MAPLE: Multilingual Evaluation of Parameter Efficient Finetuning of Large Language Models

Divyanshu Aggarwal† Ashutosh Sathe†
Ishaan Watts Sunayana Sitaram

†Equal Contribution

Motivation

- **GAP in Multilingual Performance of LLMs:** We have seen from past work that there is a large gap between the performance of LLMs on English and performance on other languages [1].
- **Parameter Efficient Finetuning of LLMs:** Parameter efficient fine-tuning is a promising solution to improve the performance of pretrained LLMs when you don't have resources to do full fine-tuning. [2, 3]
- **Multilingual Instruction Finetuning:** Multilingual Instruction Datasets like Bactrian-X and Multi-Alpaca are enabling finetuning of Open Source LLMs. [4, 5]
- **Studying Effects of Quantisation and Rank in Finetuning Stage:** We wanted to explore how far we could go in terms of multilingual performance with PEFT techniques, and also experiment with different factors such as LoRA rank and quantisation.

Contributions

- We benchmark effects of various ranks and quantisation with LLaMA-2-7B and Mistral-7B models finetuned on MultiAlpaca and Bactrian-X-22 dataset.
- We analyse the effects of % of trainable parameters and quantisation on 6 various tasks and 40 languages.
- We study efficacy of finetuning by comparing results with non-finetuned models of similar sizes.
- We analyse the effects of multilingual PEFT on English performance to check for degradations due to forgetting.
- We experiment with the choice of instruction finetuning dataset to study any variations in model performance on our downstream tasks.

Experiments

- **Models:** Mistral, LLaMA-2
- **Datasets:** MultiAlpaca, Bactrian-X-11, Bactrian-X-22
- **Ranks:** 8, 16, 32, 64, 128
- **Quantisations:** 4, 8, 16
- **Evaluation tasks:** XNLI, XCOPA, XQUAD, MLQA, Belebele, XLSUM, Alpaca Eval

Detailed Task Wise Performance

model	finetuning dataset	xnli	xcopa	xquad	belebele	mlqa	xlsum	Model Average
GPT-4	NA	0.75	0.90	0.69	0.85	0.67	**0.25**	0.69
Mistral-7B-Instruct	NA	0.38	0.53	0.23	0.44	0.24	NA	0.37
Llama-2-70b-chat	NA	0.48	0.39	0.07	0.61	0.24	0.08	0.31
PaLM2	NA	**0.76**	**0.96**	0.70	**0.87**	0.39	0.07	0.62
Llama-2-7b	MultiAlpaca	0.35	0.58	0.64	0.28	0.41	0.10	0.39
	Bactrian-X-22	0.35	0.58	0.63	0.28	0.44	0.08	0.39
	Bactrian-X-11	0.35	0.59	0.63	0.28	0.44	0.07	0.39
	alpaca	0.35	0.58	0.63	0.28	0.35	0.07	0.38
Mistral-7b	MultiAlpaca	0.53	0.59	**0.79**	0.43	**0.70**	0.14	0.53
	Bactrian-X-22	0.52	0.59	**0.79**	0.42	**0.70**	0.14	0.53
	Bactrian-X-11	0.53	0.60	**0.79**	0.42	**0.70**	0.10	0.52
	alpaca	0.53	0.59	0.78	0.45	**0.70**	0.10	0.52

Table 1. Detailed Task Wise Performance Comparison between GPT-4, PaLM-2, LLaMA-70B-chat, Mistral-7B-Instruct and finetuned models with best rank quantisation. Baseline numbers are referred from [1].

Key Takeways

- Crosslingual transfer DOES happen even in parameter efficient finetuning. Alpaca is comparable to MultiAlpaca and Bactrian-X-22 in multilingual downstream task performance.
- Having more languages in the finetuning datasets does not necessarily mean significantly better multilingual performance if the dataset sizes are comparable.
- There are no significant differences on the downstream tasks when the models are finetuned on translated or LLM generated training datasets.
- Quality and abilities of the base model far outweigh the dataset or training method for parameter efficient multilingual instruction finetuning.
- Higher capacity adapters (i.e. higher ranks or better quantisations) are better at maintaining English performance along with multilingual downstream task performance.

Belebele Rank And Quantisation Analysis

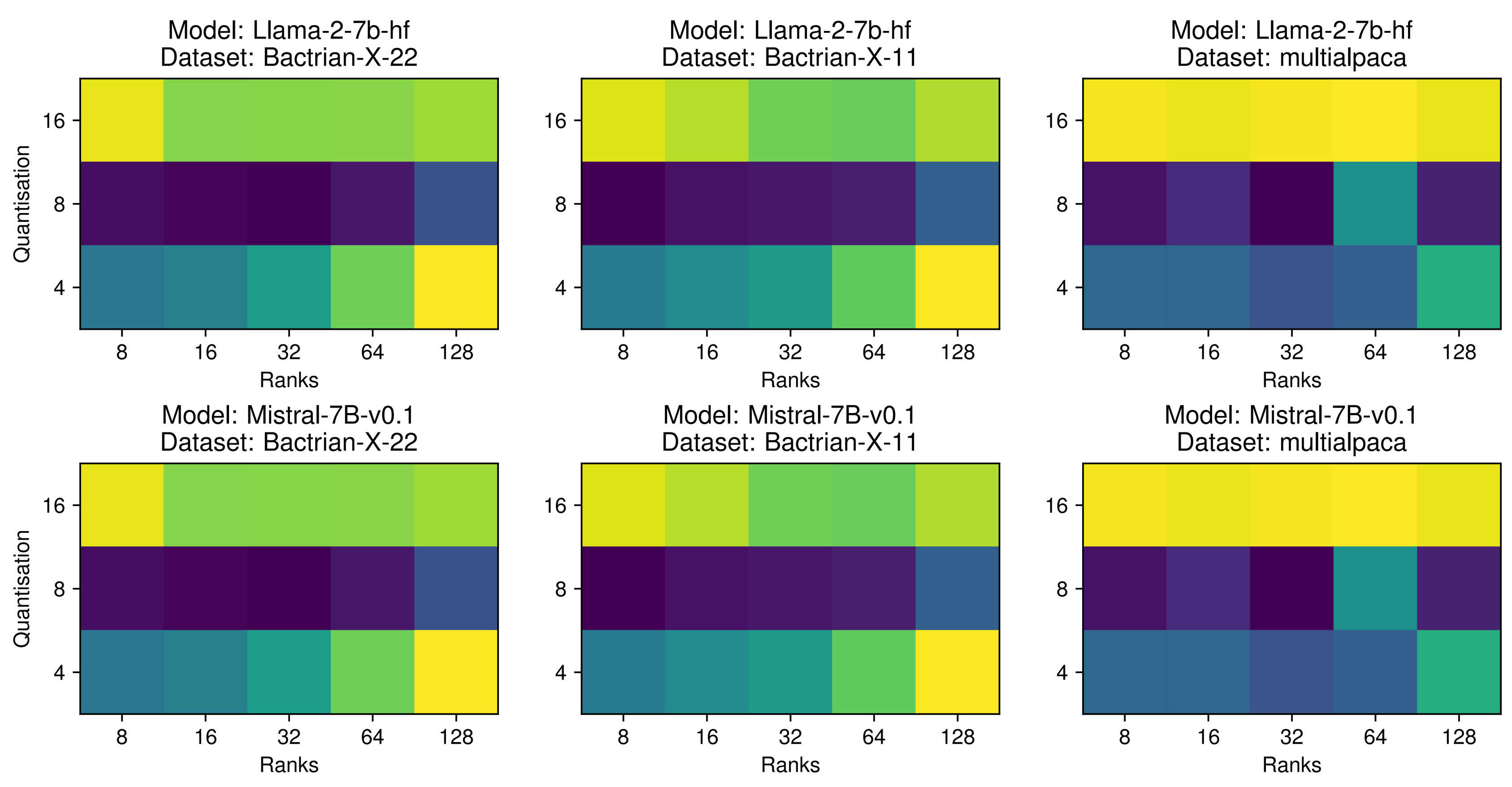

Figure 1. Average model performance of LLaMA-2-7B and Mistral-7B finetuned on Bactrian-X-22, Bactrian-X-11 and MultiAlpaca across tasks on all rank-quantisation configurations.

Task Wise Best Model

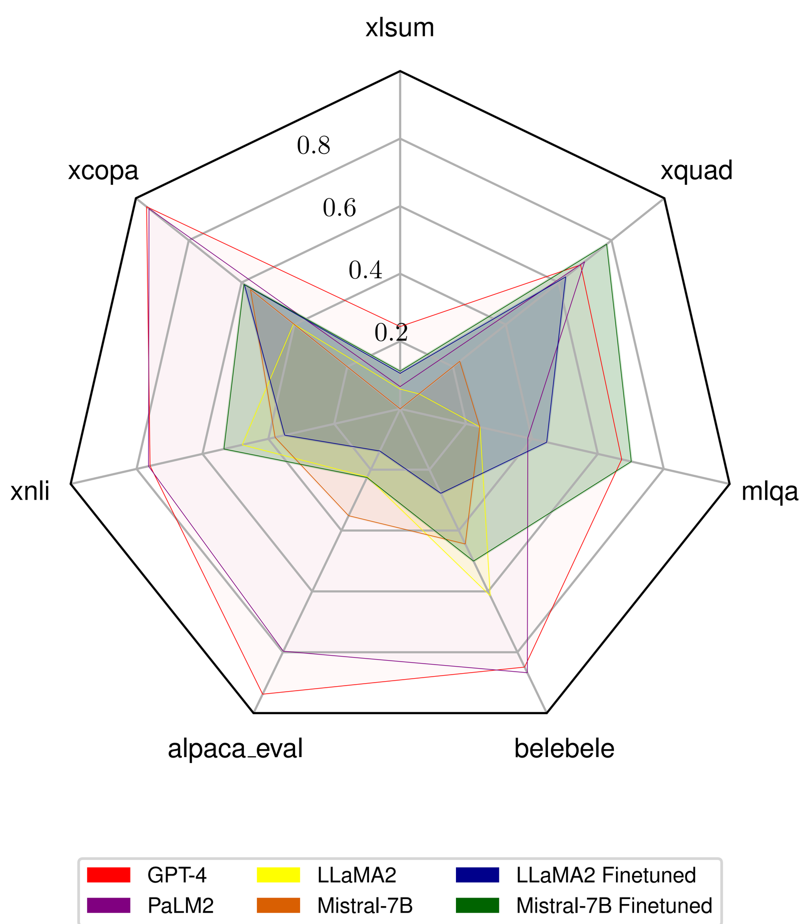

Figure 2. **Comparison of best parameter efficient instruction finetuned models with other off the shelf LLMs.** Notably, the best Mistral instruction finetuned model is able to outperform even GPT-4 and PaLM2 on "MLQA" and "XQUAD" tasks.

Effects of Number of Languages in Traning Data

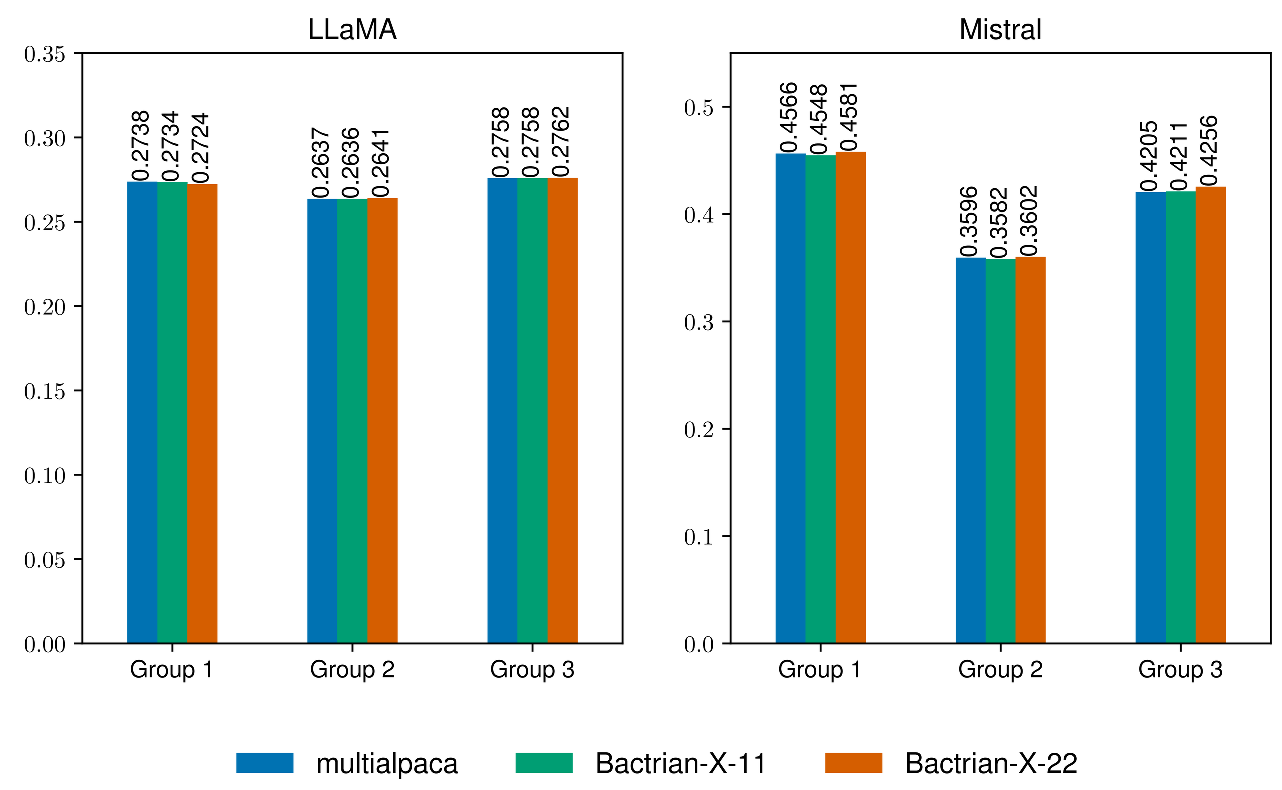

Figure 3. **Effect of diversity of languages in fine-tuning on downstream task (belebele).** Here Group 1 is the set of 11 languages from MultiAlpaca, Group 2 is the set of 11 languages in Bactrian-X-22 but not in MultiAlpaca and Group 3 contains 13 languages present in neither. We find that both models trained on either datasets perform very similar to each other across all 3 groups.

Alpaca Eval Scores

model				winrate
GPT-4				**93.78**
PaLM2				79.66
Llama-70B-Chat				22.36
Mistral-7B-Instruct				35.12
model	**dataset**	**rank**	**quantisation**	**winrate**
Llama-2-7B	Alpaca	128	16	13.28
	Bactrian-X-22	64	16	13.73
	Bactrian-X-11	16	16	**13.83**
	MultiAlpaca	128	16	**13.73**
Mistral-7B	Alpaca	64	8	**24.47**
	Bactrian-X-22	16	8	22.07
	Bactrian-X-11	128	16	22.57
	MultiAlpaca	32	8	22.45

Table 2. Best AlpacaEval Scores for each model, dataset, rank and quantisation configuration and GPT-4, PaLM-2, LLaMA-70B-Chat and Mistral-7B-Instruct baselines.

Language Wise XQUAD Analysis

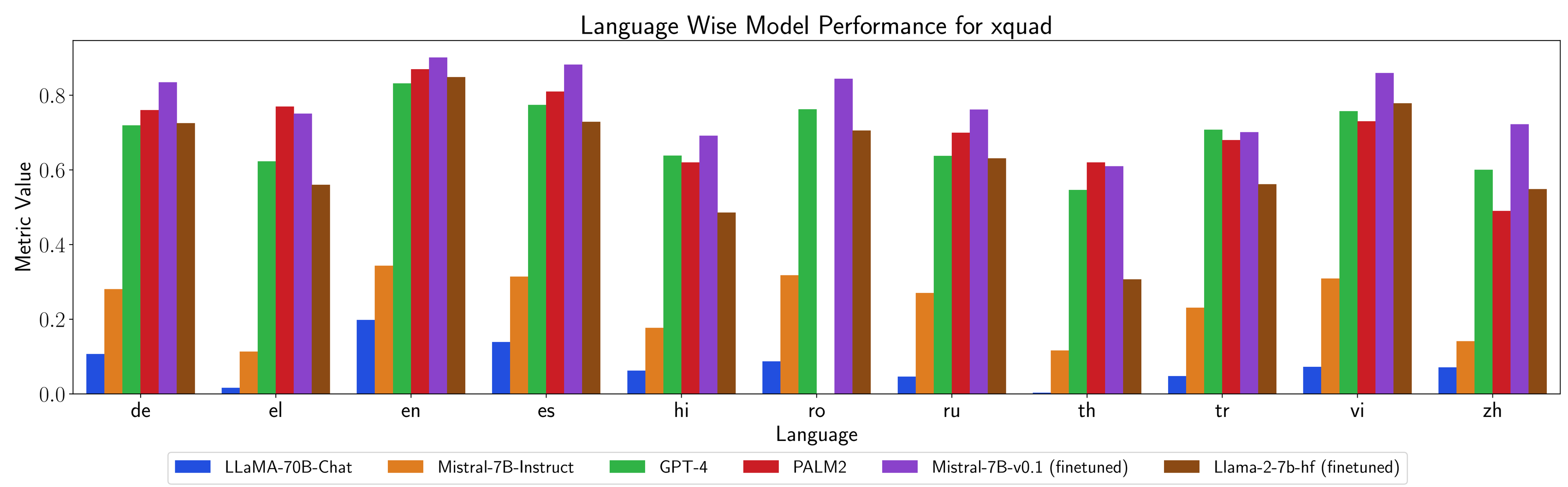

Figure 4. Detailed language-wise comparison of our finetuned and models with other baselines [1] on Arabic, German, Greek, English, Spanish, Hindi, Romanian, Russian, Thai, Turkish and Vietnamese for XQUAD.

References

[1] Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Axmed, M., Bali, K., and Sitaram, S. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks, 2023.

[2] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms, 2023.

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

[4] Li, H., Koto, F., Wu, M., Aji, A. F., and Baldwin, T. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation, 2023.

[5] Wei, X., Wei, H., Lin, H., Li, T., Zhang, P., Ren, X., Li, M., Wan, Y., Cao, Z., Xie, B., Hu, T., Li, S., Hui, B., Yu, B., Liu, D., Yang, B., Huang, F., and Xie, J. Polylm: An open source polyglot large language model, 2023.