Topic-XICL: Demonstration Selection with Topic Inference for Cross-lingual In-context Learning

Anonymous ACL submission

Abstract

Cross-lingual in-context learning (XICL) shows promise for adapting large language models (LLMs) to low-resource languages. Previous methods rely on off-the-shelf retrievers or task-specific retrievers based on feedback signals from LLM for demonstration selection, they often overlook important factors beyond semantic similarity or can be resource-costly. To address these challenges, we propose a novel approach called Topic-XICL, which leverages a latent topic model to select demonstrations across languages. We assume that latent topic variables incorporate additional information beyond semantics, such as syntax and task structure. By training this topic model on rich-resource language data with a small parameter LLM, we obtain more informative demonstrations by topic inference and utilize them for in-context learning across various LLMs. Our method is tested on three multilingual tasks (XNLI, XCOPA, and TydiQA-GoldP) using three different-size BLOOMZ models and three models with approximately 7 billion parameters (BLOOM, XGLM, and Llama2). Comparative evaluations against random selection, semantic similarity selection, and clustering-based selection baselines show consistent improvements in multilingual average performance with our approach.

1 Introduction

Large Language Models (LLMs) have exhibited exceptional natural language understanding capabilities across diverse NLP tasks. However, their training data is predominantly English-centric, posing challenges for cross-lingual generalization (Lai et al., 2023; Bang et al., 2023; Zhang et al., 2023). In-context learning (ICL) (Brown et al., 2020) presents a promising solution for LLMs in low-resource language settings, as demonstrated by the strong ICL performances of models like BLOOM (Scao et al., 2022) and XGLM (Lin et al., 2022) in various multilingual tasks.

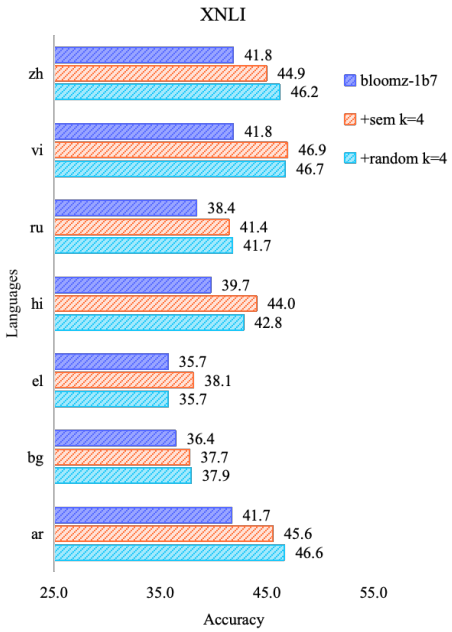

Figure 1: Accuracy scores for 7 languages from the XNLI dataset (Conneau et al., 2018) based on the BLOOMZ-1b7 model (Muennighoff et al., 2023) (with 1.7 billion parameters). "sem" denotes semantic-based demonstration selection, while "random" denotes random selection. k represents the number of demonstrations for ICL.

The impressive comprehension abilities of LLMs in English have sparked interest in Cross-lingual In-Context Learning (XICL). This approach utilizes demonstrations from rich-resource languages to guide learning tasks in low-resource languages. However, the effectiveness of XICL depends heavily on the selection of demonstration examples (Zhao et al., 2021; Perez et al., 2021). Researchers have proposed two main approaches to select demonstration: leveraging off-the-shelf retrievers (Nie et al., 2023; Chang and Fosler-Lussier, 2023; Winata et al., 2023; Li et al., 2023), such as BM25 or Sentence-BERT (Reimers and Gurevych, 2019), and training task-specific retrievers (Shi

et al., 2022) by a specially designed task signal, such as the feedback signals from LLMs. The latter approaches may yield better results for specific LLMs, but they often require access to model parameters or detailed output distributions, which can be costly and are typically unavailable for black-box LLMs (Sun et al., 2022). In contrast, the former methods can lightweightly exploit semantic similarity input-label pairs, but they overlook task-specific information or diversity.

As noted in Qin et al. (2023), the choice between similarity and diversity in demonstrations varies depending on the task: diversity suits tasks like commonsense reasoning question answering, while similarity is preferable for text classification. Fig.1 demonstrates the challenge of balancing these two dimensions across different languages. Semantically similar examples lead to better results for Hindi (hi) and Greek (el), while randomly selected diversity examples lead to better performance for Chinese (zh) and Arabic (ar). When selecting demonstrations across languages, it is crucial to consider not only semantic similarity but also factors such as syntactic structure, task structure, and domain information. We collectively refer to these factors as topic information, which is multidimensional and may enhance demonstration choices for cross-lingual in-context learning.

Xie et al. (2022) examined in-context learning from a Bayesian Inference perspective, while Wang et al. (2023) treated LLMs as topic models, which proved efficacious in demonstration selection for classification tasks. Inspired by this, we extended Wang et al. (2023)'s approach to cross-lingual in-context learning, proposing a demonstration selection algorithm based on topic inference (Topic-XICL). It comprises a **latent topic learning** phase and a **demonstration selection** phase. In the latent topic learning phase, data from a rich-resource language are clustered into several topics by the K-means algorithm with multilingual representations, and a topic model trained based on LLM by absorbing nuanced topic information. Specifically, we introduce c new tokens for each topic to enrich the LLM's vocabulary. These tokens, concatenated with the input, are used to predict the output and guide the LLM in updating the embedding of these new tokens. Before demonstration selection, for each target language input, we identify its topic by calculating semantic similarity with training data. Then, the latent topic model is used to predict the probability that source language examples contain the topic of target input, with the top-k examples by probability serving as demonstrations.

We trained the latent topic model on BLOOMZ-1b7 (Muennighoff et al., 2023) (with 1.7 billion parameters) and conducted cross-lingual demonstration selection on two multilingual sentence-level tasks and one cross-lingual reading comprehension task. The constructed dataset is designed to generalize across various LLMs. Our contributions are summarized as follows:

- We propose a cross-lingual demonstration selection algorithm based on topic inference (Topic-XICL), which extends Bayesian inference theory to practical applications of cross-lingual in-context learning.
- We compare our method with three demonstration selection baselines on six LLMs for three cross-lingual tasks (XNLI, XCOPA, and TydiQA-GoldP). The results demonstrate that our demonstration selected by topic information significantly outperforms existing strong baselines.

2 Related Work

Cross-lingual In-context learning The cross-lingual nature of multilingual language models further enables the possibility of learning from a different language in-context without parameter updates, such as the XICL method (Winata et al., 2021; Lin et al., 2022). Winata et al. (2021) first show that, given a few English examples as context, multilingual pre-trained language models (like GPT (Radford et al., 2019) and T5 (Raffel et al., 2020)) can predict not only English test samples but also non-English ones. (Lin et al., 2022) also found their XGLM demonstrates strong cross-lingual capability where using English prompts together with non-English examples yields competitive zero- and few-shot learning performance.

Cross-lingual Demonstration Selection Different choices of rich-resource language demonstrations can yield varying outcomes for the target languages. Existing approaches for cross-lingual retrieval-augmented demonstrations generally fall into two categories: those based on off-the-shelf multilingual representations and those leveraging feedback signals from LLMs. For example, Nie et al. (2023) conducts cross-lingual retrieval from labeled or unlabeled high-resource languages, which is based on the semantic similarity of multilingual embedding.

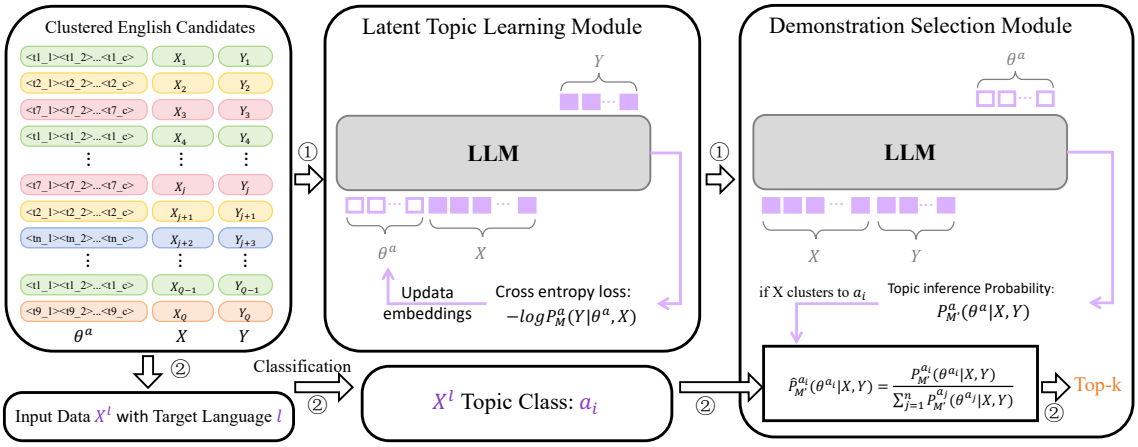

Figure 2: An overview of our proposed cross-lingual demonstration selection algorithm with topic inference.① Latent topic embedding is learned for the clustered English candidates using LLM, and probabilities of inferring n topics are calculated for each candidate example. ② For each target input, its topic classification is determined, denoted as a_i. Then, compute topic inference conditional probabilities for candidate examples classified as a_i and output the top-k examples with high probability for in-context learning in any generative LLM.

Li et al. (2023) extended the PARC framework to focus exclusively on zero-shot settings, revealing that it may not always be optimal for more complex generation tasks. Tanwar et al. (2023) augmented prompts with cross-lingual semantic similarity demonstration and additional task-specific details. Additionally, Winata et al. (2023) emphasized semantic similarity in XICL by selecting the nearest example from various sub-datasets to construct a prompt. In contrast, Shi et al. (2022) proposed a novel retrieve-rerank framework for cross-lingual Text-to-SQL, utilizing a bi-encoder architecture to identify relevant exemplars initially and then training a retriever by distilling the LLM's scoring function.

Training retrievers on specific task data and LLMs may seem advantageous, but dealing with inaccessible parameters of black-box models can be challenging. Our proposed method trains solely based on accessible LLMs. Moreover, semantic similarity demonstration alone may not suffice for complex tasks, we integrate semantics, syntax, and structure into "latent topics". Leveraging LLMs to mine the latent topic information, and select demonstrations from topic dimension to enhance cross-lingual in-context learning.

In-Context Learning with Bayesian inference Xie et al. (2022) presented a significant finding by providing a latent topic interpretation for explaining in-context learning. They demonstrated that the in-context learning predictor approaches the Bayes optimal predictor as the number of demonstrations approaches infinity, assuming both pre-training and task-specific data distribution follow Hidden Markov Models (HMM). However, the Markovian assumption about the data-generating process raises questions about its applicability to natural language and confines empirical validation to synthetic data with toy models. To bridge the theoretical understanding and real-world LLM algorithms, Wang et al. (2023) developed a practical demonstration selection algorithm for real-world LLMs. Our method extends Wang et al. (2023) to an XICL setting. Unlike their approach, which treats each classification data as a topic and learns topic variables, we conduct semantic clustering on each task's data to obtain topics. This allows our approach to apply to a wider range of tasks.

3 Method

Based on the theoretical understanding and practical algorithm of Bayesian inference in in-context learning, we proposed a cross-lingual demonstration selection algorithm with topic inference to improve the performance of XICL. First, we introduce the notations of problem setting and theoretical analysis of the problem. Then we describe the pipeline to learn latent topic embedding in Section 3.2 and the algorithm of demonstration selection in Section 3.3. The framework of our method is shown in Fig. 2.

3.1 Notations and Problem Setting

In cross-lingual in-context learning, the prompt $w_{1:t}$ comprises several rich-resource language demonstrations and a low-resource target language

test input. For the only-decoder model, the generated tokens $w_{t+1:T}$ represent the prediction of test input X, and the gold truth is $Y \in \mathbf{Y}$. $\mathbf{Y}$ is the space of gold truth, and for the generation-form task, it is the space of all possible token sequences. $\theta \in \Theta$ is a high dimensional latent topic variable continuously distributed over Θ, where Θ is the space of the variable. Following Wang et al. (2023), we posit the existence of an underlying causal relation between X, Y, and θ, directly named as $X \rightarrow Y \leftarrow \theta$, which can be represented mathematically as the following structural equation:

$$Y = \mathrm{f}(X, \theta, \epsilon), \tag{1}$$

where ϵ is an independent noise variable.

To perform in-context learning with an LLM (denoted by M), we condition on a fixed set of k demonstrations $(X_1^a, Y_1^a), (X_2^a, Y_2^a), ..., (X_k^a, Y_k^a)$ for a topic $a \in \mathbf{T}$, where $\mathbf{T}$ is the space of all topics in a task. The generation process of an instance in topic a can be written as:

$$Y_i^a = \mathrm{f}(X_i^a, \theta^a, \epsilon), \tag{2}$$

where $\theta^a \in \Theta$ is the value of the topic variable corresponding to topic a. The in-context learning output probability of LLM for an input $X^{a,l}$ classified to a topic in target language l can be denoted by $P_M^{a,l}$, and the solution can be defined as:

$$\arg\max_{y \in \mathbf{Y}} P_M^{a,l}(Y^{a,l} = y | X_1^a, Y_1^a, ..., X_k^a, Y_k^a, X^{a,l}). \tag{3}$$

It always lower or equal to the Bayes optimal decoder $\arg\max_{y \in \mathbf{Y}} P_M^{a,l}(Y^{a,l} = y | \theta^a, X^{a,l})$. Equality only holds when

$$P_M^{a,l}(\theta^a | X_1^a, Y_1^a, ..., X_k^a, Y_k^a, X^{a,l}) = 1 \tag{4}$$

Following Wang et al. (2023), we focus on estimating an optimal value of θ corresponding to a topic a. Then, we will discuss how to select an optimal set of demonstrations by using the learned optimal latent concept variable value.

3.2 Latent Topic Learning

As shown in Fig.2, we first cluster the source language task dataset into several topics $\{a_i | i = 1, 2, ..., n\}$ by the multilingual embedding with K-means algorithm, the number of topic n is a hyper-parameter. For a topic a_i, we assume that $X \rightarrow Y \leftarrow \theta$ and $\arg\max_{y \in \mathbf{Y}} P_M^{a_i}(Y = y | \theta^{a_i}, X)$ is the Bayes optimal decoder to minimize $\mathbb{E}_{X,Y,a_i}[-\log P_M^{a_i}(Y | \theta^{a_i}, X)]$.

In practice, we try to align θ^a to the token embedding space by adding new tokens to the vocabulary of LLM. Then, the learned new tokens of θ^a are used as regular tokens in the vocabulary. Specifically, to represent each specific topic a_i, c new topical tokens (denoted as $\hat{\theta}^{a_i}$) are added to the original vocabulary. c is also a hyper-parameter, and corresponding c topical tokens are appended to the input X as demonstrated, like "<t1_1><t1_2>...<t1_c>X" for the topic a_1. The new topical token can be anything as long as it does not overlap with the original vocabulary of LLM.

Subsequently, the embedding of these new tokens $E(\hat{\theta}^{a_i})$ is fine-tuned while freezing the remaining parameters of LLM. The fine-tuning objective is to minimize loss:

$$\mathcal{L}(\hat{\theta}^{a_i}) = \mathbb{E}_{X,X}[-\log P_M^{a_i}(Y | \hat{\theta}^{a_i}, X)] \tag{5}$$

and the fine-tuned LLM denoted as M'. To obtain the topical tokens for all topics in a task, we fine-tune all data together with the loss $\sum_{i=1}^{n} \mathcal{L}(\hat{\theta}^{a_i})$.

For the setting of n, the estimated conditional probability of $\hat{\theta}^{a_i}$ would be:

$$\hat{P}_{M'}^{a_i}(\hat{\theta}^{a_i} | w_{1:t}) = \frac{P_{M'}^{a_i}(\hat{\theta}^{a_i} | w_{1:t})}{\sum_{j=1}^{n} P_{M'}^{a_j}(\hat{\theta}^{a_j} | w_{1:t})} \tag{6}$$

3.3 Demonstration Selection

About the topic of target instance (X^l, Y^l), we embed the input X^l and measure its semantic similarity with all source input embeddings. Then, we statistic the topic category of the top-10 semantic similar source examples and choose the most frequent topic as the target language topic a.

According to the analysis in Section 3.1, for the target instances with topic a, our goal becomes selecting demonstrations that can best infer the topic for all inputs:

$$\arg\max_{X_1^a, Y_1^a, ..., X_k^a, Y_k^a} \mathbb{E}_X[P_M^a(\theta^a | X_1^a, Y_1^a, ..., X_k^a, Y_k^a, X)] \tag{7}$$

As test examples are sampled independently of the demonstrations and each demonstration is also sampled independently, the goal can be:

$$\begin{aligned} &\arg\max_{X_1^a, Y_1^a, ..., X_k^a, Y_k^a} P_M^a(\theta^a | X_1^a, Y_1^a, ..., X_k^a, Y_k^a) \\ &= \frac{\prod_{i=1}^{k} P_M^a(\theta^a | X_i^a, Y_i^a)}{P_M^a(\theta^a)^{k-1}} \end{aligned} \tag{8}$$

Assuming that θ has a uniform prior, then our goal becomes finding the top k demonstrations that maximize $\hat{P}_{M'}^a(\hat{\theta}^a | X_i^a, Y_i^a)$.

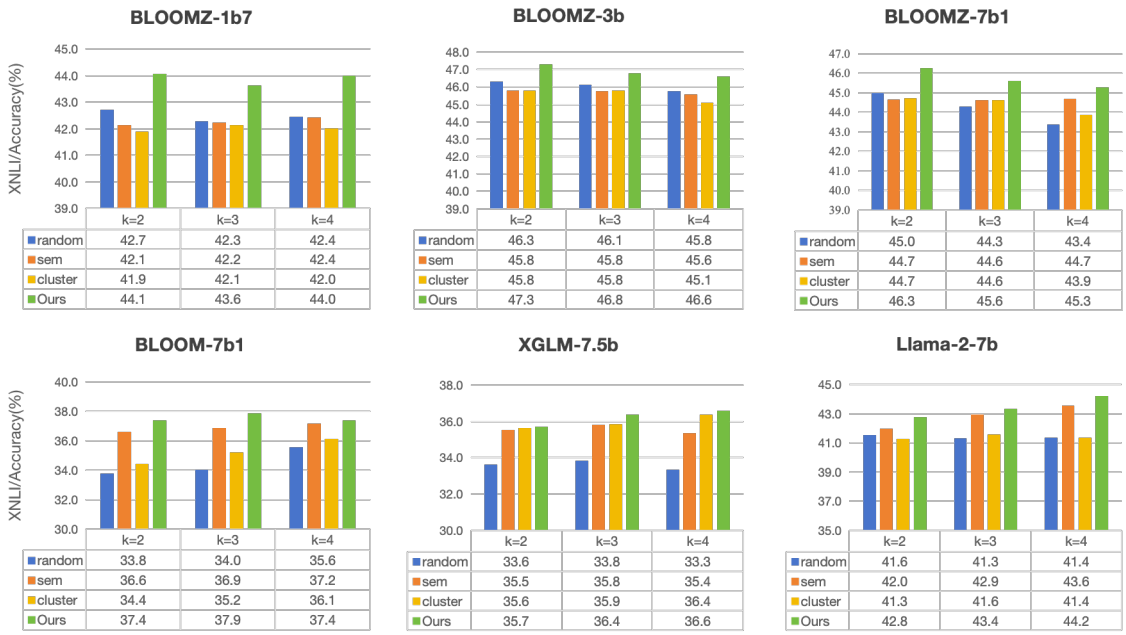

Figure 3: Average accuracy scores of XNLI (Conneau et al., 2018).

We mainly focus on the fundamental effects of topic inference on multilingual demonstration selection, without discussion on the mutual influence between demonstrations and the impact of order.

4 Experiments

4.1 Dataset

This paper presents experiments conducted on three datasets: XNLI (Conneau et al., 2018), XCOPA[1], and TyDiQA-Gold (Clark et al., 2020). Cross-lingual Natural Language Inference dataset (XNLI) is a **sentence-pair classification** task involving 15 languages, which is translated from English SNLI (Bowman et al., 2015) dataset. XCOPA is designed to assess the **causal commonsense reasoning** capabilities of multilingual language models. It is an extension and re-annotation of the English COPA dataset (Gordon et al., 2012) where the validation and test set examples are carefully translated to and annotated in 11 typologically diverse languages. Since existing work mainly discusses demonstration selection methods on classification tasks, we added exploration on **Question Answer** task in our experiments. TyDiQA-Gold is the gold passage task in TyDiQA (Clark et al., 2020) covering 9 typologically diverse languages, a challenging multilingual machine reading comprehension benchmark.

4.2 Experimental Setting

For each dataset, the English training set $\mathcal{D}$ serves as the pool of candidate demonstrations, evaluated across all test sets in each language. We employ the K-means algorithm with random initial center points to cluster the training set $\mathcal{D}$, utilizing three seed values $[32, 44, 100]$ for experimentation and reporting the average results per language or dataset. The representation of each training data is obtained using multilingual Sentence-BERT (Reimers and Gurevych, 2019). To accommodate various context length settings for different LLMs, we set $k = [2, 3, 4]$. As for hyperparameters, the number of cluster classes $n = 20$ and the length of each topic token sequence $c = 10$ are used for XNLI, and $n = 20$ and $c = 15$ are for TyDiQA-Gold, while $n = 5$ and $c = 15$ are set for XCOPA (with only 500 English training dataset). We leverage the Bloomz-1b7[2] model to learn the topic token embeddings and compute the probability of each candidate demonstration example. The same set of demonstrations selected by BLOOMZ-1b7 is used for all other LLMs. Greedy Search is employed for decoding answers in each task. The utilized prompt is detailed in Appendix A.

Base Model } The BLOOMZ-1b7 (Muennighoff et al., 2023) is a multilingual supervised fine-tuning version of BLOOM, which may be more efficient

[1]https://github.com/cambridgeltl/xcopa

[2]https://huggingface.co/bigscience/bloomz-1b7

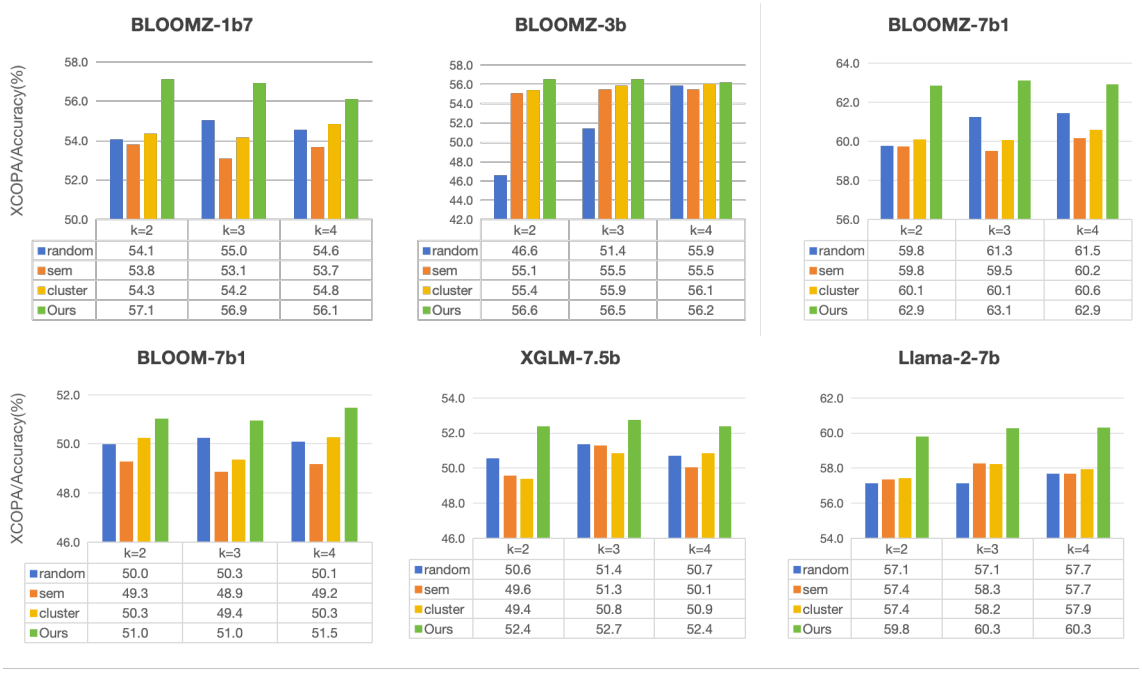

Figure 4: Average accuracy scores of XCOPA.

for learning the topic of a task than the original version. We use the same set of demonstrations selected by BLOOMZ-1b7 to three sizes of BLOOMZ models, including BLOOMZ-1b7, BLOOMZ-3b, and BLOOMZ-7b1. Moreover, we test the demonstrations selected by our method on more LLMs (BLOOM, XGLM, and LLama-2) with about 7 billion parameter sizes.

4.3 Baselines

We consider the following baselines:

Random: We random select k demonstrations from $\mathcal{D}$ for each test example. We also set three seeds to obtain the average results.

Semantic Similarity (sem): Demonstrations semantically similar to the test example would help. We apply the same multilingual Sentence-BERT (Reimers and Gurevych, 2019) as we used to calculate the cosine similarity between the inputs of the source language and target language. We choose the top k similar demonstrations from $\mathcal{D}$ for each test example.

Cluster: As our method initially clusters $\mathcal{D}$ and subsequently selects demonstrations, we randomly sample k instances from each category of the clustered data as demonstrations for all test examples within that category. This also serves as an ablation baseline for our approach.

4.4 Main Results

Fig.3, 4, and 5 show our main results for three datasets averaged over all languages on six LLMs. The detailed results can be found in Appendix B.

For sentence-level tasks, XNLI and XCOPA, our method almost always significantly outperforms baselines on six models. Notably, on XNLI, our approach achieves an average improvement of 2.8% over the best baseline across three different scales of BLOOMZ models with just two demonstrations.

For the XCOPA dataset, the performance improvement is more pronounced on the four 7B-parameter models, with average improvements of 2.26% and 1.67% on Llama-2-7b and BLOOMZ-7b, respectively. From a monolingual perspective (see Table 7 in the appendix), our method achieves the best scores based on the Llama-2-7b model, with improvements of 1.4% and 2.2% on the low-resource languages Haitian Creole (ht) and Tamil (ta) compared to the best baseline.

Our method also shows significant improvements in average performance for more complex QA tasks. However, in the BLOOMZ series of models, the improvement mainly comes from several low-resource languages. For instance, on the BLOOMZ-7b1 model, our best results in Finnish (fi) and Korea (ko) surpass the best baseline by 6.4% and 11%, respectively. It may be the case that the TydiQA task training set is already encompassed within the instruction tuning process

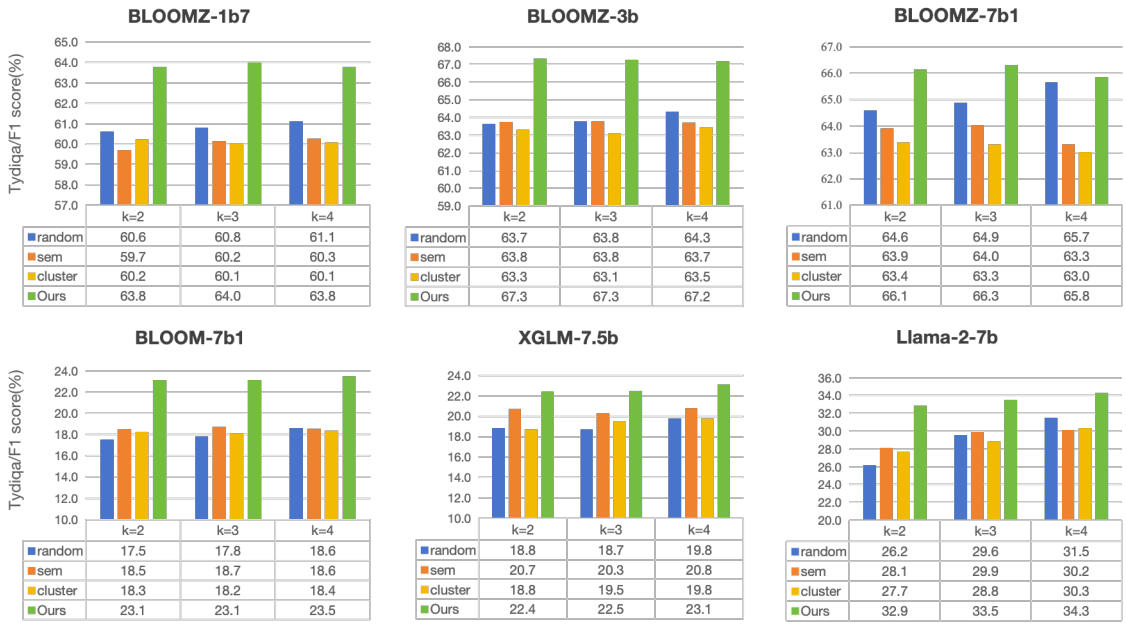

Figure 5: Average F1 scores of Tydiqa-GoldP.

of BLOOMZ, hence requiring only a few random demonstrations to stimulate its learning capacity. Nonetheless, for low-resource languages exhibiting poorer performance, a more tailored demonstration selection remains imperative, underscoring the significance of our method. Our approach notably enhances performance across the other three 7B models as well, particularly on BLOOM-7b1, where the mean improvement is 4.6%.

Experimental results demonstrate that although our method only trains the topic model on BLOOMZ-1b7 and performs demonstration selection, choosing appropriate contextual data can lead to performance improvements across LLMs of different sizes or architectures. Our method consistently outperforms the cluster baseline, indicating that our approach's superiority isn't solely derived from simple semantic clustering.

5 Analysis

In theory, our method aims to select optimal demonstrations based on the semantic cluster distribution. However, our approach employs a straightforward K-means clustering method, where parameters like the number of clusters and topic sequence lengths can influence demonstration selection. We conduct a basic exploration (on XNLI) by experimenting with different numbers of topics and topic tokens sequence lengths and assessing if our approach effectively captures useful topic information.

5.1 The number of topics

Based on the experiences of Wang et al. (2023), we initially set the length of the topic sequence to 10 and conducted ablation experiments on the number of topic categories n. We cluster the training data into 5, 10, 15, 20, and 30 topic categories, then train the topic model and select demonstrations for the target data. The results on BLOOMZ-1b7 are shown in the table 1. The results indicate that the optimal performance is achieved when the number of topics gradually increases to 20. Further increases of n may result in overly similar characteristics within each topic category, leading to a loss of diversity of demonstrations and a decrease in performance.

5.2 The length of topic token sequence

The topic token sequence contains information about the topics. Intuitively, longer sequences can capture more topic information, but excessively long sequences may introduce irrelevant information. Assuming we have determined the number of topics to be 20, we set c to four different lengths: 5, 10, 15, and 20, for learning the topic sequence. Then, we concatenate the topic token sequence directly with the input and predict the results. Their experimental results are shown in table 2, which yields better results than the original input based on BLOOMZ-1b7. It indicates that the token sequence has learned useful topical information and

XNLI (acc.)																	
BLOOMZ-1b7		en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
k=2	n=5	52.22	45.74	37.00	39.38	37.31	46.36	46.90	43.03	41.39	39.28	34.46	35.11	39.42	45.64	45.98	41.95
	n=10	52.02	46.67	38.20	39.48	37.47	47.22	47.17	42.69	40.52	39.39	34.20	34.64	39.64	45.48	45.54	42.02
	n=15	52.64	46.51	38.59	39.96	36.99	47.91	47.65	44.06	41.98	39.76	35.35	35.23	39.66	46.56	46.62	42.63
	n=20	**53.17**	46.83	**40.20**	**41.14**	**40.04**	**47.92**	**48.72**	**45.09**	**42.46**	**43.87**	**36.85**	**36.39**	**42.81**	**48.84**	**46.73**	**44.07**
	n=30	52.04	**46.87**	38.34	39.70	39.14	47.27	47.72	44.31	41.78	40.62	33.89	35.21	40.32	46.75	45.99	42.66
k=3	n=5	51.55	45.72	37.59	39.33	36.43	47.34	47.31	42.70	40.70	40.86	33.96	34.85	39.62	45.75	45.73	41.96
	n=10	51.69	46.85	38.15	39.14	36.52	46.99	47.55	42.55	41.74	40.59	33.73	34.74	40.37	47.12	45.87	42.24
	n=15	52.70	46.50	38.66	39.91	37.68	47.42	47.76	43.60	42.14	40.74	34.01	34.95	41.24	48.17	46.14	42.78
	n=20	**52.71**	**46.97**	**39.44**	**41.08**	38.22	**48.36**	**49.06**	43.95	**42.63**	**42.32**	**35.05**	**37.05**	**41.76**	**48.88**	**46.89**	**43.62**
	n=30	52.04	46.79	38.48	39.70	**39.42**	47.56	47.98	**45.29**	41.80	41.54	33.87	35.15	41.44	48.62	46.39	43.07
k=4	n=5	52.05	46.44	37.96	39.16	36.93	46.83	47.69	42.75	41.02	38.34	34.14	34.79	39.98	46.40	45.93	42.03
	n=10	51.60	46.90	38.17	40.03	36.76	47.41	48.01	44.40	42.20	40.84	34.06	34.54	39.83	46.77	46.46	42.53
	n=15	52.70	46.73	38.79	39.76	37.21	47.70	48.11	44.14	42.52	42.12	34.90	35.41	42.00	48.69	46.87	43.18
	n=20	**53.25**	**47.17**	**39.96**	**41.40**	**38.42**	**48.68**	**49.20**	**44.85**	**42.85**	**42.97**	**35.05**	**36.73**	**42.34**	**49.68**	**47.37**	**43.99**
	n=30	52.34	47.09	39.04	40.82	37.35	47.54	47.92	44.43	42.36	42.10	34.09	35.05	41.76	48.58	46.67	43.14

Table 1: In-context learning accuracy of our method in XNLI with different topices n based on BLOOMZ-1b7 model.

XNLI (acc.)																
BLOOMZ-1b7	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
	48.5	41.7	36.4	37.9	35.7	42.7	42.4	39.7	38.4	36.8	33.7	34.4	36.7	41.8	41.8	39.2
c=20	51.7	44.2	**38.6**	39.9	37.1	44.9	44.6	40.5	40.2	38.2	33.9	**35.1**	37.3	42.9	44.0	40.9
c=15	51.1	45.2	38.1	39.9	37.4	44.9	44.5	41.8	40.2	**39.0**	34.3	**35.1**	38.4	**45.5**	43.7	41.3
c=10	**52.3**	**45.4**	**38.6**	**40.7**	**37.5**	**46.0**	**45.7**	**42.2**	**40.9**	**39.0**	**34.4**	**35.1**	**38.9**	44.7	**44.5**	**41.7**
c=5	51.6	44.2	37.6	40.1	37.2	45.2	45.0	39.8	40.2	37.8	34.2	34.9	37.1	43.1	43.5	40.8

Table 2: Accuracy of our method in XNLI with different length of topic token sequence c based on BLOOMZ-1b7 model.

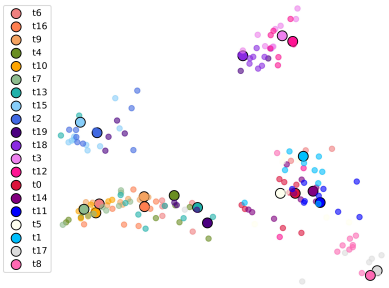

Figure 6: t-SNE plot of the learned topic tokens for XNLI task.

$c = 10$ is a suitable choice for XNLI.

5.3 Visualization of topic token embedding

From Table 2, it is evident that the topic tokens indeed learn information beneficial for in-context learning. As the topic categories are obtained through clustering, to learn about the relationships between each category, we visualize the embeddings of the topic tokens. As shown in Fig.6, in the topic model trained on the XNLI dataset, the embeddings of topic tokens are distributed in several regions, where semantically similar topics are concentrated in the same area (e.g., the centroids of t2 and t15 are very close). This indicates that the topic tokens indeed capture similarities or characteristics among different topic categories, providing more diverse information for demonstration selection.

6 Conclusion

In this work, we explore cross-lingual demonstration selection from a more informative latent topic perspective. We propose a demonstration selection algorithm based on topic inference (Topic-XICL) for cross-lingual in-context learning. Our approach requires learning the latent topic model on less parameters LLMs and selecting appropriate rich-resource language demonstrations for each topic of the target input by computing topic inference probabilities. One-time demonstration selection for a task can be generalized across various LLMs. We validate the effectiveness of our method on three task categories and six models and analyze that the latent topic variables indeed capture useful diversity information for cross-lingual in-context learning.

Limitations

Due to the computation constraints, we were not able to experiment with our framework on larger LLMs or on other task. The experiments confirm that different clustering parameter choices yield diverse outcomes. However, as we did not prioritize exploring the selection of clustering methods, we leave it for future iterations of our method to delve into and explore this aspect further.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. The 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*.

Shuaichen Chang and Eric Fosler-Lussier. 2023. Selective demonstrations for cross-domain text-to-sql. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14174–14189.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. 2020. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012*, pages 394–398. The Association for Computer Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *CoRR*, abs/2304.05613.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. From classification to generation: Insights into crosslingual retrieval augmented ICL. *CoRR*, abs/2311.06595.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP2022*, pages 9019–9052.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 15991–16111. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual retrieval augmented prompt for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pages 11054–11070.

Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. 2023. In-context learning with iterative demonstration selection. *CoRR*, abs/2310.09881.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, , and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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*, pages 3980–3990.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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*, pages 5248–5259.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 20841–20855. PMLR.

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*, pages 6292–6307. Association for Computational Linguistics.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *CoRR*, abs/2301.11916.

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.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust GPT when your question is not in english. *CoRR*, abs/2305.16339.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Prompt Template

Table 3 shows the prompt template we used for three tasks.

Dataset	Prompt
XNLI	<premise> question: <hypothesis>. True, False, or Inconclusive? Answer: [True/False/Inconclusive]
XCOPA	Question: What might be the cause of / What might have happened as a result of "<premise>"? Options: 1-<Choice1> 2-<Choice2> Answer: [1/2]
TydiQA-GoldP	Passage: <passage> question: <question> Answer: [a span in passage]

Table 3: Prompt template for three tasks.

B Detailed Results

		XNLI(acc.)															
Model		en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
BLOOMZ-1b7		48.5	41.7	36.4	37.9	35.7	42.7	42.4	39.7	38.4	36.8	33.7	34.4	36.7	41.8	41.8	39.2
k=2	random	52.9	46.8	38.5	40.8	37.1	**48.2**	48.1	42.7	42.2	40.8	34.1	35.4	40.3	46.5	46.3	42.7
	sem	51.5	45.5	38.0	40.1	37.8	46.2	46.9	43.3	41.1	40.7	34.1	35.2	39.9	46.7	45.0	42.1
	cluster	52.0	45.9	37.6	40.1	36.4	47.1	47.3	42.2	41.1	39.9	33.7	35.0	39.2	45.9	45.0	41.9
	Ours	**53.2**	**46.8**	**40.2**	**41.1**	**40.0**	47.9	**48.7**	**45.1**	**42.5**	**43.9**	**36.9**	**36.4**	**42.8**	**48.8**	**46.7**	**44.1**
k=3	random	52.6	46.5	38.0	40.3	36.2	47.7	47.9	42.4	41.4	40.2	33.9	35.3	39.9	46.2	45.8	42.3
	sem	51.4	45.4	37.8	39.7	37.8	46.4	46.9	43.7	41.3	41.1	34.2	35.4	40.6	47.0	44.9	42.2
	cluster	51.9	46.2	37.7	40.1	35.9	47.3	47.8	42.6	41.1	40.4	33.7	35.1	40.0	46.6	45.5	42.1
	Ours	**52.7**	**47.0**	**39.4**	**41.1**	**38.2**	**48.4**	**49.1**	**44.0**	**42.6**	**42.3**	**35.1**	**37.1**	**41.8**	**48.9**	**46.9**	**43.6**
k=4	random	52.5	46.6	37.9	40.6	35.7	47.8	48.1	42.8	41.7	40.6	33.8	35.4	40.3	46.7	46.2	42.4
	sem	51.6	45.6	37.7	40.1	38.1	46.3	47.3	44.0	41.4	41.4	34.2	35.6	41.1	46.9	44.9	42.4
	cluster	52.2	45.9	37.3	40.0	36.0	47.3	48.0	42.6	41.0	39.9	33.8	35.0	39.8	46.3	45.1	42.0
	Ours	**53.3**	**47.2**	**40.0**	**41.4**	**38.4**	**48.7**	**49.2**	**44.9**	**42.9**	**43.0**	**35.1**	**36.7**	**42.3**	**49.7**	**47.4**	**44.0**
BLOOMZ-3b		54.8	47.8	38.6	41.0	38.4	50.6	50.4	46.5	43.4	40.1	37.3	34.9	42.3	47.7	49.1	44.2
k=2	random	56.2	50.8	40.2	43.1	40.6	53.3	52.8	47.7	44.7	43.8	38.7	35.1	45.0	50.6	52.0	46.3
	sem	55.5	50.4	40.0	42.8	40.2	52.5	52.1	47.6	44.0	43.0	37.8	35.3	44.3	50.3	51.2	45.8
	cluster	55.8	50.5	40.0	42.7	40.2	52.9	52.2	47.4	44.7	42.4	37.6	35.0	44.0	50.2	51.3	45.8
	Ours	**56.9**	**51.7**	**41.0**	**44.1**	**41.9**	**53.9**	**53.6**	**49.1**	**45.9**	**45.1**	**39.8**	**36.2**	**45.8**	**51.9**	**52.9**	**47.3**
k=3	random	56.1	50.7	40.3	42.9	40.6	53.2	52.4	47.5	45.1	42.8	38.8	34.9	44.4	50.6	51.8	46.1
	sem	55.6	50.3	40.2	42.8	40.3	52.5	52.0	47.5	43.9	42.8	37.6	35.4	44.4	50.4	50.8	45.8
	cluster	55.8	50.4	40.1	42.8	40.3	52.7	52.3	47.2	44.5	42.8	37.3	35.0	44.0	50.6	51.4	45.8
	Ours	**57.0**	**51.2**	**40.8**	**43.4**	**41.1**	**53.6**	**53.1**	**48.3**	**45.8**	**43.9**	**38.8**	**36.3**	**45.1**	**51.5**	**52.3**	**46.8**
k=4	random	55.8	50.4	40.0	42.3	40.3	52.5	52.2	47.8	44.6	42.4	37.5	35.3	44.5	50.0	50.9	45.8
	sem	55.1	50.1	40.0	42.4	40.3	52.3	51.9	47.6	43.8	**42.9**	37.3	35.6	44.3	49.6	50.6	45.6
	cluster	55.2	49.9	39.7	42.1	39.4	52.0	51.7	46.5	44.1	41.6	36.2	35.1	43.2	49.5	50.4	45.1
	Ours	**56.6**	**51.2**	**40.9**	**43.3**	**40.8**	**53.7**	**53.2**	**48.2**	**45.8**	42.7	**38.6**	**36.2**	**45.0**	**50.7**	**52.2**	**46.6**
BLOOMZ-7b1		57.6	49.0	26.0	42.2	13.4	50.3	50.4	47.5	30.6	41.1	8.3	32.5	43.3	48.6	50.2	39.4
k=2	random	57.9	49.4	37.8	45.3	35.4	52.0	52.6	47.7	43.2	41.1	34.2	35.4	43.5	49.2	50.1	45.0
	sem	57.1	48.7	37.9	44.4	35.1	51.0	51.4	48.4	42.8	41.3	33.1	35.4	43.8	49.9	49.5	44.7
	cluster	57.3	48.9	37.8	44.2	35.7	51.3	51.8	47.8	42.5	41.4	34.5	35.4	43.4	49.5	49.5	44.7
	Ours	**58.2**	**50.3**	**39.7**	**45.7**	**37.6**	**52.6**	**53.1**	**49.1**	**44.8**	**43.6**	**36.0**	**36.4**	**45.1**	**51.0**	**50.9**	**46.3**
k=3	random	57.2	48.4	37.3	44.6	35.0	51.1	51.6	47.1	42.6	40.2	34.1	35.2	42.7	48.5	49.0	44.3
	sem	56.3	48.6	38.0	44.4	35.7	50.7	51.0	**48.3**	42.7	41.4	34.0	**35.6**	43.9	49.9	49.0	44.6
	cluster	56.7	48.7	37.8	44.1	35.8	50.9	51.2	48.1	42.6	41.4	**34.6**	35.3	43.4	49.9	49.1	44.6
	Ours	**58.1**	**49.9**	**38.2**	**45.4**	**36.9**	**52.2**	**52.7**	**48.3**	**44.9**	**42.8**	**34.6**	**35.6**	**44.4**	**50.1**	**50.2**	**45.6**
k=4	random	56.0	47.2	36.6	43.6	34.7	49.7	50.0	45.8	41.6	39.2	34.0	34.9	42.0	47.6	47.7	43.4
	sem	56.2	48.5	**38.1**	44.4	36.0	50.5	50.8	**48.6**	42.9	41.5	34.5	35.5	43.9	50.1	49.0	44.7
	cluster	56.0	48.0	37.3	43.5	35.3	50.0	50.4	46.8	42.0	40.3	34.3	35.1	42.6	48.7	48.1	43.9
	Ours	**57.8**	**49.3**	38.0	**44.9**	**36.6**	**51.5**	**51.8**	48.4	**43.6**	**42.6**	**34.6**	**35.6**	**44.4**	**50.1**	**50.1**	**45.3**

Table 4: Accuracy of XNLI in 15 languages based on BLOOMZ models

XNLI(acc.)																	
Model		en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
BLOOM-7b1		34.1	33.6	33.7	33.1	33.4	35.8	36.5	31.0	33.4	32.9	21.2	33.6	33.3	33.1	32.7	32.8
k=2	random	34.2	33.4	33.7	33.9	34.4	34.7	33.9	33.4	33.9	34.5	33.3	33.5	33.3	33.5	33.3	33.8
	sem	37.9	35.9	36.3	35.8	**36.1**	**38.0**	**37.6**	36.2	36.2	38.8	35.3	**36.5**	34.7	**38.6**	35.1	36.6
	cluster	35.7	33.8	34.9	34.3	35.0	35.3	35.5	32.5	35.5	36.1	33.9	33.7	33.2	33.5	33.7	34.4
	Ours	**38.7**	**38.1**	**37.8**	**37.0**	35.2	37.0	37.1	**36.8**	**39.0**	**39.2**	**36.7**	36.0	**37.4**	37.6	**36.8**	**37.4**
k=3	random	35.8	32.7	34.4	33.8	37.8	35.3	35.3	32.7	33.4	33.3	34.2	33.2	33.0	32.6	32.8	34.0
	sem	38.3	**37.6**	36.7	35.7	36.6	37.6	37.7	**36.4**	37.3	37.7	34.1	**36.6**	36.2	38.1	36.2	36.9
	cluster	36.4	35.6	35.7	35.2	34.2	36.2	36.3	34.9	35.9	38.0	33.4	34.5	34.1	33.8	33.8	35.2
	Ours	**41.1**	35.2	**37.2**	**36.8**	**36.7**	**39.8**	**39.9**	35.8	**37.7**	**41.1**	**37.8**	34.1	**37.8**	**39.8**	**37.2**	**37.9**
k=4	random	35.0	35.2	33.9	35.2	33.7	34.1	34.5	38.8	36.7	34.7	33.4	33.5	35.5	40.5	38.6	35.6
	sem	**38.9**	37.8	35.8	36.3	**36.5**	**39.0**	**39.3**	**36.3**	37.3	**38.1**	34.1	**36.1**	36.6	38.1	37.4	37.2
	cluster	36.6	36.3	35.1	36.0	33.9	36.7	36.7	37.7	36.4	36.2	33.8	34.0	36.6	39.2	36.9	36.1
	Ours	37.6	**40.6**	**37.2**	**36.5**	35.7	38.3	37.5	34.6	**37.6**	36.5	**34.6**	35.4	**38.5**	**40.7**	**39.2**	**37.4**
XGLM-7.5b		32.1	37.1	34.8	34.3	32.4	33.1	32.4	31.8	32.8	31.8	30.5	28.3	33.2	31.8	28.6	32.3
k=2	random	33.3	35.6	33.5	34.3	33.4	33.6	33.3	33.4	33.5	33.4	33.3	33.3	34.4	33.3	33.2	33.6
	sem	**35.5**	**39.0**	34.6	**37.8**	34.6	**37.2**	**37.1**	32.9	**37.8**	33.4	32.7	37.0	34.5	33.8	34.8	35.5
	cluster	34.8	38.2	34.0	36.4	36.4	35.9	35.6	33.6	34.9	33.2	37.0	**38.7**	33.9	**36.4**	35.7	35.6
	Ours	34.9	38.1	**36.5**	37.4	**36.7**	35.4	34.7	**34.4**	35.1	**33.9**	**35.8**	35.1	**35.6**	35.0	**37.1**	**35.7**
k=3	random	33.1	36.5	33.9	35.1	33.9	33.7	33.3	33.3	33.7	33.3	33.2	33.3	34.1	33.6	33.8	33.8
	sem	**35.3**	38.9	**35.6**	37.7	**37.5**	**38.1**	37.2	33.0	37.3	**34.4**	33.5	36.9	32.3	34.9	34.7	35.8
	cluster	34.4	39.0	34.3	37.4	36.4	37.0	36.8	33.4	36.4	33.2	35.6	**38.8**	34.0	35.4	35.8	35.9
	Ours	34.5	**39.5**	33.8	**38.7**	36.4	37.2	**38.3**	**35.0**	**37.4**	34.1	**36.3**	36.5	**35.5**	**35.9**	**36.4**	**36.4**
k=4	random	30.3	35.4	34.0	35.0	34.6	34.0	33.5	32.8	33.7	33.4	32.9	33.0	30.5	33.5	33.3	33.3
	sem	**36.0**	39.0	34.8	38.0	37.0	37.9	36.8	32.0	37.0	32.2	33.0	**37.5**	30.8	32.8	35.5	35.4
	cluster	35.2	38.8	**36.1**	**38.3**	**38.1**	37.3	36.2	**34.6**	37.0	34.0	**34.7**	37.4	35.3	**35.8**	**36.6**	36.4
	Ours	35.8	**39.7**	34.6	38.2	37.7	**39.7**	**36.9**	34.3	**37.2**	**35.0**	34.0	37.4	**36.8**	35.1	36.3	**36.6**
Llama-2-7b		48.1	37.2	41.9	41.0	37.1	43.6	42.1	37.8	43.3	32.2	34.4	37.0	35.9	40.2	41.8	39.6
k=2	random	51.9	38.5	43.2	45.2	37.4	46.3	47.2	39.4	44.9	32.1	35.6	38.1	36.1	43.2	44.5	41.6
	sem	52.3	38.6	**44.8**	46.5	37.5	47.3	47.8	38.4	**46.5**	32.5	35.7	38.2	36.0	44.1	43.8	42.0
	cluster	50.2	37.7	43.6	46.1	36.1	46.1	**47.8**	39.1	44.9	33.3	34.3	37.8	34.8	43.4	43.7	41.3
	Ours	**52.7**	**38.7**	43.5	**46.6**	**37.8**	**47.9**	**47.8**	**43.0**	45.4	**34.3**	**36.2**	**38.6**	**38.4**	**46.5**	**44.3**	**42.8**
k=3	random	50.8	38.1	44.3	46.2	37.1	46.0	48.0	38.6	44.9	32.7	34.6	37.4	36.1	43.0	41.9	41.3
	sem	53.2	**39.7**	**45.9**	**47.8**	**38.3**	**49.8**	49.6	38.4	46.9	32.8	**36.7**	39.0	36.2	45.0	44.8	42.9
	cluster	50.4	39.3	44.7	47.3	37.5	46.2	47.2	38.5	44.6	**33.6**	34.4	38.2	36.1	43.6	42.5	41.6
	Ours	**54.1**	39.5	45.6	46.3	37.6	49.3	**49.9**	**40.9**	**48.4**	**33.6**	36.5	**39.2**	**36.9**	**46.5**	**46.4**	**43.4**
k=4	random	51.1	37.1	43.9	47.2	37.2	46.7	48.0	39.0	45.5	32.5	34.8	37.4	35.7	42.5	41.9	41.4
	sem	54.0	40.7	**46.7**	**48.6**	38.7	**50.3**	50.6	38.4	47.9	33.0	37.1	**40.2**	36.6	45.3	45.7	43.6
	cluster	51.5	38.1	44.6	46.6	37.0	46.9	48.0	37.8	45.5	32.6	34.8	37.1	35.0	42.2	42.8	41.4
	Ours	**54.4**	**40.0**	46.1	47.6	**38.7**	50.1	**51.0**	**42.4**	**49.3**	**34.6**	**37.6**	39.8	**39.3**	**45.8**	**46.6**	**44.2**

Table 5: Accuracy of XNLI in 15 languages based on BLOOM-7b1, XGLM-7.5b and Llama-2-7b models.

XCOPA(acc.)													
Model		en	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	AVG
BLOOMZ-1b7		37.2	25.4	35.6	55.6	30.8	26.8	37.8	60.6	50.0	25.8	58.0	54.8	41.5
k=2	random	62.6	49.8	50.0	59.9	46.1	48.5	52.0	58.1	**50.0**	47.9	59.5	64.3	54.1
	sem	61.2	50.6	46.6	61.8	47.6	46.6	51.4	57.0	**50.0**	**50.0**	55.2	67.6	53.8
	cluster	61.9	**51.7**	48.2	61.3	47.4	47.8	51.4	57.7	49.9	49.1	59.8	65.9	54.3
	Ours	**67.6**	51.4	**51.8**	**62.8**	**52.8**	**50.2**	**53.0**	**62.4**	**50.0**	49.6	**64.0**	**69.8**	**57.1**
k=3	random	65.0	51.1	48.1	**61.9**	48.0	49.1	51.7	58.6	50.5	48.7	61.5	66.4	55.0
	sem	59.6	**52.6**	**51.0**	60.0	44.8	44.6	51.8	56.2	49.8	48.8	57.6	60.4	53.1
	cluster	60.9	51.3	49.4	59.5	49.3	48.4	51.5	59.0	**51.1**	46.9	60.4	62.1	54.2
	Ours	**69.2**	51.2	50.4	61.6	**52.8**	**50.6**	**52.4**	**62.4**	50.4	**49.4**	**63.0**	**69.6**	**56.9**
k=4	random	62.1	49.3	49.1	59.1	49.1	49.5	50.7	58.9	49.8	49.3	60.7	**67.1**	54.6
	sem	58.0	49.6	50.6	61.4	49.2	45.8	**52.2**	59.2	**50.8**	**51.0**	58.6	57.6	53.7
	cluster	63.2	49.7	48.9	59.5	**50.9**	**50.5**	50.8	59.6	50.3	50.1	61.0	63.5	54.8
	Ours	**66.8**	**53.0**	**52.0**	**61.6**	50.4	49.8	51.4	**60.6**	50.6	49.4	**61.2**	66.4	**56.1**
BLOOMZ-3b		42.2	10.8	32.8	59.2	27.2	22.2	31.6	53.2	51.2	16.0	59.8	80.8	40.6
k=2	random	47.3	32.1	40.8	60.7	37.6	32.1	42.5	58.5	47.7	31.4	60.5	67.8	46.6
	sem	68.2	48.0	47.2	63.2	**48.0**	47.6	49.4	56.6	50.8	**48.4**	62.2	71.2	55.1
	cluster	65.9	49.1	48.8	64.4	47.3	48.7	50.3	59.7	**50.7**	46.8	63.5	70.1	55.4
	Ours	**69.4**	**49.6**	**51.0**	**65.0**	47.6	**48.8**	**51.2**	**60.8**	50.0	46.4	**66.8**	**72.2**	**56.6**
k=3	random	59.5	38.9	43.6	63.4	43.2	38.5	46.3	60.7	49.3	38.9	63.1	71.5	51.4
	sem	68.0	49.2	50.0	**64.6**	46.8	48.2	52.0	58.2	48.4	**47.4**	63.8	69.2	55.5
	cluster	68.5	48.8	50.0	63.3	46.5	**50.5**	51.4	59.3	**51.1**	45.8	64.6	70.7	55.9
	Ours	**68.9**	**50.0**	**51.2**	**64.6**	**47.0**	50.0	**53.2**	**59.6**	51.0	46.2	**65.2**	**71.6**	**56.5**
k=4	random	**70.9**	49.1	49.5	63.1	48.3	48.8	50.4	59.3	50.9	48.0	63.6	68.8	55.9
	sem	68.0	**49.2**	**53.0**	**63.8**	48.2	46.0	51.0	58.0	48.0	**49.8**	**64.6**	66.4	55.5
	cluster	70.0	48.8	50.4	63.4	**49.6**	**50.1**	51.5	57.5	49.5	48.9	63.7	69.9	56.1
	Ours	68.4	49.2	52.6	62.6	46.8	49.4	**51.6**	**59.8**	**52.8**	46.8	64.0	**70.6**	**56.2**
BLOOMZ-7b1		60.2	49.4	48.6	74.0	48.2	46.8	55.0	72.2	49.8	44.2	71.0	77.2	58.1
k=2	random	66.3	48.7	49.3	71.8	54.7	51.1	57.3	69.1	51.2	49.5	71.1	77.4	59.8
	sem	73.4	48.4	48.4	72.6	53.4	50.2	58.0	67.2	50.6	50.4	68.6	75.8	59.8
	cluster	76.8	48.3	49.0	72.0	54.1	51.1	54.6	66.5	51.2	50.4	71.5	75.7	60.1
	Ours	**78.4**	**50.8**	**51.2**	**73.6**	**57.6**	**52.8**	**59.8**	**71.2**	**52.2**	**52.4**	**74.4**	**79.9**	**62.9**
k=3	random	73.1	48.6	49.0	72.9	56.5	52.0	57.7	70.2	51.1	50.7	73.3	**80.2**	61.3
	sem	75.6	50.4	49.4	69.4	54.2	52.2	54.8	66.8	50.0	46.2	71.0	74.2	59.5
	cluster	77.3	49.7	49.0	72.3	56.0	50.0	54.3	66.7	50.7	49.4	69.9	75.6	60.1
	Ours	**79.8**	**51.8**	**51.2**	**73.2**	**58.4**	**53.2**	**57.6**	**70.8**	**52.3**	**54.8**	**74.6**	79.8	**63.1**
k=4	random	78.7	48.0	48.1	69.8	56.6	51.4	**58.9**	**71.1**	51.5	50.5	72.9	79.9	61.5
	sem	74.8	50.2	49.4	70.6	55.8	49.8	54.6	68.0	48.8	52.0	72.4	75.6	60.2
	cluster	77.2	49.5	49.8	71.1	56.3	52.3	56.1	67.5	50.5	50.9	70.5	75.3	60.6
	Ours	**81.0**	**50.9**	**52.0**	**74.6**	**57.3**	**52.8**	57.1	70.0	**53.6**	**51.8**	**72.6**	**81.6**	**62.9**

Table 6: Accuracy of XCOPA in 12 languages based on BLOOMZ models

XCOPA(acc.)														
Model		en	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	AVG
BLOOM-7b1		30.6	56.4	47.8	50.2	49.8	49.2	49.8	49.4	11.8	59.4	44.6	46.8	45.5
k=2	random	**50.5**	49.5	49.6	49.8	50.1	50.9	49.6	49.3	50.3	49.9	50.9	49.6	50.0
	sem	49.2	49.0	44.8	**52.2**	47.0	**52.6**	50.2	48.4	50.2	49.6	49.2	49.0	49.3
	cluster	50.3	49.5	**50.5**	50.5	49.8	49.9	50.5	50.3	50.1	**50.1**	51.0	50.5	50.3
	Ours	50.4	**51.0**	50.0	50.6	**50.8**	52.4	**51.2**	**51.0**	**51.0**	49.6	**52.2**	**52.3**	**51.0**
k=3	random	49.5	**51.1**	49.4	49.9	50.3	51.9	**50.4**	50.5	50.3	50.4	49.5	49.9	50.3
	sem	47.4	50.0	**51.0**	**52.8**	46.8	48.2	47.8	**51.4**	48.2	48.6	46.6	47.6	48.9
	cluster	49.1	48.6	49.7	49.3	48.2	49.4	50.3	48.9	50.2	49.5	50.7	48.5	49.4
	Ours	**51.8**	49.6	50.8	50.8	**50.6**	**52.0**	50.2	50.6	**51.8**	**51.2**	**50.8**	**51.2**	**51.0**
k=4	random	50.4	49.6	49.1	49.2	**50.8**	50.9	50.0	50.3	50.5	50.7	49.8	49.9	50.1
	sem	45.0	49.0	**51.4**	**51.8**	49.0	46.6	49.2	50.8	48.2	51.8	**50.0**	47.4	49.2
	cluster	50.2	49.7	50.3	50.1	50.0	**51.3**	50.3	50.4	50.1	50.3	49.9	50.6	50.3
	Ours	**52.0**	**50.6**	50.4	50.6	**50.8**	51.0	**52.6**	**51.4**	**51.6**	**53.3**	**50.0**	**53.4**	**51.5**
XGLM-7.5b		34.2	45.4	48.6	48.6	49.4	47.4	50.2	47.0	51.6	50.0	50.0	49.2	47.6
k=2	random	49.8	50.5	49.7	50.3	**54.6**	50.1	49.2	50.3	49.7	**51.7**	50.5	50.5	50.6
	sem	48.6	52.2	46.0	51.2	49.2	**53.0**	**50.4**	46.2	51.4	50.2	46.2	50.4	49.6
	cluster	49.2	49.1	49.2	50.4	50.3	48.9	46.5	48.9	49.9	50.6	50.3	49.5	49.4
	Ours	**50.2**	**54.8**	**51.6**	**54.6**	53.6	51.4	51.6	**54.0**	**51.6**	50.8	**51.6**	**53.0**	**52.4**
k=3	random	50.4	52.9	49.0	50.9	58.9	50.0	48.7	**51.1**	49.8	53.4	50.5	50.8	51.4
	sem	48.6	53.2	52.0	**53.8**	53.2	49.6	47.4	54.6	49.2	54.0	**50.6**	49.4	51.3
	cluster	50.5	51.5	50.1	51.3	56.2	50.2	48.9	48.7	49.7	53.5	49.7	49.8	50.8
	Ours	**53.2**	**55.6**	**52.2**	53.4	**56.6**	**50.8**	**51.4**	50.6	**51.0**	**55.0**	50.0	**53.1**	**52.7**
k=4	random	49.8	52.8	50.0	50.1	54.7	50.3	48.9	49.7	50.3	51.7	50.0	50.3	50.7
	sem	47.6	50.6	51.6	52.4	52.0	47.0	50.6	**51.4**	48.4	52.6	49.4	47.0	50.1
	cluster	51.8	51.9	48.9	50.3	55.5	50.0	49.1	49.8	50.6	52.6	49.3	50.3	50.9
	Ours	52.0	**54.0**	**52.0**	**54.6**	**55.8**	**51.4**	**51.0**	**51.4**	**50.8**	**53.2**	**51.5**	**51.0**	**52.4**
Llama-2-7b		48.6	18.4	16.0	15.4	29.0	23.2	20.6	23.8	39.0	15.6	15.8	26.2	24.3
k=2	random	82.0	49.0	48.3	61.1	68.8	50.3	49.4	48.6	51.5	54.4	57.8	64.2	57.1
	sem	79.6	50.2	46.8	59.4	68.2	48.8	49.4	48.8	**54.6**	54.0	**63.2**	65.4	57.4
	cluster	80.7	50.0	50.6	59.8	69.0	50.4	50.3	48.9	52.1	53.3	57.8	66.1	57.4
	Ours	**84.0**	**51.0**	**51.8**	**63.4**	**72.6**	**52.0**	**52.4**	**51.2**	54.4	**55.6**	62.0	**67.2**	**59.8**
k=3	random	77.6	48.9	49.9	62.4	68.4	50.5	48.7	47.5	52.1	55.3	60.2	64.0	57.1
	sem	78.8	50.6	52.2	62.4	71.0	50.6	47.6	49.2	51.4	56.0	**62.2**	67.2	58.3
	cluster	81.9	48.8	51.3	63.0	70.3	49.8	49.9	49.1	**54.3**	54.0	59.3	67.1	58.2
	Ours	**84.4**	**52.0**	**53.6**	**64.0**	**72.8**	**51.8**	**51.6**	**51.4**	53.8	**56.1**	62.0	**69.8**	**60.3**
k=4	random	79.3	50.1	51.2	60.2	69.3	50.4	51.9	49.0	53.5	54.0	59.0	64.1	57.7
	sem	80.8	**52.0**	47.0	61.6	69.6	51.2	51.4	47.4	51.2	54.0	60.6	65.4	57.7
	cluster	81.4	50.5	49.5	60.9	69.9	51.7	50.5	48.5	52.5	54.5	59.0	66.1	57.9
	Ours	**84.4**	51.6	**52.4**	**64.6**	**72.2**	**52.2**	**52.8**	**49.6**	**54.2**	**54.6**	**64.4**	**70.6**	**60.3**

Table 7: Accuracy of XCOPA in 12 languages based on BLOOM-7b1, XGLM-7.5b and Llama-2-7b models.

Model		ar	bg	en	fi	id	ko	ru	sw	te	AVG
BLOOMZ-1b7		76.1	80.5	65.9	4.0	79.0	4.6	27.4	76.2	85.8	55.5
k=2	random	81.4	**86.7**	71.2	8.2	**83.0**	6.4	39.9	81.1	87.6	60.6
	sem	79.0	84.8	71.8	7.7	82.1	5.8	39.9	79.2	87.2	59.7
	cluster	**81.5**	85.5	71.2	8.4	82.3	7.0	39.1	79.6	87.5	60.2
	Ours	79.8	85.7	**72.5**	**21.6**	80.9	**21.7**	**42.1**	**81.6**	**88.1**	**63.8**
k=3	random	**81.0**	**87.3**	71.9	9.6	**83.0**	5.8	39.9	81.3	87.5	60.8
	sem	80.3	86.4	71.2	8.0	81.8	6.7	39.8	80.2	87.1	60.2
	cluster	80.8	86.5	71.1	8.2	81.7	6.3	38.8	79.6	**87.6**	60.1
	Ours	79.6	87.1	**73.0**	**21.9**	81.1	**22.2**	**42.2**	**81.4**	87.4	**64.0**
k=4	random	**81.2**	**87.2**	73.1	10.2	**83.2**	6.4	39.8	81.2	87.7	61.1
	sem	79.3	87.1	72.1	6.9	82.6	7.4	39.7	79.9	87.4	60.3
	cluster	80.7	86.4	70.8	9.1	81.8	6.5	39.0	79.2	87.4	60.1
	Ours	79.8	85.7	**72.5**	**21.6**	80.9	**21.7**	**42.1**	**81.6**	**88.1**	**63.8**
BLOOMZ-3b		79.6	88.4	74.5	10.5	81.0	11.0	36.8	76.5	88.3	60.7
k=2	random	82.6	89.0	77.4	13.3	**83.7**	12.1	43.1	**82.7**	89.0	63.7
	sem	82.1	89.0	77.8	14.2	83.0	12.7	43.1	82.6	89.4	63.8
	cluster	82.4	88.3	76.9	13.5	82.7	11.8	42.7	82.4	89.1	63.3
	Ours	**82.8**	**90.7**	**78.6**	**27.0**	83.2	**27.2**	**45.0**	82.1	**89.5**	**67.3**
k=3	random	**82.8**	**88.5**	78.0	14.9	**83.0**	11.7	43.5	**82.4**	89.3	63.8
	sem	82.6	88.4	78.8	13.9	82.4	12.7	43.9	82.2	89.3	63.8
	cluster	82.2	88.3	77.1	13.7	82.1	11.4	42.3	81.7	89.2	63.1
	Ours	**82.8**	**89.7**	**78.9**	**27.5**	82.7	**27.1**	**45.2**	82.2	**89.5**	**67.3**
k=4	random	82.5	89.0	77.8	16.1	**83.5**	13.1	44.3	**83.5**	89.3	64.3
	sem	82.1	88.7	**78.8**	14.3	83.5	12.5	42.3	81.8	**89.4**	63.7
	cluster	82.3	89.0	77.5	13.8	82.7	12.2	42.0	82.4	89.2	63.5
	Ours	**82.6**	**89.4**	78.5	**27.8**	82.5	**27.4**	**45.8**	81.5	89.2	**67.2**
BLOOMZ-7b1		77.6	86.8	70.8	8.8	67.8	16.1	34.8	70.3	85.8	57.6
k=2	random	**82.1**	88.7	**78.2**	18.3	**76.0**	18.2	**47.7**	**84.0**	88.2	64.6
	sem	81.4	**89.2**	77.1	20.0	75.6	18.7	44.0	81.1	88.2	63.9
	cluster	81.9	88.8	71.8	18.7	74.4	18.7	45.8	82.1	88.2	63.4
	Ours	79.5	87.7	77.9	**30.2**	73.2	**30.0**	46.5	82.1	**88.3**	**66.1**
k=3	random	**81.8**	88.4	**79.1**	21.4	**75.7**	18.2	**46.8**	**84.2**	**88.3**	64.9
	sem	80.0	89.1	78.5	20.1	74.4	18.7	45.9	81.8	87.8	64.0
	cluster	81.6	**89.4**	72.1	19.6	73.9	18.1	45.1	81.9	88.1	63.3
	Ours	80.2	88.5	77.2	**30.1**	74.1	**30.5**	46.6	81.5	88.1	**66.3**
k=4	random	**82.3**	**88.9**	**80.3**	23.9	**76.0**	20.0	**47.4**	**84.4**	87.9	65.7
	sem	79.3	88.0	77.3	19.4	74.2	17.2	45.2	81.0	**88.2**	63.3
	cluster	81.3	88.0	71.9	19.1	73.4	18.3	45.6	81.5	88.0	63.0
	Ours	79.2	86.7	78.1	**30.2**	72.6	**31.0**	46.6	80.6	87.8	**65.8**

Table 8: F1 scores of TydiQA-GoldP in 9 languages based on BLOOMZ models.

TydiQA-GoldP(F1)											
Model		ar	bg	en	fi	id	ko	ru	sw	te	AVG
BLOOM-7b1		26.7	21.3	20.9	5.9	26.9	3.8	12.9	20.0	16.4	17.2
k=2	random	29.0	24.6	22.5	5.5	28.2	3.8	12.6	18.7	12.7	17.5
	sem	30.0	24.1	23.9	5.6	**29.3**	3.9	15.7	**20.7**	13.3	18.5
	cluster	**30.3**	24.2	23.7	5.5	28.9	3.7	15.3	19.9	12.9	18.3
	Ours	30.2	**28.7**	**26.8**	**19.0**	28.8	**19.3**	**18.2**	19.7	**17.7**	**23.1**
k=3	random	27.8	24.0	22.5	6.7	27.3	3.2	13.2	20.4	15.5	17.8
	sem	**30.5**	26.5	23.9	5.4	**28.9**	3.9	15.9	20.2	13.3	18.7
	cluster	30.0	24.3	23.6	5.5	28.2	3.8	14.6	20.0	13.4	18.2
	Ours	30.4	**28.4**	**26.6**	**18.8**	28.4	**19.2**	**18.2**	**20.5**	**17.6**	**23.1**
k=4	random	30.4	25.9	23.2	5.5	**29.2**	4.1	15.8	20.0	13.4	18.6
	sem	**31.0**	24.5	24.0	5.6	28.7	3.7	15.6	**20.9**	13.1	18.6
	cluster	30.4	24.9	23.8	5.3	28.9	3.8	14.8	20.4	13.3	18.4
	Ours	30.7	**29.4**	**27.3**	**19.5**	28.5	**19.6**	**18.5**	19.9	**18.2**	**23.5**
XGLM-7.5b		23.6	18.7	8.5	12.6	10.8	8.7	7.9	25.2	25.8	15.8
k=2	random	26.1	20.3	13.2	15.6	18.8	14.1	11.6	21.7	27.8	18.8
	sem	27.0	21.6	17.1	17.6	**21.5**	16.1	13.4	**23.8**	**28.2**	20.7
	cluster	26.7	17.6	15.4	16.6	18.7	13.6	12.1	20.9	27.3	18.8
	Ours	**27.4**	**24.6**	**21.9**	**20.0**	20.3	**19.1**	**17.6**	22.7	28.1	**22.4**
k=3	random	25.9	20.0	13.6	16.7	18.5	13.7	12.0	20.9	27.3	18.7
	sem	**26.4**	19.5	18.0	19.1	**21.1**	15.3	12.6	**23.5**	27.2	20.3
	cluster	26.4	19.2	16.7	18.5	19.4	13.4	12.2	22.1	**28.0**	19.5
	Ours	25.6	**24.4**	**21.9**	**20.8**	21.0	**19.9**	**18.3**	22.9	27.4	**22.5**
k=4	random	**26.7**	20.1	16.1	18.4	20.4	14.1	12.9	21.4	**27.7**	19.8
	sem	25.7	20.0	20.2	19.2	**21.9**	15.9	12.6	**24.4**	27.4	20.8
	cluster	26.4	20.0	18.3	18.1	20.0	13.9	12.5	22.1	26.6	19.8
	Ours	26.6	**25.5**	**22.8**	**21.6**	21.2	**20.3**	**18.5**	24.1	27.2	**23.1**
Llama-2-7b		7.1	4.6	40.7	36.2	30.8	7.5	20.2	17.3	0.4	18.3
k=2	random	11.3	2.2	60.3	43.9	43.9	16.5	28.2	24.5	4.6	26.2
	sem	14.7	2.1	61.7	45.4	43.9	21.7	29.5	27.4	6.5	28.1
	cluster	15.8	2.5	63.2	44.1	44.4	20.1	29.6	26.7	3.2	27.7
	Ours	**16.9**	**3.2**	**69.2**	**50.3**	**51.7**	**21.9**	**34.9**	**37.7**	**10.4**	**32.9**
k=3	random	13.8	2.5	64.1	46.6	47.7	19.0	29.9	32.4	**10.1**	29.6
	sem	**15.4**	3.5	65.2	47.0	47.1	20.2	31.8	30.6	8.5	29.9
	cluster	14.8	2.2	65.9	45.4	45.2	20.4	30.4	27.8	7.5	28.8
	Ours	16.6	**4.6**	**68.8**	**52.1**	**52.2**	**24.5**	**35.2**	**39.4**	8.6	**33.5**
k=4	random	13.9	2.4	**66.9**	49.5	50.1	21.8	32.7	36.0	10.1	31.5
	sem	14.3	3.3	66.6	48.4	47.2	18.2	32.7	29.9	11.0	30.2
	cluster	14.6	2.8	66.0	48.3	47.4	21.8	31.3	32.2	8.6	30.3
	Ours	**16.8**	**5.5**	64.9	**52.3**	**54.6**	**28.9**	**33.7**	**39.8**	**12.0**	**34.3**

Table 9: F1 score of TydiQA-GoldP in 9 languages based on BLOOM-7b1, XGLM-7.5b and Llama-2-7b models.