §A). The selected layers for Glot500 and MaLA500 are 11 and 21, respectively.

Off-the-shelf Retriever. We also employ three off-the-shelf retrievers trained on parallel data: SBERT (Reimers and Gurevych, 2019), LaBSE (Feng et al., 2022), and Multilingual E5 (Wang et al., 2024).[2]

We set the number of shots as the number of classes and evaluate under two settings: the label-aware setting, where one shot is provided per class, and the label-agnostic setting, where the most similar examples are retrieved regardless of their labels.

	SIB200		MasakhaNEWS	
	Label-Aware	Label-Agnostic	Label-Aware	Label-Agnostic
Random	65.24	61.68	72.32	72.39
Glot500	66.60	68.55	73.35	73.01
MaLA500	66.75	66.25	73.39	71.58
SBERT	67.13	66.59	73.24	72.8
LaBSE	68.51	73.69	72.54	73.29
Multilingual E5	69.09	74.61	73.63	72.61
XAMPLER	**70.18**	**75.91**	**75.02**	**73.85**

Table 1: Average macro-accuracy across the evaluated languages on SIB200 and MasakhaNEWS using XAMPLER compared to the baselines.

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. Multilingual language models, which retrieve examples based on semantic similarity learned during pretraining, slightly outperform the random baseline. Leveraging an off-the-shelf retriever further improves performance, with Multilingual E5 emerging as the top performer among them. Among all the methods, XAMPLER achieves the highest performance in in-context learning. Specifically, on SIB200, XAMPLER surpasses Multilingual E5 by 1.09% in the label-aware setting and 1.30% in the

[2] https://huggingface.co/intfloat/multilingual-e5-large

label-agnostic setting, and by 1.93% and 1.24% on MasakhaNEWS, respectively.

Our two established practices to reduce resource requirements, i.e., selecting the top-k similar examples as candidates using existing off-the-shelf retrievers and using a smaller base model for retriever training, ensures that XAMPLER operates efficiently. For example, on the SIB200 benchmark (701 samples, 10 candidates per sample) using an NVIDIA GeForce GTX 1080 Ti (11GB), retriever fine-tuning took just 1.5 hours, and retrieving similar examples added only 0.1 seconds per query during inference.

3.3 Ablation Study

To further assess the effectiveness of XAMPLER, we conduct the following ablation studies:

Method	Accuracy (%)
XLT (Glot500)	69.51
XLT (MaLA500)	69.90
MT	74.50
KNN	72.85
XAMPLER	**75.91**

Table 2: Performance comparison of XAMPLER with ablation methods on the SIB200 benchmark.

Cross-lingual Transfer (XLT). Cross-lingual transfer is another approach that utilizes English data. In this method, the multilingual language model is fine-tuned on English data and then evaluated across target languages. Both Glot500 and MaLA500 are included, with their respective cross-lingual transfer methods denoted as XLT (Glot500) and XLT (MaLA500). For Glot500, we perform full-parameter fine-tuning with a batch size of 16 and a learning rate of 1e-5. For MaLA500, which is trained by incorporating LoRA (Hu et al., 2022) into LLaMA 2-7B (Touvron et al., 2023), we update only the LoRA parameters with prompt tuning. The learning rate is set to 1e-3, weight decay is 0.1, the maximum sequence length is 128, and the batch size is 16. The optimizer used is AdamW.

Machine Translation (MT). We translate the examples retrieved by XAMPLER from English to the target language and use these translated examples as few-shot examples for in-context learning. For translation, we use the distilled 600M variant of NLLB-200 (Costa-jussà et al., 2022).

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

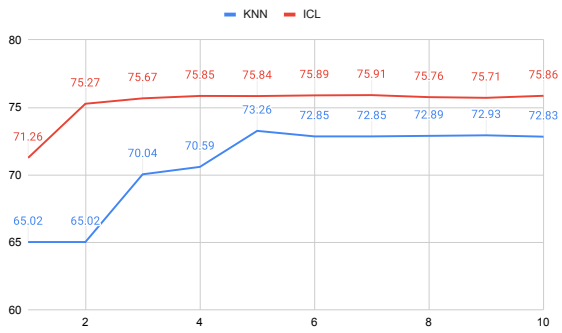

Figure 2: KNN (K-Nearest Neighbors) vs. ICL (In-Context Learning) with different number of shots. X-axis: number of shots. Y-axis: Macro-average accuracy.

The results of the ablation studies are shown in Table 2. XAMPLER outperforms XLT (Glot500) and XLT (MaLA500) by 6.40% and 6.01%, respectively, demonstrating that in-context learning, when combined with informative cross-lingual few-shot examples, is superior to traditional cross-lingual transfer methods via fine-tuning. Translating English examples into target languages results in slightly worse results than XAMPLER by 1.41%. It demonstrates that XAMPLER benefits more from English in-context examples than in-language examples. A comparison between XAMPLER and KNN shows that XAMPLER performs better by 3.06%. We also compared 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.

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 (Winata et al., 2021; Shi et al., 2022; Winata et al., 2022; Nie et al., 2022; Winata et al., 2023; Tanwar et al., 2023; Cahyawijaya et al., 2024) 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. Experiments on SIB200 and MasakhaNEWS show that XAMPLER outperforms 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.

Acknowledgements

This work was funded by DFG (SCHU 2246/14-1), the European Research Council (DECOLLAGE, ERC-2022-CoG #101088763), EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research, and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

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

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba O. Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Sabah Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo Olufunke Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Afolabi Abeeb, Aremu Anuoluwapo, Samuel Olanrewaju, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede Guge, Sakayo Toadoum Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Patrick Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos G. Nugussie, Abdulmejid Tuni Johar, Shafie Abdi Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023b. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 144–159. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 405–433. Association for Computational Linguistics.

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.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. 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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

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.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *CoRR*, abs/2109.07684.

Genta Indra 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, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 777–791. Association for Computational Linguistics.

A 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.

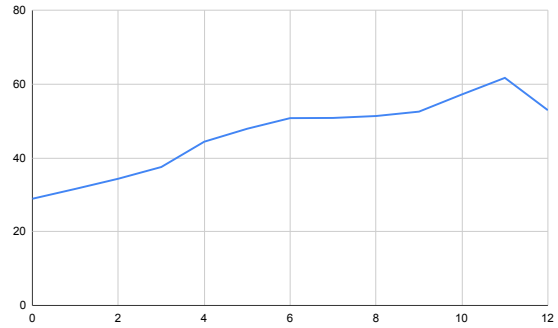

Figure 3: Results of 10-shot KNN (K-Nearest Neighbors) with Glot500 as retriever across layers.

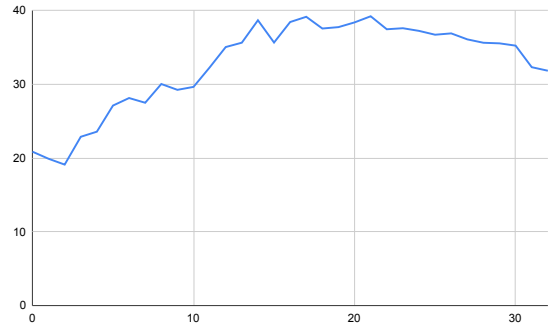

Figure 4: Results of 10-shot KNN (K-Nearest Neighbors) with MaLA500 as retriever across layers.

B 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.

C Detailed Results

The language list of SIB200 and the results of XAMPLER and the compared baselines are shown in Table 3 and Table 4. The language list of

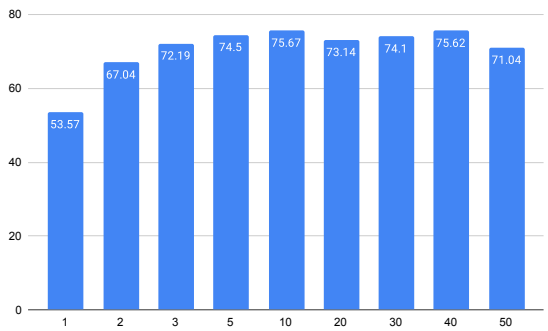

Figure 5: In-context learning with XAMPLER with different k.

MasakhaNEWS and the results of XAMPLER and the compared baselines are shown in Table 5.

	Label-Aware							Label-Agnostic						
	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER
ace_Latn	72.55	75.00	72.06	74.51	72.55	74.51	72.06	63.73	71.57	70.10	70.10	76.47	80.39	76.96
acm_Arab	62.25	64.71	65.20	70.10	66.18	68.63	75.00	61.76	71.08	65.69	78.92	75.98	75.98	79.90
afr_Latn	77.45	76.47	77.94	79.90	80.39	81.86	79.41	70.59	79.41	78.43	83.82	88.24	81.86	87.25
ajp_Arab	64.71	67.65	67.16	68.14	69.61	68.14	74.51	64.71	72.06	64.71	77.94	76.47	77.94	84.31
als_Latn	74.51	74.51	77.94	78.92	77.45	79.41	75.00	68.63	75.98	75.00	80.88	83.82	82.84	84.31
amh_Ethi	56.37	59.31	60.29	60.29	61.76	60.78	68.63	57.35	61.27	60.78	60.29	67.16	67.16	71.08
apc_Arab	63.73	67.65	65.20	68.63	68.14	71.08	75.00	61.76	72.06	62.25	78.43	79.90	76.47	85.78
arb_Arab	65.69	69.61	68.63	69.61	72.55	72.06	77.45	65.20	74.02	70.10	76.96	81.37	75.98	83.82
ary_Arab	58.82	65.20	63.24	65.69	64.71	66.18	74.51	57.84	71.57	60.78	71.57	71.57	73.53	80.39
arz_Arab	63.24	66.18	64.71	66.18	67.65	65.20	73.53	62.25	69.61	61.76	76.47	78.92	76.96	82.84
asm_Beng	70.59	72.06	73.04	72.06	75.98	75.49	73.53	69.12	76.47	72.55	64.22	78.92	76.96	83.82
ast_Latn	79.90	76.47	81.37	79.90	79.41	80.39	82.84	74.02	78.43	81.37	85.78	84.31	82.84	89.71
ayr_Latn	39.71	45.59	43.14	46.08	44.61	45.59	42.16	44.12	47.55	47.06	46.08	46.57	53.43	52.45
azb_Arab	47.55	50.49	50.98	50.98	52.45	54.41	64.71	48.04	61.27	49.51	46.57	60.29	61.76	70.59
azj_Latn	77.45	76.47	77.45	77.94	78.92	82.35	77.94	71.57	79.41	76.47	78.92	82.84	83.82	84.80
bak_Cyrl	69.12	75.49	71.57	72.55	72.06	74.51	75.00	66.18	73.53	69.61	77.94	75.00	78.92	80.88
bam_Latn	44.12	46.08	48.04	47.06	49.51	50.49	52.94	43.14	47.55	44.61	46.08	50.00	57.35	49.02
ban_Latn	75.00	74.51	76.47	77.94	78.92	78.43	79.90	69.61	73.53	78.43	75.98	83.33	84.31	82.35
bel_Cyrl	76.96	76.96	75.00	78.43	77.94	78.92	76.47	71.08	77.94	74.02	76.47	84.80	81.86	85.78
bem_Latn	50.49	52.94	52.45	49.51	52.45	54.41	53.43	47.55	58.82	50.98	50.49	54.41	65.69	62.25
ben_Beng	72.06	72.06	67.65	69.61	73.53	72.55	79.41	65.20	75.49	68.14	66.18	77.94	77.94	81.86
bjn_Latn	71.57	73.04	72.55	72.06	73.53	73.53	76.47	71.08	73.04	74.51	70.10	80.39	79.90	83.33
bod_Tibt	53.92	50.49	49.51	50.98	56.86	50.98	50.98	48.53	51.96	51.96	36.76	56.37	50.98	59.31
bos_Latn	78.43	78.43	75.98	78.92	81.37	81.86	79.90	72.06	82.35	77.94	80.39	82.84	82.84	89.22
bul_Cyrl	77.94	78.92	77.45	79.41	78.92	77.94	78.92	72.06	78.92	78.43	81.86	85.78	80.88	85.78
cat_Latn	78.43	77.45	80.88	83.82	80.39	83.33	81.86	74.02	81.86	79.41	84.31	86.76	84.31	88.24
ceb_Latn	75.98	75.98	76.96	76.47	78.92	76.47	79.90	71.08	75.98	76.96	74.51	84.80	81.37	86.27
ces_Latn	75.49	76.96	77.94	76.96	78.43	79.41	77.94	69.61	77.45	76.96	78.92	85.78	82.35	88.73
cjk_Latn	44.12	44.12	44.12	44.12	47.06	46.57	43.14	40.69	46.08	49.02	44.61	50.00	55.88	47.06
ckb_Arab	69.12	71.08	66.67	69.61	69.12	72.06	75.00	63.24	73.53	67.16	66.18	63.24	75.98	81.86
cmn_Hani	76.96	77.45	81.86	79.90	79.41	79.90	86.27	69.12	81.37	77.94	82.84	79.90	80.88	89.71
crh_Latn	71.08	69.12	67.65	72.06	74.02	72.55	72.55	62.25	74.02	69.61	73.04	75.98	75.49	75.00
cym_Latn	77.45	76.96	75.98	75.98	80.39	78.43	77.94	71.57	78.43	79.90	70.59	85.78	84.31	78.92
dan_Latn	81.37	82.84	80.88	82.35	79.90	83.82	84.80	75.49	82.35	82.84	82.35	86.76	82.84	89.71
deu_Latn	80.39	80.88	83.82	83.82	82.35	83.33	83.33	74.51	78.43	84.31	85.78	86.27	85.29	86.76
dyu_Latn	51.96	51.96	50.98	50.98	51.47	56.37	49.02	47.55	49.02	49.51	48.04	54.41	57.84	50.49
dzo_Tibt	31.86	39.22	35.78	33.33	40.69	37.75	50.00	34.31	43.14	37.25	22.06	48.53	34.80	54.90
ell_Grek	74.02	75.00	78.92	77.45	77.45	79.41	76.47	72.06	76.47	74.02	75.98	78.92	80.39	83.82
eng_Latn	82.84	84.31	85.29	85.78	84.31	84.80	86.27	73.04	84.80	85.78	86.27	85.29	87.75	91.67
epo_Latn	73.53	76.96	75.00	74.51	76.96	77.45	79.90	71.08	78.43	77.94	76.96	86.27	84.80	82.84
est_Latn	68.63	73.04	71.08	71.57	74.02	75.49	75.00	67.16	75.98	71.57	72.55	79.90	80.39	79.90
eus_Latn	59.80	70.10	69.12	69.61	69.61	73.04	75.49	63.73	75.49	68.63	69.61	79.90	86.27	83.82
ewe_Latn	42.65	45.59	44.61	47.55	48.04	47.06	48.04	41.67	42.16	43.14	47.55	49.51	54.90	53.43
fao_Latn	62.75	66.18	68.14	64.71	69.61	69.61	73.53	59.31	70.10	63.24	65.69	77.94	77.45	83.82
fij_Latn	43.14	42.65	46.57	45.10	45.59	48.04	50.49	44.12	49.51	47.55	47.55	54.41	60.29	53.43
fin_Latn	75.49	75.00	75.49	74.51	75.98	77.94	77.94	67.16	78.92	74.51	74.51	82.84	80.39	83.33
fon_Latn	43.14	46.57	46.57	46.57	50.98	52.45	42.16	41.18	48.53	42.65	46.08	50.49	57.35	48.04
fra_Latn	78.92	78.43	78.43	80.88	81.37	81.86	83.82	72.06	79.41	81.86	80.39	86.76	82.84	89.22
ful_Latn	45.59	47.06	47.55	50.49	52.94	50.98	43.63	44.61	46.08	50.00	51.96	57.35	55.88	52.94
fur_Latn	69.61	68.63	71.08	69.61	70.59	73.04	75.49	63.73	68.14	70.59	75.98	74.51	80.88	79.41
gla_Latn	63.24	57.84	64.71	65.20	68.63	68.14	62.75	60.29	63.24	68.63	59.80	75.49	74.51	67.16
gle_Latn	65.20	68.14	71.08	71.08	71.08	72.55	66.18	66.18	72.55	67.65	64.71	80.88	81.37	74.02
glg_Latn	79.41	80.88	79.41	79.90	77.94	81.37	84.80	71.57	83.33	81.86	86.76	84.80	87.75	86.76
grn_Latn	63.73	64.22	64.22	67.16	65.69	69.12	70.10	61.76	67.16	66.18	69.61	71.08	75.00	77.94
guj_Gujr	73.53	73.04	70.10	69.61	77.45	73.04	79.90	69.61	73.04	67.65	62.25	77.45	78.92	85.29
hat_Latn	70.10	72.55	75.00	73.53	75.00	76.96	76.47	65.20	77.45	73.53	73.04	82.84	81.37	82.35
hau_Latn	68.14	65.69	67.65	63.73	67.16	68.14	69.12	58.82	68.63	66.67	61.27	76.47	74.02	69.12
heb_Hebr	47.55	50.00	45.10	49.02	52.45	54.41	62.75	47.55	61.76	45.59	50.00	65.69	68.63	77.45
hin_Deva	68.63	71.08	69.61	68.63	73.53	76.47	78.92	69.12	78.43	70.59	63.24	80.39	79.90	85.29
hne_Deva	67.16	74.02	69.12	68.63	75.00	77.45	74.51	63.24	75.00	70.59	62.75	81.37	80.39	80.88
hrv_Latn	79.41	80.88	78.92	77.94	78.92	82.35	80.39	73.04	80.88	76.47	80.39	84.80	82.35	86.76
hun_Latn	76.47	75.00	77.45	77.45	77.45	76.96	80.88	71.08	76.96	76.96	77.94	86.27	78.43	87.25
hye_Armn	74.02	75.49	72.06	74.02	74.51	75.98	76.47	66.67	74.51	71.57	70.59	79.90	80.39	84.80
ibo_Latn	71.08	73.53	72.06	73.53	73.53	74.02	75.00	66.18	71.57	69.61	71.57	79.90	83.33	80.88
ilo_Latn	66.67	69.12	71.08	69.61	73.53	75.49	74.51	62.75	71.57	71.57	74.51	76.47	83.33	78.92
ind_Latn	79.41	81.86	80.88	80.88	81.37	82.35	84.80	74.51	80.88	83.33	83.33	84.80	85.78	90.69
isl_Latn	69.12	69.12	69.61	70.59	71.57	73.04	74.02	65.20	68.63	67.65	69.61	78.92	72.55	81.37
ita_Latn	80.88	79.90	82.35	82.84	83.33	83.33	84.80	73.53	82.35	83.33	86.27	87.25	86.27	90.69
jav_Latn	72.06	74.02	72.55	74.51	75.00	75.49	77.45	69.61	75.98	74.02	76.47	77.45	78.43	85.78
jpn_Jpan	78.43	78.43	80.88	80.88	83.82	81.86	83.33	74.51	77.94	81.86	79.90	86.76	80.88	86.27
kab_Latn	34.31	36.27	37.75	37.75	38.73	39.71	34.31	33.33	38.24	33.33	33.82	40.69	40.20	35.78
kac_Latn	40.20	41.67	41.18	39.71	45.10	44.12	42.65	39.71	40.69	42.65	47.06	47.06	54.90	46.08
kam_Latn	41.18	43.63	40.69	44.12	47.55	47.06	44.12	40.69	50.98	44.61	44.61	51.47	55.39	49.51
kan_Knda	69.61	72.06	66.18	69.61	71.08	73.53	74.02	65.20	74.51	68.63	64.71	76.96	77.45	77.45
kat_Geor	76.47	75.49	77.45	74.51	77.45	77.45	78.92	74.02	79.41	76.96	64.71	83.82	78.92	83.82
kaz_Cyrl	72.55	73.04	73.04	73.53	75.00	75.49	77.45	64.22	75.49	70.59	73.53	80.88	79.41	82.35
kbp_Latn	43.14	44.12	48.04	48.04	47.55	48.53	49.51	43.63	42.16	46.57	47.55	44.61	50.98	47.55
kea_Latn	74.51	75.98	75.00	77.94	75.00	76.96	79.41	68.14	77.45	75.49	82.84	81.37	82.35	78.92
khm_Khmr	76.96	76.47	75.49	74.51	76.47	77.94	80.88	71.57	77.45	74.51	75.00	78.92	79.41	85.29
kik_Latn	50.00	53.92	55.39	52.94	54.90	55.39	58.82	48.04	54.90	52.94	53.43	57.35	64.71	58.82
kin_Latn	49.51	51.96	51.96	50.49	51.47	51.96	56.37	49.02	55.39	50.49	48.04	64.71	72.06	62.75
kir_Cyrl	68.14	69.61	69.61	70.59	72.06	74.51	72.06	64.22	73.04	64.71	69.61	75.98	77.94	78.43
kmb_Latn	40.20	41.67	42.16	43.14	44.61	46.08	43.63	43.63	47.06	44.12	43.14	48.53	53.43	50.98
kmr_Latn	58.33	61.76	61.76	61.27	65.20	66.18	67.16	55.88	69.61	61.27	57.84	74.02	76.96	69.61
kon_Latn	58.33	64.71	61.27	63.24	65.69	67.16	63.24	56.37	61.76	59.80	61.76	59.31	74.51	64.22
kor_Hang	76.96	78.92	80.88	83.82	81.37	81.86	81.37	75.98	80.39	81.86	87.25	86.27	85.29	88.73
lao_Laoo	72.55	70.59	74.51	72.06	73.04	70.59	75.98	66.67	73.53	74.02	70.59	76.47	78.43	84.31
lij_Latn	73.04	73.04	72.06	74.02	74.51	75.49	75.98	67.16	72.55	73.53	78.43	76.96	78.92	80.39
lim_Latn	77.45	76.47	75.98	79.41	77.94	77.45	75.98	68.14	78.92	76.96	81.86	83.82	82.35	78.43
lin_Latn	59.31	57.84	60.78	63.24	60.29	59.31	62.75	52.45	63.24	60.78	65.20	61.27	67.65	66.67

Table 3: 7-shot accuracy with on SIB200.

	Label-Aware							Label-Agnostic						
	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER
lit_Latn	66.67	68.63	67.65	69.12	71.57	68.63	74.02	66.18	72.55	70.10	70.10	78.43	80.39	86.27
lmo_Latn	70.10	72.06	73.04	72.06	72.55	73.53	75.98	66.18	75.00	73.53	74.02	79.41	78.92	77.45
ltz_Latn	74.02	72.55	74.51	73.53	77.94	75.49	76.47	69.61	78.92	78.43	76.47	83.82	83.82	80.88
lua_Latn	48.53	48.53	50.00	49.02	51.96	50.98	49.51	48.04	50.49	50.00	54.41	57.35	62.25	53.92
lug_Latn	43.14	47.55	49.51	45.10	47.06	48.04	46.57	44.61	49.51	46.57	45.59	53.43	61.27	56.37
luo_Latn	45.10	46.08	47.55	48.04	49.51	51.96	48.04	44.61	43.63	47.55	49.51	54.90	56.37	58.82
lus_Latn	55.39	54.41	56.86	56.37	59.31	59.31	58.33	51.47	53.43	57.84	57.84	64.71	63.73	69.12
lvs_Latn	67.16	67.65	72.06	70.10	73.04	72.55	75.98	63.73	74.51	70.59	75.98	79.41	80.88	79.41
mai_Deva	62.25	67.16	68.63	66.18	72.06	72.55	72.55	58.82	72.55	61.27	60.78	79.41	76.47	83.33
mal_Mlym	65.69	66.18	66.18	63.73	68.14	69.12	75.49	62.25	69.61	65.20	54.90	74.02	75.00	80.39
mar_Deva	69.12	72.06	71.57	72.06	74.51	74.02	75.49	66.18	72.55	70.10	67.65	79.41	76.96	80.88
min_Latn	76.47	75.98	75.00	76.96	76.47	78.43	75.49	69.12	75.00	75.49	75.49	81.86	78.92	81.37
mkd_Cyrl	74.51	75.49	77.45	76.47	76.47	78.43	77.94	70.10	78.92	77.94	77.45	82.35	81.37	82.35
mlt_Latn	77.45	75.98	79.41	78.92	77.45	81.86	79.90	71.57	76.96	79.41	78.92	81.37	84.80	86.76
mon_Cyrl	71.57	70.59	71.08	72.55	74.02	73.04	78.92	65.69	75.00	68.63	64.22	76.96	80.39	82.84
mos_Latn	45.59	45.10	47.55	46.08	46.57	44.61	43.14	41.67	44.61	45.59	47.55	50.98	50.49	51.47
mri_Latn	62.25	61.27	62.25	63.24	67.65	66.18	62.25	59.31	56.37	64.22	58.82	72.55	74.51	70.59
mya_Mymr	63.73	59.31	61.27	64.22	63.24	67.65	71.08	59.80	59.80	54.90	59.80	70.10	67.16	76.96
nld_Latn	80.39	81.37	81.37	85.29	84.31	85.29	84.80	74.51	83.82	83.82	84.80	85.78	86.76	89.22
nno_Latn	76.47	76.96	77.94	80.39	79.41	79.41	83.33	74.02	75.49	78.43	81.86	85.78	82.35	90.69
npi_Deva	72.55	73.04	70.59	71.57	75.98	75.49	77.45	69.61	78.43	71.57	63.73	82.35	81.37	87.25
nso_Latn	50.98	53.43	51.47	56.37	55.88	57.35	55.88	48.53	56.37	50.49	53.92	59.80	64.71	59.80
nya_Latn	54.90	55.39	58.33	53.92	60.29	57.84	64.22	52.45	62.25	55.88	54.90	69.61	71.08	69.61
oci_Latn	79.41	77.94	78.43	80.39	77.94	80.39	79.41	68.63	76.47	78.43	80.88	83.33	80.88	85.29
orm_Latn	38.24	38.24	40.20	37.75	41.18	40.69	36.76	37.25	39.22	39.71	37.75	43.14	55.39	46.08
ory_Orya	62.75	65.20	60.78	57.84	60.78	64.71	72.55	59.31	67.65	58.33	58.33	69.61	68.14	80.88
pag_Latn	67.16	68.63	68.14	71.08	72.06	70.10	75.00	60.29	71.08	68.63	75.98	72.55	76.96	80.39
pan_Guru	61.27	67.16	63.73	64.22	65.20	65.20	72.06	59.80	67.65	62.25	58.82	73.04	74.51	78.43
pap_Latn	74.02	75.49	76.47	75.49	75.98	78.43	78.43	70.59	75.49	76.96	78.43	79.90	84.31	80.39
pes_Arab	75.00	75.98	76.47	75.00	77.45	75.98	83.33	73.53	77.45	75.49	73.53	82.35	82.35	87.75
plt_Latn	57.84	65.20	60.29	58.82	62.25	63.73	55.88	58.82	64.71	60.29	58.33	72.55	75.00	63.73
pol_Latn	77.94	80.88	78.43	81.37	78.92	79.90	81.86	74.51	79.41	80.39	83.82	82.35	81.37	87.25
por_Latn	77.94	81.37	81.37	83.82	82.35	82.84	85.29	75.49	83.33	81.86	85.78	88.24	86.27	91.18
prs_Arab	74.02	75.49	73.04	73.53	75.00	75.49	82.84	68.63	80.88	74.02	69.12	81.37	84.31	87.25
pus_Arab	54.90	54.90	57.35	58.33	59.31	59.31	66.18	53.92	60.29	55.88	53.92	66.67	68.63	71.57
quy_Latn	53.43	56.37	56.86	57.35	57.35	56.37	63.73	53.43	55.88	56.37	60.29	61.27	65.69	62.75
ron_Latn	76.47	75.98	78.92	76.96	76.47	77.45	81.86	71.08	75.49	79.41	84.31	80.88	81.37	84.31
run_Latn	48.04	49.02	52.94	47.55	52.94	53.92	54.41	47.06	48.53	49.51	50.98	67.16	70.10	65.69
rus_Cyrl	79.41	79.90	78.92	83.82	81.37	81.86	82.35	70.59	80.88	79.90	85.78	85.78	85.29	90.20
sag_Latn	52.94	50.00	53.43	52.45	53.43	52.94	57.84	49.51	53.92	49.51	50.49	51.47	54.90	59.31
san_Deva	57.35	61.27	64.71	59.80	64.71	62.75	66.18	59.31	62.75	59.31	56.86	69.12	69.61	71.08
scn_Latn	75.98	77.45	78.43	78.92	80.88	81.86	79.41	70.59	73.04	77.94	79.90	80.39	81.37	80.88
sin_Sinh	69.61	72.55	66.18	70.10	72.06	72.55	73.53	66.18	70.59	69.12	67.65	77.94	75.49	81.86
slk_Latn	73.53	75.00	75.49	75.49	78.43	77.94	78.92	69.12	79.90	75.49	76.47	83.82	83.33	84.31
slv_Latn	74.02	75.49	75.00	74.02	77.45	77.45	75.98	68.63	73.53	76.47	74.51	83.33	80.88	84.80
smo_Latn	61.27	62.25	67.16	65.20	68.14	69.61	66.67	60.78	65.69	64.71	64.22	73.04	71.57	73.53
sna_Latn	54.41	54.90	55.39	52.94	55.39	55.39	50.00	50.00	54.41	50.98	51.47	60.29	70.10	63.24
snd_Arab	48.04	50.00	50.00	53.43	54.41	57.84	56.37	50.49	59.31	50.00	53.92	63.24	65.69	66.67
som_Latn	49.02	50.49	51.47	52.45	53.43	53.43	49.02	46.08	61.27	49.51	46.57	67.65	66.18	57.84
sot_Latn	58.33	62.25	61.76	62.75	65.20	62.25	59.80	54.41	62.25	58.82	50.00	64.71	68.63	69.61
spa_Latn	78.92	79.90	78.43	82.84	80.39	81.37	84.80	73.04	77.94	79.41	83.33	85.29	83.33	89.71
srd_Latn	75.49	71.57	75.98	73.53	75.00	72.06	74.02	69.12	74.51	77.94	78.43	78.43	80.88	77.45
srp_Cyrl	76.47	80.39	77.94	78.43	81.37	77.45	77.45	72.55	81.86	79.41	81.37	83.33	83.33	81.86
ssw_Latn	50.98	52.45	56.37	53.43	55.88	59.31	60.29	53.43	58.82	55.39	49.51	65.20	69.12	63.73
sun_Latn	77.94	75.49	75.98	79.41	80.39	79.41	80.88	74.02	77.45	76.96	78.92	83.33	80.88	86.76
swe_Latn	75.49	76.47	76.47	79.41	79.90	79.41	83.82	71.08	76.47	78.92	75.98	81.86	82.84	86.76
swh_Latn	65.20	62.75	64.71	63.24	66.18	65.20	68.63	61.27	65.69	60.78	59.80	73.53	72.06	77.45
szl_Latn	72.55	73.04	75.49	74.02	75.98	75.49	76.96	67.65	71.08	71.57	75.49	78.92	81.86	75.00
tam_Taml	68.14	67.16	68.63	68.14	67.65	69.12	74.02	62.75	67.16	68.14	57.35	75.98	75.98	78.92
tat_Cyrl	72.06	75.00	72.55	72.55	74.51	75.98	76.47	67.65	75.98	73.53	70.59	82.84	77.45	81.37
tel_Telu	66.18	69.61	66.18	64.22	72.06	70.10	75.00	59.31	73.04	64.22	65.69	75.98	75.98	81.86
tgk_Cyrl	69.61	70.10	70.10	71.08	74.02	70.10	73.53	63.24	70.59	68.63	68.14	79.90	75.98	81.37
tgl_Latn	76.47	79.41	79.41	78.92	78.92	79.90	81.37	71.57	78.92	80.39	74.02	84.80	84.80	87.75
tha_Thai	76.96	78.43	75.49	76.47	76.47	78.43	78.43	70.59	78.43	72.06	65.69	79.41	79.90	85.78
tir_Ethi	47.55	48.53	50.98	51.96	46.08	48.04	51.96	47.55	52.45	46.57	47.06	53.43	58.82	57.35
tpi_Latn	76.47	78.43	79.90	79.41	79.90	80.39	85.78	72.55	77.45	75.00	78.92	80.88	82.35	85.29
tsn_Latn	54.90	52.45	56.86	55.88	58.82	57.84	55.88	50.49	52.94	53.92	52.94	64.71	66.67	61.27
tso_Latn	54.90	55.39	56.37	57.35	54.90	57.35	57.84	50.49	56.37	54.41	50.00	55.88	68.63	63.24
tuk_Latn	61.76	66.18	65.20	67.16	70.10	66.18	69.61	60.78	72.55	61.27	66.18	75.49	75.98	75.98
tum_Latn	50.00	53.43	53.43	50.98	52.94	55.88	53.43	50.00	55.88	49.02	49.51	62.75	70.59	67.16
tur_Latn	77.45	75.98	76.47	78.92	75.98	78.43	81.86	67.65	79.41	74.51	84.80	83.33	79.90	85.29
uig_Arab	44.12	49.02	43.63	45.59	50.00	53.43	62.25	45.10	50.98	42.16	42.65	58.33	53.43	66.67
ukr_Cyrl	75.00	76.47	73.53	76.96	79.41	78.92	78.92	71.08	80.39	78.43	80.39	83.33	82.35	86.27
umb_Latn	33.33	41.67	39.22	38.73	39.71	42.65	40.20	35.78	42.65	39.22	37.25	46.57	56.37	50.49
urd_Arab	66.67	69.12	68.14	64.71	66.67	66.67	74.02	58.33	68.63	65.69	60.29	77.94	72.55	75.49
uzb_Latn	69.61	70.59	68.14	69.61	71.57	71.57	72.55	65.20	75.49	67.16	65.69	81.37	77.45	77.94
vec_Latn	77.45	76.96	78.43	79.41	80.88	80.39	78.43	73.53	76.47	78.43	81.37	81.86	80.88	81.37
vie_Latn	81.37	79.90	81.86	79.90	82.84	82.84	84.80	74.51	79.41	82.35	69.61	85.78	85.78	87.25
war_Latn	74.51	77.45	76.47	76.96	82.35	79.90	83.33	69.61	78.92	77.94	80.88	82.35	84.80	83.82
wol_Latn	53.92	53.43	54.41	55.39	56.86	58.33	54.90	49.51	49.51	51.47	57.84	58.82	64.71	58.33
xho_Latn	57.84	61.76	61.76	59.80	61.27	65.20	63.24	58.82	62.75	62.25	55.39	68.63	71.57	66.18
yid_Hebr	47.06	48.04	49.51	53.43	49.02	54.41	51.47	47.06	50.00	49.02	45.59	57.35	60.78	62.75
yor_Latn	47.55	50.49	51.47	49.51	52.94	57.35	50.00	45.59	48.53	50.00	51.47	62.25	62.75	57.84
yue_Hani	78.43	81.37	83.82	82.84	83.82	86.27	84.31	71.08	81.86	80.88	83.82	85.29	81.86	85.29
zsm_Latn	76.96	79.41	78.43	80.39	81.37	79.41	81.37	72.55	81.37	79.41	80.88	85.78	82.84	89.71
zul_Latn	61.27	64.22	61.76	65.69	66.67	66.18	70.59	55.88	67.65	59.31	57.35	74.02	72.55	76.96
Avg	65.24	66.60	66.75	67.13	68.51	69.09	70.18	61.68	68.55	66.25	66.59	73.69	74.61	75.91

Table 4: 7-shot accuracy with on SIB200.

	Label-Aware							Label-Agnostic						
	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER	Random	Glot500	MaLA500	SBERT	LaBSE	Multilingual E5	XAMPLER
amh	76.06	76.33	78.46	78.19	77.13	77.66	81.38	78.99	78.72	72.34	81.65	75.00	77.39	79.52
eng	76.05	82.17	81.96	81.12	81.22	81.12	82.81	75.32	81.43	81.54	81.86	83.02	83.02	86.50
fra	85.31	84.60	85.07	83.18	83.18	83.18	85.31	84.12	81.75	83.89	80.57	81.99	79.62	80.81
hau	63.94	63.20	62.45	63.01	63.20	64.31	65.43	61.15	61.52	59.48	62.08	63.94	64.87	66.91
ibo	82.40	85.87	83.73	85.33	81.60	85.33	83.73	83.73	84.80	85.87	85.60	82.67	81.87	78.13
lin	82.29	86.29	81.14	84.57	84.00	81.71	81.14	82.86	80.57	77.71	82.29	86.29	81.14	83.43
lug	70.10	67.16	67.65	66.18	65.20	67.16	66.18	67.16	66.67	64.71	66.18	65.20	64.71	69.12
orm	59.08	60.62	61.54	59.69	59.69	62.46	67.69	60.92	61.54	60.00	61.54	59.08	58.15	68.62
pcm	93.44	94.10	94.43	94.10	93.44	95.08	95.41	93.11	94.10	93.11	93.77	93.44	93.11	93.11
run	78.83	80.78	79.15	80.13	79.48	80.78	79.48	79.48	80.46	78.18	79.15	80.78	80.78	80.13
sna	69.11	65.58	66.67	65.85	62.87	63.96	62.06	63.69	63.96	61.25	63.96	61.79	65.31	59.08
som	62.01	62.37	64.52	63.08	62.37	63.80	63.80	60.57	62.37	63.44	61.29	64.87	63.08	58.42
swa	74.34	72.90	71.94	72.42	73.38	72.66	76.02	70.26	72.90	70.26	72.18	75.30	73.14	75.30
tir	51.84	55.88	56.25	58.82	56.25	57.72	65.44	57.35	61.03	54.41	61.76	61.40	59.56	65.07
xho	48.15	50.51	54.21	52.53	53.20	55.89	57.91	52.86	50.51	55.22	46.80	54.55	53.54	55.56
yor	84.15	85.30	85.01	83.57	84.44	85.30	86.46	86.74	85.88	83.86	84.15	83.29	82.42	81.84
Avg	72.32	73.35	73.39	73.24	72.54	73.63	75.02	72.39	73.01	71.58	72.80	73.29	72.61	73.85

Table 5: 6-shot accuracy with on MasakhaNEWS.