DParT: Transferring knowledge between languages changing a few weights

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have significantly impacted both research and business domains, automating tasks previously unattainable by artificial intelligence. However, the primary focus on English and European languages presents a barrier in adapting and applying LLMs to other languages due to the challenges involved in data collection, pre-processing, and model training. To overcome this issue, we propose a Double Partial Tuning (DParT) strategy. It involves modifying the structure of the training data in the first stage and employing low rank adapters (LoRA) in the second stage, leading to knowledge transfer between languages and low computational efforts in terms of trainable parameters and data quantity. Tests on Arabic and Russian languages demonstrate the superiority of DParT over other training methods, potentially expanding the application of LLMs in various languages and further revolutionizing research and business fields. We selected Arabic and Russian languages, as they originate from distinct language families and utilize two different non-Latin scripts, in order to demonstrate the effectiveness of the proposed approach. Code and datasets will be made publicly available.

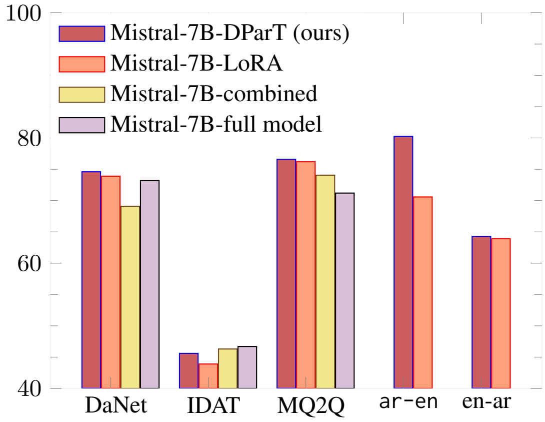

Figure 1: A comparison between Mistral models fine-tuned using our approach, LoRA and a combined approach on a 30k instruction dataset in Arabic. `en-ar` and `ar-en` are the translation COMET scores from English to Arabic and vice versa.

1 Introduction

Large language models have gained significant attention in the last years due to their importance in accelerating automation processes in various aspects and their ability to function as a valuable human assistant in different tasks. However, one of the major challenges faced is the lack of data for some languages, which has led most open-source models to primarily focus on English or other popular languages such as Chinese or European languages. In order to address this problem, organizations and individuals must gather large amounts of data in the desired language to train LLMs, which is expensive in terms of time and resources. Using automated processes to collect data may result in noisy data, leading to poor model performance. In our research, we propose DParT a two-stage training method to overcome the issue with minimal data requirements. In the first training phase, our aim is to enhance the model's ability to accommodate the new language. We propose a technique to transfer knowledge between languages by fine-tuning the embedding layer to be representative of the new language while maintaining its primary trained language (typically English) as a basis. The second stage of training focuses on training the model to generate answers in the target language more effectively using basic adapter training. In conducting our experiments, we utilized various models for different languages. For Arabic, we trained on two distinct models - Mistral (Jiang et al., 2023), featuring a limited portion of Arabic data during training, and mGPT (Shliazhko et al., 2022), which was trained with a substantial amount of Arabic data. In contrast, for Russian, we explored the Mistral and Llama2 (Touvron et al., 2023) models, both of which were trained on Russian data but with a relatively small quantity compared to the English counterpart. To summarize the contributions of our paper are the following: (i) Present a novel two-stage training method (DParT) that facilitates knowledge transfer between languages, even when they are under-resourced; (ii) Develop and release an open-source translated Arabic dataset of 30,000 instructions and a translated benchmark based on the DaNetQA Russian benchmark.

2 Related Work

Modeling non-English languages. Training models on specific data became common solution for adaptation to low-resource languages, especially with the emergence of LLM: InternLM (2023) authors used large multilingual dataset with an emphasis on Chinese language, Alabi et al. (2022) proposed an approach to adapt multilingual PLMs with African datasets, BLOOM (2022) was trained on dataset of 46 natural languages. In our work we chose to train on Arabic and Russian data to show efficiency of our method on distinct language families.

Recently, Zhao et al. (2024) used mainstream models such as Llama and shows that they can be comparable to state-of-the-art transfer models in understading non-English languages by extending the pretraining data by tiny amount.

Cross-Lingual Pretraining. The idea of learning embeddings as a preliminary step for a better understanding of multilingual tasks (Cohn et al., 2017; Artetxe and Schwenk, 2019; Artetxe et al., 2020). Another common idea is using bilingual input exemplars during training. Tang et al. (2020) present a finetuning method for translation, where model is trained on many directions at the same time with collection of multilingual bitexts. Nguyen et al. (2023) propose to collect exemplars from a diverse set of different languages to prompt the LLMs to translate into English. Just like in our work, they use these prompts to create intra-lingual exemplars to perform tasks in the target languages. However, we draw fundamentally different conclusions since one of our main contributions is the multistage training.

Training methods. Addressing size of LLM, full parameter fine-tuning usually requires enormous computational resources. PEFT (Xu et al., 2023) has emerged as a viable solution to compensate for it. We compare our multistage training method to other parameter-efficient fine-tuning methods, particularly, LoRA (Hu et al., 2022).

3 Method

In this section, we begin by offering a concise overview of Large Language Models (LLMs), subsequently diving into the difficulties encountered when attempting to fine-tune these models for languages that they were not initially trained on. Then, we introduce DParT an enhanced version of the LoRAadapters training method, which subsequently overcome the shortcomings associated with fine-tuning LLMs for low-resource languages.

Large language models (LLMs) are neural networks trained on enormous amounts of text data to produce text that appears to have been written by a human, based on the input they receive. These models are typically pre-trained in multiple languages and can be fine-tuned for specific tasks by altering their parameters according to additional data specifically relevant to the task at hand. Unfortunately, when these models are fine-tuned for languages that they were not originally trained on, they often exhibit poor performance as they struggle to adapt to the unique characteristics and complexities of that particular language.

One of the primary issues faced by LLMs is their inability to effectively adapt to a variety of languages, which is largely due to the limited availability of open-source data for those languages. Fine-tuning such models on languages that are rarely included in the original training set often results in disappointing outcomes.

DParT method comprises two distinct stages, each serving a specific purpose in helping the model comprehend and communicate effectively in the target language. During the first stage, our focus is on transferring knowledge between English and the target language, providing the model with the foundational skills needed to understand the language. This stage is crucial to establish a stable base for the model's language proficiency. In order to achieve this, during the initial stage of the process, we fine-tune only the embedding layer of the model, ensuring that all other components remain untouched. Then, we present pairs of questions to the model, with the first question being in the target language and the second question being in English but prompted specifically to elicit a response in the target language. Consequently,

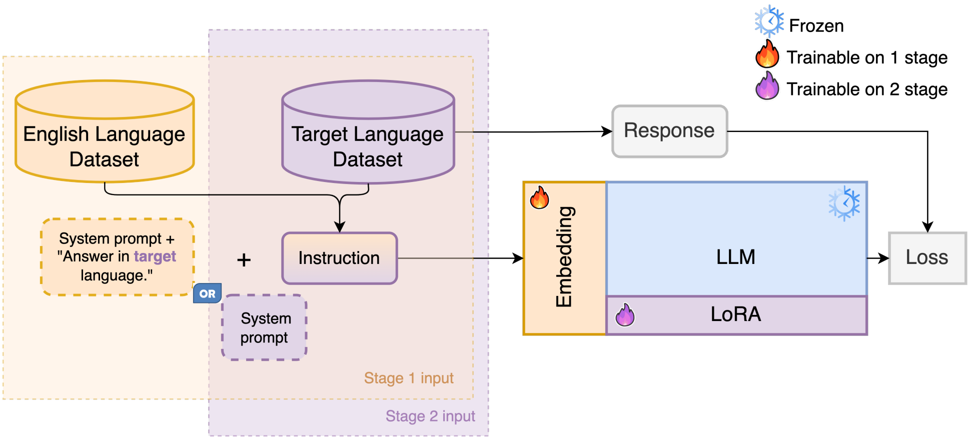

Figure 2: DParT tunes the embeddings a the first stage with special structured data, then uses LoRA at the second phase to get the model aligned with the new language

the desired outcome for both queries should be the same, drawing from responses based on the truth grounding solely within the target language Fig.2. Once the first stage is complete, we progress to the second stage where we fine-tune the model using LoRA to further hone its language-specific capabilities. In this stage, the model learns to adopt the unique characteristics and nuances of the target language, ultimately enabling it to speak the language fluently and accurately. By combining the knowledge transfer from the first stage with the specialized language learning in the second stage, our method ensures that the model is well-prepared to handle the complexities and intricacies of the target language. We provide a geometric explanation supports our hypothesis on the supplementary material.

4 Experiments

We start with describing the datasets employed for benchmarking our approach and examine the metrics utilized to assess the performance. Our proposed training method demonstrates superior results compared to LoRA and full finetuning across various scenarios, with substantial improvement in performance for certain models.

Datasets and metrics. We used Alpaca-cleaned (gururise, 2023), as a higher-quality version of the original, and OpenOrca (Mukherjee et al., 2023) datasets for training, because these are high quality datasets defacto standard in research. We translated these datasets into Arabic with YandexTranslate, and corpora will be released upon publication. We took all 50k samples from Alpaca and 30k from larger OpenOrca dataset for both languages. Evaluation metrics for Russian language tasks were from RussianSuperGLUE (Shavrina et al., 2020) (DaNetQA, RUSSE, TERRa). For Arabic, ALUE benchmark (Seelawi et al., 2021) (MQ2Q, IDAT) was chosen. We also translated DaNetQA into Arabic and plan to make it publicly available. Refer to supplementary material for examples of prompts for these tasks.

We measure the performance using accuracy (whether or not the model answered correctly) and so-called rate. This measure counting if the model answered either “yes” or “no” to a question, since we generate an answer instead of choosing it. We count the answer into if it starts with “yes” or “no” only. These metrics show how the model understands the context and can follow instructions.

To validate machine translation, we used the Tatoeba (Tiedemann, 2020) and Flores200 (Costa-jussà et al., 2022) datasets. These datasets are considered high quality as they consist of hand-written or native speaker-collected translations. For evaluation we used COMET (Rei et al., 2020) metric.

Comparison with LoRA methods. We experiment on different pretrained and instruction-tuned models and on two different languages (Arabic and Russian). We compare our results with the same training hyper-parameters and datasets with LoRA against DParT our proposed training method. DParT outperforms LoRA on majority of experiments for both instruction and generative models. Tables 1 and 2 illustrates our results on Arabic, Russian benchmarks in order. We have conducted multiple identical experiments on Mistral models with different seed values. the statistical significance of 5% shows the superiority and stability of DParT. For full model training results comparison and number of trainable parameters refer to ablation study on the supplementary material.

Table 1: Comparison between DParT and LoRA training on Arabic benchmarks (mean average) in zero-shot. CI stands for Confidence Interval

Model	Training Method	Accuracy↑	Rate↑	CI
Mistral (Dolphin)	LoRA	64.67	99.96	$63.89_{\pm 0.83}$
Mistral (Dolphin)	DParT	**65.6**(+1)	**100**	$64.42_{\pm 1.18}$
Mistral-7B-v0.1	LoRA	59.33	99.9	$58.36_{\pm 0.86}$
Mistral-7B-v0.1	DParT	**60.96**(+1.63)	**100**	$59.03_{\pm 1.84}$
mGPT	LoRA	48.03	98.93	
mGPT	DParT	**48.26**(+0.23)	98.93	
gpt-3.5-turbo	-	63.57	99.4	

Table 2: Comparison between DParT and LoRA training on Russian benchmarks (mean average) in zero-shot. CI stands for Confidence Interval

Model	Training Method	Accuracy↑	Rate↑	CI
Mistral (Dolphin)	LoRA	**70.48**	**99.7**	
Mistral (Dolphin)	DParT	70.29(-0.19)	99.57	
Llama2-7b-chat	Lora	55	99.9	
Llama2-7b-chat	DParT	**56.57(+1.5)**	99.9	
Mistral-7B-v0.1	Lora	66.9	99.6	$66.34_{\pm 0.53}$
Mistral-7B-v0.1	DParT	**67.2**(+0.3)	99.6	$67.10_{\pm 0.27}$
Llama2-7b	LoRA	59.26	99.42	
Llama2-7b	DParT	**59.38**(+0.12)	**99.67**	
gpt-3.5-turbo	-	64.5	-	

Machine Translation. We evaluate instruction-tuned versions on two tracks, the first one for translating between Arabic and English and the second for Russian and English. Models such as Mistral or Llama2 were used for fine-tuning and were subsequently compared with their foundation versions and other models such as gpt-3.5-turbo.

Table 3: **Results on machine translation task.** Presented metric is COMET, values are in percentage.

tgt	**Models**	**Tatoeba**		**Flores-200**	
		tgt-en	en-**tgt**	**tgt**-en	en-**tgt**
ar	Mistral-Dolphin	54.53	39.58	75.42	45.11
	+LoRA	70.64	66.70	70.52	61.12
	+DParT	79.52	68.20	80.96	60.39
	gpt-3.5-turbo	86.30	85.73	87.21	87.03
ru	Mistral-Dolphin	83.78	67.00	83.11	67.51
	+LoRA	82.30	83.71	82.08	83.38
	+DParT	82.31	83.66	81.79	83.30
	Llama2-7b-chat	48.10	57.04	54.86	58.68
	+LoRA	74.58	56.40	70.89	73.25
	+DParT	73.94	52.53	69.53	65.30
	gpt-3.5-turbo	87.98	90.32	87.63	91.05

In Table 3 we notice that DParT yields better results in Arabic-to-English translation when compared with LoRA, which shows how embedding training stage helps the model to better understand syntactic and semantic properties of Arabic. Although Mistral trained with our method still falls behind gpt-3.5-turbo, the gap is relatively small when translating to English. When translating from English to Arabic the impact of additional embedding training is lesser. However, the metric values are still high compared to the original instructive model, which attests to the consistency of our cross-lingual training method. Different conclusions can be drawn for Russian translation track: the original Mistral model performs better while translating to English; however, we managed to improve its performance in translation to Russian. Likewise, the Llama-7b-chat greatly benefits from further training with Double Partial Tuning.

We can see how translations to non-English languages fall short compared to the opposite case, confirming the existence of a significant bias for English language in LLM. When translating from Russian to English and vice versa the metrics are noticeably higher than for a similar track with Arabic. We assume that the foundation Mistral model had a larger share of Russian-language text in the training set and the embeddings of Cyrillic tokens have been generally updated more often than Arabic during all training stages, which results in deeper understanding of the Russian language.

5 Conclusion

In this study, we introduced a novel two-stage approach for knowledge transfer between languages in the context of Large Language Models (LLMs). The presented method is designed to enhance the information and reasoning abilities of LLMs by leveraging minimal amounts of data. Our proposed approach was evaluated against existing techniques across different benchmarks, languages (Arabic and Russian), and various categories of instruction-tuned/generation models. The results showed that the new method significantly outperforms its predecessors, thus advancing the field of multilingual LLM development. This innovative technique marks a crucial step towards bridging the gap in open-source data availability for less-resourced languages, ultimately increasing the overall effectiveness of LLMs in diverse linguistic environments.

Limitations

In the proposed technique DParT, there has been a significant improvement observed in diverse performance metrics. This strategy provides a simplistic yet effective way to fine-tune Large Language Models (LLMs) in numerous languages with limited training data. However, this two-stage method necessitates more resources for training and results in a more intricate training pipeline. Our internal studies have demonstrated that adding new data to the training process leads to substantially better outcomes. Nevertheless, the expense involved in acquiring fresh data remains a point of concern.

Ethics Considerations

Our approach is aimed to simplify the adaptation of LLMs to new languages. The models we are considering in our experiments were not trained on the languages we used. We enrich them with knowledge of these languages. We expect that minority groups will benefit from the adaptation of the existing large language models to their languages. We should state that our method is not guaranteed to work on all the existing languages. Thus its application on some languages could lead to poor model quality on these languages. Our method on the one hand is not intended to incorporate a bias into a model, but also on another hand it is not targeted to correct the pre-existing in the model it is applied onto. Our method is not intended to collect any data from any person, although it is not guaranteed from such a data usage being fed with one.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Trevor Cohn, Steven Bird, Graham Neubig, Oliver Adams, and Adam J. Makarucha. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

gururise. 2023. Alpacadatacleaned. https:https://github.com/gururise/AlpacaDataCleaned.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *CoRR*, abs/2306.11372.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Anton A. Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4717–4726. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *CoRR*, abs/2204.07580.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1174–1182. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *CoRR*, abs/2312.12148.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *CoRR*, abs/2401.01055.

A Appendix

In the ensuing appendix, we provide an exhaustive examination aimed at contrasting our novel training technique with existing counterparts. A set of experiments utilizing Alpaca Arabic and Russian corpora serve to assess the efficacy and versatility of our proposed method. These experiments were instrumental in establishing the generalizability of our approach within diverse linguistic realms.

A.1 Geometric explanation of the proposed method

The proposed technique entails converting each question in the target language into English. This permits us to input instructions in English into the system while appending a request to the prompt to compel the response to be in the target language. By leveraging the answer provided in the target language as ground truth, we can compare it to the instruction given in the target language and in English. The embedding in this approach represents the hyperspace of the languages, where the English language hyperspace acts as the dominating space. Our objective is to connect the hyperspace of the target language with the dominating one, thus allowing us to map any question in the target language to the dominating hyperspace. Consequently, the model is capable of understanding questions in the target language if their answers are present in the dominating hyperspace.

$$\Omega_t \subset \Omega \ and \ \Omega_e \subset \Omega \tag{1}$$

where Ω is the hyperspace, Ω_t is the hyperspace of the target language, and Ω_e is the hyperspace of the dominating language which is English on our case. we will denote the projection of a query on the hyperspace as $\Omega(q)$. Let us assume that we have a query q_t in our target language and q_e in English, then our training goal is minimize the difference of the projected similar queries (in different languages) into the hyperspace 2 :

$$min \ ||\Omega(q_t) - \Omega(q_e)||^2 \tag{2}$$

where $\Omega(q_t) \in \Omega_t$ and $\Omega(q_e) \in \Omega_e$. Some models already trained over a sufficient data of different languages therefore the boost of the results could vary, according to the model initial training datasets. Our hypothesis suggests that the resulting model will possess improved capabilities in understanding target languages. To evaluate this assumption, we selected a set of GLUE tasks that measure the ability to analyze instructions and answer with a "yes" or "no". Zero-shot translation tasks also provide valuable information about the model's ability to understand languages. a detailed comparison and ablation provided on the experiments section.

A.2 Datasets

Training datasets For the Arabic and Russian languages we used Alpaca-cleaned (gururise, 2023) and the first 30K instructions from OpenOrca (Mukherjee et al., 2023) dataset.

The Alpaca-cleaned dataset is a revised version of the original Alpaca Dataset released by Stanford. It addresses various issues found in the original dataset, such as hallucinations, merged instructions, empty outputs, and missing code examples. Additionally, it removes instructions for generating images and addresses inconsistent input fields. The Alpaca-cleaned dataset consists of 51,760 examples. Each example contains instruction, input and output. The instruction describes the task that the model is required to perform, and each instruction is unique among the 51,760 examples. The input is an optional field that provides context or input for the task. The output contains the answer to the instruction, generated by the text-davinci-003 model.

The OpenOrca dataset is a open-source collection of sub-collections, each containing multiple tasks and queries. The dataset focuses on zero-shot queries and includes the CoT, NiV2, T0 and Flan 2021 sub-collections. The dataset consists of 5 million examples. Each example contains three parts: a system message, a user query, and the response from LFM. The system message is provided at the beginning of the prompt and includes important context and guidelines. There are 16 different handcrafted system prompts. The user query specifies the task we want the LFM to perform.

Evaluation datasets For the Russian language, we chose three tasks from the RussianSuperGLUE benchmark (Shavrina et al., 2020), specifically DaNetQA, RUSSE, and TERRa. The task of all benchmarks is to classify whether the answer to each question is true or false.

DaNetQA is a question answering dataset that focuses on yes/no questions. The dataset contains triplets of (question, passage, answer), with the option for an additional title for context. The dataset is unique because the questions are generated in natural and unconstrained settings, rather than being prompted by annotators. We evaluated on validate data, which contains 821 examples.

Russian WiC - RUSSE Dataset is designed as a benchmark for evaluating context-sensitive word embeddings. It addresses the limitation of mainstream static word embeddings by providing dynamic representations of words that can adapt based on context. The problem was whether the word has the same meaning in two sentences. The evaluated split contains 8505 examples.

TERRa dataset consists of text fragments that are used for Textual Entailment Recognition. The task is to determine whether the meaning of one text can be inferred from another text. The dataset includes pairs of text fragments, where each pair consists of a premise and a hypothesis. The label indicates whether the premise entails the hypothesis or not. This dataset is used for sentence pair classification, specifically for recognizing textual entailment. We evaluated 307 examples.

For the Arabic language, we chose MQ2Q and IDAT from ALUE benchmark (Seelawi et al., 2021) and translated DaNetQA to Arabic, which we will make it publicly available.

The IDAT focuses on detecting irony in Arabic tweets. It uses a dataset of approximately 1,006 tweets, each of which is classified as "1" if it contains irony, satire, parody, sarcasm, or if the intended meaning is opposite to the literal one. Tweets without these characteristics are labeled as "0".

The task of MQ2Q in Arabic aims to determine the level of similarity between pairs of questions based on their semantic meaning and answer. In this task, a pair of questions is considered semantically similar if they share the same answer and meaning, which is labeled as "1". If the questions do not meet this criteria, they are labeled as "0". There are 11,997 pairs with an equal distribution of "0"s and "1"s.

For validation on machine translation task we decided to compute metrics on Tatoeba (Tiedemann, 2020) and Flores200 (Costa-jussà et al., 2022) datasets.

The first metric for evaluation is chrF++ (Popović, 2017), which uses character n-grams for comparing machine translation output with reference translations. This method is especially useful for high-morphology languages, unlike metrics based on word n-grams. The second metric

COMET (Rei et al., 2020) is a neural framework with state-of-the-art levels of correlation with human judgments.

A.3 Ablation

In this section, we present various experiments designed to showcase the efficiency of the initial stage training (knowledge transfer) and to highlight the advantages of our proposed method over full model training. Our results indicate that the proposed approach outperforms the conventional method and emphasize the significance of the data structure in the early stages of the training process. Through these experiments, we aim to demonstrate the effectiveness of our method and underscore the importance of implementing the appropriate data structure during the first stage of training.

Our method against full model training The purpose of this study is to evaluate the efficacy of our approach, which we implemented by conducting two different training experiments on the orca-arabic 30K dataset using the Mistral-Dolphin chat model. The first experiment involved employing our method, which entails knowledge transfer at the initial stage and proceeding with LoRA finetuning thereafter. The second experiment consisted of training the entire model on the same dataset. Both experiments were conducted with the same number of epochs and hyperparameters, except for the learning rate. For the full model training, we utilized the same learning rate as that used to train the embedding in the first stage, but it was 10 times smaller than the learning rate assigned for the LoRA training. Within the presented study, as demonstrated in Table .4, a comprehensive comparison is performed between the various experiments conducted on the Glue Arabic benchmarks.

Table 4: Comparison between our proposed training method and full model training on Arabic benchmarks in zero-shot

Model	Type	Training Method	Accuracy↑	Rate↑
Mistral	chat(Dolphin)	full-model	63.15	100
Mistral	chat(Dolphin)	DPart	**65.6**(+2.45)	100

Data structure importance for knowledge transfer The objective of our experiment was to highlight the significance of the chosen data structure in facilitating knowledge transfer, and to emphasize that we are comparing our findings against a single stage training process with equivalent trainable parameters. In this particular instance, the Mistral-Dolphin model's embedding and LoRA adapters, along with the embedding layer, were unfreezed during a single stage of training, utilizing identical hyper-parameters on the Orca-Arabic 30K dataset. The results presented in Table 5 serve as a compelling demonstration of the superiority of our approach in enhancing performance on Arabic Glue Tasks.

Table 5: Comparison between our proposed training method and the two stages combined, on Arabic ORCA dataset

Model	Type	Training Method	Accuracy↑	Rate↑
Mistral	chat(Dolphin)	2 stages combined	63.7	100
Mistral	chat(Dolphin)	DPart	**65.6**(+1.9)	100

Table 6: Comparison between the number of trainable parameters according to each training method

Training method	Trainable parameters in millions			
	Mistral	Llama2-7B	Falcon 7B	mGPT
LoRA	54.52	67.1	37.74	288.35
DParT	185.6	198.18	333.21	800.36
Full-model	7296	6805	7254	13396

A.4 Detailed ORCA experiments

we present detailed outcomes relating to the primary findings documented in the study discussing the achievements in two languages, Arabic and Russian, for various adhesive tasks. References to Tables 7 and 8 provide specific numerical data. Previously, we explained converting tasks into yes/no queries to gauge accuracy rates. To enhance comprehension of this conversion process, we will analyze each task individually and supply the corresponding evaluation prompt employed.

A.4.1 Arabic Benchmark

DaNetQA The given trial data originates as a translated adaptation from the Russian DaNet collection. It initially presents certain facts in textual format, subsequently accompanied by a query pertaining to the information. To initiate our approach, we implemented the following system prompt: أنت مساعد شخصي أجب عن الأسئلة ب نعم أو لا Following this, both the textual content and the query are processed.

IDAT In the realm of recognized Arabic tasks from Alue benchmark, we direct the artificial

Table 7: Comparison between our proposed training method and LoRA training on Arabic benchmarks in zero-shot. all models trained on ORCA 30K dataset.

Model	Type	Training Method	DaNet		IDAT		MQ2Q	
			Accuracy	Rate	Accuracy	Rate	Accuracy	Rate
Mistral	dolphin	LoRA	73,9	100	43,9	99,9	76,2	100
		DParT	**74,6**	100	**45,6**	100	**76,6**	100
Mistral-7B-v0.1	Generation	LoRA	66,99	100	46,5	99,7	64,5	100
		DParT	**70,2**	100	**48**	100	**64,7**	100
mGPT	Generation	LoRA	44,3	99,6	44,9	97,3	54,9	99,9
		DParT	44	99,7	**45,8**	97,2	**55**	99,9
gpt-3.5-turbo	chat	-	80,63	99,39	42,8	98,9	67,3	99,9

Table 8: Comparison between our proposed training method and LoRA training on Russian benchmarks in zero-shot. all models trained on ORCA 30K dataset.

Model	Type	Training Method	DaNet		RUSSE		TERRA	
			Accuracy	Rate	Accuracy	Rate	Accuracy	Rate
Mistral	dolphin	LoRA	82,46	99,87	43,33	99,27	85,66	100
		DParT	82,46	99,75	42,43	98,98	**85,99**	100
Llama2	chat	LoRA	72,59	99,87	44,02	99,98	48,53	100
		DParT	**72,95**	99,75	43,04	99,98	**53,74**	100
Mistral-7B-v0.1	Generation	LoRA	79,9	99	39,5	99,8	81,4	100
		DParT	**79,9**	99	**41,3**	99,8	80,4	100
Llama2	Generation	LoRA	70,4	98,78	45,84	99,81	61,56	99,67
		DParT	**71,25**	99,14	44,69	99,87	**62,21**	100
gpt-3.5-turbo	chat	-	79	-	39.8	46.6	74,6	86.97

intelligence model to ascertain whether a given sentence represents a factual statement. If so, it will lack any irony, satire, parody, or sarcasm through implementing this system prompt:

أنت مساعد شخصي اقرأ الجمل التالية

أجب ب »نعم« في حال صحة المعلومة أو »لا«

في حال عدم صحتها

Following this process, we treat the sentence as a user query.

MQ2Q In addition to the previously mentioned assignment from Alue's benchmark, we encounter a collection of paired data. Our objective with this task is to determine whether two given sentences share the same semantic significance through applying this system prompt:

أنت مساعد شخصي اقرأ الجملة الأولى

وتأكد إن كانت ذات المعنى في الجملة الثانية.

أجب ب »نعم« أو »لا«

following that, we pass the sentences as a user request.

A.4.2 Russian Benchmark

In the given segment, we outline the approach adopted for RussianSuperGLUE assignments involving DaNetQA, RUSSE, and TERRa. Our objective is to transform these tasks into 'yes' or 'no' instructions. It should be highlighted that we used the same system prompt for all of them.

DaNetQA DaNetQA represents a question answering dataset focusing on yes/no queries. Comprising triads of information, it includes a question, a text excerpt (passage), and the corresponding response. To utilize this resource, models are given directions in the format of 'passage' and 'question', accompanied by the following instruction "Ответь да или нет".

RUSSE The RUSSE dataset serves to assess context-dependent word representations. It offers variable word depictions capable of adjusting according to surrounding text, mitigating the constraint of static word embeddings. We construct the query format as such: Имеет ли слово 'word' одинаковый смысл в следующих двух предложениях? 'sentence1' 'sentence2' Ответь да или нет.

TERRa The TERRa dataset comprises segments of text designed for identifying textual inference recognition. It comprises combinations of premise statements and hypotheses, accompanied by labels detailing whether the premise logically leads to the hypothesis. To accomplish this undertaking, we employed three distinct prompts, chosen at random with an equal likelihood of selection.

- 'premise' Следует ли из этого что 'hypothesis'? Ответь да или нет.
- 'premise' Верно ли что 'hypothesis'? Ответь да или нет.
- 'premise' 'hypothesis'? Ответь да или нет.

A.5 Alpaca experiments

In order to establish the reliability and versatility of the suggested training approach, it is crucial to conduct training on various datasets to avoid obtaining outcomes exclusive to a specific dataset. To achieve this, we undertook training on the Alpaca dataset, which has more noise and lower context compared to ORCA. The resulting data displayed in Tables 9, 10, and **??** align with the findings obtained when training on the ORCA 30K dataset. Notably, there is an enhanced improvement compared to training with LoRA alone.

A.6 Machine Translation

We constructed a prompt to help models understand task by adding a correct input-output exemplar to instruction (one-shot prompting). Before training our instruction models we add special tokens **<|im_start|>** and **<|im_end|>** to the tokenizer dictionary. For models that we didn't train or the ones without mentioned tokens in tokenizer we used special tokens which correspond to models' conversation template.

Following prompts were used while translating to English:

- **Arabic**
  أنت مساعد شخصي اقرأ النص بإمعان وترجمه الى الانكليزية. مثال: ألبرت آينشتاين كان من أفضل علماء الفيزياء النظرية في ١٩٠٥ قام بطرح نظرية النسبية العامة والخاصة.
  ###Albert Einstein was one of the best theoretical physicists in 1905, he put forward the theory of general and special relativity.

- **Russian**
  Ты - полезный помощник, прочитай текст и переведи его на английский. Пример: ### Альберт Эйнштейн был одним из лучших физиков-теоретиков, в 1905 году он выдвинул теорию общей и специальной теории относительности.### Albert Einstein was one of the best theoretical physicists in 1905, he put forward the theory of general and special relativity.

Prompts for translation in the opposite direction were constructed by analogy.

Table 9: Comparison between our proposed training method and LoRA training on Arabic benchmarks in zero-shot. all models trained on Alpaca cleaned dataset.

Model	Type	Training Method	DaNet		IDAT		MQ2Q	
			Accuracy	Rate	Accuracy	Rate	Accuracy	Rate
Mistral	dolphin	LoRA	62,2	100	47	100	70,8	99,8
		DParT	**63**	100	**47**	100	**71,4**	99,7
Mistral-7B-v0.1	Generation	LoRA	61,75	100	50	98,9	62,09	100
		DParT	**67,2**	100	**51,3**	98,4	57,5	99,9
mGPT	Generation	LoRA	43,3	85,62	32,4	66,4	0	0
		DParT	**49,8**	99,1	**50,7**	96	**30,08**	64,4
Falcon 7B	Generation	LoRA	6,4	16,8	6,3	15,7	9,1	3,1
		DParT	6,33	15,4	**20,87**	38,2	0,1	0,2

Table 10: Comparison between our proposed training method and LoRA training on Russian benchmarks in zero-shot. all models trained on Alpaca cleaned dataset.

Model	Type	Training Method	DaNet		RUSSE		TERRA	
			Accuracy	Rate	Accuracy	Rate	Accuracy	Rate
Mistral	dolphin	LoRA	76,12	99,75	37,43	99,97	71,33	100
		DParT	75,88	99,75	**37,74**	99,97	**72,63**	100
Llama2	chat	LoRA	73,69	95,49	39,97	99,71	60,26	97,77
		DParT	**75,63**	96,34	38,37	99,91	**61,23**	97,06
Mistral-7B-v0.1	Generation	LoRA	64,67	98,53	40,11	99,97	73,61	99,67
		DParT	**65,52**	98,9	38,85	99,97	**74,59**	99,67
Llama2	Generation	LoRA	64,19	93,42	48,1	98,93	56,02	92,5
		DParT	58,46	87,21	45,05	97,95	52,44	90,87

Languages	**Models**	**Tatoeba**		**Flores-200**	
		chrF++	COMET	chrF++	COMET
Arabic → English	Mistral-Dolphin-2.1	32.98	54.53	51.61	75.42
	+LoRA	6.55	60.34	19.96	60.62
	+DParT	51.02	76.28	51.57	77.05
	gpt-3.5-turbo	62.97	86.30	61.85	87.21
English → Arabic	Mistral-Dolphin-2.1	9.02	39.58	21.79	45.11
	+LoRA	28.44	64.37	29.48	61.00
	+DParT	27.61	64.33	28.38	61.72
	gpt-3.5-turbo	47.67	85.73	50.10	87.03
Russian → English	Mistral-Dolphin-2.1	62.75	83.78	59.24	83.11
	+LoRA	52.75	81.29	56.95	82.33
	+DParT	52.93	81.31	57.35	82.51
	Llama2-7b-chat	16.32	48.10	28.15	54.86
	+LoRA	56.30	80.78	54.52	80.90
	+DParT	55.04	80.18	50.20	78.57
English → Russian	Mistral-Dolphin-2.1	35.91	67.00	41.90	67.51
	+LoRA	47.43	84.97	47.20	84.22
	+DParT	47.49	85.02	47.29	84.16
	Llama2-7b-chat	23.81	57.04	28.61	58.68
	+LoRA	35.04	76.48	36.67	74.37
	+DParT	20.43	58.42	32.39	68.16

Table 11: **Results on machine translation task.** All our models were fine-tuned on Alpaca cleaned dataset, metric values are in percentage.