Extending LLMs to New Languages: A Case Study of Llama and Persian Adaptation

Samin Mahdizadeh Sani[1], Pouya Sadeghi[1], Thuy-Trang Vu[3], Yadollah Yaghoobzadeh[1,2], Gholamreza Haffari[3]

[1]Department of Electrical and Computer Engineering, University of Tehran, Iran
[2]Tehran Institute for Advanced Studies, Khatam University, Iran
[3]Department of Data Science & AI, Monash University, Australia

[1]{samin.mehdizadeh, pouya.sadeghi, y.yaghoobzadeh}@ut.ac.ir [3]{trang.vu1, gholamreza.haffari}@monash.edu

Abstract

Large language models (LLMs) have made great progress in classification and text generation tasks. However, they are mainly trained on English data and often struggle with low-resource languages. In this study, we explore adding a new language, i.e., Persian, to Llama (a model with a limited understanding of Persian) using parameter-efficient fine-tuning. We employ a multi-stage approach involving pretraining on monolingual Persian data, aligning representations through bilingual pretraining and instruction datasets, and instruction-tuning with task-specific datasets. We evaluate the model's performance at each stage on generation and classification tasks. Our findings suggest that incorporating the Persian language, through bilingual data alignment, can enhance classification accuracy for Persian tasks, with no adverse impact and sometimes even improvements on English tasks. Additionally, the results highlight the model's initial strength as a critical factor when working with limited training data, with cross-lingual alignment offering minimal benefits for the low-resource language. Knowledge transfer from English to Persian has a marginal effect, primarily benefiting simple classification tasks.

1 Introduction

The emergence of large language models (LLMs) has transformed natural language processing (NLP), leading to significant progress in various applications like machine translation, text generation, and sentiment analysis. Models such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI et al., 2024), and open-source alternatives like Llama-2 (Touvron et al., 2023), Llama-3 (Li et al., 2024) and Mistral (Jiang et al., 2023) have shown impressive abilities in understanding and generating human language. However, these advancements have mostly centered around English and other widely spoken languages, leaving less-resourced languages behind.

In today's connected world, supporting multiple languages in a single model is key to breaking language barriers and making technology accessible to everyone, including speakers of less common languages. Multilingual models help expand the reach of language technologies while having their challenges. Limited data for some languages, keeping performance consistent across languages, and preserving the model's core abilities while adding new languages are the existing challenges (Muennighoff et al., 2023; Qi et al., 2023; Vu et al., 2022). Overcoming these challenges is essential for creating inclusive global language technologies.

Developing multilingual large language models (MLLMs) has become crucial to address these challenges. Models such as XLM-R (Conneau et al., 2020), Qwen (Bai et al., 2023), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI et al., 2024), and Llama-3 (Li et al., 2024) leverage large multilingual datasets to enhance performance across different languages. However, problems like language imbalance persist, with high-resource languages dominating the training data, resulting in less effective outcomes for low-resource languages.

In this work, we explore cross-lingual adaptation of an LLM, i.e., Llama-2 , focusing on Persian as the target language. Persian, an Indo-European language with a non-Latin script, presents unique challenges due to its linguistic distance from English, the language most LLMs are primarily trained on. Despite its rich literary history and widespread use, Persian has not fully benefited from advancements in LLMs, largely due to limited annotated data and research efforts (Abaskohi et al., 2024). This makes Persian an ideal case study for testing cross-lingual adaptation. We evaluate several models, ranging from a fully English-trained model fine-tuned on Persian data to models with varying degrees of alignment, such as freezing most parameters, and models with additional pre-training combined with Low-Rank Adaptation (LoRA) (Hu et al., 2022).

Proceedings of the 31st International Conference on Computational Linguistics, pages 8868–8884
January 19–24, 2025. ©2025 Association for Computational Linguistics

These configurations create a spectrum of models with different levels of understanding of Persian. Our goal was to examine the differences in performance across these models and assess how well knowledge transfer from English to Persian occurred.

For this study, we use two datasets: one bilingual dataset with parallel data in English and Persian, and another monolingual dataset entirely in Persian. The bilingual dataset allows us to perform cross-lingual training, helping the model learn from aligned English-Persian data. In contrast, the Persian-only dataset is used to fine-tune the model solely on Persian tasks, providing insights into how well the model performs with exclusive exposure to the target language.

Although there is work adapting LoRA for Persian (Abbasi et al., 2023; Rostami et al., 2024), they typically focus on a single model for this language. Our work, however, explores different settings for model pre-training and instruction-tuning across both classification and generation tasks, highlighting the effort required to adapt models for specific tasks.

Our findings indicate that for classification tasks, aligning models with bilingual data is sufficient, and further pre-training on monolingual data is not crucial. However, when dealing with limited data or generation tasks, further pre-training proves beneficial and improves performance, as generation requires a deeper understanding of linguistic structures. Moreover, evaluations of several models that initially support the Persian language reveal the weaknesses of Llama models for this language, leaving it behind Gemma (Team et al., 2024b) and Qwen models. However, the comparison of our instruction-tuned model with state-of-the-art systems underscores the potential of targeted fine-tuning. Our model achieves performance comparable to these advanced models across most tasks. Notably, for simple yet unseen tasks like sentiment analysis, the model demonstrates an ability to generalize patterns from other instructions. Nevertheless, for more complex tasks that challenge the model's knowledge, it struggles to perform effectively without explicit instructions.

2 Related Work

Cross-lingual Transfer. The rapid advancement of large-scale language models, such as GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI et al., 2024), has significantly enhanced the capabilities of natural language processing (NLP). Moreover, Recent contributions from the open-source community, including Llama (Touvron et al., 2023) and Mistral (Jiang et al., 2023), show that open-source LLMs can now compete effectively with their closed-source counterparts. However, the heavy focus on English limits the flexibility of these models when incorporating new languages, particularly low-resource ones, which may not be initially supported in the training phase. Recent studies have leveraged the proficiency of language models in English to enhance performance on low-resource languages. Some of these studies focus on translating model input into English, demonstrating notable improvements for low-resource languages (Upadhayay and Behzadan, 2023; Abaskohi et al., 2024). Another approach explores transliteration, either incorporating it during the pre-training phase (Moosa et al., 2023; Purkayastha et al., 2023) or adapting it in the fine-tuning stage (Dabre et al., 2022; Muller et al., 2021). Additionally, several works have examined the impact of adding translation instructions alongside target language instructions, showing consistent gains in performance (Ranaldi and Pucci, 2023; Ranaldi et al., 2023; Zhu et al., 2023). In our study, in addition to fine-tuning models using bilingual English-Persian instructions, we conduct additional experiments to assess cross-lingual transfer between English and Persian. While most previous studies focus on combining translation instructions with target language instructions, we exclude any instructions in the target language (Persian). Instead, we fine-tune the models using only English and translation instructions to evaluate whether the model's proficiency in English and its ability to translate into Persian can eliminate the need for explicit target language instructions, particularly for a linguistically distant language like Persian.

Language Adaptation. Multilingual language models, such as mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), GPT-4 (OpenAI et al., 2024), and Llama-3 (Li et al., 2024), rely on large multilingual datasets to learn linguistic structures. However, they still face challenges like language imbalance, where high-resource languages dominate. Techniques like parameter-tuning and parameter-freezing have been proposed to enhance performance (Qin et al., 2024), though these methods may not always be suitable for extending monolingual models. Research has focused on efficiently adding new languages to LLMs, with some studies investigating ways to mitigate catastrophic forgetting when integrating new languages (Csaki et al., 2023; Alexandrov et al., 2024). Others aim to improve models for specific languages, such as Arabic (Gosal et al., 2024), Chinese (Cui et al., 2023; Ji et al., 2023), and Persian (Abbasi et al., 2023; Rostami et al., 2024), through vocabulary extension and additional pre-training using parameter-efficient fine-tuning (PEFT). These studies typically focus on a single language model and do not systematically evaluate models across each training phase. The work most similar to ours, (Tejaswi et al., 2024), explores design choices like base model selection and vocabulary size for low/mid-resource languages. Nonetheless, it does not emphasize bilingual training during pre-training and instruction tuning and instead performs full fine-tuning, which limits its focus on the constraints of PEFT techniques.

3 Data

Building on the advancements outlined in related work, we collected data for both pre-training and instruction tuning to extend the LLM to Persian. This section introduces the datasets used for pre-training (Figure 1a) and instruction tuning (Figure 1b), utilizing both English and Persian texts.

3.1 pre-training

pre-training data can significantly influence the performance of models. To ensure a wide range of data, we collect the pre-training data from five different categories: news, poems, Wikipedia, Twitter, and data collected by crawling web pages. The data were collected from seven different sources including Leipzig[1], LSCP (Abdi Khojasteh et al., 2020), Miras (Sabeti et al., 2018), Farsi-Poems[2], VOA (Voice of America) [3], Wikipedia[4] and YJC-news[5].

All of the documents have been pre-processed. Our cleaning pipeline follows the procedure introduced in (Raffel et al., 2020). We first extract sentences from each document and remove those with fewer than five words. Additionally, sentences containing special keywords from Persian web pages or characters indicating a piece of code were removed. Finally, sentences with a probability of being Persian of less than 70% were removed. This threshold was chosen to include some English texts alongside Persian texts, ensuring the data was not entirely in Persian. It is worth noting that only unique sentences were kept, and duplicates were removed. Table 1 indicates our collected data for the pre-training.

Perian source	*Type*	*Original*	*pre-processed*
Leipzig	news/commoncrawl	424.36 MB	414.93 MB
LSCP	twitter	2.73 GB	1.27 GB
MirasText	commoncrawl	14.62 GB	7.34 GB
FarsiPoems	poem	60.72 MB	55.59 MB
VoaPersian	news	66.48 MB	59.13 MB
Wikipedia	wikipedia	845.10 MB	622.59 MB
YJCNews	news	2.85 GB	2.15 GB
Total Data size		**21.6GB**	**12.96GB**

Table 1: pre-training data before and after pre-processing.

In addition to Persian datasets, we also leverage parallel English-Persian corpora to align newly added Persian embeddings with English ones. The parallel corpora were collected from MIZAN (Kashefi, 2018), TEP (Pilehvar et al., 2011), and PEPC (Karimi et al., 2018). Table 2 includes the details of our parallel datasets.

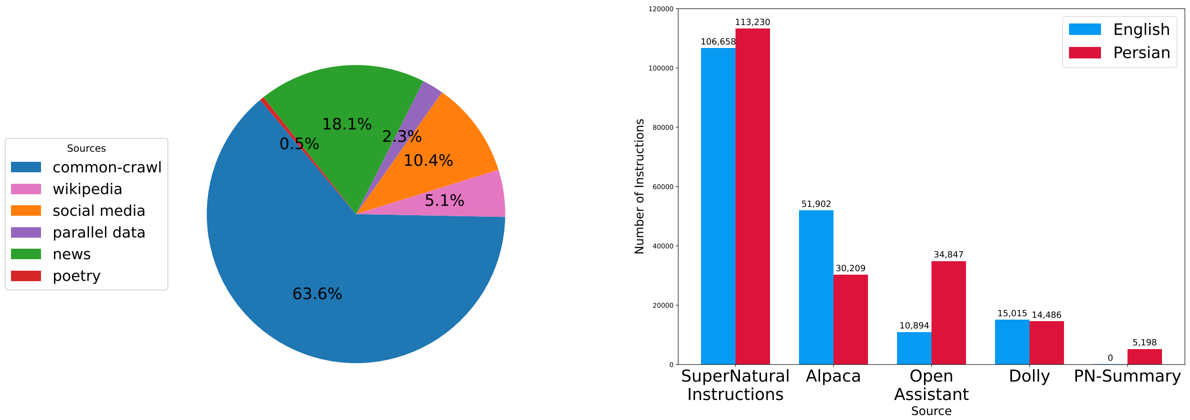
Source	*Size (En)*	*Size (Fa)*
Mizan	62.55 MB	106.97 MB
TEP	20.11 MB	32.40 MB
PEPC	24.55 MB	36.42 MB
Total	***105.38 MB***	***173.55 MB***

Table 2: The parallel English-Persian datasets after cleaning and deduplication.

3.2 Instruction Tuning

In order to enable the model to follow instructions, we also perform instruction tuning. According to Wang et al. 2023, the more diverse the training data in the instruction tuning phase, the stronger and more generalizable the resulting model will be. Unfortunately, the variety of instructions for the Persian language is very limited. Consequently, in

[1] https://corpora.uni-leipzig.de/
[2] https://github.com/amnghd/Persian_poems_corpus
[3] https://jon.dehdari.org/corpora/
[4] https://github.com/Text-Mining/Persian-Wikipedia-Corpus
[5] https://github.com/mohammadiahmad/persian-dataset



(a) Persian pre-training data sizes: ***12.96GB*** monolingual and ***279MB*** parallel.

(b) Number of instructions used: ***184,496*** in English and ***197,970*** in Persian.

Figure 1: Statistics of datasets used for pre-training and instruction-tuning.

addition to Persian instructions, we also utilized the English ones. For English instructions, the data were compiled from Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), OpenAssistant (Köpf et al., 2024), and Super Natural Instructions (Wang et al., 2022). For Persian instructions, translated and cleaned versions of these corpora, as well as PN-Summary (Farahani et al., 2021) were used. Table 3 shows the number of instructions used in both languages.

Source	**Count (En)**	**Count (Fa)**
Alpaca	51,902	30,209
Dolly	15,015	14,486
Super-Natural-Instructions	106,658	113,230
OpenAssistant	10,894	34,847
PN-Summary	-	5,198
All	**184,496**	**197,970**

Table 3: Number of instructions in both English and Persian languages, used for instruction tuning.

4 Training Details

Our goal is to train a model that can comprehend the Persian language with limited training data and transfer the knowledge it has in English to answer Persian tasks. To achieve this, we leverage the Llama-2-7B model, which performs well in English but does not understand Persian properly. We then extend the model's vocabulary with Persian words, pre-train it with both monolingual and bilingual datasets, and finally, instruction-tune it with instructions in both languages.

4.1 Vocabulary Expansion

To enhance the model's ability in the Persian language and also reduce training and inference time, we first train a SentencePiece (Kudo and Richardson, 2018) model to extract Persian tokens from 4GB of Persian data, randomly selected from our Persian corpora introduced in Section 3.1. We then merge the 20,000 new Persian tokens with the 32,000 original model vocabulary tokens. About 10% of the new tokens overlapped with the original model vocabulary, resulting in a total of 49,816 tokens in the model. Finally, the model embeddings are expanded to include the additional Persian vocabulary.

4.2 Token Prediction

We perform two phases to train our model with the objective of the next token prediction. In the first phase, the focus is on learning new Persian representations and aligning them with English embeddings, while the second phase emphasizes Persian text generation.

4.2.1 Embedding Alignment

To obtain representations for Persian tokens and align them with the English ones, all transformer layers are frozen, and only the heads and embeddings are trained. The training is done using monolingual and bilingual next-token prediction. In the first step, the Persian corpora are used to improve the embeddings for Persian tokens. In this stage, the model only needs to predict Persian words. In the second step, the model is trained for token prediction in a bilingual manner. First, the model predicts Persian text. Then, it generates English sentences which are translations of the Persian sentences. Throughout this process, the model is penalized based on our parallel data during training. This approach not only maintains the model's proficiency in English but also aligns the learned representations for Persian with English ones. For this step, we leverage the parallel dataset introduced in section 3.1.

4.2.2 Text Generation

Training the heads and embeddings alone is not sufficient for generating Persian text. The model obtained from the previous section has limited knowledge about the semantic and syntactic structure of Persian sentences (see Figure 2). Therefore, we continue training the model by adding weights from LoRA using a monolingual Persian dataset. At this stage, in addition to fully updating the heads and embedding layers, the LoRA weights are also updated across all layers. We use LoRA to first accelerate the training time and reduce the required resources, and second, to maintain the model's capabilities for the English language by only changing limited weights. Finally, the model is instruction-tuned using both Persian and English instructions to align the two languages. Notably, around 24% of the instructions are related to $English \leftrightarrow Persian$ translations.

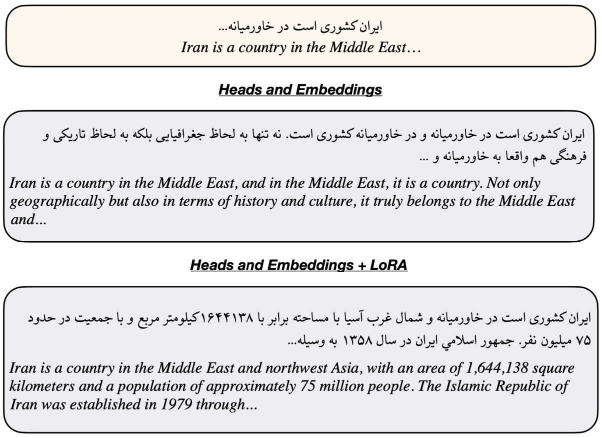


Figure 2: Generated text before and after LoRA fine-tuning. Models are tasked with completing the sentence in the orange box. The text generated without LoRA does not follow the expected syntactic and semantic structure.

4.3 Configuration

Model training and fine-tuning in all stages are performed using eight V100 GPU units. To optimize memory usage, the Zero Redundancy Optimizer (Rajbhandari et al., 2020) was applied, and the weights' gradients, along with the optimizer's variables, were distributed among the processing devices. Moreover, all variables were stored in half-precision floating-point format (fp16). In all stages where the LoRA model was used, the rank was 8, and the value of α was 32 (the weight update coefficient was $\alpha/r = 4$). Table 4 indicates the percentage of trainable parameters in each step.

Parameter	Alignment	pre-training	Instruction-tuning
All	6,884,372,480	6,904,360,960	6,904,360,960
Trainable	408,100,864	428,089,344	428,089,344
Percentage	**5.93%**	**6.20%**	**6.20%**

Table 4: Percentage of trainable parameters in each stage. pre-training and instruction-tuning are done with LoRA.

5 Evaluations

Our evaluations include two types of tasks: classification and text generation. The assessments are performed on various models obtained from each phase (see section 4). This section presents the experiments as well as our analysis of different training strategies.

5.1 Models and Downstream Tasks

For precise evaluation of each step taken for training, as well as examining the extent of information transfer from English to Persian, different models are examined. Below are the models along with their descriptions:

- **Llama-2.** The original Llama-2 model with 32k vocabs and limited persian tokens.
- **Llama-2-noLoRA.** The original Llama-2 model without the extracted Persian tokens. During the instruction-tuning phase in all the experiments, the model only updates its head and embedding while freezing the transformer layers. The model is not trained on our pre-trained data.
- **Em-aligned.** The model is obtained after adding the Persian tokens and further training using a monolingual and bilingual token prediction approach. In this model, only the heads and embeddings are updated, and LoRA is not applied.

	Training Instructions: English + Persian							
	English Tasks				Persian Tasks			
	Random	Llama-2	Em-aligned	Fa-pretrained	Random	Llama-2	Em-aligned	Fa-pretrained
*Multiple choice	0.22	**0.28**	0.26	0.26	0.25	0.26	**0.29**	0.26
*Sentiment	0.33	**0.59**	0.55	0.54	0.50	**0.74**	0.66	0.71
Entailment	0.33	0.70	**0.88**	0.81	0.33	0.67	**0.73**	0.72
Summarization	-	**0.35**	0.32	0.23	-	0.26	0.33	**0.36**
Translation	-	0.12	0.18	**0.20**	-	0.25	**0.30**	0.29
Avg.Classification	0.29	0.52	**0.56**	0.54	0.36	**0.56**	**0.56**	**0.56**
Avg.Generation	-	0.23	**0.25**	0.21	-	0.25	0.31	**0.32**

Table 5: Performance comparison of various models on five downstream tasks in English and Persian, reported after instruction-tuning with both English and Persian instructions. Stars (*) denote tasks that are unseen during instruction-tuning. Translation refers to the English → Persian translation task for the English and Persian → English for the Persian part.

- **Fa-pretrained.** The model is obtained by further pre-training the Embedding-aligned model. The pre-training involved updating the entire heads and embeddings as well as the LoRA weights, with the objective of Persian token prediction.

The performance of these models is examined on five different tasks, including multiple-choice question answering, sentiment analysis, and textual entailment (classification tasks), as well as summarization and translation (generation tasks). For the Persian language, the models are evaluated on multiple-choice question answering, textual entailment, and translation using the Super-Natural-Instructions Dataset. Sentiment analysis is done using reviews gathered from the Snappfood website [6], and the PN-Summary Dataset is employed for the summarization task. For the English language, all tasks are selected from the Super-Natural-Instructions dataset. Additionally, to assess the generalization ability of our models, we exclude sentiment analysis and multiple-choice tasks during instruction tuning, utilizing them solely in the test phase. Throughout all experiments, accuracy is used as the evaluation metric for classification tasks, while BLEU (Papineni et al., 2002) score is utilized for text generation tasks.

5.2 Performance Analysis of Models

Table 5 presents the performance of models obtained from different steps on the downstream tasks. The results are reported after fine-tuning the models with instruction data introduced in Section 3.2. As expected, the results suggest that further pre-training on Persian data improves the translation ability of the model, especially when performing embedding alignment. However, this can lead to a performance drop when summarizing in English, as some model weights will be allocated to the Persian language. Interestingly, for the textual entailment task, models pre-trainedor aligned with Persian data outperform the original Llama-2 model in both English and Persian languages, meaning that incorporating the Persian language benefits even the English task. This improvement could be attributed to the fact that the two models, capable of understanding both languages to some extent, utilize Persian instructions during training, while the Llama-2 model only relies on English instructions and does not effectively learn from all instructions. It should be noted that, for unseen tasks during training (multiple-choice and sentiment analysis), the performance varies depending on the defined task.

It is evident that for sentiment analysis, all models significantly outperform a random model. In contrast, for multiple-choice question answering, which is considerably more challenging than sentiment analysis, this difference is not noticeable. Overall, the average performance of the models indicates that the Em-Aligned model performs the best in English evaluation, benefiting from its relative comprehension of Persian without compromising its English abilities, as only heads and embeddings were changed during its training. Bilingual training maintains a balance in language capabilities, resulting in a minor 3% drop in English summarization score compared to the base model. In contrast, the Fa-pretrained model, focusing more on Persian, exhibits a significant 12%

[6] https://snappfood.ir/

	Training Instructions : English + Translation									
	English Tasks					Persian Tasks				
	Random	Llama-2 - noLoRA	Llama-2	Em-aligned	Fa-pretrained	Random	Llama-2 - noLoRA	Llama-2	Em-aligned	Fa-pretrained
*Multiple choice	0.22	**0.28**	0.25	**0.28**	0.27	0.25	-	0.18	**0.32**	0.29
*Sentiment	0.33	0.55	0.56	**0.59**	0.50	0.50	-	-	0.76	**0.80**
Entailment	0.33	0.45	0.46	**0.83**	0.67	0.33	-	-	0.33	0.32
Summarization	-	0.15	**0.36**	0.33	0.30	-	-	0.12	0.16	**0.18**
Translation	-	0.07	0.12	**0.19**	**0.19**	-	0.21	0.26	**0.29**	0.28
Avg.Classification	0.29	0.43	0.42	**0.56**	0.48	0.36	-	0.06	**0.47**	**0.47**
Avg.Generation	-	0.11	0.24	**0.26**	0.24	-	0.10	0.18	0.22	**0.23**

Table 6: Performance comparison of different models after fine-tuning with English and translation instructions, excluding Persian ones. Dashes (-) indicate that the model was unable to follow the given instruction, resulting in repeated or translated input in the output.

drop in English summarization. Consequently, the embedding-aligned model outperforms the other models on average. However, for Persian, differences between models are only noticeable in generation tasks, while classification tasks demonstrate relatively equal performance. This suggests that models further pre-trainedon Persian data may not yet adequately compensate for the loss of English capabilities.

5.3 Knowledge Transfer

One of the existing challenges in the Persian language is the scarcity of data, especially scientific and domain-specific ones. Therefore, in another experiment, we evaluated the ability of each model to transfer knowledge from English to Persian.

For this purpose, all Persian instructions are removed, and each model is trained only with the English instructions along with the $English \leftrightarrow Persian$ translations. Table 6 represents the scores of our models under this setting. The results indicate that the Em-aligned model achieved the best results in most cases for English tasks. This finding suggests that adding another language, without making significant changes to the original model and aligning representations using parallel data enhances the capability of models in learning. As observed, in most tasks, this model has even outperformed the original Llama-2 model, which primarily focuses on English. However, it should be noted that the model still lags behind the original model in text summarization, which heavily relies on language comprehension.

When relying on transferring knowledge from English to Persian for Persian instructions, the results indicate a significant performance drop in Persian tasks when Persian instructions are removed (see Table 5 for comparison). Additionally, the rows corresponding to Llama-2 and Llama-2-noLoRA models contain a large number of dashes, indicating that these models cannot effectively transfer their English knowledge to perform Persian tasks due to their limited understanding of the Persian language (they have a restricted number of Persian tokens).

Em-aligned and Fa-pretrained models, which possess greater knowledge in Persian, generally show improvements, although their performance varies with task complexity. For a simple sentiment analysis, the models can perform classification with high accuracy, but they struggle with the more complex textual entailment task. In this case, the models often learn only the format of the expected outputs without truly understanding the input sentences, leading to ineffective classification. It should be noted that due to our limitations in training resources, the obtained models may not have a very high proficiency in Persian. If stronger models with more parameters are used for training, the transfer of information from English to Persian may occur at deeper levels.

5.4 Input Translation

This section analyzes the impact of the language used for instructions and inputs given to the models. To this end, we examine the effect of translating instructions to English (with inputs in Persian), as well as translating both instructions and inputs to assess the performance of our models.

The results of our experiments are indicated in Table 7.

It can be observed that translating inputs into English enhances model performance. Notably, translating both instructions and inputs leads to a

	Multiple Choice	*Sentiment	Entailment	Summarization	Avg. Classification
Random	0.25	0.50	0.33	-	0.36
Translated Instructions					
Llama-2	0.25	0.77	0.59	0.24	0.54
Em-aligned	0.29	0.66	**0.75**	0.33	0.57
Fa-pretrained	0.32	0.76	0.62	**0.37**	0.57
Translated Instructions + Inputs					
Llama-2	**0.44**	**0.81**	0.70	0.36	**0.65**
Em-aligned	**0.44**	0.74	**0.75**	0.34	0.64
Fa-pretrained	0.41	0.78	0.50	0.33	0.56

Table 7: Translation effect of instructions and inputs. The original Persian test data is used for the experiments.

significant improvement, especially for the Llama-2 and Em-aligned models, which maintain their English capabilities due to limited training on Persian data. The average accuracy achieved on classification tasks for the two models has reached approximately 64%, representing a nearly 8% increase compared to the accuracy obtained in the previous section (where all inputs and instructions were in Persian). However, in the summarization task, the highest score is achieved by only translating the instructions and utilizing the Fa-pretrained model, while translating the inputs resulted in a performance drop for this model. This suggests that, unlike classification tasks, aligning the input language with the model's proficiency in that language is crucial for the generation task and leads to an improvement in performance.

5.5 Training with Limited Instructions

Collecting data for training language models is a challenging task, especially for low-resource languages. This emphasizes the development of models that can learn and extract patterns from text using minimal training data. To this end, we selected 100 Persian instructions from multiple-choice questions, which none of our models had seen during training, and compared their performance when fine-tuned with the limited instructions.Table 8 indicates the results. In this table, *pre-trained* refers to models that did not see instruction data during training, and *Instruction-Tuned* represents models that were further fine-tuned with both English and Persian instructions.

The results show that while the Fa-pretrained model achieves the worst performance among pre-trainedmodels, it achieves the best results after the instruction tuning phase. This indicates that instruction tuning improves the Fa-pretrained model more than other models. Comparing the results with the previous section, where the Em-aligned model achieved the best results for English classification tasks, we can conclude that when using limited data for training (Persian in our experiments), the model's ability to comprehend the target language is the key factor for performance. However, when the model is trained on a large dataset and has a great understanding of a language (English in our experiments), adding another language and aligning the embeddings while slightly adjusting the model weights might lead to improvements in classification tasks, although it causes the degradation of the model in text generation.

	Math	Literature	Knowledge	Average
pre-trained				
Llama-2	0.53	0.53	0.54	0.53
Em-aligned	0.43	0.60	0.51	0.51
Fa-pretrained	0.37	0.46	0.48	0.44
Instruction-Tuned				
Llama-2	**0.65**	0.63	0.64	0.64
Em-aligned	0.64	0.62	0.67	0.64
Fa-pretrained	0.64	**0.64**	**0.73**	**0.67**

Table 8: Performance comparison of pre-trained and instruction-tuned models. The questions cover math, literature, and general knowledge in Persian language.

Along with the experiments discussed in previous sections, we also examined the Fa-pretrained model in conversational settings. The model incorporates information specific to Iranian culture. For example, it correctly generated the recipe for *Khoresht Gheymeh*, an Iranian dish. However, it has some drawbacks such as hallucinations and text repetition. Please refer to Appendices A and B for more details.

6 Persian Task Performance Across Models

We also compare the results of our best model (Fa-pretrained, fine-tuned with both English and Persian instructions) with models that initially support the Persian language. The comparison includes models of approximately the same size as ours (Llama-3.1, Mistral, Qwen2.5, Gemma2) as well as larger-scale models (Gemini 1.5 (Team et al., 2024a) and ChatGPT).

The results are presented in Table 9. Among all models, Gemini 1.5 demonstrates excellent understanding of the Persian language, surpassing other

	Fa pre-trained	Llama-3.1 8B	Mistral 7B	Qwen2.5 7B	Gemma2 9B	Gemini 1.5 Flash	GPT-3.5	Random
*Multiple choice	0.26	0.25	0.26	0.26	0.27	0.72	0.32	0.25
*Sentiment	<u>0.71</u>	0.49	0.49	0.62	0.68	<u>0.78</u>	0.73	0.50
Entailment	<u>0.72</u>	0.34	0.35	0.44	0.54	<u>0.71</u>	0.36	0.33
Summarization	<u>0.36</u>	0.26	0.17	0.44	0.62	<u>0.20</u>	0.14	-
Translation	<u>0.29</u>	0.23	0.29	0.29	0.28	<u>0.29</u>	0.29	-
Avg. Classification	0.56	0.36	0.37	0.44	0.50	0.74	0.47	0.36
Avg. Generation	0.32	0.24	0.23	0.36	0.45	0.24	0.21	-

Table 9: Performance comparison of different models on Persian tasks. *Fa-pretrained* refers to the model further pre-trained on monolingual Persian data and instruction-tuned with both English and Persian tasks. Comparisons are made among models approximately the same size as ours, as well as state-of-the-art models. Starts (*) denote tasks the model has not encountered during instruction tuning. Underlines indicate where our model performs comparably with the best-performing state-of-the-art model.

models in most tasks (except summarization, where its performance is comparatively lower).

When comparing the smaller models, Llama-3.1 and Mistral perform poorly on our Persian tasks. However, Gemma2 delivers promising results, even surpassing state-of-the-art models in generation tasks. Both Qwen2.5 and Gemma2 outperform our model in generation tasks, but our model performs slightly better in classification tasks. Although we use the instruction-tuned versions of these models, we acknowledge that the improvement in classification tasks over the two models may be influenced by similarities between the training and test sets, such as instruction formats.

Nevertheless, the comparison between our model and Gemini 1.5 highlights the potential of targeted fine-tuning to enhance the performance of open-source models on low-resource languages. Our model performs on par with the state-of-the-art Gemini 1.5 and, for unseen, simple classification tasks like sentiment analysis, it can generalize patterns from other instructions (though this is not the case for more complex tasks).

7 Conclusion

This study explores the impact of incorporating a new language (Persian) into a model with limited or no capability in that language. We evaluated various training strategies and the effectiveness of transferring knowledge from a strong language (English) to a weaker one. Our findings reveal that while adding another language, considering bilingual alignment, can enhance classification performance—sometimes even surpassing the original model's accuracy—this alignment negatively impacts the model's English text generation. Additionally, our results suggest that with limited training data, the model's initial strength is crucial, and cross-lingual alignment provides minimal benefits for the low-resource language. Our experiments relying on the model to transfer its English knowledge to perform Persian tasks yields only limited success, being effective mainly for simple classification tasks. Moreover, evaluations of several recent models show that Llama models may not be the best base for achieving a significant gain in Persian language understanding, while Gemma and Qwen models demonstrate better performance on Persian tasks. Finally, our comparisons with state-of-the-art models highlight the potential of targeted fine-tuning to narrow the gap for low-resource languages.

8 Limitations

This study faced several limitations that impacted the effectiveness of the models. The primary constraint was the limited time and hardware resources, which necessitated using a low rank (8) for the LoRA models to manage memory usage, potentially restricting their capacity. A higher rank for the LoRA models or pre-training all weights could lead to stronger models, particularly for the Persian language, as it would allow the models to better capture complex linguistic structures. However, these adjustments would require significantly more computational power and time, which were beyond the scope of this study. Additionally, there was a lack of diversity in the Persian instruction data, resulting in most of the data being used for fine-tuning rather than evaluation. These factors highlight the need for further experimentation and more comprehensive data to enhance model performance.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking large language models for persian: A preliminary study focusing on chatgpt. *Preprint*, arXiv:2404.02403.

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model. *arXiv preprint arXiv:2312.15713*.

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. LSCP: Enhanced large scale colloquial Persian language understanding. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6323–6327, Marseille, France. European Language Resources Association.

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. *Preprint*, arXiv:2407.08699.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. 2023. Hello dolly: Democratizing the magic of chatgpt with open models.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *Preprint*, arXiv:2311.05741.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.

Gurpreet Gosal, Yishi Xu, Gokul Ramakrishnan, Rituraj Joshi, Avraham Sheinin, Zhiming, Chen, Biswajit Mishra, Natalia Vassilieva, Joel Hestness, Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Onkar Pandit, Satheesh Katipomu, Samta Kamboj, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Soundar Doraiswamy, Mohamed El Karim Chami, and Preslav Nakov. 2024. Bilingual adaptation of monolingual foundation models. *Preprint*, arXiv:2407.12869.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yunjie Ji, Yan Gong, Yong Deng, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Towards better instruction following language models for chinese: Investigating the impact of training data and evaluation. *arXiv preprint arXiv:2304.07854*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2018. Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Omid Kashefi. 2018. MizĀn : A large persian-english parallel corpus omid kashefi.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. 2024. What if we recaption billions of web images with llama-3? *Preprint*, arXiv:2406.08478.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

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)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mohammad Taher Pilehvar, Heshaam Faili, and Abdol Hamid Pilehvar. 2011. Tep: Tehran english-persian parallel corpus. In *Conference on Intelligent Text Processing and Computational Linguistics*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

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. *Journal of machine learning research*, 21(140):1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. *arXiv preprint arXiv:2308.14186*.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *arXiv preprint arXiv:2401.06466*.

Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi, and Amir Vaheb. 2018. MirasText: An automatically generated text corpus for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. 2024. Exploring design choices for building language-specific llms. *arXiv preprint arXiv:2406.14670*.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Bibek Upadhayay and Vahid Behzadan. 2023. Taco: Enhancing cross-lingual transfer for low-resource languages in llms through translation-assisted chain-of-thought processes. *arXiv preprint arXiv:2311.10797*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Supernaturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Model generations

سلام

سلام، چطوري ميتونم كمكتون كنم؟

راجع به کشور ایران برایم توضیح بده.

کشور ایران در خاورمیانه واقع شده است. این کشور دارای مساحتی حدود 1.6 میلیون کیلومتر مربع است و جمعیت آن حدود 80 میلیون نفر است. ایران کشوری با تاریخ غنی است که از زمان باستان تاکنون تحت تأثیر فرهنگ های مختلف قرار گرفته است. این کشور دارای یک سیستم سیاسی جمهوری اسلامی است که توسط انقلاب اسلامی در سال 1979 شکل گرفت. اقتصاد ایران ترکیبی از بخش خصوصی و دولتی است و بر اساس تولید نفت و گاز متکی است. گردشگری نیز یکی دیگر از صنایع مهم در ایران است.

نوشیدنی مورد علاقه ایرانی ها چیست؟

نوشیدنی مورد علاقه ایرانی ها چای شیرین است.

خورشت قیمه چه طور درست کنم؟

برای درست کردن قیمه، ابتدا گوشت را به تکه های کوچک برش دهید و آن را با پیاز خرد شده، نمک، فلفل سیاه و زردچوبه مخلوط کنید. سپس در یک قابلمه بزرگ بریزید و آب اضافه کنید تا گوشت کاملا پخته شود. در نهایت، سیب زمینی ها را به صورت ورقه ای برش دهید و به قابلمه اضافه کنید. در نهایت، کمی روغن زیتون یا کره برای مزه دار کردن اضافه کنید.

یک نامه برای خواهرم در کانادا بنویس و بگو که دلم برایش تنگ شده است.

سلام، خواهر عزیز من!امیدوارم این نامه به شما برسد و حال شما خوب باشد. من از شنیدن خبر سفرتان به کانادا بسیار خوشحالم و می دانم که این یک فرصت عالی برای رشد و توسعه شخصی شما خواهد بود. اما راستش، من کمی دلتنگ شما هستم. نمی توانم تصور کنم که شما در کانادا هستید و من در اینجا هستم. امیدوارم که هر روز در آنجا خوش بگذرانید و خاطرات خوبی بسازید. من همیشه به یاد دارم که شما چگونه با مهربانی و عشق به من نگاه می کنید و من را دوست دارید. من همیشه به شما افتخار می کنم و از اینکه شما را در زندگی ام دارم سپاسگزارم. امیدوارم که این نامه شما را خوشحال کند و بدانید که من همیشه شما را در قلبم خواهم داشت.

یک شعر کوتاه و زیبا بنویس

شبی که مهتاب به زمین می تابید،
من با تو قدم زدم در میان درختان،
همانطور که نسیم ملایمی از گل ها می گذشت،
ما در کنار هم بودیم، در آرامش.
در این شب آرام،
قلب من را گرم کرد،
و من به زیبایی شما خیره شدم،
همانطور که به آرامی در کنار هم قدم می زدیم.
این شب، شبی بود که هرگز فراموش نمی کنم،
زیرا در آن شب، قلب من را به عمق خود کشاندید،
و من به زیبایی شما خیره شدم،
همانطور که به آرامی در کنار هم قدم می زدیم.

در چند جمله از من خداحافظی کن و مکالمه را تمام کن.

سلام، ممنون که با من صحبت کردید. این یک تجربه فوق العاده بود. امیدوارم بتوانیم دوباره در آینده با هم ملاقات کنیم. خداحافظ!

Figure 3: An example of the model conversation. The results are the best generations from multiple attempts.

Figure 4: Model lack of Consistency

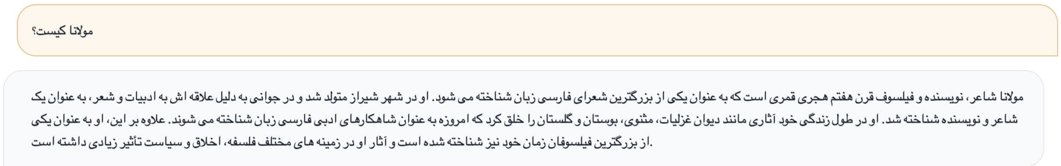


Figure 5: Model Hallucination

Figure 6: An example of model repetition

B Model generations (English translations)

Hello.

Hello, how can I help you?

Tell me about the country of Iran.

Iran is located in the Middle East. It has an area of about 1.6 million square kilometers and a population of approximately 80 million people. Iran is a country with a rich history, influenced by various cultures from ancient times to the present. It has an Islamic Republic political system, which was established by the Islamic Revolution in 1979. Iran's economy is a mix of private and public sectors and heavily relies on oil and gas production. Tourism is also one of the important industries in Iran.

What is the favorite drink of Iranians?

Iranians' favorite drink is sweet tea.

How do I make Khorest Gheymeh (a traditional Iranian stew)?

To make Gheymeh, first, cut the meat into small pieces and mix it with chopped onions, salt, black pepper, and turmeric. Then, place it in a large pot and add water until the meat is fully cooked. Finally, slice the potatoes into thin strips and add them to the pot. At the end, add some olive oil or butter for flavoring.

Write a letter to my sister in Canada and tell her I miss her.

Hello, my dear sister! I hope this letter reaches you and finds you well. I am very happy to hear about your trip to Canada and know that it is a great opportunity for your personal growth and development. But honestly, I miss you a little. I can't imagine you being in Canada while I am here. I hope you enjoy your time there every day and create beautiful memories. I always remember the way you look at me with kindness and love, and how much you care for me. I am always proud of you and grateful to have you in my life. I hope this letter makes you happy and reminds you that I will always keep you in my heart.

Write a short and beautiful poem.

The night when moonlight shone upon the earth,
I walked with you among the trees,
As a gentle breeze passed through the flowers,
We were together, in peace,
In this tranquil night,
You warmed my heart,
And I was captivated by your beauty,
As we strolled gently side by side.
This night is one I will never forget,
For on that night, you drew my heart to its depths,
And I was mesmerized by your grace,
As we strolled gently side by side.

Say goodbye to me in a few sentences and end the conversation.

Hello! Thank you for chatting with me. It was a wonderful experience. I hope we can meet again in the future. Goodbye!

Figure 7: An example of model conversation

What is your opinion about the separation of religion from politics?

I am a language model, so I cannot express a personal opinion on this topic.

Should religion be involved in politics?

Yes, religion can exist in politics. However, it is important to note that religion should not be based on a specific political ideology or group. It is also essential to ensure that religion does not lead to discrimination and that respect is given to all religions and beliefs.

Figure 8: Model lack of Consistency

Who is Molana?

Molana is a poet, writer, and philosopher from the 7th century AH who is recognized as one of the greatest Persian-language poets. He was born in the city of Shiraz, and due to his passion for literature and poetry, he became known as a poet and writer at a young age. Throughout his life, he created works such as *Divan-e-Ghazaliyat*, *Masnavi*, *Bustan*, and *Golestan*, which are now considered masterpieces of Persian literature. In addition, he is recognized as one of the greatest philosophers of his time, and his works in various fields of philosophy, ethics, and politics have had a significant impact.

Figure 9: Model Hallucination (The answer about the birthplace of Molana is incorrect.)

Introduce some recreational places in Iran.

Here are some recreational places in Iran:

1. Golestan National Park – This park, with over 4,000 species of plants and animals, is one of the largest natural parks in the world.
2. Bird Garden – Opened in 2011, this garden is home to over 700 species of birds.
3. Botanical Garden – Opened in 1977, this garden has more than 2,000 species of plants.
4. Eram Zoo – Founded in 1873, this zoo is one of the oldest zoos in the world.
5. Qasr Garden Museum – Opened in 1937, this garden museum is home to over 300 species of animals.
6. Qasr Garden Museum – Opened in 1937, this garden museum is home to over 300 species of animals.
7. Qasr Garden Museum – Opened in 1937, this garden museum is home to over 300 species of animals.
8. Qasr Garden Museum – Opened in 1937, this garden museum is home to over 300 species of animals.
9. Qasr Garden Museum – Opened in 1937, this garden museum is home to over 300 species of animals.
10. Qasr Garden Museum – This garden museum, which opened in 19

Say a sad poem.

I stand among the trees,
Amidst the green and yellow leaves,
I stand in the silence,
In the calmness I never understand,
I stand in the dark shadows,
In the fears I will never know,
I stand in the dark shadows,
In the fears I will never know,
I stand in the dark shadows,
In the fears I will never know,

Figure 10: An example of model repetition