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 or task-specific retrievers based on LLM feedback signals for demonstration selection. However, these approaches often neglect factors beyond semantic similarity and can be resource-intensive. To address these challenges, we propose a novel approach called Topic-XICL, which leverages a latent topic model to select demonstrations for XICL. We assume that latent topic variables encapsulate information that more accurately characterizes demonstrations. By training this topic model on rich-resource language data with a small-parameter LLM, we obtain more informative demonstrations through 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) and three models with approximately 7 billion parameters, including two multilingual LLMs (BLOOM and XGLM), and an English-centric model, Llama2. Comparative evaluations against baselines of random selection, semantic similarity selection, and clustering-based selection show consistent improvements in multilingual 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.

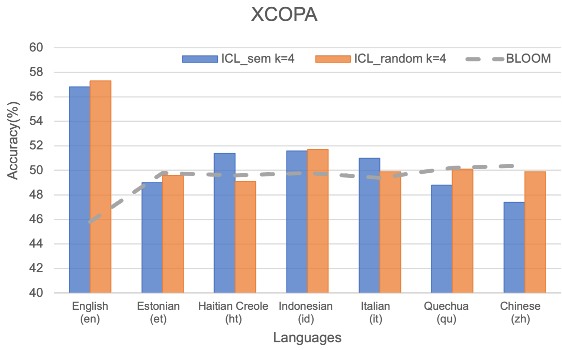

Figure 1: Accuracy scores for 7 languages from the XCOPA dataset (Gordon et al., 2012) using direct inference (dashed line) or 4-shot in-context learning (ICL) with the BLOOM model (Scao et al., 2022) (7.1 billion parameters). k represents the number of demonstrations. "sem" denotes semantic-based selection, while "random" denotes random selection.

The impressive comprehension abilities of LLMs in English have sparked interest in Cross-lingual In-Context Learning (XICL)(Winata et al., 2021; Lin et al., 2022; Asai et al., 2023; Cahyawijaya et al., 2024; Zhang et al., 2024). 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; Qin et al., 2023; Cahyawijaya et al., 2024). 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; Cahyawijaya et al., 2024), 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 Haitian Creole (ht) and Italian (it), while randomly selected diversity examples lead to better performance for Quechua (qu) and Chinese (zh). 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 latent 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, and Wang et al. (2023) treated LLMs as topic models to apply the theory, which proved productive in demonstration selection for classification tasks. Inspired by this, we extended Wang et al. (2023)'s approach to cross-lingual in-context learning and more tasks, proposing a demonstration selection algorithm based on topic inference (Topic-XICL), as shown in Fig. 2. It comprises a **latent topic learning** phase and a **demonstration selection** phase. In the latent topic learning phase, demonstration candidates 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 cluster the candidate data for a task into n topics. For each topic, we introduce c new tokens to enrich the LLM's vocabulary. These tokens are concatenated with the input to predict the output, enabling the LLM to update the embeddings of these new tokens. During the demonstration selection phase, we perform topic inference on the candidate data, selecting the k most representative examples for each topic. For each target language input, we determine its topic by calculating semantic similarity with the candidate data and using the corresponding representative examples as the context.

We trained the latent topic model on BLOOMZ-1b7 (Muennighoff et al., 2023) (with 1.7 billion parameters) and conducted cross-lingual ICL on two multilingual sentence-level tasks and one cross-lingual reading comprehension task.

Our contributions are summarized as follows:

- We propose a cross-lingual demonstration selection algorithm based on topic inference (Topic-XICL), extending Bayesian inference theory to practical applications in cross-lingual ICL.
- Intuitively, the Bayesian theorem is primarily suited for classification tasks. To our knowledge, we are the first to apply it to non-classification tasks on XICL, and we have experimentally validated its effectiveness.
- We compared our method with three demonstration selection baselines using three LLMs (BLOOM, XGLM, and Llama2) on three cross-lingual tasks (XNLI, XCOPA, and TyDiQA-GoldP). The results show that our topic-based demonstration selection 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, as demonstrated by the XICL method (Winata et al., 2021; Lin et al., 2022). Winata et al. (2021) first showed that, given a few English examples as context, multilingual pre-trained language models (such as 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 that their XGLM demonstrates strong cross-lingual capability, where using English prompts together with non-English examples yields competitive zero- and few-shot learning performance. Cahyawijaya et al. (2024) extensively studied XICL on some low-resource languages from four aspects: cross-lingual alignment, alignment formatting, label configuration, and cross-lingual retrieval, highlighting the importance of advancing ICL research. Our research mainly focuses on the aspect of cross-lingual retrieval to select demonstrations for XICL.

Cross-lingual Demonstration Selection Different rich-resource language demonstrations yield vary-

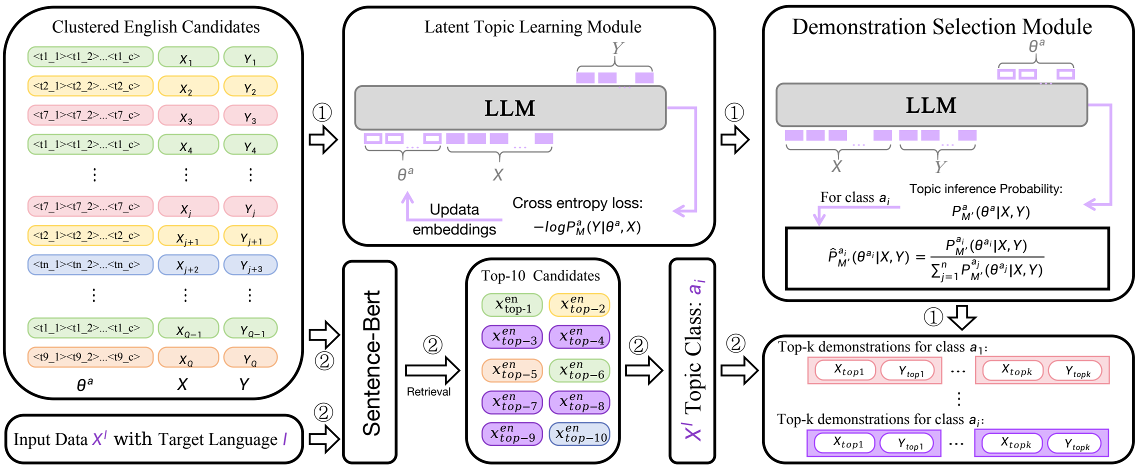

Figure 2: An overview of our proposed cross-lingual demonstration selection framework with topic inference.① Latent topic embeddings are learned for the clustered English candidates using LLMs, and probabilities of inferring to n topics are calculated for each candidate. The top-k representative demonstrations for each topic are then obtained. ② For each target input, the semantic relationship with the candidates is calculated. The most frequent topic in the top-10 examples is used as its classification topic, denoted as a_i. The k most representative examples in the a_i topic are used as the context for the target input, which can be used for ICL in any generative LLM.

ing XICL outcomes for target languages. Current cross-lingual retrieval methods fall into two categories: using off-the-shelf multilingual representations and leveraging LLM feedback signals. For example, Nie et al. (2023) conducts cross-lingual retrieval from labeled or unlabeled high-resource languages based on the semantic similarity of multilingual embeddings. Li et al. (2023) extended this to focus on zero-shot settings, revealing limitations for complex generation tasks. Tanwar et al. (2023) augmented prompts with cross-lingual semantic similarity demonstrations and in-context label alignment, but Cahyawijaya et al. (2024) identified shortcomings and introduced translation pairs for alignment. Additionally, Winata et al. (2023) emphasized semantic similarity by selecting the nearest examples from various sub-datasets for classification tasks. In contrast, Shi et al. (2022) proposed a retrieve-rerank framework for cross-lingual Text-to-SQL, using a bi-encoder to identify relevant exemplars, and then training a retriever by distilling the LLM's scoring function.

Training retrievers on specific task data and LLMs can be advantageous, but managing inaccessible parameters of black-box models is challenging. Our method trains using only accessible LLMs. Semantic similarity alone may not suffice for complex tasks, so we expect to integrate richer information into "latent topics," such as article types in question-answering tasks, question types, and the structural relationship between answers and articles. We use LLMs to mine this latent topic information and select demonstrations to enhance cross-lingual in-context learning.

In-Context Learning with Bayesian inference Xie et al. (2022) provided a latent topic interpretation to explain in-context learning, showing that the in-context learning predictor approaches the Bayes optimal predictor as the number of demonstrations increases, assuming both pre-training and task-specific data follow Hidden Markov Models (HMM). However, the Markovian assumption about data generation limits empirical validation to synthetic data and toy models, raising questions about its applicability to natural language.

To bridge the gap between theoretical understanding and real-world LLM algorithms, Wang et al. (2023) developed a practical demonstration selection algorithm. Our method extends Wang et al. (2023) to an XICL setting. Unlike their approach, which treats each classification data as a topic, we perform semantic clustering on each task's data to obtain topics, making our approach applicable to a wider range of tasks. To our knowledge, this is the first attempt to use Bayesian theory for demonstration selection beyond classification.

3 Method

Based on the theoretical understanding and practical algorithm of Bayesian inference in ICL, we proposed a cross-lingual demonstration selection framework (as shown in Fig. 2) with topic inference to improve the performance of XICL for various tasks. 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.

3.1 Notations and Problem Setting

In cross-lingual in-context learning, the prompt comprises k rich-resource language demonstrations $(X_1, Y_1), (X_2, Y_2), ..., (X_k, Y_k)$ and a low-resource target language test input X, and the gold truth is $Y \in \mathbf{Y}$. For the generation-form task based on decoder-only LLMs, $\mathbf{Y}$ is the space of all possible token sequences. Similar to that of the topic model, a simplified assumption can be made for LLM (denoted by M):

$$P_M(Y|X) = \int_{\Theta} P_M(Y|\theta) P_M(\theta|X) d\theta, \quad (1)$$

$\theta \in \Theta$ 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 $X \to Y \leftarrow \theta$, which can be represented mathematically as the following structural equation:

$$Y^a = \mathrm{f}(X^a, \theta^a, \epsilon), \quad (2)$$

where ϵ is an independent noise variable. a is the topic of (X, Y), and $\theta^a \in \Theta$ is the value of the topic variable corresponding to the 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:

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

It is 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

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

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, the objection of Bayes optimal decoder is 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:

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

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 $\sum_{i=1}^{n} \mathcal{L}(\hat{\theta}^{a_i})$.

3.3 Demonstration Selection

About the topic of target instance (X^l, Y^l), we embed the input X^l and measured its semantic similarity with all source input embeddings by Sentence-BERT (Reimers and Gurevych, 2019). 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:

$$\arg\max_{X_1^a, Y_1^a, ..., X_k^a, Y_k^a} \mathbb{E}_X[P_M^a(\theta^a | X_1^a, Y_1^a, ..., X_k^a, Y_k^a, X)] \quad (6)$$

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

$$\begin{aligned} &\arg\max_{X_1^a, Y_1^a, ..., X_k^a, Y_k^a} P_M^a(\theta^a | X_1^a, Y_1^a, ..., X_k^a, Y_k^a) \\ &= \frac{\prod_{i=1}^{k} P_M^a(\theta^a | X_i^a, Y_i^a)}{P_M^a(\theta^a)^{k-1}} \end{aligned} \quad (7)$$

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

For the setting of n, the estimated conditional probability of $\hat{\theta}^{a_i}$ for instance (X, Y) would be:

$$\hat{P}^{a_i}_{M'}(\hat{\theta}^{a_i}|(X,Y)) = \frac{P^{a_i}_{M'}(\hat{\theta}^{a_i}|(X,Y))}{\sum_{j=1}^{n} P^{a_j}_{M'}(\hat{\theta}^{a_j}|(X,Y))} \quad (8)$$

We mainly focus on the fundamental effects of topic inference on multilingual demonstration selection, without discussion of 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-GoldP (Clark et al., 2020). The Cross-lingual Natural Language Inference dataset (XNLI) is a **sentence-pair classification** task involving 15 languages, translated from the English SNLI (Bowman et al., 2015) dataset. Since existing work mainly discusses demonstration selection methods for classification tasks, we also explored the multilingual **causal commonsense reasoning** task XCOPA and the **Question Answering** (QA) task in our experiments. XCOPA is an extension and re-annotation of the English Choice of Plausible Alternatives (COPA) dataset (Gordon et al., 2012), with validation and test examples translated and annotated in 11 typologically diverse languages. TyDiQA-GoldP is the gold passage task in TyDiQA (Clark et al., 2020), covering 9 typologically diverse languages and serving as a challenging multilingual QA benchmark.

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 list the English training set volume, 24 target languages, and their test set sizes in Table 4. The XCOPA test set is a combination of the official open-source 100 validation sets and 400 test sets. Due to the large size of the XNLI training dataset (392,701 instances in total), we only used the first 10,000 instances.

4.2 Experimental Setting

We employ the K-means algorithm with random initial center points to cluster the training set $\mathcal{D}$, using three seed values $[32, 44, 100]$ and reporting the average results and standard deviation per language for $k = [2, 3, 4]$. Each training data representation is obtained using multilingual Sentence-BERT[2]. As for hyper-parameters, 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). The guidelines for the hyper-parameters section can be seen in A.

We leverage the Bloomz-1b7[3] model to learn the topic token embeddings and compute the probability of each candidate. BLOOMZ-1b7 (Muennighoff et al., 2023) is a multilingual supervised fine-tuning version of BLOOM, which may be more efficient for learning the topic of a task. Greedy Search is employed for decoding answers in each task. For XCOPA, the gold output is changed to "1" or "2". For two-sentence tasks, we set the output length to 1 to obtain the answer label. For the QA task, the maximum output length is 30, and the metric is F1. The prompts used for each task are detailed in Appendix B.

4.3 Baselines

We use the same set of demonstrations for three LLMs, each with about 7 billion parameters, including BLOOM, XGLM, and Llama-2. We consider the following demonstration selection methods as baselines:

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

ICL_sem: We use the same sentence-BERT to calculate the cosine similarity between the inputs of the source and target language. We select the top k demonstrations from $\mathcal{D}$ for each test example.

Cluster: Since 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

Table 1 presents our main results for the three datasets averaged over all languages baseline on

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

[2] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

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

Model	Method	XNLI (accuracy, %)			XCOPA (accuracy, %)			TidyQA-GoldP (F1, %)		
		k=2	k=3	k=4	k=2	k=3	k=4	k=2	k=3	k=4
BLOOM	Zero-shot		32.8			49.6			20.3	
	ICL_random	35.3(0.004)	34.8(0.014)	34.3(0.059)	51.3(0.040)	51.4(0.033)	51.3(0.059)	26.8(0.001)	27.9(0.001)	29.9(0.001)
	ICL_sem	36.6(0.000)	36.9(0.000)	37.2(0.000)	50.7(0.160)	50.4(0.250)	51.5(0.056)	29.3(0.001)	29.4(0.001)	29.3(0.001)
	ICL_cluster	34.4(0.031)	35.2(0.003)	36.1(0.001)	51.7(0.027)	51.0(0.128)	51.9(0.036)	28.6(0.001)	27.9(0.001)	28.4(0.001)
	Topic-XICL(ours)	**37.4(0.000)**	**37.9(0.000)**	**37.4(0.000)**	**53.9(0.000)**	**54.5(0.000)**	**54.4(0.000)**	**36.2(0.000)**	**34.6(0.000)**	**35.7(0.000)**
XGLM	Zero-shot		32.3			49.7			15.8	
	ICL_random	34.4(0.002)	35.0(0.000)	35.8(0.000)	50.8(0.079)	51.6(0.041)	50.9(0.074)	18.8(0.010)	18.7(0.015)	19.8(0.008)
	ICL_sem	35.5(0.000)	35.8(0.000)	35.4(0.001)	50.5(0.169)	52.2(0.002)	52.2(0.000)	20.7(0.002)	20.3(0.004)	20.8(0.004)
	ICL_cluster	35.2(0.000)	35.8(0.000)	36.0(0.000)	50.5(0.088)	51.9(0.005)	52.1(0.002)	18.8(0.023)	19.5(0.007)	19.8(0.009)
	Topic-XICL(ours)	**35.7(0.000)**	**36.4(0.000)**	**36.6(0.000)**	**53.1(0.000)**	**53.5(0.000)**	**53.1(0.000)**	**24.8(0.000)**	**24.4(0.001)**	**24.5(0.001)**
Llama2	Zero-shot		39.6			50.6			24.1	
	ICL_random	41.6(0.000)	41.3(0.001)	41.4(0.002)	57.1(0.005)	57.1(0.002)	57.7(0.001)	28.2(0.043)	31.1(0.005)	33.1(0.001)
	ICL_sem	42.0(0.000)	42.9(0.000)	43.6(0.000)	57.4(0.004)	58.3(0.002)	57.7(0.003)	29.0(0.019)	31.0(0.006)	32.1(0.003)
	ICL_cluster	41.1(0.001)	42.1(0.000)	42.5(0.000)	57.4(0.003)	58.2(0.002)	57.9(0.002)	31.3(0.010)	32.4(0.005)	33.7(0.001)
	Topic-XICL(ours)	**42.8(0.000)**	**43.4(0.000)**	**44.2(0.000)**	**60.0(0.001)**	**60.4(0.000)**	**60.6(0.001)**	**41.4(0.000)**	**42.2(0.000)**	**42.7(0.000)**

Table 1: Average performance across languages for three tasks with different numbers of demonstrations. Parentheses contain the p-values from the statistical significance analysis of the ICL methods and zero-shot baseline results, with those greater than 0.05 marked with a gray background. We also calculated the standard deviation over 3 seeds for ICL_random, ICL_cluster, and Topic-XICL, as shown in Appendix D.

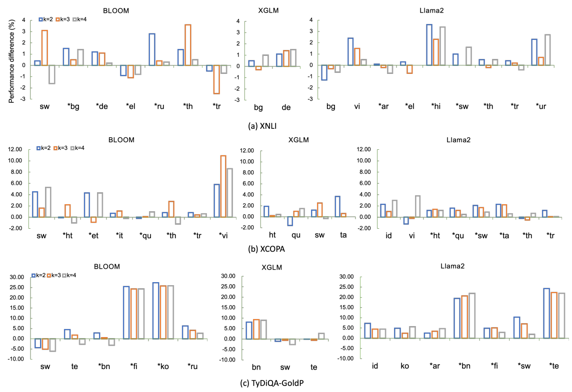

Figure 3: Performance difference between 4-shot Topic-XICL and best baseline results for individual languages in Three datasets."*" represents the language is unseen for the models

three LLMs, along with the p-values from significance analysis of the ICL methods and the zero-shot. Detailed results can be found in Appendix D. Across all three datasets, our method consistently outperforms the baselines on three models with different lengths of demonstrations. Figure 3 illustrates the performance difference between Topic-XICL and the best baseline results for individual low-resource languages across the three datasets, and languages marked with a "*" signal are unseen languages for the models. Please refer to Appendix C for definitions of the languages.

For classification task XNLI, our method can achieve significant gains when $k = 3$, such as the average performance of our method improves by 1.0% over the best baseline on the BLOOM model. In other cases, although the overall improvement is not significant, our method shows substantial improvements for low-resource languages, as shown in Figure 3(a). Specifically, our method achieves improvements of 3.1% and 3.6% in Swahili (sw) and Thai (th) over the best baseline on the BLOOM model with $k = 3$ respectively.

For the XCOPA dataset, the performance improvement is more pronounced, with average improvements of 2.8%, 1.6%, and 2.5% on BLOOM, XGLM, and Llama2, respectively. Moreover, our method achieves significant improvements, especially on multilingual models like BLOOM and XGLM. As shown in Figure 3(b), our model achieves improvements in low-resource languages, with a 10.9% improvement in the unseen language Vietnamese (vi) compared to the best baseline based on BLOOM.

Our method also shows significant improvements in average performance for more complex QA tasks TyDiQA-GoldP. In BLOOM, the improvement mainly comes from several low-resource languages. For instance, our best results in unseen languages Finnish (fi) and Korean (ko) surpass the best baseline by 25.5% and 27.4%, respectively. Our approach notably enhances performance across the other two models as well, particularly on the English-centric LLM Llama2, where the mean improvement is 9.6%.

Experimental results show that training the topic model on BLOOMZ-1b7 and selecting appropriate contextual data can improve performance across different LLM architectures. From a task-level perspective, our method achieves greater improvements in relatively complex reasoning and question-answering tasks. It indicated our method makes successful use of the Bayesian theorem for non-

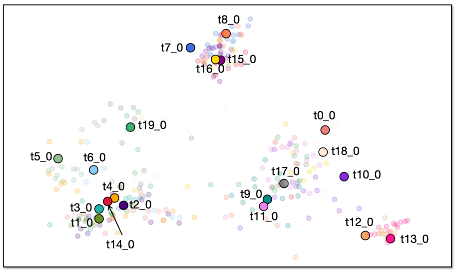

Figure 4: t-SNE plot of the learned topic tokens for TyDiQA-GoldP. "tx_0" represents the first token of the xth topic.

classification tasks ICL. Topic-XICL consistently outperforms the cluster baseline, indicating that our approach's superiority isn't solely derived from simple semantic clustering.

5 Analysis

The experimental results show that our topic model has effectively learned latent information beneficial for in-context learning. We visualized the embeddings of the topic tokens to understand the relationships between each category. Through case studies, we observed the characteristics of representative demonstrations for a topic. Furthermore, we explored our method in terms of model scale and source language.

5.1 Visualization of topic token embedding

As shown in Figure 4, the embeddings of the 20 topics in the topic model trained on the TyDiQA-GoldP dataset are distributed in about three to four distinct regions. This clustering indicates that our topic model can recognize the similarities between different topics. For example, the twelfth topic "t12" and the thirteenth topic "t13" belong to different clusters but are close in the token sequence space. This demonstrates that even if the initial clustering is not very precise, our topic model can still effectively identify and group similar topics.

Therefore, our model can adapt to different seed settings of initial clustering, resulting in a lower standard deviation, as shown in Figure 4. For non-classification tasks, where topic classification is inherently ambiguous, our method shows adaptability. This illustrates that our framework can extend the application of Bayesian theory in context sample selection to a wider range of tasks.

5.2 Case Study

We observed the characteristics of representative examples from different topics in TyDiQA-GoldP.

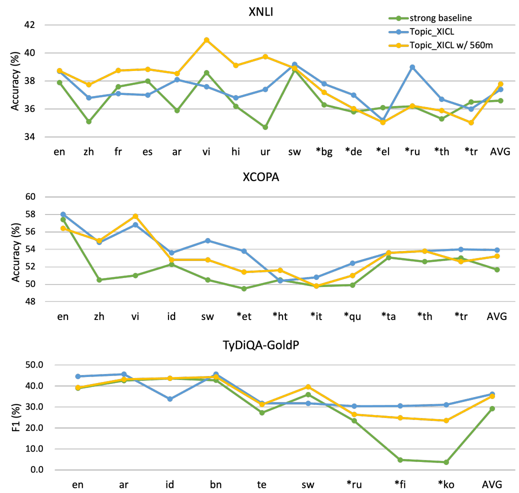

Figure 5: The 2-shot performance of BLOOM in three tasks based on the Topic-XICL model trained with fewer parameters (BLOOMZ-560m).

For instance, examples from the ninth topic "t9" mainly consist of paragraphs introducing an item or concept; those from the fourth topic "t14" relate to population themes; and examples from the third topic "t3" have longer answers, not just a single noun or short phrase. These samples show that our topic inference method incorporates more information than just semantic similarity. Details are provided in Appendix E.

5.3 Results with Less Parameter Topic Model

Since the cluster boundaries of source language candidates may not be very clear, we primarily conducted experiments on the BLOOMZ model with 1.7 billion parameters and also experimented with a smaller BLOOMZ model with 560 million parameters (BLOOMZ-560m). Fig. 5 shows the ICL results on the BLOOM model for three datasets with $k = 2$. Our method consistently outperforms the strongest baseline in terms of mean performance on the three tasks. As shown in the figure, using the BLOOMZ-560m model to learn the latent topic model improves performance on tasks in visible languages in the XNLI task, but the advantage is not significant for unseen languages. On XCOPA and TyDiQA-GoldP, the topic model based on BLOOMZ-560m also lags behind the BLOOMZ-1b7 model, primarily in unseen languages.

5.4 Results with Other Source Languages

For multilingual LLMs, besides English, other languages like Chinese and Italian have significant pre-training data. We translated the English

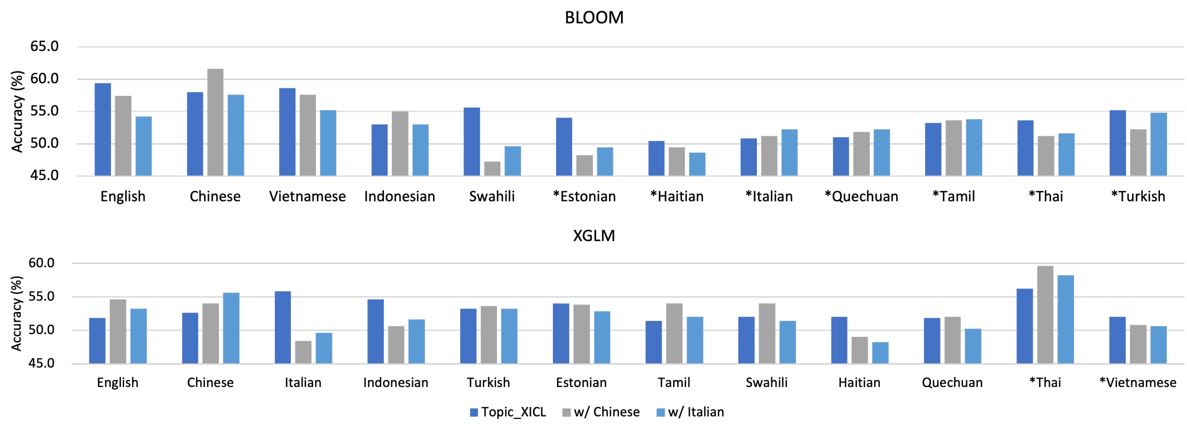

Figure 6: Results of 4-shot ICL for Individual Languages in XCOPA by the Topic-XICL model trained with Chinese and Italian.

XCOPA training data into Chinese and Italian using the Google Translation API and conducted experiments with these translations as source language data. The results are shown in Table 2, and performance in various languages is detailed in Figure 6. Since the Chinese have relatively more pre-training data than other languages in BLOOM and XGLM, the ICL performance of Topic-XICL demonstrations in it consistently outperforms the strongest baselines.

However, Italian also has a substantial amount of training data in XGLM, but the average performance of Topic-XICL in it is worse than the English-based baseline. Nonetheless, Topic-XICL based on Italian showed significant improvements in Chinese and unseen languages like Thai (non-Latin script) on XGLM. On BLOOM, using Italian as the context language for unseen languages also yielded good results. For non-English contexts, it is difficult to predict performance based on the amount of training data or language similarity, and the conclusions can vary across different models.

Zhang et al. (2024) conducted a multidimensional study on ICL for low-resource languages and found that the effectiveness of demonstration samples varies significantly across different models, tasks, and languages. This is similar to our conclusions. They also found that carefully designed templates can completely eliminate the benefits of demonstration samples for some tasks and languages. In our experiments, we also observed that for a few languages, changing the prompt could yield greater benefits than ICL. However, this phenomenon is not consistent across all languages, posing a challenge for automatic multilingual prompt design. Our primary focus is on comparing the performance of ICL sample selection, and prompt selection will be reserved for future work.

Model	method	k=2	k=3	k=4
BLOOM	best basline	51.67	51.43	51.87
	Topic_XICL	**53.92**	**54.50**	**54.41**
	Topic_XICL w/ Chinese	52.98	52.85	53.03
	Topic_XICL w/ Italian	52.40	52.83	52.68
XGLM	best basline	50.84	52.23	52.22
	Topic_XICL	53.07	**53.53**	**53.12**
	Topic_XICL w/ Chinese	**53.18**	53.18	52.87
	Topic_XICL w/ Italian	51.78	51.87	52.22

Table 2: The average accuracy of the Topic-XICL model trained with Chinese and Italian.

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 fewer 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 three 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 other tasks. 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

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: benchmarking large language models for few-shot cross-lingual transfer. *CoRR*, abs/2305.14857.

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.*

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *CoRR*, abs/2403.16512.

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.

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.

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.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O. Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. *CoRR*, abs/2402.12976.

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 Empirical guidelines For Hyper-parameter Selection

Regarding the choice of the number of topics (n) and tokens (c), there are empirical guidelines. For tasks with a large amount of English candidate data (greater than or equal to 2000), the number of clustering categories is set to $n = 20$, and for tasks with other data sizes, it is selected from (5, 10, 15), such as XCOPA with only 500 training data, which chooses $n = 5$. As for the topic tag sequence length, it is set to $c = 10$ for general classification tasks, and $c = 15$ for tasks that require reasoning or understanding of longer texts.

B 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> You should tell me the choice number 1 or 2. Answer: [1/2]
TyDiQA-GoldP	Passage: <passage> question: <question> Answer: [a span in passage]

Table 3: Prompt template for three tasks.

C Low-resource Languages

All 24 languages in the three datasets are not always pre-trained on the three baseline LLMs. Based on the language distribution in the pre-training data for each model, we selected some languages as low-resource or unseen languages, as shown in Table **??**. For BLOOM (Scao et al., 2022), English training data accounts for 30.4% of the total, with pre-training data covering 46 natural languages. We define languages accounting for less than 0.1% as low-resource languages, and languages without training data are unseen languages. In XGLM (Lin et al., 2022), with 7.5 billion parameters, English tokens constitute 48.99%. It is pre-trained in 30 natural languages, including all 24 languages we evaluate. We define languages with a token ratio of less than 0.1% as low-resource languages. Llama2 (Touvron et al., 2023) is an English-centric LLM, with English training data making up 89.7% and covering 27 natural languages. Its language classification standards are the same as BLOOM's.

Dataset	Task	Languages	Train num.	Dev num.
XNLI	natural language inference	English(en), German(de), Russian(ru), French(fr), Spanish(es), Chinese(zh), Vietnamese(vi), Turkish(tr), Arabic(ar), Greek(el), Thai(th), Bulgarian(bg), Hindi(hi), Urdu(ur), Swahili(sw)	10,000	5010
XCOPA	commonsense reasoning	Chinese(zh), Italian(it), Vietnamese(vi), Indonesian(id), Turkish(tr), Thai(th), Estonian(es), Tamil(ta), Swahili(sw), Haitian(ht), Quechua(qu)	500	500
TyDiQA-GoldP	TyDiQA-GoldP	English(en), Russian(ru), Indonesian(id), Korean(ko), Arabic(ar), Finnish(fi), Bengali(bn), Telugu(te), Swahili(sw)	3,695	113-921

Table 4: The detailed information of datasets.

D Detailed Results

As shown in Figures 7, 8, and 9, we visualized the results for each language in the 4-shot setting, including the mean and standard deviation, except for the semantic similarity method. All results are reported in Tables 6, 7, and 8.

Model	Dataset	low-resource languages	extremly low-resource languages
BLOOM	XNLI	Swahili(sw)	German(de), Russian(ru), Turkish(tr), Greek(el), Thai(th), Bulgarian(bg)
	XCOPA	Swahili(sw)	Italian(it), Turkish(tr), Thai(th), Estonian(es), Haitian(ht), Quechua(qu)
	TyDiQA-GoldP	Telugu(te), Swahili(sw)	Russian(ru), Korean(ko), Finnish(fi), Bengali(bn)
XGLM	XNLI	Urdu(ur), Swahili(sw)	
	XCOPA	Tamil(ta), Swahili(sw), Haitian(ht), Quechua(qu)	
	TyDiQA-GoldP	Bengali(bn), Telugu(te), Swahili(sw)	
Llama2	XNLI	Vietnamese(vi), Bulgarian(bg)	Turkish(tr), Arabic(ar), Greek(el), Thai(th), Hindi(hi), Urdu(ur), Swahili(sw)
	XCOPA	Vietnamese(vi), Indonesian(id)	Turkish(tr), Thai(th), Tamil(ta), Swahili(sw), Haitian(ht), Quechua(qu)
	TyDiQA-GoldP	Indonesian(id), Korean(ko)	Arabic(ar), Finnish(fi), Bengali(bn), Telugu(te), Swahili(sw)

Table 5: Classification of languages for three datasets (XNLI, XCOPA, TyDiQA-GoldP) across three LLMs (BLOOM, XGLM, LLama2).

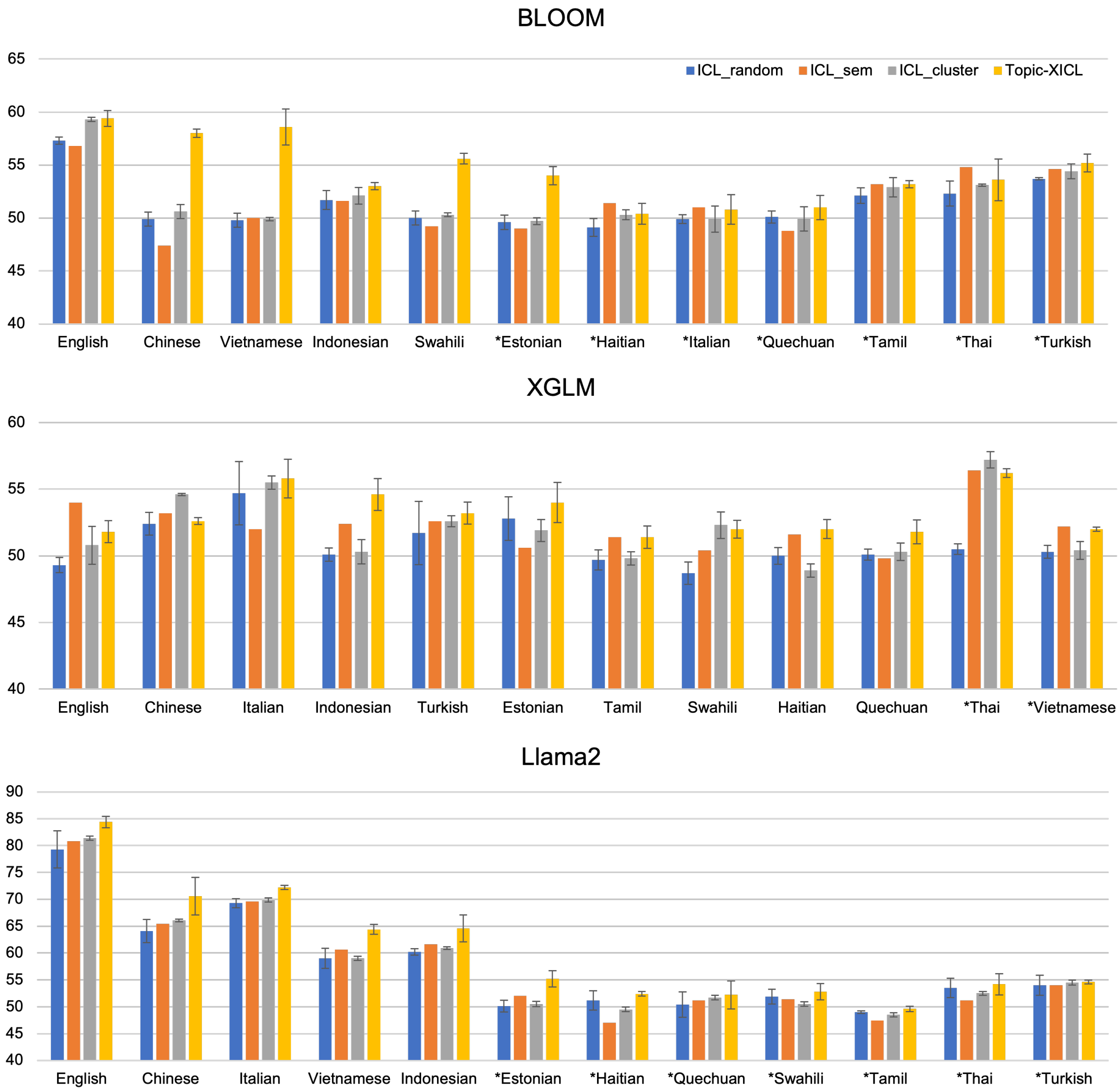

Figure 7: The 4-shot performance of individual languages in XCOPA.

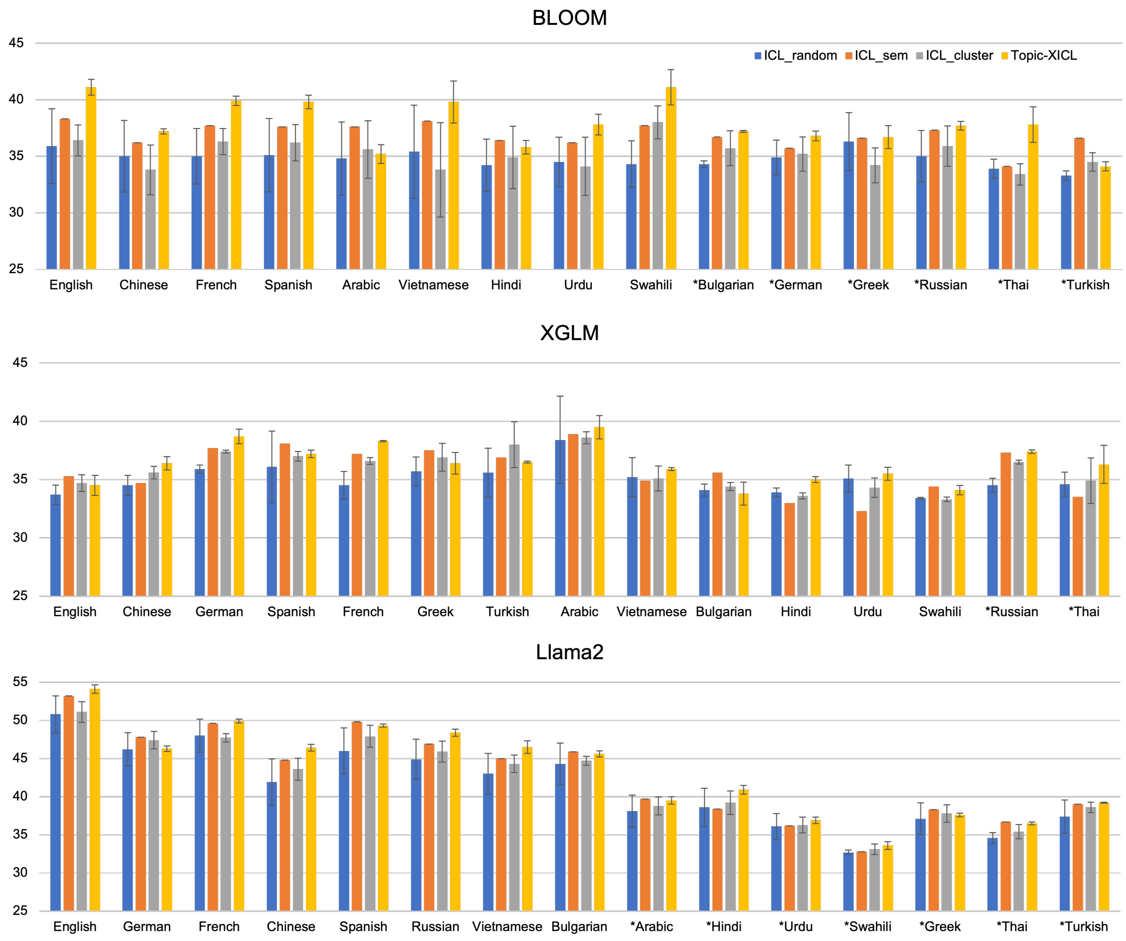

Figure 8: The 4-shot performance of individual languages in XNLI.

		XNLI(acc.)															
	Model	en	ar	bn	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	AVG
	BLOOM	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	ICL_random	37.8±5.13	35.1±2.47	34.7±1.49	34.5±0.94	34.5±0.27	**38.3±5.14**	**37.9±5.62**	33.8±0.65	34.0±0.42	36.2±2.65	34.9±2.34	34.6±1.58	34.4±1.51	34.6±1.52	34.2±1.33	35.3±2.15
	ICL_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
	ICL_cluster	35.7±1.63	33.8±1.33	34.9±0.2	34.3±1.1	35.0±1.3	35.3±1.49	35.5±1.58	32.5±1.09	35.5±0.35	36.1±0.47	33.9±1.4	33.7±0.28	33.2±0.93	33.5±0.83	33.7±1.41	34.4±0.92
	Topic-XICL(ours)	**38.7±0.11**	**38.1±0.08**	**37.8±0.41**	**37.0±0.07**	35.2±1.08	37.0±0.15	37.1±0.03	**36.8±0.62**	**39.0±0.44**	**39.2±0.38**	**36.7±1.84**	36.0±0.46	**37.4±0.72**	37.6±1.37	**36.8±0.09**	**37.4±0.33**
k=3	ICL_random	35.9±3.29	34.8±3.24	34.3±0.3	34.9±1.54	36.3±2.56	35.1±3.23	35.0±2.47	34.2±2.31	35.0±2.28	34.3±2.06	33.9±0.83	33.3±0.41	34.5±2.18	35.4±4.11	35.0±3.18	34.8±1.56
	ICL_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
	ICL_cluster	36.4±1.38	35.6±2.53	35.7±1.54	35.2±1.53	34.2±1.54	36.2±1.59	36.3±1.16	34.9±2.77	35.9±1.8	38.0±1.46	33.4±0.94	34.5±0.81	34.1±2.57	33.8±4.16	33.8±2.19	35.2±1.67
	Topic-XICL(ours)	**41.1±0.69**	35.2±0.82	**37.2±0.09**	**36.8±0.42**	**36.7±1.02**	**39.8±0.61**	**39.9±0.41**	35.8±0.6	**37.7±0.38**	**41.1±1.55**	**37.8±1.57**	34.1±0.4	**37.8±0.91**	**39.8±1.85**	**37.2±0.22**	**37.9±0.25**
k=4	ICL_random	33.6±2.17	34.9±2.34	33.4±0.64	35.0±0.82	33.4±0.53	33.1±0.72	33.5±0.91	35.1±2.63	35.8±0.95	34.6±0.76	33.1±0.38	33.2±0.61	34.1±1.19	35.9±3.53	35.3±2.39	34.3±1.13
	ICL_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
	ICL_cluster	36.6±1.61	36.3±1.79	35.1±1.83	36.0±1.63	33.9±0.33	36.7±1.48	36.7±0.83	**37.7±1.51**	36.4±1.84	36.2±2.68	33.8±1.71	34.0±0.51	36.6±2.61	39.2±3.22	36.9±2.07	36.1±1.59
	Topic-XICL(ours)	37.6±0.33	**40.6±0.51**	**37.2±0.2**	36.5±0.45	35.7±0.47	38.3±0.65	37.5±0.78	34.6±2.59	**37.6±0.25**	36.5±1.12	**34.6±0.11**	35.4±0.38	**38.5±0.94**	**40.7±3.14**	**39.2±0.65**	**37.4±0.52**
	XGLM	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	ICL_random	33.7±0.53	35.8±2.29	33.7±0.21	35.3±1.27	34.2±0.93	33.8±0.28	33.5±0.16	34.0±0.76	34.0±0.52	33.5±0.15	**36.5±4.3**	35.4±2.73	34.7±1.96	33.7±0.18	34.9±1.62	34.4±0.77
	ICL_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
	ICL_cluster	34.7±0.43	37.7±1.5	33.9±0.38	36.3±0.89	35.6±1.99	35.6±0.84	35.1±1.1	33.5±0.35	34.6±0.59	33.2±0.16	35.8±1.87	**37.3±2.35**	33.9±0.35	35.3±2.82	35.4±0.97	35.2±1.05
	Topic-XICL(ours)	34.9±0.84	38.1±0.61	**36.5±0.51**	37.4±0.73	**36.7±0.47**	35.4±0.54	34.7±0.24	**34.4±0.69**	35.1±0.61	**33.9±0.8**	35.8±0.61	35.1±0.87	**35.6±1.18**	**35.0±0.8**	**37.1±1.07**	**35.7±0.69**
k=3	ICL_random	33.7±0.83	38.4±3.74	34.1±0.51	35.9±0.33	35.7±1.25	36.1±3.06	34.5±1.2	33.9±0.37	34.5±0.62	33.4±0.07	34.6±1.03	35.6±2.09	35.1±1.14	35.2±1.68	34.5±0.86	35.0±0.69
	ICL_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
	ICL_cluster	34.7±0.72	38.6±0.51	34.4±0.35	37.4±0.13	36.9±1.19	37.0±0.42	36.6±0.29	33.6±0.25	36.5±0.16	33.3±0.2	34.9±1.96	**38.0±1.97**	34.3±0.83	35.1±1.07	35.6±0.54	35.8±0.26
	Topic-XICL(ours)	34.5±0.86	**39.5±1**	33.8±0.98	**38.7±0.62**	36.4±0.94	37.2±0.31	**38.3±0.07**	**35.0±0.24**	**37.4±0.15**	34.1±0.39	**36.3±1.63**	36.5±0.08	**35.5±0.56**	**35.9±0.13**	**36.4±0.58**	**36.4±0.22**
k=4	ICL_random	33.7±1.31	38.4±4.34	35.0±1.71	37.5±0.88	**37.8±2.37**	37.6±3.23	35.9±2.74	**34.8±1.41**	37.1±1.4	33.8±0.94	**34.9±0.4**	37.2±4.33	33.7±1	34.4±2.43	35.4±2.41	35.8±1.42
	ICL_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
	ICL_cluster	35.2±0.73	38.8±0.21	**35.7±0.56**	38.1±0.6	37.4±1.02	37.4±0.21	36.8±0.76	33.9±1.01	36.8±0.45	33.9±0.24	33.7±1.57	37.4±0.58	34.5±1.47	**35.2±0.87**	35.8±1.45	36.0±0.51
	Topic-XICL(ours)	35.8±0.78	**39.7±0.68**	34.6±0.9	**38.2±0.65**	37.7±0.21	**39.7±0.56**	**36.9±0.35**	34.3±0.42	**37.2±0.15**	**35.0±0.74**	34.0±2.2	37.4±0.4	**36.8±1.65**	35.1±0.19	**36.3±1.03**	**36.6±0.31**
	Llama2	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	ICL_random	51.9±4.52	38.5±1.54	43.2±2.3	45.2±2.6	37.4±1.09	46.3±2.96	47.2±2.91	39.4±0.69	44.9±3	32.1±0.59	35.6±1.13	38.1±1.87	36.1±0.59	43.2±2.35	**44.5±3.09**	41.6±2.01
	ICL_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
	ICL_cluster	50.0±2.26	38.2±1	43.1±1.58	45.8±1.18	36.4±0.63	46.4±1.5	46.7±1.4	39.2±1.16	45.0±1.89	32.4±1	34.3±0.15	37.5±1.18	35.4±1.04	42.8±1.05	43.3±0.96	41.1±1.09
	Topic-XICL(ours)	**52.7±0.56**	**38.7±0.5**	43.5±0.33	**46.6±0.33**	**37.8±0.22**	**47.9±0.35**	**47.8±0.28**	**43.0±0.96**	45.4±0.36	**34.3±0.55**	**36.2±0.16**	**38.6±0.11**	**38.4±0.65**	**46.5±1.07**	44.3±0.43	**42.8±0.45**
k=3	ICL_random	50.8±2.4	38.1±2.11	44.3±2.74	46.2±2.2	37.1±2.07	46.0±3.01	48.0±2.16	38.6±2.5	44.9±2.64	32.7±0.3	34.6±0.71	37.4±2.16	36.1±1.71	43.0±2.65	41.9±3.03	41.3±2.1
	ICL_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
	ICL_cluster	51.1±1.35	38.8±1.16	44.7±0.6	47.4±1.14	37.8±1.13	47.9±1.45	47.7±0.56	39.2±1.54	45.9±1.39	33.1±0.69	35.4±0.93	38.6±0.66	36.3±1.05	44.3±1.14	43.6±1.43	42.1±1
	Topic-XICL(ours)	**54.1±0.56**	39.5±0.5	45.6±0.4	46.3±0.36	37.6±0.23	49.3±0.22	**49.9±0.27**	**40.9±0.56**	**48.4±0.47**	**33.6±0.51**	36.5±0.17	**39.2±0.07**	**36.9±0.42**	**46.5±0.83**	**46.4±0.44**	**43.4±0.37**
k=4	ICL_random	51.1±1.71	37.1±1.4	43.9±1.88	47.2±1.96	37.2±1.75	46.7±1.98	48.0±1.99	39.0±2.41	45.5±2.22	32.5±0.3	34.8±1.08	37.4±1.49	35.7±1.69	42.5±2.53	41.9±1.45	41.4±1.67
	ICL_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
	ICL_cluster	51.9±0.83	39.2±0.94	45.4±1.03	47.3±0.84	37.5±0.42	48.3±1.12	48.8±0.76	39.6±1.4	46.5±0.91	33.2±0.93	35.8±0.81	39.0±1.41	36.5±1.45	44.1±1.63	43.9±1.12	42.5±1
	Topic-XICL(ours)	**54.4±0.32**	40.0±0.12	46.1±0.1	47.6±0.17	**38.7±0.13**	50.1±0.17	**51.0±0.29**	**42.4±0.42**	**49.3±0.1**	**34.6±0.11**	**37.6±0.16**	39.8±0.13	**39.3±0.33**	**45.8±0.38**	**46.6±0.07**	**44.2±0.05**

Table 6: Accuracy of XNLI in 15 languages based on BLOOM-7b1, XGLM-7.5b and Llama-2-7b models.

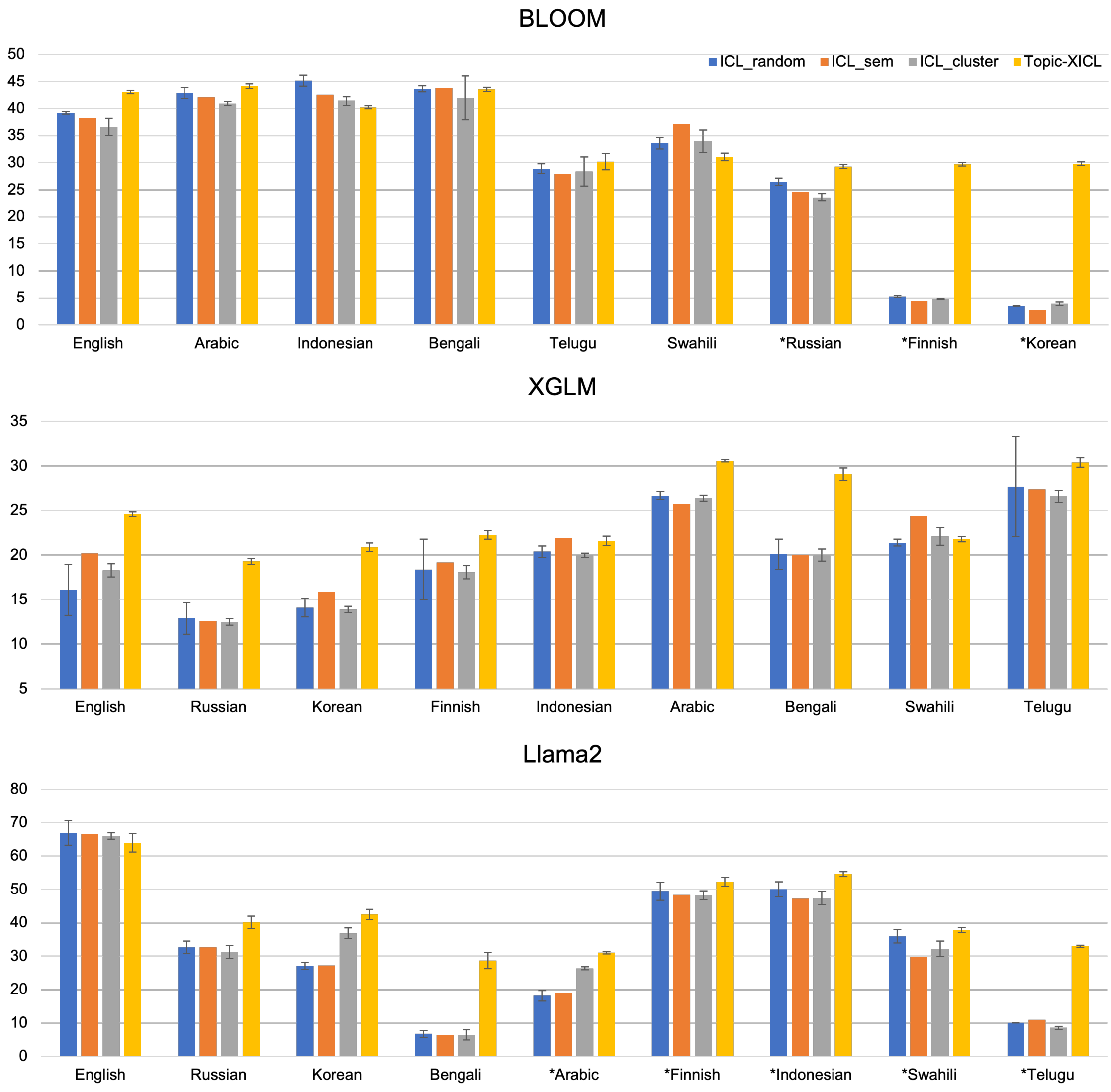

Figure 9: The 4-shot performance of individual languages in TyDiQA-GoldP.

	XCOPA(acc.)												
	Model	en	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	AVG
	BLOOM	45.8	49.8	49.6	49.8	49.4	50.2	49.6	49.4	50.4	50.0	51.0	50.4	49.6
k=2	ICL_random	56.4±0.93	49.5±0.34	49.6±1.31	51.6±0.9	50.1±0.33	50.9±1.33	49.6±1.75	51.6±0.75	52.8±0	53.2±0.68	50.9±0.82	49.6±0.43	51.3±0.4
	ICL_sem	55.2	49.0	44.8	52.4	47.0	**52.6**	50.2	**54.2**	53.0	52.4	49.2	49.0	50.8
	ICL_cluster	57.4±0.65	49.5±0.82	**50.5±0.9**	52.3±0.57	49.8±0.29	49.9±0.84	50.5±0.29	53.1±0.82	52.6±1.02	53.0±1.72	51.0±1.72	50.5±0.62	51.7±0.09
	Topic-XICL(ours)	**58.0±0**	**53.8±0.82**	50.4±0.66	**53.6±1.36**	**50.8±0.9**	52.4±0.19	**55.0±1.32**	53.6±1.43	**53.8±0.93**	**54.0±1.71**	**56.8±1**	**54.8±0.62**	**53.9±0.13**
k=3	ICL_random	56.6±0.34	**51.1±0.49**	49.4±0.93	51.9±0.81	50.3±0.47	51.9±1.52	50.4±0.52	50.5±0.78	52.4±0.98	53.3±0.34	49.5±0.19	49.9±0.62	51.4±0.21
	ICL_sem	56.0	50.0	51.0	52.2	46.8	48.2	47.8	51.4	53.6	53.8	46.6	47.6	50.4
	ICL_cluster	58.2±0.29	48.6±0.43	49.7±0.52	52.6±0.16	48.2±0.47	49.4±1.16	50.3±0.68	48.9±0.84	52.9±0.96	53.5±0.75	50.7±0.57	48.5±0.38	51.0±0.26
	Topic-XICL(ours)	**58.6±0.66**	50.2±0.9	**53.2±0.34**	**53.2±0.9**	**51.4±0.56**	**52.0±0.84**	**52.0±0.66**	**54.2±1.73**	**56.4±1.57**	**54.2±1**	**61.6±1.14**	**57.0±0.57**	**54.5±0.09**
k=4	ICL_random	57.3±0.34	49.6±0.68	49.1±0.85	51.7±0.9	49.9±0.41	50.1±0.57	50.0±0.65	52.1±0.73	52.3±1.18	53.7±0.1	49.8±0.66	49.9±0.65	51.3±0.29
	ICL_sem	56.8	49.0	**51.4**	51.6	**51.0**	48.8	49.2	**53.2**	**54.8**	54.6	50.0	47.4	51.5
	ICL_cluster	59.3±0.19	49.7±0.33	50.3±0.47	52.1±0.78	49.9±1.23	49.9±1.14	50.3±0.19	52.9±0.9	53.1±0.1	54.4±0.71	49.9±0.16	50.6±0.66	51.9±0.17
	Topic-XICL(ours)	**59.4±0.75**	**54.0±0.87**	50.4±1	**53.0±0.34**	50.8±1.39	**51.0±1.14**	**55.6±0.49**	**53.2±0.34**	53.6±1.96	**55.2±0.85**	**58.6±1.7**	**58.0±0.41**	**54.4±0.16**
	XGLM	50.4	50.8	49.0	47.0	49.6	49.2	48.4	49.4	55.2	46.4	50.0	50.4	49.7
k=2	ICL_random	50.8±0.43	50.5±2.42	49.7±0.9	50.3±0.25	**54.6±2.44**	50.1±0.68	49.2±1.11	50.3±1.09	49.9±0.66	**51.7±3.56**	50.5±2.08	52.5±2.69	50.8±0.92
	ICL_sem	53.0	52.2	46.0	51.2	49.2	**53.0**	50.4	46.2	55.2	50.2	46.2	53.0	50.5
	ICL_cluster	50.7±0.73	49.1±0.47	49.2±0.65	50.4±0.1	50.3±0.1	48.9±0.87	46.5±0.5	48.9±0.25	**57.3±0.34**	50.6±0.5	50.3±0.1	53.5±0.84	50.5±0.08
	Topic-XICL(ours)	**54.2±3**	**54.8±0.47**	**51.6±2.77**	**54.6±0.92**	53.6±1.73	51.4±1.15	**51.6±0.47**	**54.0±0.92**	54.6±0.23	50.8±0.92	**51.6±0.69**	**54.0±0.81**	**53.1±0.1**
k=3	ICL_random	49.3±1.22	52.9±1.07	49.0±0.68	50.9±1.22	**58.9±2.34**	49.8±0.43	48.7±0.57	50.6±0.85	51.3±0.94	53.4±2.16	50.8±0.56	53.8±0.91	51.6±0.71
	ICL_sem	52.8	53.2	52.0	**53.8**	53.2	51.0	47.4	49.6	55.6	54.0	51.0	53.2	52.2
	ICL_cluster	50.1±1.14	51.5±0.78	50.1±0.66	51.3±0.73	56.2±0.66	49.9±0.34	48.9±0.56	49.3±0.41	**57.7±0.41**	53.5±0.1	50.7±0.57	53.9±0.85	51.9±0.22
	Topic-XICL(ours)	**53.4±1.23**	**55.6±0.99**	**52.2±1**	53.4±1.39	56.6±0.16	**52.0±1.32**	**51.4±1.73**	**51.2±0.16**	56.2±0.57	**55.0±0.62**	**51.4±1.09**	**54.0±0.84**	**53.5±0.24**
k=4	ICL_random	49.3±0.57	52.8±1.63	50.0±0.62	50.1±0.5	54.7±2.38	50.1±0.41	48.7±0.84	49.7±0.75	50.5±0.41	51.7±2.38	50.3±0.47	52.4±0.85	50.9±0.49
	ICL_sem	**54.0**	50.6	51.6	52.4	52.0	49.8	50.4	**51.4**	56.4	52.6	**52.2**	53.2	52.2
	ICL_cluster	50.8±1.42	51.9±0.82	48.9±0.5	50.3±0.91	55.5±0.5	50.3±0.66	**52.3±0.99**	49.8±0.5	**57.2±0.62**	52.6±0.41	50.4±0.66	**54.6±0.1**	52.1±0.22
	Topic-XICL(ours)	51.8±0.82	**54.0±1.52**	**52.0±0.71**	**54.6±1.2**	**55.8±1.45**	**51.8±0.9**	52.0±0.68	**51.4±0.84**	56.2±0.33	**53.2±0.82**	52.0±0.16	52.6±0.25	**53.1±0.11**
	Llama2	57.8	44.8	48.2	51.8	52.4	46.8	49.0	49.0	49.6	52.4	50.0	55.0	50.6
k=2	ICL_random	82.0±2.08	49.0±1.88	48.3±2.27	61.1±1.25	68.8±1.91	50.3±0.96	49.4±1.22	48.6±1.25	51.5±0.43	54.4±1.79	57.8±0.25	64.2±0.82	57.1±0.36
	ICL_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
	ICL_cluster	80.7±0.41	50.0±0.5	50.6±1.06	59.8±1.27	69.0±0.78	50.4±0.73	50.3±0.78	48.9±0.66	52.1±0.19	53.3±1.14	57.8±1.32	66.1±2.14	57.4±0.3
	Topic-XICL(ours)	**84.0±0.96**	**53.2±0.93**	**51.8±0.96**	**63.4±1.45**	**72.6±1.71**	**52.0±3.21**	**52.4±1.56**	**51.2±0.75**	54.4±1.63	**55.6±1.8**	**62.0±1.36**	**67.2±0.81**	**60.0±0.18**
k=3	ICL_random	77.6±1.8	48.9±2.62	49.9±2.29	62.4±2.26	68.4±2.69	50.5±2.78	48.7±0.71	47.5±1.23	52.1±1.06	55.3±1.72	60.2±1.55	64.0±2.29	57.1±1.09
	ICL_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
	ICL_cluster	81.9±0.96	48.8±0.1	51.3±0.16	63.0±1.11	70.3±0.25	49.8±0.68	49.9±0.9	49.1±0.81	**54.3±0.16**	54.0±1.09	59.3±0.9	67.1±0.62	58.2±0.26
	Topic-XICL(ours)	**84.4±1.31**	**54.0±3.69**	**53.6±0.19**	**64.0±1.31**	**72.8±1.73**	**51.8±1.14**	**51.6±1.09**	**51.4±0.25**	53.8±1.82	**56.1±1.8**	62.0±1.09	**69.8±2.34**	**60.4±0.17**
k=4	ICL_random	79.3±3.44	50.1±1.11	51.2±1.79	60.2±0.59	69.3±0.87	50.4±2.39	51.9±1.37	49.0±0.25	53.5±1.8	54.0±1.85	59.0±1.89	64.1±2.16	57.7±0.42
	ICL_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
	ICL_cluster	81.4±0.41	50.5±0.49	49.5±0.41	60.9±0.25	69.9±0.41	51.7±0.41	50.5±0.43	48.5±0.38	52.5±0.34	54.5±0.43	59.0±0.41	66.1±0.25	57.9±0.14
	Topic-XICL(ours)	**84.4±1.05**	**55.2±1.52**	**52.4±0.43**	**64.6±2.53**	**72.2±0.38**	**52.2±2.62**	**52.8±1.55**	**49.6±0.5**	**54.2±1.98**	**54.6±0.33**	**64.4±0.93**	**70.6±3.51**	**60.6±0.22**

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

E Case Study

Table 9 shows the representative examples selected from some topics in TyDiQA-GoldP.

	TyDiQA-GoldP(F1)										
	Model	ar	bg	en	fi	id	ko	ru	sw	te	AVG
	BLOOM	28.8	28.9	29.0	4.1	28.1	2.6	12.3	27.9	21.5	20.3
k=2	ICL_random	38.1±0.92	42.3±0.56	32.8±1.01	4.8±0.3	39.9±1.76	3.3±0.19	21.9±1.94	31.7±2.18	26.5±0.42	26.8±0.91
	ICL_sem	42.7	42.8	39.0	4.8	43.7	3.7	23.5	36.0	27.3	29.3
	ICL_cluster	40.9±1.24	40.8±1.05	38.1±0.89	5.0±0.2	**44.1±1.85**	3.2±0.26	24.2±0.6	**36.1±3.07**	25.2±1.98	28.6±0.54
	Topic-XICL(ours)	**45.7±0.78**	**45.7±0.96**	**44.7±0.6**	**30.5±0.14**	33.8±0.23	**31.1±0.16**	**30.5±0.45**	31.8±1.16	**31.8±1.7**	**36.2±0.23**
k=3	ICL_random	39.5±1.54	42.8±2.64	34.3±1.75	4.8±0.01	41.3±2.15	3.6±0.06	22.9±1.28	33.9±2.6	28.1±0.8	27.9±1.37
	ICL_sem	41.9	42.3	39.5	4.6	**43.2**	3.2	24.6	**37.7**	27.4	29.4
	ICL_cluster	40.1±1.34	40.7±2.03	36.5±1.11	4.9±0.31	41.9±0.85	3.2±0.32	23.4±2.25	33.2±0.79	27.1±0.56	27.9±0.6
	Topic-XICL(ours)	**43.6±0.82**	**43.3±0.86**	**42.5±0.89**	**29.2±0.05**	31.9±0.56	**29.4±0.4**	**28.7±0.9**	32.6±0.39	**29.9±0.77**	**34.6±0.36**
k=4	ICL_random	42.9±1.01	**43.7±0.54**	39.2±0.28	5.3±0.17	**45.2±1.01**	3.5±0.03	26.5±0.65	33.6±1.05	28.9±0.9	29.9±0.4
	ICL_sem	42.1	43.8	38.2	4.4	42.6	2.7	24.6	**37.2**	27.9	29.3
	ICL_cluster	40.9±0.35	42.0±4.06	36.6±1.57	4.8±0.14	41.4±0.82	3.9±0.32	23.6±0.69	34.0±2.06	28.4±2.69	28.4±0.24
	Topic-XICL(ours)	**44.2±0.41**	43.6±0.4	**43.1±0.29**	**29.7±0.33**	40.2±0.28	**29.8±0.36**	**29.3±0.35**	31.1±0.7	**30.2±1.47**	**35.7±0.3**
	XGLM	23.6	18.7	8.5	12.6	10.8	8.7	7.9	25.2	25.8	15.8
k=2	ICL_random	26.1±0.69	20.3±1.49	13.2±1.83	15.6±0.88	18.8±1.18	14.1±0.91	11.6±0.72	21.7±0.15	27.8±0.36	18.8±0.58
	ICL_sem	27.0	21.6	17.1	17.6	21.5	16.1	13.4	**23.8**	**28.2**	20.7
	ICL_cluster	26.7±0.43	17.6±1.31	15.4±1.08	16.6±0.91	18.7±0.34	13.6±0.14	12.1±0.24	20.9±0.25	27.3±0.51	18.8±0.1
	Topic-XICL(ours)	**31.5±0.46**	**29.7±0.82**	**25.0±0.77**	**22.7±0.78**	**22.0±0.43**	**21.5±0.29**	**19.9±0.43**	22.7±0.4	28.1±0.19	**24.8±0.25**
k=3	ICL_random	25.9±0.24	20.0±1.54	13.6±1.34	16.7±1.89	18.5±0.16	13.7±0.92	12.0±1.22	20.9±0.59	27.3±0.8	18.7±0.55
	ICL_sem	26.4	19.5	18.0	19.1	21.1	15.3	12.6	**23.5**	27.2	20.3
	ICL_cluster	26.4±0.3	19.2±0.57	16.7±0.91	18.5±0.18	19.4±0.92	13.4±0.32	12.2±0.33	22.1±0.87	**28.0±0.63**	19.5±0.11
	Topic-XICL(ours)	**30.9±0.19**	**29.3±0.82**	**24.8±0.16**	**22.2±0.49**	**21.5±0.45**	**20.9±0.33**	**19.3±0.22**	22.9±0.15	27.4±0.27	**24.4±0.08**
k=4	ICL_random	26.7±0.47	20.1±1.68	16.1±2.86	18.4±3.4	20.4±0.62	14.1±1.01	12.9±1.78	21.4±0.38	27.7±5.63	19.8±0.96
	ICL_sem	25.7	20.0	20.2	19.2	**21.9**	15.9	12.6	**24.4**	27.4	20.8
	ICL_cluster	26.4±0.37	20.0±0.68	18.3±0.76	18.1±0.73	20.0±0.24	13.9±0.38	12.5±0.37	22.1±0.99	26.6±0.69	19.8±0.14
	Topic-XICL(ours)	**30.6±0.13**	**29.1±0.69**	**24.6±0.24**	**22.3±0.49**	21.6±0.52	**20.9±0.49**	**19.3±0.33**	21.8±0.3	**30.4±0.53**	**24.5±0.24**
	Llama2	15.4	1.1	45.3	38.9	33.6	21.4	29.7	31.4	0.5	24.1
k=2	ICL_random	17.7±1.39	4.3±0.87	60.3±5	43.9±7.1	43.9±2.97	26.5±3.18	28.2±3.91	24.5±3.04	4.6±0.12	28.2±2.49
	ICL_sem	17.4	4.9	61.7	45.4	43.9	24.5	29.5	27.4	6.5	29.0
	ICL_cluster	25.4±0.58	6.5±2.16	63.2±2.3	44.1±2.39	44.4±3.04	38.8±0.42	29.6±1.41	26.7±1.85	3.2±0.22	31.3±0.77
	Topic-XICL(ours)	**27.9±0.6**	**26.0±1.15**	**69.2±1.91**	**50.3±0.4**	**51.7±1.04**	**43.7±2.31**	**35.8±1.09**	**37.7±1.76**	**30.8±0.09**	**41.4±0.24**
k=3	ICL_random	17.2±1.06	4.2±0.51	64.1±2.29	46.6±1.87	47.7±2.55	28.0±0.87	29.9±2.31	32.4±1.33	10.1±0.37	31.1±1.16
	ICL_sem	18.0	4.5	65.2	47.0	47.1	26.8	31.8	30.6	8.5	31.0
	ICL_cluster	25.8±0.82	6.2±2.06	65.9±0.51	45.4±1.72	45.2±0.69	37.5±0.37	30.4±0.69	27.8±2.34	7.5±0.02	32.4±0.36
	Topic-XICL(ours)	**29.2±0.14**	**26.9±1.71**	**68.8±2.44**	**52.1±1.44**	**52.2±0.97**	**39.9±1.48**	**38.9±1.43**	**39.4±1.6**	**32.4±0.02**	**42.2±0.27**
k=4	ICL_random	18.2±1.61	6.8±1	**66.9±3.67**	49.5±2.73	50.1±2.17	27.2±1.06	32.7±1.89	36.0±2.02	10.1±0.05	33.1±1.45
	ICL_sem	19.0	6.5	66.6	48.4	47.2	27.3	32.7	29.9	11.0	32.1
	ICL_cluster	26.4±0.42	6.5±1.54	66.0±0.94	48.3±1.32	47.4±2.05	36.9±1.58	31.3±1.89	32.2±2.31	8.6±0.48	33.7±0.29
	Topic-XICL(ours)	**31.1±0.33**	**28.7±2.41**	64.0±2.76	**52.3±1.38**	**54.6±0.77**	**42.5±1.54**	**40.1±1.85**	**37.9±0.77**	**33.0±0.33**	**42.7±0.55**

Table 8: F1 score of TyDiQA-GoldP in 9 languages based on BLOOM-7b1, XGLM-7.5b and Llama-2-7b models.

Topic	Top-4 Examples Passage	Question	Answer
t3	Magnetoreception (also magnetoception) is a sense which allows an organism to detect a magnetic field to perceive direction, altitude or location. This sensory modality is ...	Do birds use magnetoreception?	**Magnetoreception (also magnetoception) is a sense which allows an organism to detect a magnetic field to perceive direction, altitude or location**
	In most countries, the academic year begins in late summer or early autumn and ends during the following spring or summer. In Northern Hemisphere countries, this means that the academic year lasts from August, September, or October to May, June, or July...	How long does a school year last in Germany?	**from August, September, or October to May, June, or July**
	Cuneiform law refers to any of the legal codes written in cuneiform script, that were developed and used throughout the ancient Middle East among the Sumerians, Babylonians, Assyrians, Elamites, Hurrians, Kassites, and Hittites.[1] The Code of...[1]	What is cuneiform law?	**any of the legal codes written in cuneiform script, that were developed and used throughout the ancient Middle East among the Sumerians, Babylonians, Assyrians, Elamites, Hurrians, Kassites, and Hittites**
	The view of most scholars (see organology) is that the term "brass instrument" should be defined by the way the sound is made, as above, and not by whether the instrument is actually made of brass. Thus one finds brass instruments made of wood, like...	Why are they called brass instruments?	**the term "brass instrument" should be defined by the way the sound is made, as above, and not by whether the instrument is actually made of brass**
t9	**Ice wine** (or icewine; German: Eiswein) is a type of dessert wine produced from grapes that have been frozen while still on the vine...	What makes an ice wine an ice wine?	produced from grapes that have been frozen while still on the vine
	Earth's magnetic field, also known as the geomagnetic field, is the magnetic field that extends from the Earth's interior out into space, where it meets the solar wind...	What is the magnetic force of the Earth?	Earth's magnetic field
	The lux (symbol: lx) is the SI derived unit of illuminance and luminous emittance, measuring luminous flux per unit area.[1] It is equal to one lumen per square metre...	What is the unit of measurement for light brightness?	lux
	General speed limits in Germany are set by the federal government. All limits are multiples of 5km/h. There are two default speed limits: 50km/h (31mph) inside built-up areas and 100km/h (62mph) outside built-up areas. While parts of the autobahns and many other freeway-style highways have posted limits up to 130km/h (81mph) based on accident experience, congestion and other factors, many rural sections have no general speed limit...	How fast can you drive on the Autobahn?	130km/h
t14	The demography of France is monitored by the Institut national d'études démographiques (INED) and the Institut national de la statistique et des études économiques (INSEE). As of 1 January 2018, 67.19 million **people** lived in France (67,186,638), including all the five overseas departments (2,141,000), but excluding the overseas collectivities and territories (604,000).[1] 65,017,000 of these lived in Metropolitan France, which is mainland France located in Europe.	How many people live in France?	67.19 million
	The Balkans are usually said to comprise Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Kosovo,[a] the Republic of Macedonia, Montenegro, Romania, Serbia and Slovenia, while Greece and Turkey are often excluded. Its total area is usually given as 666,700 square km (257,400 square miles) and the **population** as 59,297,000 (est. 2002).[38][39]	What countries are on the Balkan Peninsula?	Albania, Bosnia and Herzegovina, Bulgaria, Croatia, Kosovo,[a] the Republic of Macedonia, Montenegro, Romania, Serbia and Slovenia
	In United States, the poverty thresholds are updated every year by Census Bureau. The threshold in United States are updated and used for statistical purposes. In 2015, in the United States, the poverty threshold for a **single person** under 65 was an annual income of US$11,770; the threshold for a **family group** of four, including two children, was US$24,250...	What's the poverty line in America?	24250
	The metropolis is an alpha global city as listed by the Globalization and World Cities Research Network. In 2011, the **population** of the city of Johannesburg was 4,434,827, making it the most populous city in South Africa.[4] In the same year, ...	How large is Johannesburg's population?	4434827

Table 9: The top-4 representative samples of some topics in TyDiQA-GoldP selected by our Topic-XICL model.