PersianMind: A Cross-Lingual Persian-English Large Language Model

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti

{pedram.rostami, alisalemi, mjdousti}@ut.ac.ir
University of Tehran

Abstract

Large language models demonstrate remarkable proficiency in various linguistic tasks and have extensive knowledge across various domains. Although they perform best in English, their ability in other languages is notable too. In contrast, open-source models, such as LLaMa, are primarily trained on English datasets, resulting in poor performance in non-English languages. In this paper, we introduce PersianMind, an open-source bilingual large language model which demonstrates comparable performance to closed-source GPT-3.5-turbo in the Persian language. By expanding LLaMa2's vocabulary with 10,000 Persian tokens and training it on a dataset comprising nearly 2 billion Persian tokens, we show that our approach preserves the model's English knowledge and employs transfer learning to excel at transferring task knowledge from one language to another.

Keywords: Large Language Model, LLaMa, Persian language

1 Introduction

Large language models (LLMs) have been the most significant development in the field of natural language processing in recent years, reintroducing the concept of employing a model as an *artificial general intelligence* (AGI) (Bubeck et al., 2023). These transformer-based decoder-only (Vaswani et al., 2017) models are distinguished by their considerable scale and training on extensive textual datasets. LLMs are versatile tools for various language tasks in natural language processing. Furthermore, there are instances that LLM-based chatbots can replace traditional information retrieval systems (Zhao et al., 2023). When combined with multi-modal capabilities, they can also be used in computer vision, as demonstrated by the emergence of visual chatbots (C. Wu et al., 2023).

Prominent commercial LLMs such as ChatGPT (OpenAI, 2022), GPT-4 (OpenAI et al., 2023), PaLM2 (Anil et al., 2023), and Claude (Anthropic, 2023) demonstrate excellent performance across a diverse range of tasks, including text generation, summarization, and code generation. Additionally, these models exhibit promising results when applied to non-English languages. However, it is important to note that these LLMs are proprietary and come with certain limitations for fine-tuning and access to their original models is restricted. On the other hand, although open-source LLMs like LLaMa2 have demonstrated impressive results in the English language, their performance significantly degrades when applying to other languages (Touvron, Martin, et al., 2023). This disparity can be attributed to the fact that their training dataset consist of mostly English texts, which limits their ability to understand and generate contents in other languages.

To address the poor performance of open-source LLMs in the Persian language, we introduce PersianMind[1], an open-source Persian-English LLM that achieves comparable results to GPT-3.5-turbo (OpenAI, 2022) in reading comprehension benchmark. In this paper, we employed a Persian Byte-Pair Encoding tokenizer comprising 10,000

[1]The model can be downloaded from `https://huggingface.co/universitytehran/PersianMind-v1.0`.

tokens, which has been trained on a cleaned Persian Wikipedia corpus. These tokens are added to LLaMa2's vocabulary, and the model's embeddings are subsequently expanded. We utilize the LoRA technique to train our model on a 2-billion-token Persian corpus and subsequently fine-tune the model using various instruction tuning datasets to enhance its performance on natural language processing tasks. Due to limited instruction tuning datasets for the Persian language, we refined our model by fine-tuning it on high-quality Persian machine-translated datasets.

Our key contributions in this paper are as follows:

- Introduced PersianMind, an open-source Persian-English large language model trained using a cost-aware approach which utilizes the LoRA technique and data parallelism.

- Achieved state-of-the-art results on Persian subset of the Belebele benchmark and the ParsiNLU multiple-choice QA task.

- Attained performance comparable to GPT-3.5-turbo in a Persian reading comprehension task.

- Alleviated catastrophic forgetting resulting from extensive training on Persian datasets by training on Persian-English parallel datasets and employing the LoRA technique.

- Demonstrated that PersianMind can generate high-quality sentence embeddings, surpassing the performance of previous masked language models. Additionally, we showed that sentence embeddings generated by PersianMind exhibit cross-linguality.

- Showed the efficacy of multilingual transfer learning on PersianMind, evidencing that fine-tuning the model with Persian data notably enhances its performance on the corresponding English task.

The rest of this paper is organized as follows: Section 2 reviews open-source LLMs and parameter efficient fine-tuning methods. Section 3 details our training approach, while Section 4 compares PersianMind's performance across various tasks with other competitors. In Section 5, we discuss the carbon footprint associated with training PersianMind . Finally, Section 6 concludes the paper.

2 Related Work

2.1 Open LLMs

While GPT-3.5-turbo demonstrated excellent proficiency in natural language generation, the LLaMa model (Touvron, Lavril, et al., 2023) was the first LLM to claim achievement of similar performance in various English tasks. LLaMa family of models are one of the most popular open foundation large language models, ranging in scale from 7B to 65B parameters. Smaller LLaMa models are trained on 1T tokens, while larger ones are trained on 1.4T tokens. In both cases, training data is predominantly consist of English text and code. Only 4.5% of their dataset is multilingual, including either Latin or Cyrillic scripts. The Mosaic Pretrained Transformers (MPT) model (MosaicML, 2023) has 7B parameters and is trained on 1T tokens of English text and code. This model increases the context length of inputs from 2k to 65k in its storyteller model. LLaMa2 (Touvron, Martin, et al., 2023) models are an updated version of LLaMa models, comprising a collection of LLMs with 7B, 13B, and 70B parameters. These models are pretrained on a larger and higher quality dataset compared to the first version of LLaMa models. Their pretraining dataset primarily consists of English text and code, with less than 2% of the text in other languages.

Falcon (Tii, 2023) models are a set of LLMs with 1.3B, 7.5B, 40B, and (closed-source) 180B parameters, trained on the RefinedWeb dataset (Penedo et al., 2023) — a curated, high-quality web-based dataset. Although the RefinedWeb dataset is multilingual and includes many languages, such as Persian, open-source Falcon models are trained on European languages, particularly English. Yi (Yi, 2023) models are a collection of Chinese-English LLMs with 6B and 34B parameters. These models are trained on a 3T token dataset of English and Chinese and outperform previous models on English and, especially, Chinese benchmarks.

While newer models like Falcon and Yi demonstrate better performance on English benchmarks, we decided to use Llama2 model due to training on more multilingual datasets. We specifically opted to fine-tune the LLaMa2-7B-chat variant because loading it with the `fp16` data type requires only 14GB of GPU memory, making it easily loadable on a consumer GPU with 24GB of memory. Loading our model on a single GPU allows us to avoid the overhead of model parallelism. Furthermore, in a multi-GPU setup, this approach leads to faster training by leveraging data parallelism.

2.2 Parameter Efficient Fine-Tuning

While fine-tuning LLMs with billions of parameters can be an expensive task, *parameter-efficient fine-tuning* (PEFT) techniques aim to reduce training costs by fine-tuning a small number of parameters. Adapter tuning (Houlsby et al., 2019) was one of the first PEFT techniques. In adapter tuning, small adapter layers are inserted after the multihead attention and feed forward layers of each transformer block, with only these layers undergoing training. Adamix (Y. Wang et al., 2022) suggested leveraging adapters in a mixture-of-expert fashion, while SparseAdapters (He et al., 2022) acknowledges the redundancy in many adapter parameters. By pruning these redundant parameters during initialization and subsequently fine-tuning, SparseAdapter achieves better results.

Unlike adapter-based approaches, which often involve fine-tuning large models with the addition of small trainable parameters, a set of methods proposes a different approach — fine-tuning only a small subset of the existing model. BitFit (Ben Zaken et al., 2022) proposes fine-tuning only the biases of the model. FishMask (Sung et al., 2021) recommends selecting parameters with the highest Fisher information value for training. While Freeze and Reconfigure (FAR) (Vucetic et al., 2022) suggests freezing less important columns and only training the crucial ones.

Low-rank adaptation (LoRA) (Hu et al., 2022) approaches represent another category of PEFT methods. LoRA introduces the concept of freezing the pretrained model and adding small, tunable weights into specific layers. These weights take the form of the rank decomposition matrices derived from the pretrained model weights. We refer interested readers to (Lialin et al., 2023) which provides a comprehensive survey of PEFT approches.

Given the small number of trainable parameters and its great performance across models of various sizes (ranging from 125M to 175B), we choose to fine-tune our model utilizing the LoRA technique. Moreover, the Bacterian-X (H. Li et al., 2023) models are a notable example of the efficacy of LoRA in fine-tuning LLMs for learning an additional language.

3 PersianMind Model

3.1 Bilingual Tokenizer and Expansion of Embeddings

The LLaMa2's tokenizer has 32,000 tokens, including only 55 Arabic and Persian tokens. It specifically covers Persian letters and does not include any additional Persian subwords. In this situation, fine-tuning the model with a Persian corpus can take a long time because every word is tokenized into letters. Therefore, we decide to train a new Persian tokenizer. Augmenting the LLaMa2's tokenizer with Persian subwords enable fine-tuning our model on more extensive corpora within the constraints of the same computational budget.

Hence, our approach involve training a byte-pair encoding (BPE) (Sennrich et al., 2016) tokenizer with 10,000 tokens on a 1GB Persian Wikipedia corpus. Furthermore, we enhance the LLaMa2 tokenizer by incorporating Persian subwords. The combined Persian-English tokenizer has 41,510 tokens. Although training a larger BPE tokenizer can capture more named entity tokens, it would increase the input and output embeddings' size, resulting in more trainable parameters, which requires more computational resources. Furthermore, we expand the input and output embeddings by the size of our tokenizer. The newly added embeddings are randomly initialized within the space of LLaMa2's embeddings.

3.2 Training Details

Our model training approach is focused on optimizing training on constrained computational budget. To achieve this, we leverage the 7B-chat variant of LLaMa2 models. We avoid model parallelism due to significant overhead in the initial release of PersianMind and will consider it for the future releases. Instead, we embrace data parallelism to reduce training time. Consequently, the LoRA technique is implemented by incorporating LoRA weights across all layers of the transformer architecture, utilizing a LoRA rank of 8 and a dropout rate of 0.05. Therefore, our training parameters encompass LoRA weights, input, and output embeddings.

Our training methodology comprises two key phases. First, we fine-tune the embeddings and LoRA weights of the model through training on a plain Persian text corpus. Subsequently, we perform supervised fine-tuning (SFT) of the model on instruction datasets. During both phases, our training objective is casual about 2B tokens for pretrain

	Fine-Tuning		Instruction-Tuning	
	Dataset	Num. Tokens	Dataset	Num. Intructions
Step 1	Persian Wikipedia	222M	–	–
Step 2	CC100	1.147B	Alpaca (Fa)	52,000
			TED2020	50,000
			ParsQuad	5,000
			PersianQA	900
			Total	107,900
Step 3	CC100	600M	TED2020	300,000
			MMLU auxiliary (Fa)	100,000
			CoT (Fa)	74,000
			PN Summary	20,000
			ParsQuad	5,000
			Internal dataset	1,241
			ParsiNLU Mutiple-Choice QA	1,200
			PersianQA	900
			ParsiNLU (Reading Comprehension)	600
			Total	502,941

Table 1. Training and instruction tuning datasets for each step. Instruction dataset names with (Fa) indicate that they are Persian machine-translated datasets.

language modeling. To ensure a thorough evaluation, we divide the training into three distinct steps.

Step 1: In the first step, we fine-tune our model for two epochs on the Persian Wikipedia corpus, which contains 111 million tokens. During this phase, we intentionally skip fine-tuning on instructions to explore the model's capabilities. Despite the less-than-ideal perplexity, the model still manage to produce acceptable results.

Step 2: In the second step, we fine-tune our model on a 10GB subset of CC100 (Conneau et al., 2020) dataset, which includes 1.147 billion tokens. In this phase, we fine-tun our model on PersianQA (Ayoubi and Davoodeh, 2021) and ParsQuad (Abadani et al., 2021) datasets to improve its performance in the question-answering task. Additionally, we perform fine-tuning on a segment of the English-Persian TED2020 (Reimers and Gurevych, 2020) parallel dataset for better translations, and fine-tune on machine-translated Alpaca (Taori et al., 2023) dataset, a comprehensive instructions dataset covering various tasks and questions.

Step 3: In the third step, we fine-tune our model on an additional 5GB subset of the CC100 dataset, encompassing 600 million tokens. This phase prioritize fine-tuning on a variety of instructions datasets. Notably, to enhance our model's translation capabilities, we fine-tune on the entire English-Persian TED2020 parallel dataset. PersianQA and ParsQuad are once again employed to augment model's capabilities in the question-answering task. To improve proficiency in multiple-choice QA and reading comprehension tasks, we fine-tune our model on ParsiNLU multiple-choice QA (Khashabi et al., 2021) , ParsiNLU reading comprehension (Khashabi et al., 2021), and the machine-translated auxiliary dataset of MMLU (Hendrycks et al., 2021). Additionally, 20,000 instructions from the PN summary dataset (Mehrdad Farahani, Gharachorloo, and Manthouri, 2021) are incorporated to strengthen our model's summarization capabilities. Natural language reasoning capabilities are refined by fine-tuning on the machine-translated ORCA's Chain of Thought dataset (Mukherjee et al., 2023). Furthermore, a few manually created instructions related to Persian food recipes and proverbs are introduced to diversify the model's understanding of natural language contexts.

The model trained in this step exhibits proficiency in generating Persian text, with a perplexity score superior to that of the second-step model. Additionally, there is a slight improvement in the perplexity score of the model when applied to the English text. Notably, fine-tuning on the ORCA's Chain of Thought dataset contributed to enhancing the model's ability to analyze and solve problems systematically and step by step.

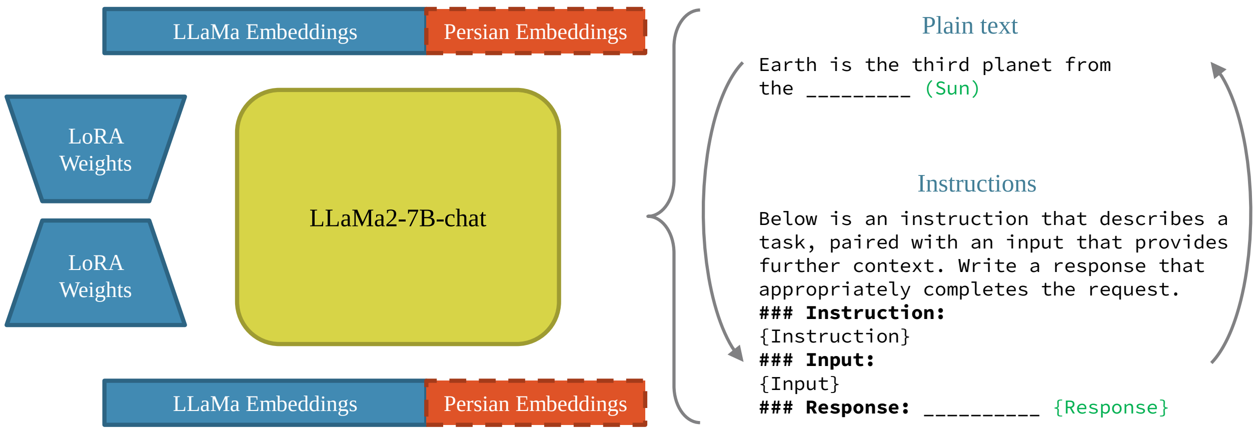

Figure 1. Our Approach for Training PersianMind: We expand LLaMa2's tokenizer and embeddings with 10,000 Persian subwords. Next, we employ the LoRA technique to reduce memory usage during training, with trainable components including input and output embeddings, as well as LoRA weights. The model is trained on Persian plain text and instructions iteratively, evaluating at each step. The training objective is focused on causal language modeling, wherein the model predicts the next token based on previously observed ones.

	After Fine-Tuning		After Instruction-Tuning	
	English	Persian	English	Persian
Step 1	15.10	34.09	-	-
Step 2	24.18	26.56	14.89	16.95
Step 3	29.04	20.35	12.30	14.55

Table 2. Perplexity scores for Persian and English evaluation datasets across training steps after fine-tuning and instruction-tuning. Note that LLaMa2-7B-chat achieves a perplexity score of 8.82 on the English evaluation dataset.

Figure 1 provides an overview of our training approach, while Table 1 presents the datasets used for training, along with their respective token counts. Additionally, it provides information on the datasets used for instruction tuning in each step, along with the corresponding number of instructions.

Table 2 displays the perplexity scores observed in Persian and English datasets after each step of our model's training, before and after SFT. The English perplexity score was computed on an English corpus comprising 24 articles from The New York Times[2] and The New Yorker[3], covering various subjects and containing 54,000 tokens. For the Persian perplexity score, we selected nine articles from Tarjomaan[4] and Faradars[5], spanning various domains and containing 58,000 tokens. For both English and Persian datasets, we selected recently published articles to ensure that our model was not trained on them.

The results indicate that SFT has a significant impact on perplexity scores for both Persian and English languages. In the second step, training our model on more than 1 billion plain Persian text only resulted in an 8-point improvement in perplexity score. However, performing SFT on 170,000 instructions led to a 10-point improvement.

Notably, perplexity scores show that training on plain Persian text could lead to catastrophic forgetting of the model's English knowledge. However, performing SFT on Persian-English parallel dataset could help mitigate this effect.

4 Evaluations

4.1 Multiple-Choice QA and Reading Comprehension

We evaluated our model using the test subset of ParsiNLU multiple-choice QA and Belebele multiple-choice reading comprehension dataset (Bandarkar et al., 2023). The ParsiNLU dataset consists of Persian questions with multiple candidates, where one of them is the correct answer. Notably, there is no context paragraph for these questions. In contrast, the Belebele

[2] https://www.nytimes.com/
[3] https://www.newyorker.com/
[4] https://tarjomaan.com/
[5] https://blog.faradars.org/

	Model	Lit.	Com. Know.	Math	All
Zero-shot	PersianMind	35.7	<u>43.1</u>	29.4	36.1
Fine-tuned	mT5(l)-fa	32.6	27.1	**38.9**	32.9
	mT5(xl)-fa	33.7	27.7	**38.9**	33.4
	mT5(l)-en	27.4	33.1	25.4	28.6
	mT5(xl)-en	28.3	38.6	22.0	29.6
	mT5(l)-fa&en	30.6	28.9	38.6	32.7
	mT5(xl)-fa&en	<u>38.0</u>	33.7	38.0	<u>36.6</u>
	PersianMind	**39.7**	**45.4**	<u>38.8</u>	**41.3**

Table 3. Comparison of PersianMind's performance in ParsiNLU multiple-choice QA dataset across literature, common knowledge, and math & logic Categories.

Model	English	Persian
GPT-3.5-tubro	**87.7**	61.8
LLaMa2-7b-chat	43.9	25.6
XLM-V large	77.8	<u>70.8</u>
PersianMind	<u>86.4</u>	**73.9**

Table 4. Accuracy comparison of PersianMind with the original LLaMa2-7B-chat, the multilingual masked language model XLM-V, and GPT-3.5-turbo on Belebele benchmark.

dataset presents questions extracted from provided paragraphs, with four candidate answers, with one being the correct response.

Table 3 shows evaluation results on the ParsiNLU multiple-choice QA dataset, categorized into *literature*, *common knowledge*, and *math & logic* question types. The results also involve comparing our model with the large and x-large variants of mT5 models (Xue et al., 2021), each trained on Persian, English, or a mixture of both training datasets. The evaluation encompasses both the PersianMind model and its fine-tuned version. The PersianMind was fine-tuned for a single epoch on the training dataset, whereas mT5 model was fine-tuned for 20k steps on the training set for at least 15 epoch (Khashabi et al., 2021).

As can be seen, PersianMind achieved comparable results to the mT5 x-large model trained on both Persian and English datasets. Remarkably, the fine-tuned PersianMind model outperformed the best-performing model by almost 5%. While it demonstrates relatively weaker performance on *math & logic* questions, it excels in the *common knowledge* category, outperforming other models in this specific domain.

We evaluated PersianMind on both the Persian and English subsets of the Belebele dataset. In Table 4, we compared the results of our model with recently released multilingual masked language model XLM-V (Liang et al., 2023), the original LLaMa2-7B-chat, and GPT-3.5-turbo. Results indicate that PersianMind enhanced the performance of LLaMa2-7B-chat by 43% and 48% on English and Persian subsets, respectively. Additionally, PersianMind outperformed the XLM-V model. Although its performance in English was 1% lower than that of GPT-3.5-turbo, it surpassed GPT-3.5-turbo by 12% in the Persian subset.

Despite PersianMind was fine-tuned exclusively on Persian multiple-choice QA datasets in step 3, its performance in English was also improved. This observation highlights the potential for multilingual transfer learning for LLMs.

4.2 Translation

We evaluated PersianMind on the Flores-200 dataset (Costa-jussà et al., 2022) for Fa→En and En→Fa translation directions. Model evaluations were performed using both BLEU (Post, 2018) and Comet (Rei et al., 2022) scores, and results are presented in Table 5. Our model was compared in various setups, including zero-shot, 2-shot, and 4-shot. Additionally, we fine-tuned PersianMind on 5,000 parallel data instances from WikiMatrix (Schwenk et al., 2021) and compared its performance with other models. The ParsiNLU mT5-large model was fine-tuned for 200k steps on Persian-English parallel datasets (Khashabi et al., 2021). Furthermore, we benchmarked our model against GPT-3.5-turbo and NLLB-MoE (Costa-jussà et al., 2022), a multilingual translation model with 54B parameters, supporting 200 languages.

Results show that while PersianMind's Fa→En translation improved in a few-shot setup, the translation quality of En→Fa declined. This observation suggests that in-context learning is not an effective approach for the translation task. Subsequently, we fine-tuned PersianMind on 5,000 translation instructions. The results demonstrate that fine-tuned PersianMind could generate results on a par with mT5-large model which was fine-tuned on a 200,000 Persian-English parallel dataset. However, fine-tuned PersianMind still generated weaker results compared to GPT-3.5-turbo.

In our comparison, we additionally evaluated LLM results against the NLLB-MoE translation model. Despite a substantial gap between LLM results and NLLB-MoE in terms of BLEU score, the gap is much

Model	Fa→En BLEU	Fa→En COMET	En→Fa BLEU	En→Fa COMET
GPT-3.5-turbo	31.7	87.37	18.2	84.66
mT5(l)-ParsiNLU (fine-tuned)	23.3	82.39	15.8	83.77
NLLB-MoE	**40.9**	**88.70**	**24.9**	**87.75**
PERSIANMIND (zero-shot)	13.1	73.51	12.7	78.75
PERSIANMIND (2-shot)	15.4	79.68	11.0	77.25
PERSIANMIND (4-shot)	15.4	79.92	10.2	77.73
PERSIANMIND (fine-tuned)	25.7	83.61	15.4	79.44

Table 5. BLEU and COMET scores of PERSIANMIND in various setups, compared against mT5-large, GPT-3.5-turbo, and NLLB-MoE for for Fa→En and En→Fa translation directions. Note that PERSIANMIND (fine-tuned) was fine-tuned only on 5,000 parallel sentences, while mT5(l)-ParsiNLU (fine-tuned) was fine-tuned on 200,000 Persian-English parallel sentences.

smaller in COMET score. This conveys the fact that while LLM translations have very similar meanings to the reference translations, there is not a high overlap between n-gram of model translations and reference translations.

4.3 Semantic Textual Similarity

To assess the quality of sentence embeddings generated by our model, we conducted evaluations on Semantic Textual Similarity (STS) benchmarks (Corley and Mihalcea, 2005). Initially, we evaluated our model's sentence embeddings on both Persian and English STS datasets independently. Subsequently, we evaluated the cross-lingual performance of semantically similar sentences in the Persian-English context.

For Persian, we utilized the FarSick dataset (Ghasemi and Keyvanrad, 2021), providing sentence pairs with relatedness scores ranging from 1.0 to 5.0. For English, we employed MTEB's STS benchmark dataset (Muennighoff et al., 2023), structured similarly to the FarSick dataset. In our evaluation process, we generated embedding for each sentence in a sentence pair separately and computed their relatedness using cosine similarity. Subsequently, we compared the model's similarity scores with the gold scores using the Spearman correlation metric.

We generated sentence embedding with our model using the *AnglE approach* (X. Li and Jing Li, 2023), utilizing the embedding of the padding token in the following prompt template: `Summarize sentence "text" in one word:` Our results were then compared to other multilingual foundation language models, including mBERT(Devlin et al., 2019), ParsBERT(Mehrdad Farahani, Gharachorloo, Marzieh

Model	FarSick Spearman Corr.	STS Spearman Corr.	Avg.
LaBSE	**66.79**	72.25	69.52
LASER3	60.57	69.77	65.17
LLaMa2-7B-chat	54.24	73.65	63.94
mBERT	52.09	50.97	51.53
ParsBERT	54.73	51.40	53.06
XLM-RoBERTa	49.45	34.49	41.97
PERSIANMIND	63.76	**75.73**	**69.74**

Table 6. Comparison of PERSIANMIND's semantic similarity scores on FarSick and STS-benchmark datasets with other foundation language and bitext mining models.

Farahani, et al., 2021), and XLM-RoBERTa (Conneau et al., 2020). The computation of sentence embedding from these models involved various approaches, such as utilizing the `[CLS]` token's embedding, mean pooling, and the AnglE-BERT approach. Subsequently, we compared the best results from other models with the results obtained from our model. Additionally, we evaluated our model's results in comparison to LaBSE (Feng et al., 2022) and LASER3 (Heffernan et al., 2022) models, which are commonly employed for bitext mining purposes.

In Table 6, we present the results of semantic similarity scores for PERSIANMIND and other models. The findings demonstrate that PERSIANMIND achieved the highest Spearman correlation score among all foundation language models, both in English and Persian. It showcased improvement over LLaMa2-7B-chat correlation scores by 9% in Persian and 2% in English. Furthermore, PERSIANMIND outperformed LASER3 sentence embeddings and achieved comparable average results with LaBSE.

Model	Avg. Cosine Similarity
LaBSE	**86.86**
LASER3	<u>74.49</u>
LLaMa2-7B-chat	58.68
mBERT	64.06
ParsBERT	48.29
PersianMind	72.36

Table 7. Cross-lingual semantic similarity comparison of PersianMind against other foundation language and bitext mining models on the English-Persian subset of the Flores Dataset.

To assess the cross-linguality of our model's sentence embedding, we utilized the Persian-English subset of the Flores-200 dataset. For evaluating the semantic similarity of parallel English-Persian sentences, we computed the cosine similarity of Persian and English sentence embedding and then averaged across all pairs. Table 7 compares PersianMind against other foundation language and bitext mining models. While PersianMind did not reach the average cosine similarity score of LaBSE, a model specialized for this task, it still achieved a notable 72%, surpassing the original model by 14% and emerging as the top-performing language model among others.

5 Carbon Footprint

PersianMind was trained on four NVIDIA RTX 3090 GPUs. Training on Persian plain text required 9 days, while performing the SFT on instructions took 1 day. Consequently, the training of PersianMind consumed a total of 960 GPU-hours. The power consumption of each NVIDIA RTX 3090 GPU is 350W. Additionally, considering a Power Usage Effectiveness (PUE) of 1.4 for our servers due to suboptimal equipment efficiency, the overall power consumption for training PersianMind, based on the formula (C.-J. Wu et al., 2022), was 470kWh.

To calculate carbon emissions, we need to estimate the CO_2 emission intensity, which is linked to the location of our data center. We utilized global electricity data (Ember, 2023) on Iran's carbon intensity of electricity, which measured 494 grams of CO_2 equivalents emitted per kilowatt-hour of electricity. Therefore, the training of PersianMind resulted in emitting 232.38 kCO_2eq.

6 Conclusion

In this paper, we introduced PersianMind, a large language model built upon LLaMa2-7B-chat as the foundation model which incorporated additional 10,000 Persian subwords and trained on an extensive dataset of almost 2 billion Persian tokens. By employing LoRA in our training, our aim was to achieve cost-effective training. We demonstrated that employing the LoRA technique and conducting SFT on English-Persian parallel datasets allows us to reduce catastrophic forgetting of English knowledge despite our model being trained on a large collection of Persian datasets. Our findings indicate that PersianMind achieved comparable results in reading comprehension multiple-choice QA datasets with GPT-3.5-turbo. By fine-tuning with 5,000 parallel sentences, it surpassed the performance of the mT5-large model, which was fine-tuned with 200,000 samples, in translation tasks. Additionally, we demonstrated that PersianMind produced highly effective sentence embedding for both English and Persian sentences. As a language model, it outperformed previous masked language models, showcasing its efficacy in various natural language processing tasks.

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohamadali Nematbakhsh, and Arefeh Kazemi (2021). "ParSQuAD: Persian Question Answering Dataset based on Machine Translation of SQuAD 2.0". In: *International Journal of Web Research.*

Rohan Anil et al. (2023). "Palm 2 technical report". In: *arXiv preprint arXiv:2305.10403.*

Anthropic (2023). *Claude.* `https://claude.ai` [Accessed: 01-07-2024].

Sajjad Ayoubi and Mohammad Yasin Davoodeh (2021). *PersianQA: a dataset for Persian Question Answering.* `https://github.com/SajjjadAyobi/PersianQA`.

Lucas Bandarkar et al. (2023). "The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants". In: *arXiv preprint arXiv:2308.16884.*

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel (2022). "BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.*

Sébastien Bubeck et al. (2023). "Sparks of artificial general intelligence: Early experiments with GPT-4". In: *arXiv preprint arXiv:2303.12712.*

Alexis Conneau et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Courtney D Corley and Rada Mihalcea (2005). "Measuring the semantic similarity of texts". In: *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pp. 13–18.

Marta R Costa-jussà et al. (2022). "No language left behind: Scaling human-centered machine translation". In: *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ember (2023). *Global Electricity Review 2023*. `https://ember-climate.org/insights/research/global-electricity-review-2023/` [Accessed: 01-07-2024].

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri (2021). "ParsBERT: Transformer-based Model for Persian Language Understanding". In: *Neural Processing Letters*.

Mehrdad Farahani, Mohammad Gharachorloo, and M. Manthouri (2021). "Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization". In: *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang (2022). "Language-agnostic BERT Sentence Embedding". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Zahra Ghasemi and Mohammad Ali Keyvanrad (2021). "FarSick: A Persian Semantic Textual Similarity And Natural Language Inference Dataset". In: *International Conference on Computer Engineering and Knowledge*.

Shwai He, Liang Ding, Daize Dong, Jeremy Zhang, and Dacheng Tao (2022). "SparseAdapter: An Easy Approach for Improving the Parameter-Efficiency of Adapters". In: *Findings of the Association for Computational Linguistics*.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk (2022). "Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages". In: *Findings of the Association for Computational Linguistics: EMNLP*.

Dan Hendrycks et al. (2021). "Measuring Massive Multitask Language Understanding". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neil Houlsby et al. (2019). "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*.

Edward J Hu et al. (2022). "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*.

Daniel Khashabi et al. (2021). "ParsiNLU: A Suite of Language Understanding Challenges for Persian". In: *Transactions of the Association for Computational Linguistics*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin (2023). "Bactrian-X: A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation". In: *arXiv preprint arXiv:2305.15011*.

Xianming Li and Jing Li (2023). "AnglE-optimized Text Embeddings". In: *arXiv preprint arXiv:2309.12871*.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky (2023). "Scaling down to scale up: A guide to parameter-efficient fine-tuning". In: *arXiv preprint arXiv:2303.15647*.

Davis Liang et al. (2023). "XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Masked Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

MosaicML (2023). *MPT*. `https://www.mosaicml.com/blog/mpt-7b` [Accessed: 01-07-2024].

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers (2023). "MTEB: Massive Text Embedding Benchmark". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Subhabrata Mukherjee et al. (2023). "Orca: Progressive learning from complex explanation traces of gpt-4". In: *arXiv preprint arXiv:2306.02707*.

OpenAI (2022). *ChatGPT*. `https://openai.com/blog/chatgpt` [Accessed: 01-07-2024].

OpenAI et al. (2023). *GPT-4 Technical Report*. arXiv: `2303.08774 [cs.CL]`.

Guilherme Penedo et al. (2023). "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only". In: *arXiv preprint arXiv:2306.01116*.

Matt Post (2018). “A Call for Clarity in Reporting BLEU Scores”. In: *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Ricardo Rei et al. (2022). “COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task”. In: *Proceedings of the Seventh Conference on Machine Translation*.

Nils Reimers and Iryna Gurevych (2020). “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán (2021). “WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Yi-Lin Sung, Varun Nair, and Colin A Raffel (2021). “Training neural networks with fixed sparse masks”. In: *Advances in Neural Information Processing Systems*.

Rohan Taori et al. (2023). *Stanford Alpaca: An Instruction-following LLaMA Model*. `https://github.com/tatsu-lab/stanford_alpaca`.

Tii (2023). *Falcon*. `https://falconllm.tii.ae/falcon.html` [Accessed: 01-07-2024].

Hugo Touvron, Thibaut Lavril, et al. (2023). “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, et al. (2023). “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*.

Danilo Vucetic et al. (2022). “Efficient fine-tuning of bert models on the edge”. In: *2022 IEEE International Symposium on Circuits and Systems*.

Yaqing Wang et al. (2022). “AdaMix: Mixture-of-Adaptations for Parameter-efficient Model Tuning”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Carole-Jean Wu et al. (2022). “Sustainable AI: Environmental implications, challenges and opportunities”. In: *Proceedings of Machine Learning and Systems*.

Chenfei Wu et al. (2023). “Visual chatgpt: Talking, drawing and editing with visual foundation models”. In: *arXiv preprint arXiv:2303.04671*.

Linting Xue et al. (2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yi (2023). *01*. `https://01.ai/` [Accessed: 01-07-2024].

Wayne Xin Zhao et al. (2023). “A survey of large language models”. In: *arXiv preprint arXiv:2303.18223*.

A Generation samples from PersianMind

We generated the following chat discussing the enhancement of a child's math skillset. To ensure deterministic results, we also set the `seed` value to 1442.

Prompt:

تو یک دستیار هوش مصنوعی هستی که در زمینه یادگیری مباحث مختلف، مشاوره میدهی.
کودکی دبستانی در یادگیری ریاضی مشکل دارد. چگونه میتوان به او کمک کرد تا در درس ریاضی پیشرفت داشته باشد؟

Response:

برای کمک به کودکتان در بهبود مهارت‌های ریاضی، می‌توانید از چند راهکار زیر استفاده کنید:
1) بازی های آموزشی: بازی های آموزشی مانند پازل ها و معماها را با فرزندتان انجام دهید. این کار نه تنها سرگرم کننده است بلکه به آنها یاد می دهد که اعداد را به صورت تصویری ببینند و درک کنند.
2) کتاب های داستان: خواندن کتاب های داستانی که شامل مسائل ریاضی هستند می تواند به کودکان علاقه مند کند و آن ها را تشویق کند که بیشتر درگیر ریاضیات شوند.
3) تمرین حل مسئله: ارائه تمرینات حل مسئله به بچه ها به آنها امکان می دهد تا مهارت های حل مسئله خود را توسعه دهند و در عین حال مفاهیم ریاضی را نیز بهتر بفهمند.
4) معلم خصوصی: اگر کودک شما هنوز هم در کلاس مدرسه مشکل دارد، ممکن است گرفتن تدریس خصوصی بتواند به او کمک کند تا مهارت های ریاضی خود را تقویت کند.
5) برنامه ریزی منظم: ایجاد یک برنامه درسی منظم برای فرزندتان می تواند به او انگیزه بدهد و به او اجازه دهد تا پیشرفت قابل توجهی در دروس ریاضی داشته باشد.
6) تشویق: تشویق کردن فرزندانتان برای تلاش در درس ریاضی بسیار مهم است. به آنها نشان دهید که چقدر این موضوع برای رشدشان اهمیت دارد و اینکه چه قدر سخت کوشانه تلاش کرده اند.

Prompt:

درباره مورد 3 بیشتر میگی؟

Response:

در مورد 3، تمرین حل مسئله، پیشنهاد من این است که هر دو والدین باید با هم همکاری کنند تا به فرزندشان فرصت بیشتری برای حل مسئله دادن داده شود. این به آنها امکان می دهد تا مهارت های حل مسئله خود را گسترش دهند و همچنین مفهوم ریاضی را عمیق تر بفهمند. علاوه بر این، آنها باید به دنبال منابع مناسب باشند که بتوانند به طور موثر به سوالات ریاضی پاسخ دهند.
در پایان، توجه به نکته ای ضروری است که حل مسئله به خودی خود کافی نیست. دانش آموزان باید به شیوه هایی مختلفی مطالب را یاد بگیرند و تمرین کنند تا به خوبی آن را حفظ کنند. بنابراین، سعی کنید روش های متفاوتی را امتحان کنید تا به بهترین شکل ممکن به آنها کمک کنید.

Prompt:

اگه والدین نتونن وقت زیادی رو صرف تمرین کودکشون کنن، از چه راه حل‌هایی میتونن استفاده کنن؟

Response:

اگر والدین نمی توانند زمان زیادی را برای تمرین اختصاص دهند، می توان از ابزارهای آنلاین یا نرم افزارها استفاده کرد. بسیاری از وب سایت ها و اپلیکیشن های موجود در اینترنت وجود دارند که به شما امکان می دهند به راحتی به سؤالات ریاضی بپردازید و به تدریج مهارت های حل مسئله خود را افزایش دهید. علاوه بر این، برخی از مدارس و موسسات آموزش خصوصی نیز دوره های آنلاین برگزار می کنند که می توانید از آنها بهره ببرید.

Prompt:

میتونی چند موردشون رو نام ببری؟

Response:

یکی از معروف ترین وب سایت های ریاضی Learning Games for Kids است که دارای مجموعه وسیعی از بازی های آموزشی است. علاوه بر این، پلتفرم های متعددی مانند Khan Academy و Duolingo وجود دارد که به شما امکان می دهد مهارت های زبانی و ریاضی خود را ارتقا دهید.

Prompt:

درباره‌ی تشویق هم میتونی بیشتر توضیح بدی؟

Response:

تشویق کردن فرزندانتان برای تلاش در درس ریاضی بسیار مهم است. به آنها نشان دهید که چقدر این موضوع برای رشدشان اهمیت دارد و اینکه چه قدر سخت کوشانه تلاش کرده اند. تشویق کردن آنها به تلاش در این زمینه به آنها احساس موفقیت می دهد و باعث می شود که به جای فرار از درس ریاضی، مشتاقانه به آن روی بیاورند.

Prompt:

چه تشویق‌هایی میتونه برای بچه‌های دبستانی مناسب باشه؟

Response:

تشویق‌های مختصر و کوتاه مدت اغلب مؤثرتر هستند زیرا به کودکان اجازه می‌دهند بدون فشار زیاد به خودشان افتخار کنند و احساس موفقیت کنند. به عنوان مثال، می‌توانید بگویید «خیلی خوب بود!» یا " این خیلی عالیه!"

Prompt:

چه بازی های آموزشی‌ای میتونه موثر باشه؟ 5 موردشون رو نام ببر.

Response:

1) پازل ها: پازل ها نوعی بازی هستند که به بازیکنان اجازه می دهد تا با حرکت دادن قطعات مختلف به ترتیب صحیح، یک هدف خاص را بدست آورند. این نوع بازی ها به کودکان کمک می کند تا الگوها و روابط بین اعداد را بهتر بشناسند.
2) معما ها: معماها نوعی بازی هستند که به بازیکنان اجازه می دهند تا با حل یک معما، یک هدف خاصی را بدست آورند. این نوع بازی ها به کودکان کمک می کند تا مهارت های حل مسئله خود را گسترش دهند و همچنین مفهوم ریاضی را عمیق تر بفهمند.
3) بازی های فکری: بازی های فکری انواع مختلفی دارند اما همگی به بازیکنان اجازه می دهند تا با فکر کردن به یک مسئله، یک جواب را پیدا کنند. این نوع بازی ها به کودکان کمک می کند تا مهارت های تفکر انتقادی خود را گسترش دهند و همچنین مفهوم ریاضی را عمیق تر بفهمند.
4) بازی های استراتژیک: بازی های استراتژیک انواع مختلفی دارند اما همه آنها به بازیکنان اجازه می دهند تا با تصمیم گیری های هوشمندانه، یک هدف خاص را بدست آورند. این نوع بازی ها به کودکان کمک می کند تا مهارت های تصمیم گیری خود را گسترش دهند و همچنین مفهوم ریاضی را عمیق تر بفهمند.
5) بازی های رومیزی: بازی های رومیزی انواع مختلفی دارند اما همگی به بازیکنان اجازه می دهند تا با حرکت دادن مهره ها، یک موقعیت خاص را بسازند. این نوع بازی ها به کودکان کمک می کند تا مهارت های حل مسئله خود را گسترش دهند و همچنین مفهوم ریاضی را عمیق تر بفهمند.