Language-emphasized Cross-lingual In-context Learning for Multilingual LLM

Anonymous ACL submission

Abstract

Cross-lingual learning, which can transfer knowledge from high-resource languages to low-resource languages, has been widely studied. With the recent rise of large language models (LLMs), in-context learning (ICL) has shown remarkable performance, eliminating the need for fine-tuning parameters and reducing the reliance on extensive labeled data. It sounds tempting to use cross-lingual ICL to solve cross-lingual tasks based on multilingual LLMs. However, the intricacies of cross-lingual ICL remain underexplored. Prior studies on cross-lingual ICL overlooked the significance of language-specific nuances, neglecting not only the intrinsic linguistic properties of sentences but also the interlingual connections between sentences in different languages. In this paper, we propose a novel cross-lingual prompt structure: **L**anguage-**E**mphasized cross-lingual **I**n-context learning (**LEI**). LEI implements language alignment of demonstrations while introducing a third language (example language) as an example of language conversion to adapt LLMs to language conversion in cross-lingual tasks. Extensive experiments validate the state-of-the-art performance of LEI on 42 cross-lingual tasks. [1]

1 Introduction

Due to substantial disparities in the quantity of publicly available labeled datasets across different languages, the ability to learn from the high-resource source context to solve tasks in low-resource targets sounds enticing (Tanwar et al., 2023), which is known as cross-lingual learning. Traditional cross-lingual pre-trained language models with transformer structure, such as multilingual BERT (mBERT), have achieved effective cross-lingual transfer and performed surprisingly well on a large number of downstream tasks (Devlin et al., 2018). But those methods require language models to be fine-tuned on much supervised data for downstream tasks to improve performance on low-resource languages (Ruder et al., 2019). With the popularity of LLMs, cross-lingual learning can also be carried out by fine-tuning multilingual LLMs. Yet for multilingual LLMs with hundreds of millions of parameters, fine-tuning consumes a lot of computing resources.

Large Language Models (LLMs) (Radford et al., 2019) (Chowdhery et al., 2023)have demonstrated the ability to adapt to target tasks during inference through few-shot demonstrations (Wei et al., 2022), also referred to as in-context learning (ICL). In ICL, LLMs require input-output pairs from training data, often referred to as demonstrations (Brown et al., 2020), with subsequent inputs for testing. In this setup, LLMs predict the next token without updating the model parameters. At the same time, ICL relies on natural language instructions that humans can understand (Dong et al., 2022), so it provides a window for humans to explore the potential of LLMs and has a wide range of application prospects. The exploration of ICL using multilingual LLMs in cross-lingual scenarios is currently limited. There have been some notable advancements in task alignment for cross-lingual tasks recently (Tanwar et al., 2023). But the current ICL structure produces near-random prediction results when predicting certain languages, this situation occurs in both multilingual and single-language ICLs for some target languages (Tanwar et al., 2023) (Webson and Pavlick, 2021) (Lin et al., 2021). In addition, those prior studies on cross-lingual ICL overlooked the significance of language-specific nuances, neglecting not only the intrinsic linguistic properties of sentences but also the interlingual connections between sentences in different languages. There is currently a need for an ICL structure that focuses on cross-lingual tasks so that multilingual LLMs can better solve cross-lingual tasks. Figure 1 shows the current difficulties encountered by

[1] We will release our code when the paper is accepted.

cross-lingual ICL.

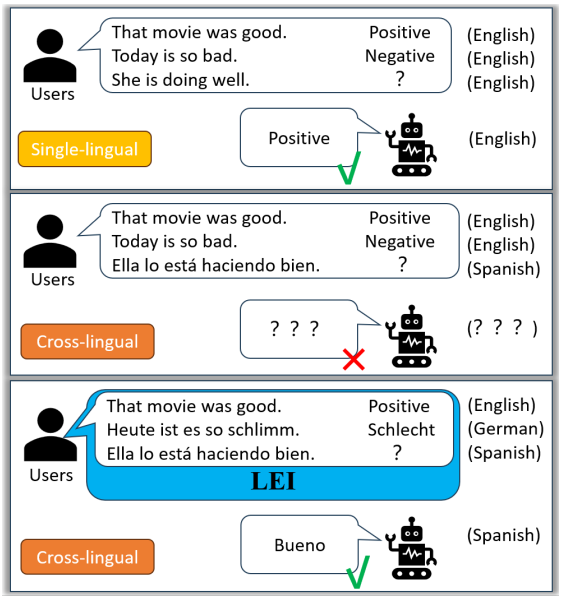

Figure 1: An illustration of cross-lingual in-context learning with two demonstrations, making a prediction between positive and negative in different languages. The German 'Heute ist es so schlimm. Schlecht' in the figure means 'Today is so bad. Negative'. The Spanish 'Ella lo está haciendo bien. Bueno' in the figure means 'She is doing well. Positive'. The first and second subfigures show that the traditional ICL format performs well in single-lingual tasks, but performs poorly in cross-lingual tasks. The third subfigure shows that after introducing the example language (German) and using the LEI structure, the LLM can achieve more accurate inference.

In this paper, we propose **LEI** (**L**anguage-**E**mphasized cross-lingual **I**n-context learning), a versatile cross-lingual ICL structure designed for adaptation to various language application scenarios. Specifically, we introduce a third language that is different from the source and target languages: the example language. Through the language conversion from the source language to the example language in the demonstration, we teach LLMs how to perform language conversion to complete the cross-lingual task from the source language to the target language. Then we add a new element 'Language' to the ICL structure suitable for cross-lingual tasks, which serves as a good language aligner. At the same time, we also introduce our task language in the instruction of ICL to improve the adaptability of the LLMs. To adapt to our structure, we modify task aligner (Tanwar et al., 2023) as a label aligner and apply it to our structure.

We evaluate the LEI on two multilingual sentiment analysis datasets. Our structure motivates LLMs to demonstrate strong adaptability to cross-lingual scenarios. LEI improves 35 of the 42 tasks compared to baselines, with an average relative improvement of a staggering **16**%. Additionally, LEI effectively addresses the issue of random prediction in some target languages and greatly improves the reasoning capabilities of ICL. In summary, we make the following contributions:

1)We are the first to introduce information about the language itself and the associations between languages to the cross-lingual ICL structure. And we identify that cross-lingual ICL is not sensitive to the number of demonstrations, but is greatly affected by the selected source language.

2) We design a cross-lingual in-context learning structure suitable for cross-lingual tasks by introducing a third language and emphasizing linguistic information.

3) Experimental results verify the state-of-the-art performance of LEI on two challenging datasets in cross-lingual in-context learning settings.

2 Related Work

2.1 In-context Learning

Brown et al. (2020) introduced in-context few-shot learning using the GPT-3 model. Many studies showed that the performance of ICL is sensitive to specific settings, including the prompting template, the selection of in-context examples, and order of examples, and so on (Zhao et al., 2021) (Lu et al., 2021). Dong et al. (2022) defined the formulation of ICL. They considered the key idea of in-context learning is to learn from analogy. Min et al. (2022a) showed that ICL primarily derives its benefits from the accurate distribution of inputs and labels, rather than the correspondence between input and label.

2.2 Cross-lingual Learning.

Early cross-lingual works required training word embeddings using multilingual datasets (Mikolov et al., 2013). Multilingual pretrained models represented by mBERT (Devlin et al., 2018) were also used to solve cross-language problems. Those language models could improve multilingual capability by augmenting data (Lin et al., 2022) or fine-tuning models (Chen et al., 2021), but they required more computing resources and labeled data. A few previous works used ICL to solve cross-lingual tasks. Zhang et al. (2021); Winata et al. (2021) only used randomly selected text-label pairs to sim-

ply explore cross-context learning. Tanwar et al. (2023) improved the cross-lingual demonstration in terms of the selection of examples and the alignment of tasks. However, their work did not focus on adapting LLMs to language conversion, which is a key point in cross-lingual ICL. Thus, we focus on the design of ICL demonstrations with language emphasis. Inspired by the essence of ICL, i.e., analogy, we use the analogy method to show the cross-lingual process to LLMs. Our experiments show that this approach inspires the cross-lingual capabilities of multilingual LLMs.

3 Our Method: LEI

In this section, we introduce **LEI**: **L**anguage-**E**mphasized cross-lingual **I**n-context learning, a versatile cross-lingual in-context learning structure designed for adaptation to various language application scenarios. In Section 3.1, we introduce the selection strategy of demonstrations. Section 3.2 describes the instruction based on the task and language. The implementation method of example language conversion is introduced in Section 3.3. Section 3.4 describes a language aligner for each input-output pair. In Section 3.5, we describe the label aligner and its application in our structure. Finally, we construct the ICL structure LEI and introduce the inference process in Section 3.6. The structure of each part of LEI is shown in Figure 2.

3.1 Similarity example selection

There is a source language s, an example language e, and a target language t. Let $L_s = \{(x_s^i, y_s^i)\}_i$ be a monolingual labeled dataset in language s, which is a collection of input example-label pairs. $x_s^i \in X_s$ and $y_s^i \in Y_s$. Similarly, there is $L_e = \{(x_e^i, y_e^i)\}_i$ for language e and $L_t = \{x_t^i\}_i$ for language t.Note that the example language is different from the source language and the target language that we introduce into the cross-lingual ICL, and we need to use the demonstration of the example language to adapt the LLMs to the language conversion. Its role is described in Section 3.3. Language s and language e are languages with more abundant labeled data, whose labeled datasets are very easy to obtain.

Our goal is to select k demonstrations from sufficient annotated data and combine them with a small number of natural language prompts as demonstrations. Tanwar et al. (2023) prove k-NN cross-lingual demonstrations can be retrieved for multi-lingual ICL to strengthen source-target language alignment. We improve their work and utilize multilingual sentencetransformers (Reimers and Gurevych, 2020) to extract the sentence embeddings of the test input $x_t \in L_t$ and the source inputs X_s and X_e. Based on the cosine similarity between the target input x_t^j and source input $x_s^i \in X_s$ with $x_e^i \in X_e$, we then extract the top k demonstrations. The specific selection method is as follows: Let $k_s = \lceil k/2 \rceil$ be the demonstration number of s, Let $k_e = k - k_s$ be the demonstration number of e. For one target language input example x_t^j, by multilingual sentence encoder θ ,we get all the embeddings m_s^i and m_e^i, $m_s^i = \theta(x_s^i)$, $m_e^i = \theta(x_e^i)$, and get $m_t^j = \theta(x_t^j)$. Then we calculate the similarity score c_s^i between m_s^i and m_t^j:

$$c_s^i = \frac{m_t^j \cdot m_s^i}{\left\|m_t^j\right\|_2 \|m_s^i\|_2}. \quad (1)$$

The similarity score c_e^i between m_e^i and m_t^j is:

$$c_e^i = \frac{m_t^j \cdot m_e^i}{\left\|m_t^j\right\|_2 \|m_e^i\|_2}. \quad (2)$$

Then we select top k_s sentences based on c_s^i: $\{x_s^1, ..., x_s^{k_s}\}$, top k_e sentences based on c_e^i: $\{x_e^1, ..., x_e^{k_e}\}$. At the same time, we also get their corresponding labels $\{y_s^1, ..., y_s^{k_s}\}$ and $\{y_e^1, ..., y_e^{k_e}\}$. We accomplish choosing k examples for each input x_t^j.

3.2 Instruction

Dong et al. (2022) define the formulation of ICL: $C = \{I, s(x_1, y_1), ...s(x_k, y_k)\}$. Task instruction I often be used to indicate specific tasks. In a cross-lingual task, we need to give instruction not only to the task but also to the language to be used. We create $L_I = \{I_{s,e,t}\}$ for a given task and source language s, example language e, target language t. Our instruction has two parts: task introduction and language introduction. Task introduction describes the task we will accomplish, language introduction emphasizes what language will be used. For example, when the source language is English, the example language is Japanese and the target language is German, and the task is emotional classification, "Sentiment classification of the rating text, with the example language being English and Japanese, and the last sentence in German." will be the Instruction.

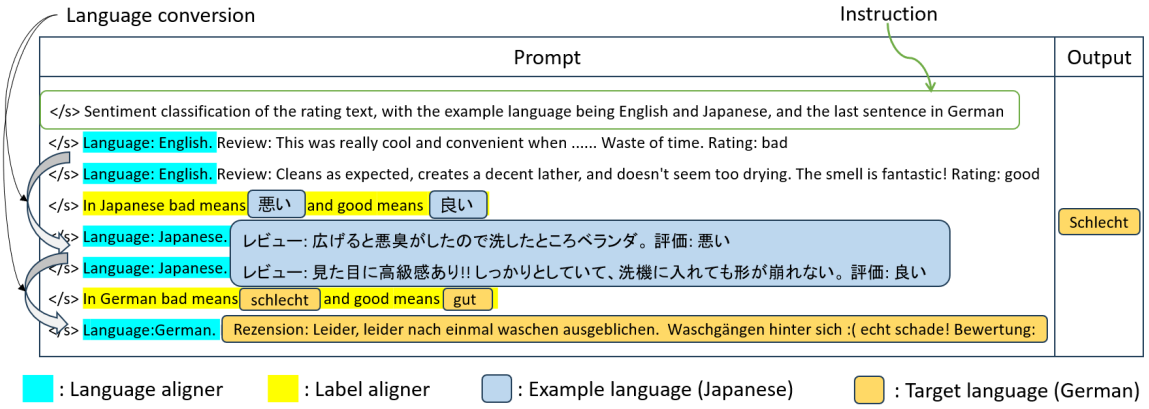

Figure 2: Explanation of LEI structure. In this figure, demonstration number k is 4. The source language is English, the example language is Japanese, and the target language is German.

3.3 Example language conversion

We believe that abrupt language conversions are an important constraint on multilingual LLMs. Like the essence of ICL: learn from analogy, we decide to use an analogy to adapt the LLM to example language conversion. So we choose an example language based on the source language. Labeled data is readily available in both languages. As we define in Section 3.1, we select k_s and k_e demonstrations from source language s and example language e, respectively, to form k demonstrations. The demonstrations are arranged in the following way: we put the demonstrations of the source language s in front of the demonstrations of the example language e. Target language input will be placed last. In this way, with the conversion example from source language to example language, the LLM can learn the language conversion from it, to better complete the final task of the target language.

3.4 Language aligner

Dong et al. (2022) have constructed the formulation of ICL. This structure has achieved good results in most of the tasks. However, in cross-lingual tasks, this structure often struggles to perform well. We add a new element 'Language' to the structure. It can explicitly linguistically align the demonstrations. Let $L_l = \{A_{s,t'}\}$ be a language aligner set. It should be noted that language s is the source language we define in Section 3.1, but language t' depends on the language of the current input. For the same example, when the source language is English, the example language is Japanese and the target language is German, there will be three different language aligners for three different inputs of language. For the source language English input, our language aligner is $A_{s,s}$ i.e. "Language: English". For the example language Japanese input, our language aligner is $A_{s,e}$ i.e. "Language: English". For the target language German input, our language aligner is $A_{s,t}$ i.e. "Language: German".

3.5 Label aligner

After completing the input language alignment, we proceed to align the labels. We improve the design of Tanwar et al. (2023). We think the task alignment in their article serves as a label alignment. We have modified it in conjunction with our cross-lingual demonstration structure. We create $L_a = B_{s,t'}$ as a collection of statements that emphasizes the difference between the labels of language s and language t'. Note that t' depends on the language of the next input when a language conversion occurs for a demonstration. For the same example, when the source language is English, the example language is Japanese and the target language is German, there will be two different label aligners for two different converts of the demonstration language. When the language is converted from source language s to example language e, our label aligner is $B_{s,e}$: "In Japanese bad means 悪い and good means 良い.", it maps the source language label (bad and good) to the example language label (悪い and 良い) of the same meaning. Similarly, when the language is converted from source language s to target language t, our label aligner is $B_{s,t}$: "In German bad means schlecht and good means gut." ('schlecht' and 'gut' in German mean 'bad' and 'good' in English.)

3.6 Inference

Based on the test input $x_t \in L_t$, we define a demonstration set C that can be adapted to cross-language tasks:

$$C = \{I_{s,e,t}, (A_{s,s}, x_s^1, y_s^1), ..., (A_{s,s}, x_s^{k_s}, y_s^{k_s}) , B_{s,e}, (A_{s,e}, x_e^1, y_e^1), ..., (A_{s,e}, x_e^{k_e}, y_e^{k_e}), B_{s,t}\}. \quad (3)$$

Then we also need to perform language alignment on the input of target language t for the test input x_t:

$$x_t' = (A_{s,t}, x_t). \quad (4)$$

Then we infer the label $y_t \in Y_t$, where Y_t is a set of candidate answers in language t corresponding to the new test input x_t':

$$\hat{y} = arg \max_{y_t \in Y_t} f_M(y_t|C, x_t'), \quad (5)$$

where $f_M(\cdot)$ is the score function of the whole input sequence with the LLM M.

4 Experimental Setup

4.1 Mutilingual LLM

After many comparative experiments with other large models, we are consistent with experiments of Tanwar et al. (2023). We experiment with a multilingual LLM XGLM (Lin et al., 2021) 7.5 billion variant. It shows better performance on multiple tasks than other large multi-language models.

4.2 Datasets

We experiment on two datasets: Multilingual Amazon Reviews Corpus (MARC) (Keung et al., 2020) and Cross-language sentiment classification (CLS) (Prettenhofer and Stein, 2010). Different from the setting of Tanwar et al. (2023), we do not choose another dataset with only two languages, because our experimental setup requires at least three languages. At the same time, we reconstruct the two datasets we adopted. (We describe the two datasets in detail in Appendix A. The languages in the data set are introduced in Appendix B)

4.3 Baselines

There are two methods we use to compare with LEI. Random Prompting is the most classic method for constructing demonstrations. This method is also used by Tanwar et al. (2023) as a contrast. The demonstration examples used in ICL are all randomly selected. X-InSTA (Tanwar et al., 2023) proposed methods of semantic alignment and task alignment. Semantic alignment is to find the k sentences that are semantically closest to the target language input within the source language data set as demonstrations. Task alignment is the label aligner we continue to use.

4.4 Implementation Details

Following the setting of X-InSTA (Tanwar et al., 2023), we set the demonstration number k to be 4, and the maximum input length to be 1024 tokens. See Appendix C for specific hyperparameter settings. If the demonstration text is too long, we will truncate the demonstration text. For each source-target language task, we select unused languages in the dataset as example languages one by one. We calculate the result of this source-target language task as the average of the results of all example language situations. We calculate the Macro-F1 scores of the source-target language task in the experimental tables as the average of the Macro-F1 scores of all example language situations. For example, in the dataset MARC, if we select English (es) as the source language and German (de) as the target language, then we will take turns selecting other languages in the data set (Spanish (es), French (fr), Japanese (ja), Mandarin (zh)) as the example language. They are used as four subtasks, and the average Macro-F1 scores of these four subtasks are used as the result of the task of English as the source language and German as the target language. Complete results are presented in the Appendix D. We construct the data set into 10 language-task pairs, 42 source-target language cross-lingual settings, and set up a total of 144 specific subtasks based on different example languages e.

5 Results and Analysis

5.1 Main results

Results of LEI on two datasets are reported in Table 1 and 2. (See Appendix D for complete experimental results.) Note that for the sake of fairness, the baseline results are reproduced in our environment according to the setting of Tanwar et al. (2023). On the MARC dataset, the macro F1 scores of different tasks increased by an average of 14.6%. On the CLS dataset, the macro F1 scores of different tasks increased by an average of 19%.

It is worth noting that in the context where the target language is German, LEI stimulates the extremely strong reasoning ability of LLM. The ran-

dom prediction of LLM is avoided. In tasks with German as the target language, the F1 score improvements reached 94.7% and 62.4% respectively.

However, the performance of LEI in certain languages (such as Mandarin) is still unsatisfactory. It may be that the large language model is still lacking in capabilities in this language and cannot complete these language tasks in a cross-lingual ICL manner. This issue has also been raised by previous works (Tanwar et al., 2023).

SRC \ TAR	de	en	es	fr	ja	zh
Random Prompting						
de	-	0.446	0.517	0.547	0.454	**0.413**
en	0.380	-	0.761	0.663	0.526	**0.362**
es	0.339	0.696	-	0.563	0.519	**0.445**
fr	0.340	0.692	0.864	-	0.479	**0.410**
ja	0.333	0.701	0.678	0.612	-	**0.678**
zh	0.333	0.632	0.836	0.402	0.521	-
AVG	0.345	0.633	0.731	0.557	0.499	**0.462**
X-InSTA(2023)						
de	-	0.721	0.666	0.865	0.718	0.337
en	0.397	-	**0.886**	0.790	0.783	0.341
es	0.348	0.857	-	0.892	0.835	0.339
fr	0.354	0.849	0.900	-	0.779	0.350
ja	0.333	0.817	**0.890**	0.808	-	0.372
zh	0.333	0.713	0.890	0.750	0.797	-
AVG	0.353	0.791	0.847	0.821	0.782	0.348
LEI(Ours)						
de	-	**0.804**	**0.866**	**0.866**	**0.867**	0.356
en	**0.809**	-	0.871	**0.883**	**0.859**	0.36
es	**0.644**	**0.888**	-	**0.906**	**0.897**	0.343
fr	**0.699**	**0.862**	**0.902**	-	**0.889**	0.377
ja	**0.599**	**0.892**	0.848	**0.809**	-	0.409
zh	**0.685**	**0.769**	**0.893**	**0.824**	**0.902**	-
AVG	**0.687**	**0.843**	**0.876**	**0.858**	**0.883**	0.369

Table 1: Comparison of Macro-F1 scores between two ICL methods on MARC. 'SRC' means source language, and 'TAR' means target language. The same meaning is represented in the table that follows.

Different demonstration number. To verify the generality of our method, we conduct experiments on different numbers of demonstrations k in two datasets ($k = 2, 3, 4$). The result is reported in Table 3 and 4. In most experimental settings, LEI is overall ahead of the baseline. We enable LLM to make accurate inferences using only two demonstrations. This also proves the effectiveness of our method. A very important finding is that the value of demonstration number k has little impact on the reasoning ability of LLM, which is different from the experimental conclusion of (Min et al., 2022b). We believe that for text in the target language, too many demonstrations in other languages will cause LLMs to become dependent on this language. This affects LLM's cross-lingual reasoning capabilities.

SRC \ TAR	de	en	fr	ja
Random Prompting				
de	-	0.517	0.597	0.618
en	0.682	-	0.412	0.609
fr	0.545	0.694	-	0.666
ja	0.344	0.595	0.475	-
AVG	0.524	0.621	0.543	0.697
X-InSTA(2023)				
de	-	0.622	0.788	0.779
en	0.588	-	0.778	0.794
fr	0.524	0.821	-	0.834
ja	0.339	0.701	0.705	-
AVG	0.483	0.715	0.757	0.802
LEI(Ours)				
de	-	**0.736**	**0.868**	**0.828**
en	**0.806**	-	**0.902**	**0.818**
fr	**0.793**	**0.845**	-	**0.862**
ja	**0.757**	**0.883**	**0.751**	-
AVG	**0.785**	**0.821**	**0.84**	**0.836**

Table 2: Comparison of Macro-F1 scores between two ICL methods on CLS.

Method \ TAR	de	en	es	fr	ja	zh
k = 2						
X-InSTA	36.3	81.7	86.9	83.9	80.7	34.0
LEI	**71.8**	**86.6**	**87.1**	**84.9**	**84.0**	**45.5**
k = 3						
X-InSTA	33.2	81.1	**86.9**	84.6	80.9	33.9
LEI	**72.8**	**85.5**	86.8	**85.3**	**85.5**	**41.5**
k = 4						
X-InSTA	35.3	79.1	84.7	82.1	78.2	34.8
LEI	**68.7**	**84.3**	**87.6**	**85.8**	**88.3**	**36.9**

Table 3: Comparison of Macro-F1 scores (%) between two ICL methods in different settings of k on MARC. The results in the table are the average of all experimental results for a target language.

Method \ TAR	de	en	fr	ja
k = 2				
X-InSTA	51.5	75.6	75.4	78.9
LEI	**78.9**	**84.7**	**83.0**	**79.7**
k = 3				
X-InSTA	50.5	73.2	75.9	80.3
LEI	**78.6**	**83.8**	**84.0**	**81.3**
k = 4				
X-InSTA	48.3	71.5	75.7	80.2
LEI	**78.5**	**82.1**	**84.0**	**83.6**

Table 4: Comparison of Macro-F1 scores (%) between two ICL methods in different settings of k on CLS.

Setting \ TAR	MARC						CLS			
	de	en	es	fr	ja	zh	de	en	fr	ja
LEI	**0.687**	**0.843**	0.876	**0.858**	**0.883**	0.369	**0.785**	**0.821**	**0.840**	**0.836**
w/o e-l-c	0.426	0.842	**0.885**	0.845	0.821	**0.370**	0.575	0.760	0.836	0.820
w/o l-a	0.655	0.769	0.874	0.848	0.856	0.361	0.762	0.749	0.829	0.794
w/o instruction	0.684	0.816	0.855	0.833	0.875	0.362	0.769	0.790	0.756	0.827

Table 5: Ablation study on the contribution of different parts of the LEI structure under two datasets. Mean macro-F1 scores are reported for every target language. (w/o means without, e-l-c means *example language conversion*, l-a means *language aligner*, inst- means *instruction*.)

5.2 Ablation Study and Analysis

Function of different parts. To investigate the effect of the proposed methods, we perform the ablation study as shown in Table 5. We find that in most cases when we independently removed certain components of the LEI, the performance dropped significantly. This suggests that all three language emphasis components contribute to the overall structure. And, using three language emphasis strategies at the same time can further improve performance, which shows that the three language emphasis strategies are complementary.

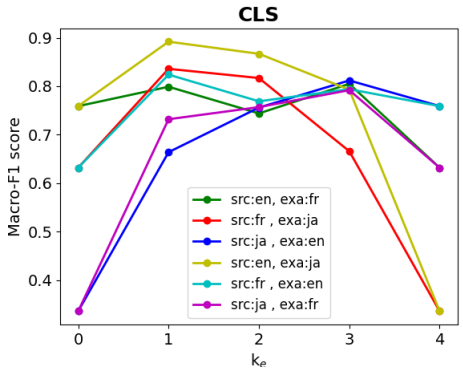

Figure 3: Line chart of changes in macro-F1 scores with k_e under different source language and example language settings, e.g., en-fr represents that the source language is English and the example language is French. The dataset we used is CLS.

Effect of Example Language Conversion. It can be seen from ablation experiments that *example language conversion* can significantly improve the reasoning capabilities of LLM in certain target language tasks. To evaluate the impact of introducing an example language on LLM's cross-lingual ICL capabilities, we take German as the target language as an example. While keeping the total number of demonstrations $k = 4$ unchanged, we use different example language demonstration numbers k_e on the two datasets to construct LEI. The result is shown in the figure 3 and 4. $k_e = 0$ means that the demonstration of the example language is not used. $k_e = 4$ means that the k demonstrations used are all in example language. We can observe that when German is used as the target language, the introduction of an example language ($k_e = 1, 2, 3$) can greatly improve the reasoning ability of LLMs. This situation is seen in other languages as well. We infer that LLM learns from language conversion of demonstrations to reason across language contexts.

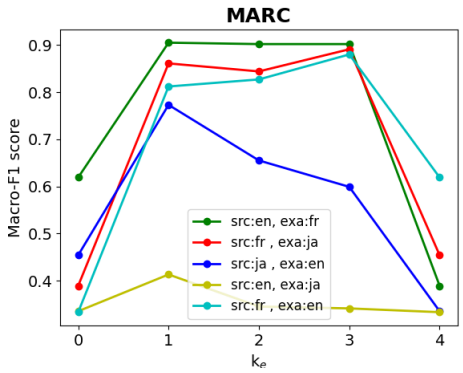

Figure 4: Line chart of changes in macro-F1 scores with k_e under different source language and example language settings, e.g., en-fr represents that the source language is English and the example language is French. The dataset we used is MARC.

SRC \ TAR	de	en	es	fr	ja	zh
de	-	0.804	0.866	0.866	0.867	0.356
en	**0.809**	-	0.871	0.883	0.859	0.36
es	0.644	0.888	-	**0.906**	0.897	0.343
fr	0.699	0.862	**0.902**	-	0.889	0.377
ja	0.599	**0.892**	0.848	0.809	-	**0.409**
zh	0.685	0.769	0.893	0.824	**0.902**	-
AVG	0.687	0.843	0.876	0.858	0.883	0.369

Table 6: Macro-F1 scores when used LEI on MARC. Bold refers to the result of the source language performing the best in this target language task.

Source Language Selection. Cross-lingual tasks are usually oriented to a language with less labeled data. However, there may be more languages that have a large amount of labeled data. How to choose a suitable language as a source language as an example of cross-lingual ICL is also worth studying. We focus on source languages that belong to the same language family as the target language. A comparison table of languages and language families can be found in the appendix B. Among them, although Japanese and Mandarin do not belong to the same language family, the characters of the two languages are relatively similar, so we also regard them as languages similar to the same language family. In Tabel 6, we find that LLMs have the best inference effect when using a language in the same language family as the target language as the source language. (German and English belong to the same language family, Spanish and French belong to the same language family.) The exception is English as the target language. The reason is that multilingual LLMs use much more English data than other languages when training. Therefore, we emphasize that when choosing a source language, you need to choose a language family that belongs to the same language family as the target language or has a relatively similar script.

The more demonstrations, the worse the effect? Section 5.1 shows the performance of LEI under different numbers of demonstrations k, which displays a counter-intuitive phenomenon. The increase in the number of demonstrations does not improve the reasoning ability of LLMs. We conduct comparative experiments on the MARC dataset between single-lingual tasks and cross-lingual tasks under the two methods, whose result is shown in the figure 5. In order to ensure the universality of the results, we remove languages similar to Chinese that would cause LLMs to generate random guesses during the experiment. Experimental results show that single-language tasks are very sensitive to the value of k, while the opposite is true for cross-language tasks. The value of k has little impact on it. Even when k is larger, it will affect the reasoning ability of LLMs. While the LEI is better than X-InSTA (Tanwar et al., 2023) in terms of performance, it can also make better use of the increased demonstrations, avoiding the negative impact of these different language demonstrations on the inference of the LLM.

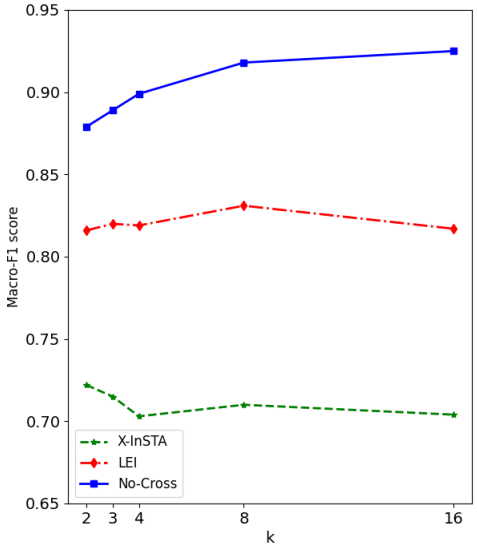

Figure 5: Line chart of changes in macro-F1 scores with k under different method and setting. Where X-InSTA is the method in the baselines. No-cross denotes a single language task. That is, the source language and the target language are the same language. The value of each point in the graph is the average of the results of all target language subtasks.

6 Conclusion

In this work, we explore the problem of LLM language conversion adaptation in cross-lingual ICL. For the first time, we find that introducing a third example language can improve cross-lingual ICL capabilities, which perform extremely well in some languages (improved from random prediction to a certain degree of accuracy). At the same time, adding a language aligner component to cross-lingual ICL can also realize the language alignment function and improve cross-lingual capabilities. Based on the above findings, we designed a cross-lingual ICL structure based on language enhancement: LEI. LEI has achieved great improvements over traditional cross-lingual ICL methods in various cross-lingual tasks. We examine the characteristics of cross-lingual ICL and find that cross-lingual ICL is not sensitive to the number of demonstrations, while the language family has a greater impact on the performance of cross-lingual ICL. These provide guidance for future cross-lingual ICL research directions.

Limitation

Poor performance in specific languages: As shown in the experimental results, LEI performs very poorly in Chinese. There are also many studies that have found this problem, that is, certain languages perform poorly in cross-lingual environments. The reason is that the segmentation of Chinese in tokenization is quite different from that of other languages, and the training during the pretraining process is insufficient. This is also one of the problems that multilingual LLMs need to solve.

Input length: Texts such as human comments are often very long, so in some situations, prompts may exceed the limit of 1024 tokens. However, summarizing long texts in multiple languages is likely to lead to a lack of semantics and thus affect the performance of cross-lingual ICL. At the same time, it also makes the operation too complicated, which is contrary to the original intention of ICL's simple demonstration. We can only truncate demonstration text that is too long. We are also continuing to work on synopsis summary structures for multiple languages.

Beyond classification: Our work only targets cross-lingual classification tasks, and cross-lingual generation tasks have not yet been explored. Our future work will also focus on cross-lingual natural language generation tasks.

Larger LLMs: Due to limitations of computing resources, we did not verify our experiments in LLMs with larger parameter scales.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. *arXiv preprint arXiv:2104.08757*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A Smith. 2020. The multilingual amazon reviews corpus. *arXiv preprint arXiv:2010.02573*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Computer ence*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *arXiv preprint arXiv:2108.13161*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Datasets

Multilingual Amazon Reviews Corpus: MARC (Keung et al., 2020) is a large-scale multilingual corpus of Amazon reviews of customers. There are six languages in this dataset: English, Spanish, German, French, Japanese, and Mandarin. Each language possesses a training dataset comprising 200K instances utilized for selecting our demonstrations, along with a test set consisting of 40,000 reviews categorized as either positive or negative.

Cross-language sentiment classification: CLS (Prettenhofer and Stein, 2010) is a multilingual corpus of four languages – German, English, French, and Japanese. The dataset comprises reviews on DVDs, music, and books, featuring a training set and a test set, each containing 2,000 sentences for every language, categorized as either negative or positive.

B Language description

The language used in the experiment and related information are shown in Table 7.

Language	Language Family	ISO 639-1 code
GERMAN	IE: GERMANIC	DE
ENGLISH	IE: GERMANIC	EN
FRENCH	IE: ITALIC	FR
SPANISH	IE: ITALIC	ES
JAPANESE	JAPANIC	JA
MANDARIN	SINO-TIBETAN	ZH

Table 7: list of languages along with their corresponding ISO codes utilized in our experiments.

C Hyperparameters

All codes were written in PyTorch. We utilized the Huggingface repository to load the LLM and used the sentence transformer to extract semantic similarity. Sklearn was used to calculate the F1 score. Table 8 shows the experimental environment and hyperparameter settings.

D Complete Result

The complete experimental data in the main experiment (including the different values of k and all the selected example languages) are shown in Tables 9, 10, 11, 12, 13, 14.

Hyperparameter	Value
Model	XGLM-7.5B
GPU	NVIDIA RTX 4090
Batch Size	2
Max length	1024
k	2,3,4,8,16

Table 8: Hyperparameters in experiments.

	TAR	de	en	es	fr	ja	zh
SRC	EXA						
de	en	-	-	0.863	0.836	0.813	0.336
de	es	-	0.871	-	0.864	0.855	0.334
de	fr	-	0.871	0.897	-	0.820	0.338
de	ja	-	0.760	0.820	0.837	-	0.617
de	zh	-	0.852	0.876	0.811	0.858	-
en	de	-	-	0.876	0.895	0.837	0.346
en	es	0.890	-	-	0.864	0.843	0.370
en	fr	0.885	-	0.881	-	0.822	0.402
en	ja	0.902	-	0.859	0.860	-	0.652
en	zh	0.747	-	0.887	0.895	0.859	-
es	de	-	0.891	-	0.877	0.824	0.334
es	en	0.874	-	-	0.892	0.862	0.341
es	fr	0.832	0.874	-	-	0.851	0.342
es	ja	0.638	0.886	-	0.872	-	0.602
es	zh	0.401	0.879	-	0.836	0.824	-
fr	de	-	0.885	0.884	-	0.827	0.342
fr	en	0.891	-	0.894	-	0.833	0.402
fr	es	0.850	0.904	-	-	0.870	0.360
fr	ja	0.662	0.888	0.860	-	-	0.672
fr	zh	0.477	0.868	0.876	-	0.812	-
ja	de	-	0.901	0.899	0.862	-	0.599
ja	en	0.639	-	0.874	0.871	-	0.555
ja	es	0.755	0.874	-	0.872	-	0.586
ja	fr	0.823	0.888	0.894	-	-	0.574
ja	zh	0.369	0.869	0.867	0.835	-	-
zh	de	-	0.788	0.833	0.787	0.807	-
zh	en	0.766	-	0.887	0.844	0.846	-
zh	es	0.740	0.843	-	0.771	0.866	-
zh	fr	0.848	0.810	0.882	-	0.835	-
zh	ja	0.367	0.822	0.854	0.828	-	-

Table 9: Macro-F1 scores of LEI on MARC. EXA in the table is the example language. Among $k = 2$.

	TAR	de	en	fr	ja
SRC	EXA				
de	en	-	-	0.819	0.799
de	fr	-	0.795	-	0.793
de	ja	-	0.768	0.842	-
en	de	-	-	0.844	0.785
en	fr	0.778	-	-	0.780
en	ja	0.877	-	0.86	-
fr	de	-	0.871	-	0.805
fr	en	0.839	-	-	0.822
fr	ja	0.760	0.868	-	-
ja	de	-	0.881	0.825	-
ja	en	0.741	-	0.787	-
ja	fr	0.837	0.896	-	-

Table 10: Macro-F1 scores of LEI on CLS. EXA in the table is the example language. Among $k = 2$.

	TAR	de	en	es	fr	ja	zh
SRC	EXA						
de	en	-	-	0.856	0.854	0.831	0.333
de	es	-	0.819	-	0.855	0.848	0.333
de	fr	-	0.841	0.896	-	0.814	0.334
de	ja	-	0.800	0.746	0.852	-	0.505
de	zh	-	0.829	0.872	0.821	0.868	-
en	de	-	-	0.849	0.898	0.845	0.334
en	es	0.903	-	-	0.875	0.855	0.335
en	fr	0.866	-	0.881	-	0.823	0.350
en	ja	0.918	-	0.831	0.888	-	0.560
en	zh	0.792	-	0.889	0.844	0.875	-
es	de	-	0.893	-	0.884	0.856	0.334
es	en	0.863	-	-	0.899	0.863	0.335
es	fr	0.830	0.883	-	-	0.867	0.335
es	ja	0.657	0.898	-	0.898	-	0.535
es	zh	0.409	0.882	-	0.833	0.831	-
fr	de	-	0.878	0.887	-	0.832	0.334
fr	en	0.886	-	0.905	-	0.849	0.339
fr	es	0.877	0.876	-	-	0.863	0.339
fr	ja	0.730	0.877	0.887	-	-	0.562
fr	zh	0.562	0.865	0.897	-	0.882	-
ja	de	-	0.896	0.897	0.849	-	0.516
ja	en	0.661	-	0.845	0.824	-	0.498
ja	es	0.748	0.877	-	0.877	-	0.539
ja	fr	0.788	0.888	0.865	-	-	0.555
ja	zh	0.377	0.856	0.859	0.815	-	-
zh	de	-	0.771	0.850	0.810	0.871	-
zh	en	0.762	-	0.887	0.864	0.854	-
zh	es	0.724	0.813	-	0.764	0.895	-
zh	fr	0.843	0.877	0.895	-	0.869	-
zh	ja	0.364	0.782	0.859	0.855	-	-

Table 11: Macro-F1 scores of LEI on MARC. EXA in the table is the example language. Among $k = 3$.

	TAR	de	en	fr	ja
SRC	EXA				
de	en	-	-	0.883	0.809
de	fr	-	0.768	-	0.793
de	ja	-	0.734	0.842	-
en	de	-	-	0.889	0.821
en	fr	0.780	-	-	0.800
en	ja	0.896	-	0.862	-
fr	de	-	0.876	-	0.820
fr	en	0.825	-	-	0.830
fr	ja	0.825	0.863	-	-
ja	de	-	0.894	0.800	-
ja	en	0.642	-	0.762	-
ja	fr	0.749	0.888	-	-

Table 12: Macro-F1 scores of LEI on CLS. EXA in the table is the example language. Among $k = 3$.

	TAR	de	en	es	fr	ja	zh
SRC	EXA						
de	en	-	-	0.867	0.873	0.875	0.333
de	es	-	0.800	-	0.850	0.859	0.333
de	fr	-	0.812	0.912	-	0.828	0.333
de	ja	-	0.782	0.784	0.875	-	0.423
de	zh	-	0.823	0.902	0.866	0.906	-
en	de	-	-	0.834	0.913	0.853	0.333
en	es	0.902	-	-	0.847	0.853	0.333
en	fr	0.859	-	0.892	-	0.832	0.337
en	ja	0.908	-	0.846	0.900	-	0.438
en	zh	0.566	-	0.910	0.872	0.897	-
es	de	-	0.900	-	0.918	0.909	0.333
es	en	0.855	-	-	0.917	0.913	0.333
es	fr	0.844	0.870	-	-	0.896	0.333
es	ja	0.534	0.880	-	0.915	-	0.374
es	zh	0.342	0.900	-	0.875	0.871	-
fr	de	-	0.872	0.903	-	0.888	0.333
fr	en	0.857	-	0.910	-	0.893	0.333
fr	es	0.875	0.857	-	-	0.866	0.334
fr	ja	0.655	0.845	0.895	-	-	0.507
fr	zh	0.407	0.875	0.898	-	0.907	-
ja	de	-	0.912	0.887	0.722	-	0.416
ja	en	0.604	-	0.864	0.801	-	0.392
ja	es	0.675	0.897	-	0.898	-	0.413
ja	fr	0.770	0.901	0.863	-	-	0.416
ja	zh	0.345	0.857	0.777	0.813	-	-
zh	de	-	0.746	0.855	0.852	0.901	-
zh	en	0.827	-	0.910	0.878	0.904	-
zh	es	0.679	0.800	-	0.694	0.904	-
zh	fr	0.874	0.764	0.912	-	0.898	-
zh	ja	0.36	0.765	0.894	0.873	-	-

Table 13: Macro-F1 scores of LEI on MARC. EXA in the table is the example language. Among $k = 4$.

	TAR	de	en	fr	ja
SRC	EXA				
de	en	-	-	0.868	0.835
de	fr	-	0.761	-	0.820
de	ja	-	0.710	0.867	-
en	de	-	-	0.904	0.804
en	fr	0.744	-	-	0.832
en	ja	0.867	-	0.899	-
fr	de	-	0.857	-	0.862
fr	en	0.769	-	-	0.862
fr	ja	0.817	0.832	-	-
ja	de	-	0.876	0.607	-
ja	en	0.756	-	0.831	-
ja	fr	0.757	0.890	-	-

Table 14: Macro-F1 scores of LEI on CLS. EXA in the table is the example language. Among $k = 4$.