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

Anonymous ACL submission

Abstract

Parameter Efficient Finetuning (PEFT) has emerged as a viable solution for improving the performance of Large Language Models (LLMs) without requiring massive resources and compute. Prior work on multilingual evaluation has shown that there is a large gap between the performance of LLMs on English and other languages. Further, there is also a large gap between the performance of smaller open-source models and larger LLMs. Finetuning can be an effective way to bridge this gap and make language models more equitable. In this work, we finetune the LLAMA-2-7B and MISTRAL-7B models on two synthetic multilingual instruction tuning datasets to determine its effect on model performance on six downstream tasks covering forty languages in all. Additionally, we experiment with various parameters, such as rank for low-rank adaptation and values of quantisation to determine their effects on downstream performance and find that higher rank and higher quantisation values benefit low-resource languages. We find that PEFT of smaller open-source models sometimes bridges the gap between the performance of these models and the larger ones, however, English performance can take a hit. We also find that finetuning sometimes improves performance on low-resource languages, while degrading performance on high-resource languages.

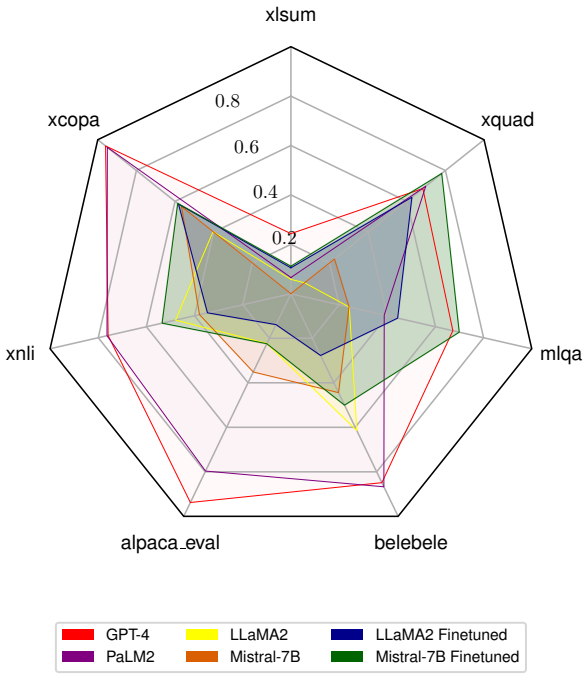

Figure 1: **Comparison of best parameter efficient instruction finetuned models with other off the shelf LLMs.** Notably, the best Mistral instruction finetuned model is able to outperform even GPT-4 and PaLM2 on "MLQA" and "XQUAD" tasks.

1 Introduction

Large Language Models (LLMs) show impressive performance on several tasks, sometimes even surpassing human performance. This has been attributed to the vast amounts of training data used during the pretraining phase, as well as various techniques used to align the models during the finetuning phase. Several variants of finetuning exist, including supervised finetuning (SFT) and instruction tuning (OpenAI, 2023), where the model is finetuned with task specific data, or instructions on how to perform tasks. However, finetuning all the parameters of the model can be expensive and time consuming, due to the increasing size of language models in recent years. Parameter Efficient Finetuning (PEFT) has emerged as a viable alternative to full finetuning (Chen et al., 2023).

Most studies on LLMs focus on training, finetuning and evaluating models in performing tasks in English. Recent work on comprehensive evaluation of LLM capabilities in non-English settings (Ahuja et al., 2023a) have shown that LLMs perform far worse on languages other than English. Studies that compare multilingual performance across different models (Ahuja et al., 2023b), show that there is a large performance gap between large, proprietary

and closed models such as GPT-4 and PaLM2 and smaller open-source models like LLAMA-2-7B and MISTRAL-7B.

PEFT techniques like LoRA (Hu et al., 2022) have been shown to strengthen the multilingual capabilities of these open-source LLMs (Zhao et al., 2024). Moreover, Adapters (Pfeiffer et al., 2020a) have been proposed to boost the capabilities of language models to newer languages. Since full finetuning of models is not always feasible due to resource and compute constraints, exploring how far PEFT techniques can take us in boosting performance on non-English languages is a promising direction. Adoption of model quantisation techniques (Dettmers et al., 2022; yang Liu et al., 2023) has also made PEFT LLMs more accessible.

Not much work has been done on analyzing the impact of different choices, configurations and settings for PEFT on multilingual downstream tasks. In this work we aim to analyse how LoRA rank and quantisation affects the performance of finetuned models across 6 downstream tasks, covering 40 languages in all. We are interested in knowing whether multilingual PEFT can lead to reasonable gains in performance, or whether full finetuning of models is required. Furthermore, we check if it is better to generate multilingual instructions from larger base models or translate existing english instruction to more languages. Lastly, we also study the effect of multilingual finetuning on English performance.

Prior works have demonstrated that LoRA finetuning is the most effective PEFT out of existing techniques so far (Zhuo et al., 2024). Hence, we keep our study limited to analysing different LoRA and QLoRA (Dettmers et al., 2023) configurations to answer our research questions. Our contributions are as follows:

- 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 or larger 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.
- We present results and an analysis of trends across these models and instruction finetuning datasets with directions for future research.

2 Related Work

Parameter Efficient Finetuning: Recently, Parameter Efficient Finetuning has gained significant attention in the NLP research community since full finetuning of LLMs is prohibitively expensive for most organizations. Following early works on adapters (Houlsby et al., 2019; Pfeiffer et al., 2020a), several finetuning techniques like LoRA (Hu et al., 2022), $(IA)^3$ (Liu et al., 2022a), P-Tuning (Liu et al., 2022b) and Prefix Tuning (Li and Liang, 2021) have been proposed. These techniques make the compute costs manageable by significantly reduce the number of trainable parameters during finetuning. Several works have used these techniques for efficient cross lingual transfer (Ansell et al., 2022), to tackle catastrophic forgetting (Vu et al., 2022) or compose multiple adapters (Pfeiffer et al., 2021) for multi-task performance.

Multilingual Instruction Finetuning: Recently, there has been a lot of interest in creating multilingual instruction finetuning datasets to enhance the reasoning capabilities of LLMs on languages other than English (Li et al., 2023a; Wei et al., 2023). Such datasets, are being used to create LLMs that can serve to larger demographic and can also be more efficient during inference time (Jiang et al., 2023). Li et al. (2023a) gained significant performance on multilingual tasks by LoRA finetuning LLaMA and BLOOM on the BACTRIAN-X dataset. These instruction datasets are generally derived by generation or translation. Wei et al. (2023) translated seed tasks of ALPACA dataset (Taori et al., 2023) to 11 languages and then prompted GPT-3.5-Turbo to generated more instructions in those languages, while Li et al. (2023a) translated ALPACA instructions to 50 other languages using google translate API and generated responses using GPT-3.5-Turbo. Moreover, efforts like (Singh et al., 2024) attempts to create crowdsourced multilinugal instruction datasets to capture better linguistic and cultural nuances.

Quantisation for Model Compression: Model quantisation is another way of reducing the overall memory footprint of the LLM. While many popular LLMs (notably LLAMA-2-7B (Touvron et al., 2023) and MISTRAL-7B (Jiang et al., 2023)) are pre-trained with weights represented in 16 bit floating point numbers (Wu et al., 2020), it is shown that finetuning with lower quantisation yields similar performance. The most popular quantisation techniques – LLM:Int8() (Dettmers et al., 2022) and 4 bit (yang Liu et al., 2023) are usually combined with LoRA (Dettmers et al., 2023) to further reduce the memory footprint of LLM finetuning.

LLM Evaluation: Principled LLM evaluation has gained significant interest with demonstrations of increasingly complex abilities of LLMs (Brown et al., 2020; Cobbe et al., 2021; Wei et al., 2022; Shi et al., 2023) on various tasks. However, many evaluations are monolingual or English-only and multilingual evaluation of LLMs (Ahuja et al., 2023a; Asai et al., 2023; Ahuja et al., 2023b) remains a challenging problem. Past work by Ramesh et al. (2023) has evaluated the effects of model compression techniques such as quantisation, distillation and pruning on LLMs performance on downstream tasks in multilingual setting.

3 Experiments

3.1 Setup

Finetuning Models: We finetune open-source, multilingual LLMs on multilingual instruction finetuning datasets. We pick models that are pretrained on multilingual data as it would be unfair to compare English-only LLMs when finetuning on multilingual data. Specifically, we explore PEFT on LLAMA-2-7B (Touvron et al., 2023) and MISTRAL-7B (Jiang et al., 2023) models.

Finetuning Dataset: We finetuned our models on MULTIALPACA (Wei et al., 2023) and BACTRIAN-X (Li et al., 2023a) datasets for all our experiments.

MULTIALPACA is a self instruct dataset which follows the same approach as (English-only) ALPACA dataset (Taori et al., 2023) by translating seed tasks to 11 languages and then using GPT-3.5-Turbo for response collection. The languages included in the dataset are Arabic, German, Spanish, French, Indonesian, Japanese, Korean, Portuguese, Russian, Thai and Vietnamese.

BACTRIAN-X is a machine translated dataset of the original alpaca-52k and dolly-15k (Conover et al., 2023) datasets. In this dataset, the instructions were translated using google translate API to 52 diverse languages and responses were generated using GPT-3.5-Turbo. In our experiments we finetune our models on a subset of 22 languages namely, Afrikaans, Arabic, Bengali, Chinese-Simplified, Dutch, French, German, Gujarati, Hindi, Indonesian, Japanese, Korean, Marathi, Portuguese, Russian, Spanish, Swahili, Tamil, Telugu, Thai, Urdu and Vietnamese. We rename this dataset to BACTRIAN-X-22. Each language in BACTRIAN-X-22 consists of 67k instructions parallel and responses whereas MULTIALPACA consists of nearly 100k instructions. To get a dataset of comparable size, we take 10% instructions (6700) from each language giving us close to 150k instructions in 22 languages. We first shuffle and partition indices 0-67k and divide them into 10 partitions. Each partition now consists of random indices from 0-67k. Then we iterate over all 22 languages assigning language i to partition $i \bmod 10$. This gives us a partition of 6700 indices from each languages which we use to form the instruction tuning dataset from each language. This means that every instruction at index from 0-67k is included in at least two languages. We study whether this enhanced sampling ensuring at least two languages per instruction helps in cross-lingual transfer.

Finetuning Techniques: We follow the LoRA (Dettmers et al., 2023; Hu et al., 2022) finetuning recipe for each finetuning run. We finetune models on various ranks and quantisations, specifically LoRA Ranks 8, 16, 32, 64 and 128 and 4bit, 8bit and 16bit quantisation.

3.2 Evaluation

We evaluate multilingual capabilities of our finetuned models on three Classification tasks, two Question-Answering tasks and one Summarisation task. We use prompts that are similar to those used in the MEGA benchmarking study (Ahuja et al., 2023a) but adapted to the Alpaca-style (Taori et al., 2023) instruction format. We study the impact of multilingual finetuning on English capabilities using Alpaca Eval (Li et al., 2023b). We use lm-eval-harness (Gao et al., 2021) for the evaluations. LM-Evaluation-Harness is a unified framework for few shot evaluation of language models.

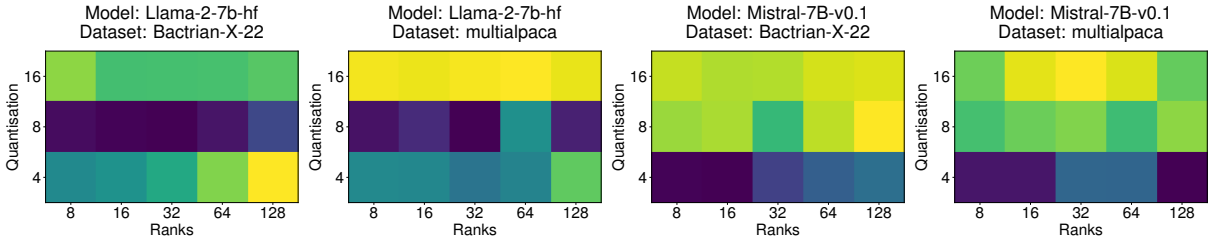

Figure 2: Average model performance of LLAMA-2-7B and MISTRAL-7B finetuned on BACTRIAN-X-22 and MULTIALPACA across tasks on all rank-quantisation configurations.

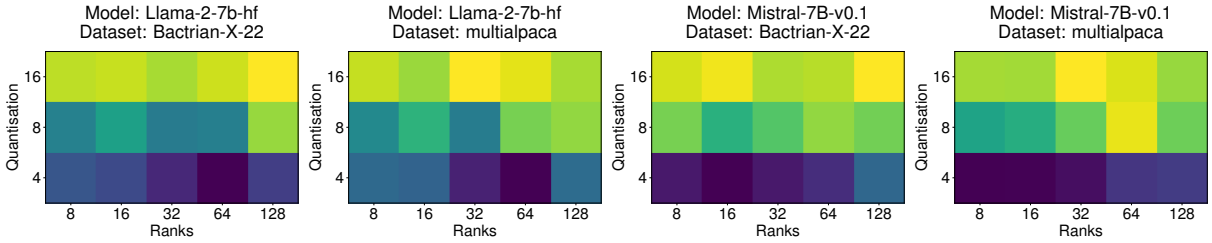

Figure 3: Belebele evaluation results of LLAMA-2-7B and MISTRAL-7B finetuned on BACTRIAN-X-22 and MULTIALPACA across tasks on all rank-quantisation configurations.

This framework standardises the inference and few shot example selection pipeline across tasks and models. We created the task configurations from MEGAVERSE (Ahuja et al., 2023b) with the ALPACA-style prompt template. We mention these prompts in detail in the Appendix Section C.

Classification Datasets: As part of our evaluation process, we benchmark our finetuned models on several datasets. The **Belebele** dataset (Bandarkar et al., 2023), which is parallel across 122 languages, is evaluated on a subset of 33 languages. We report our results on 30% of the test split in the zero-shot setting due to resource constraints. We report the results in Table 4, 5 6, 7, 8 and 9. The **XNLI** dataset (Conneau et al., 2018) consists of 122k training, 2490 validation, and 5010 test examples in 15 languages. We have evaluated our models on 1000 examples from test split with 4 in-context examples sampled from the validation split and report our results in Table 28, 29, 30, 31, 32 and 33. The **XCOPA** dataset (Ponti et al., 2020) covers 11 languages, and we evaluate our models on Estonian, Indonesian, Italian, Quechua, Thai and Vietnamese in the 4-shot setting similar to XNLI. We report our results in Table 22, 23, 24, 25, 26 and 27.

Question Answering Datasets: The **MLQA** dataset (Lewis et al., 2020) contains 5K extractive question-answering instances in 7 languages. For the interest of time, we evaluate our models for 1000 examples of the test split in a 4-shot setting and report our results in Table 16, 17, 18, 19, 20 and 21. The **XQuAD** dataset (Artetxe et al., 2020) consists of a subset of 240 paragraphs and 1190 question-answer pairs across 11 languages. We use a 4-shot setting similar to MLQA and evaluate 1000 examples of the test split. We report our results in Table 34, 35, 36, 37, 38 and 39.

Summarisation Dataset: The **XLSUM** dataset (Hasan et al., 2021) spans 45 languages, and we evaluate our models in Arabic, Chinese-Simplified, English, French, Hindi, Japanese and Spanish. We evaluate our models on 100 text-summarization pairs from the test split in a zero-shot setting and report our results in Table 40, 42, 41, 43, 44 and 45.

English Instruction Following Dataset: We also use **AlpacaEval** (Li et al., 2023b) to benchmark English proficiency. We evaluate our models against text-davinci-003 responses on 800 instructions and use GPT4 (gpt-4-32k) as the evaluator. We report our results in Tables 10, 11, 12, 13, 14 and 15.

We discuss more about these benchmark datasets in detail in Appendix Section B.

4 Analysis of Results

Analysis of Rank and Quantisation In this study we aim to analyse the trade-offs between

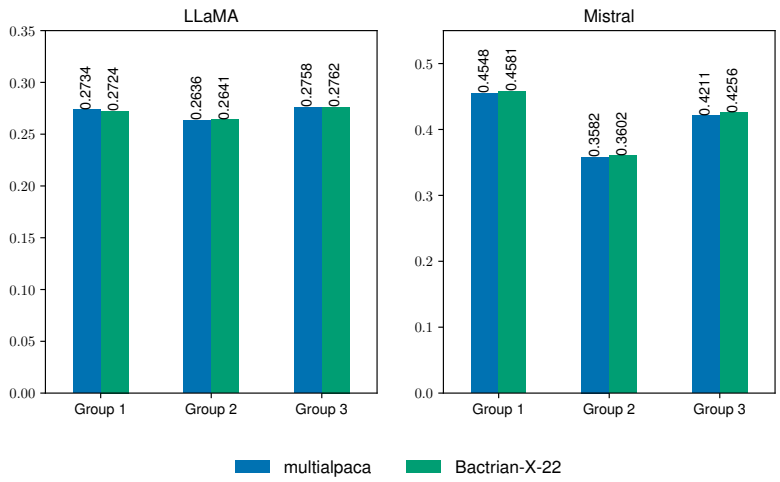

Figure 4: **Effect of diversity of languages in fine-tuning on downstream task (belebele).** Here Group 1 is the set of 11 languages from MULTIALPACA, Group 2 is the set of 11 languages in BACTRIAN-X-22 but not in MULTIALPACA and Group 3 contains 13 languages present in neither. We find that both models trained on either datasets perform very similar to each other across all 3 groups. Additional details in Tables 4 to 9.

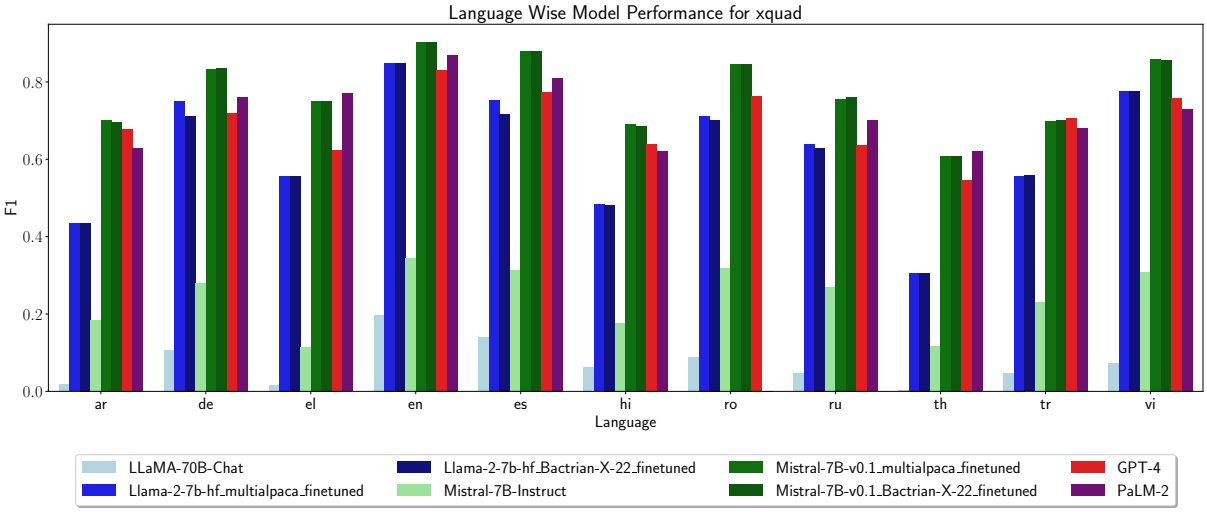

Figure 5: Detailed language-wise comparison of our finetuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Arabic, German, Greek, English, Spanish, Hindi, Romanian, Russian, Thai, Turkish and Vietnamese for XQUAD (Artetxe et al., 2020).

cost of compute and model performance. Both the LLAMA-2-7B and MISTRAL-7B models were finetuned on all rank-quantisation configurations using the MULTIALPACA and BACTRIAN-X-22 datasets resulting into 60 models.

We evaluate our finetuned models across the six benchmarking datasets mentioned in Section 3.2. We present the averaged results across these datasets in Fig 2. Lighter colours (yellow) indicate higher performance and it decreases with darker shades (blue). For MISTRAL-7B we can see a clear trend for both the finetuning datasets. Decreasing the quantisation can lead to a hit in model performance. For LLAMA-2-7B the trend is not very clear but the highest quantisation gives the best results. Additionally, higher ranks seem to give slightly better performance. According to our studies, using Rank 32 or 64 with 16bit Quantisation works the best on average. This can be inferred very clearly from our results on Belebele in Fig 3. To delve deeper, we provide a detailed task-wise performance in Fig 14.

MULTIALPACA v/s BACTRIAN-X-22 as Instruction Finetuning Dataset Here, we aim to study the model performance finetuned on multilingual instruction dataset created in 2 different settings i.e. LLM generated (MULTIALPACA) and machine-translated (BACTRIAN-X-22). In MULTIALPACA, both multilingual instructions and their responses are generated using GPT-3.5-Turbo from translated ALPACA seed instructions. While in BACTRIAN-X-22, the final set of ALPACA instructions were translated and then responses were collected using GPT-3.5-Turbo.

In our findings, we observe that models trained

on both datasets give similar performance on average across tasks and languages, implying that method of instruction finetuning data creation has little to no effect on the model performance. Rather, we observe that multilingual capabilities of base model is a good indicator of the finetuned model performance across tasks. From Ahuja et al. (2023b) we know that MISTRAL-7B is a better base multilingual model than LLAMA-2-7B and we observe that multilingual capabilities of MISTRAL-7B also reaps greater benefits of multilingual instruction finetuning.

Hence, in our experiments, we observe that for creating multilingual instruction datasets both approaches are equivalent - generating multilingual data from seed tasks or translating an existing English instruction dataset to more languages.

Effect of Number of Languages in Training Data Our finetuning datasets MULTIALPACA and BACTRIAN-X-22 have 11 and 22 languages respectively. We want to study if the additional 11 languages in BACTRIAN-X-22 help the model perform better at multilingual tasks.

We do a case study on the "belebele" task which consists of 50+ languages. We sample 35 of these and divide them into 3 groups. The first group (Group 1) consists of 11 languages from MULTIALPACA, the next group (Group 2) consists of 11 languages present in BACTRIAN-X-22 but absent from MULTIALPACA. The final group (Group 3) contains 13 languages that are not present in any of our finetuning datasets. We compute average accuracy of finetuned models across all ranks on the 3 groups and present them in Figure 4. We find that the larger number of languages in BACTRIAN-X-22 do not necessarily help. This behavior is consistent with the findings of Shaham et al. (2024). Moreover,we observe that for other tasks as well BACTRIAN-X-22 dataset has no edge over MULTIALPACA as we can see from Table 1 and Figure 6.

Effect of Multilingual Finetuning on performance on downstream tasks for High Resource v/s Low Resource Languages For XNLI and MLQA, we observe that finetuning improves performance on low-resource languages but worsens performance on high-resource languages. In Belebele, we find that finetuning worsens the performance for all languages for both models. For XCOPA, we get the same or better results for all languages with finetuning. For XQUAD, multilingual finetuning boosts performance for all languages for both LLAMA-2-7B and MISTRAL-7B and even surpasses GPT-4 as seen in Fig 5. Moreover, for XLSUM parameter efficient finetuned models does not perform at par with fully finetuned or larger LLMs due to the generative nature of the task. This shows that overall, PEFT on smaller LLMs using multilingual instruction data can prove to be beneficial and can bridge the gap between smaller open-source models and large proprietary models.

Moreover, we observe that language-wise performance does not get affected by the choice of the training dataset, i.e. MULTIALPACA and BACTRIAN-X-22 have similar performances across languages and tasks. BACTRIAN-X-22 sometimes, leads to better performance on some low-resources languages as it includes more languages in the training data.

Analysis of Performance on English In Table 2 we compare models on the English AlpacaEval benchmark. Overall, the capability of model to follow English instruction reduces drastically after multilingual finetuning. In general, finetuning on English-only ALPACA dataset or using higher capacity adapters (higher rank, better quantisation) seem to help in preserving the performance on English instruction finetuning.

More concisely, we observe that MISTRAL-7B is able to preserve more English capabilities than LLAMA-2-7B on AlpacaEval. While, BACTRIAN-X-22 and MULTIALPACAhave difference in performance in English for the respective finetuned model.

Furthermore, in a task-wise analysis we observe that multilingual finetuning leads to deterioration in English performance in Belebele, while in XNLI it deteriorates for LLAMA-2-7B and improves for MISTRAL-7BḞor question-answering tasks MLQA and XQUAD, multilingual finetuning leads to improvement in performance. In XLSUM the performance improves for LLAMA-2-7B when finetuned on multilingual data, while for MISTRAL-7B the performance decreases or remains the same.

Task Wise Performance Analysis We analyse the average performance across languages on each task for GPT-4, PaLM-2, LLaMA-2-70B-Chat, Mistral-7B-Instruct and LLAMA-2-7B and MISTRAL-7B models finetuned using MULTIALPACA, ALPACA and BACTRIAN-X-22. In Figure 6

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
	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
	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 Ahuja et al. (2023b).

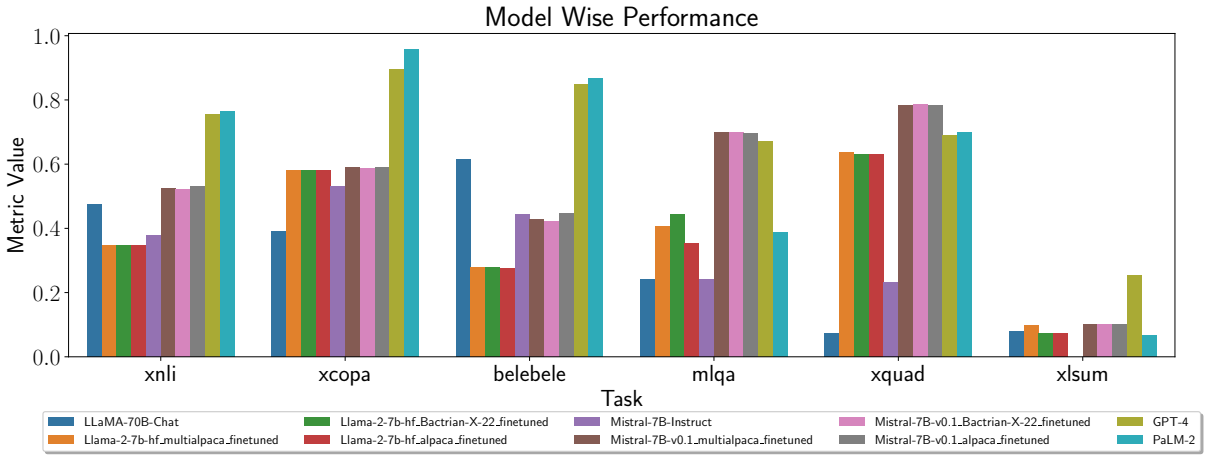

Figure 6: Task Wise Performance Comparison of Llama-2-70B-Chat, GPT-4, PaLM-2, Mistral-7B-Chat and our finetuned models averaged across languages.

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**
	MULTIALPACA	128	16	**13.73**
Mistral-7B	Alpaca	64	8	**24.47**
	Bactrian-X-22	16	8	22.07
	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.

we can see that finetuning is usually better or at par with LLaMA-70B-Chat and Mistral-7B-Instruct which are not finetuned on multilingual data. We also compare the performance of finetuned models with LLMs like GPT-4 and PaLM2 and English instruction tuned versions of these models provided in (Ahuja et al., 2023b). Table 1 shows the best model score averaged across languages per task. We observe that MISTRAL-7B beats GPT-4 and fairs as the best model on XQUAD and MLQA. While in XNLI, XCOPA and Belebele, it bridges the gap between GPT-4 and PaLM-2 by 20% on an average.

While we see some gap being bridged (20% on average) on classification tasks (XNLI, XCOPA and Belebele) using PEFT, it tends to beat larger models on question answering tasks (MLQA and XQUAD) like GPT-4. While on XLSUM there is no significant difference in performance after PEFT.

Interestingly, we observe that finetuning LLAMA-2-7B and MISTRAL-7B on ALPACA leads to comparable results with finetuning on MULTIALPACA and BACTRIAN-X-22. This can be due to the parameter efficient nature of the finetuning which prevents catastrophic forgetting and helps the model learn the instruction following ability from the English instruction data. Second

reason can be the difference in the token fertility of LLAMA-2-7B and MISTRAL-7B as shown by Ahuja et al. (2023b). We can deduce that MISTRAL-7B having higher token fertility and being a better base multilingual model can benefit greatly from English instruction dataset and show excellent cross-lingual transfer. While LLAMA-2-7B having a lower token fertility does not benefit greatly from even multilingual instruction finetuning, as most multilingual LLaMAs resort to vocabulary expansion during pre-finetuning phase (Zhao et al., 2024). We illustrate the language-wise analysis of our model results in Figure 5, 15, 16, 17, 18, 20, 21, 22, 22 and 23.

While we compare our multilingual finetuned models with models finetuned on English, we should note that we do not have complete information about instruction datasets used for Llama-70B-chat and Mistral-7B-Instruct. Hence, there may be chances of data contamination for some datasets in these models (Ahuja et al., 2023a) or the presence of multilingual instruction data in them.

5 Conclusion

In this paper we perform an extensive analysis of how rank, quantisation, finetuning dataset and base LLM effects the performance of the finetuned models on 6 multilingual tasks and AlpacaEval when finetuned in a paramater efficient manner.

- Crosslingual transfer DOES happen even in parameter efficient finetuning.
- ALPACA (English-only instruction finetuning dataset) is comparable to MULTIALPACAand Bactrian-X-22 in multilingual downstream task performance. We hypothesize that this is due to Mistral being a superior model due to its better tokenizer, and that PEFT prevents catastrophic forgetting compared to full finetuning.
- Having more languages in the finetuning datasets does not necessarily mean significantly better multilingual performance (Section 4) if the dataset sizes are comparable.
- 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.

We also beat GPT-4 on question-answering tasks (MLQA and XQUAD) using just multilingual PEFT on MISTRAL-7B showing that multilingual finetuning of 7B parameter LLMs is a promising direction for the future to bridge the gap of performance on multilingual downstream tasks.

6 Future Work

More PEFT Techniques This study explores the effects of PEFT using LoRA, while newer techniques by Ansell et al. (2024) can also be promising to study the parameter efficient techniques for multilingual instruction tuning for these LLMs. We can also work towards building better PEFT techniques for specifically multilingual settings or crosslingual transfer for LLMs like MAD-X for encoder-like models (Pfeiffer et al., 2020b).

Better Use of Adapters With LLMs becoming more popular in the NLP research, it is prohibitive to re-pretrain models on newer languages. Hence, modular techniques like Pfeiffer et al. (2020a) were introduced to finetune these models in a more parameter efficient manner. Furthermore, the works of Chen et al. (2024) introduced a novel technique of adapting language models to newer languages without further pretraining. However, such techniques can be compute intensive in LLMs and a direction of future can be pursued where similar approach can be taken in a parameter efficient manner.

Better Multilingual Instruction Datasets While the datasets used in this study are derived from ALPACA, newer instruction datasets using Chain Of Thought prompting (Mukherjee et al., 2023; Mitra et al., 2023) leads to better reasoning capabilities on smaller LLMs (7 billion parameters), which can be explored as future work. More controlled crowd-sourcing efforts like Singh et al. (2024) can also lead to better multilingual instruction datasets.

7 Limitations

Our evaluation is performed using standard benchmarks, which has known limitations. Datasets used to create benchmarks may have been seen by models during pretraining or finetuning, and due to lack of transparency about the datasets used for training

we cannot rule out test data contamination. Second, we use synthetic datasets that are created by prompting LLMs to finetune our models, this can lead to bias, which is also a limitation of the work. Finally, we compare the results obtained by our models to results from the MEGAVERSE benchmarking study while comparing the differences between finetuned models and models that are not finetuned for multilingual performance, which may have some differences in prompting and setup.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023b. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Ivan Vulić, Hannah Sterz, Anna Korhonen, and Edoardo M. Ponti. 2024. Scaling sparse fine-tuning to large language models.

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.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv cs.CL 2305.14857*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. Parameter-efficient fine-tuning design spaces.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2024. Improving language plasticity via pretraining with active forgetting.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Neural Information Processing Systems (NeurIPS)*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2021. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*,

pages 4693–4703, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

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

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

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. AlpacaEval: An Automatic Evaluator of Instruction-following Models.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022b. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason.

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

OpenAI. 2023. Gpt4 technical report.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model.

Hao Wu, Patrick Judd, Xiaojie Zhang, Mikhail Isaev, and Paulius Micikevicius. 2020. Integer quantization for deep learning inference: Principles and empirical evaluation. *ArXiv*, abs/2004.09602.

Shih yang Liu, Zechun Liu, Xijie Huang, Pingcheng Dong, and Kwang-Ting Cheng. 2023. Llm-fp4: 4-bit floating-point quantized transformers.

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

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models.

A Hyperparameters for Finetuning and Training Setup

Our code for finetuning is based on the open source `axolotl`[1] framework. We plan to release our configuration files for better reproducibility. Each finetuning experiment took ∼16-24 hours to complete on a single NVIDIA A100 GPU with 80 GB RAM. Exact hyperparameters for finetuning are mentioned below:

Hyperparameter	Value
Learning rate	1×10^{-6}
Epochs	5
Global batch size	16
Scheduler	Cosine
Warmup	Linear
Warmup steps	10
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Weight decay	0

Table 3: Hyperparameters for finetuning.

B Evaluation Dataset Details

The detailed description of the datasets that we use for evaluation are as follows:

XNLI: The XNLI (Cross-lingual Natural Language Inference) dataset (Conneau et al., 2018) is an extension of the Multi-Genre NLI (MultiNLI) corpus to 15 languages. The dataset was created

[1] https://github.com/OpenAccess-AI-Collective/axolotl

by manually translating the validation and test sets of MultiNLI into each of those 15 languages. The English training set was machine translated for all languages. The dataset is composed of 122k train, 2490 validation, and 5010 test examples. XNLI provides a robust platform for evaluating cross-lingual sentence understanding methods. We evaluated our models on the test split with 4 in-context examples sampled from the validation split. We report our results in Table 28, 29, 30, 31, 32 and 33.

XCOPA: The XCOPA (Cross-lingual Choice of Plausible Alternatives) dataset (Ponti et al., 2020) is a benchmark for evaluating the ability of machine learning models to transfer commonsense reasoning across languages. It is a translation and re-annotation of the English COPA dataset (Roemmele et al., 2011) and covers 11 languages from 11 families and several areas around the globe. The dataset is challenging as it requires both the command of world knowledge and the ability to generalize to new languages. We evaluated our models on Estonian, Thai, Italian, Indonesian, Vietnamese and Southern Quechua. We evaluated our models in the 4-shot setting similar to XNLI. We report our results in Table 22, 23, 24, 25, 26 and 27.

Belebele: Belebele (Bandarkar et al., 2023) is a multiple choice machine reading comprehension (MRC) dataset parallel across 122 languages. Each question is linked to a short passage from the FLORES-200 dataset (Team et al., 2022). The human annotation procedure was carefully curated to create questions that discriminate between different levels of language comprehension. We evaluated our models in the zero-shot setting and report results in Table Table 4, 5 6, 7, 8 and 9.

MLQA: MLQA (Lewis et al., 2020) is a multilingual question answering dataset designed for cross lingual question answering. It contains 5K extractive question answering instances. It consists of 7 languages i.e. English, Arabic, Vietnamese, German, Spanish, Hindi and Simplified Chinese. The evaluation uses a 4-shot setting similar to that of XNLI. We report our results in Table 34, 35, 36, 37, 38 and 39.

XQuAD: The XQuAD (Cross-lingual Question Answering Dataset) (Artetxe et al., 2020) is a benchmark dataset for evaluating cross-lingual question answering performance. It consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1, along with their professional translations into ten languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. As a result, the dataset is entirely parallel across 11 languages. This dataset provides a robust platform for developing and evaluating models on cross-lingual question answering tasks. For evaluation, we use a 4-shot setting similar to MLQA. We report our results in table 34, 35, 36, 37, 38 and 39.

XLSUM: XLSUM (Hasan et al., 2021) is a comprehensive and diverse dataset for abstractive summarization comprising 1 million human annotated article-summary pairs from BBC. The dataset covers 44 languages ranging from low to high-resource, for many of which no public dataset is currently available. We evaluate our models on a subset of 7 languages, namely, Arabic, Chinese-Simplified, English, Hindi, French, Japanese and Spanish in a zero-shot setting. We present our results in Table 40, 42, 41, 43, 44 and 45.

AlpacaEval: AlpacaEval (Li et al., 2023b) is an LLM based automatic evaluator for instruction following models. It consists of around 800 instructions and corresponding responses obtained from (text-davinci-003) GPT3. The benchmark compares responses from GPT3 (or any other "oracle" model) with target (finetuned) model using another LLM (typically GPT4) as an evaluator. The evaluator LLM decides which response is better and overall win rate (higher the better) is computed for the target model. For our evaluation, we use the text-davinci-003 responses from the dataset as our oracle/gold responses and use GPT4 (gpt-4-32k) as our evaluator. We report our results in Table 10, 11, 12, 13, 14 and 15.

C Evaluation Prompts

For XNLI, XCOPA, Belebele, MLQA, XQUAD, XLSUM we use the standard ALPACA system prompt **"Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request."**.

```
### Instruction:
The task is to solve Natural
Language Inference (NLI) problems.
NLI is the task of determining
the inference relation between two
(short, ordered) texts: entailment,
contradiction, or neutral. Answer
as concisely as possible in the same
format as the examples below:

{{premise}}

Question: {{hypothesis}} True,
False, or Neither?

### Response
```

Figure 7: XNLI Prompt

```
### Instruction:
The task is to perform open-domain
commonsense causal reasoning. You
will be provided a premise and two
alternatives, where the task is to
select the alternative that more
plausibly has a causal relation with
the premise. Answer as concisely as
possible in the same format as the
examples below:

Given this premise:
{{premise}}

What's the best option?
-choice1 : {{choice1}}
-choice2 : {{choice2}}

We are looking for {% if question
== "cause" %} a cause {% else %} an
effect {% endif %}

### Response:
```

Figure 8: XCOPA Prompt


```
### Instruction:
The task is to perform reading
comprehension task. Given the
following passage, query, and
answer choices, output the letter
corresponding to the correct answer.

Passage: {{flores_passage}}
Query : {{question}}
Choices :
A : {{mc_answer1}}
B : {{mc_answer2}}
C : {{mc_answer3}}
D : {{mc_answer4}}

###Response :
```

Figure 9: Belebele Prompt

```
### Instruction:
The task is to solve reading
comprehension problems. You will
be provided questions on a set
of passages and you will need to
provide the answer as it appears in
the passage. The answer should be
in the same language as the question
and the passage.

Context:{{context}}
Question:{{question}}

Referring to the passage above, the
correct answer to the given question
is

### Response:
```

Figure 10: MLQA Prompt

```
### Instruction:
The task is to solve reading
comprehension problems. You will
be provided questions on a set
of passages and you will need to
provide the answer as it appears in
the passage. The answer should be
in the same language as the question
and the passage.

Context:{{context}}
Question:{{question}}

Referring to the passage above, the
correct answer to the given question
is

### Response:
```

Figure 11: XQUAD Prompt

```
### Instruction:
The task is to summarize any given
article.  You should summarize all
important information concisely
in the same language in which you
have been provided the document.
Following the examples provided
below:

{{text}}

### Response:
```

Figure 12: XLSUM Prompt

```
Below is an instruction that
describes a task, paired with an
input that provides further context.
Write a response that appropriately
completes the request.
```

Figure 13: Alpaca Prompt

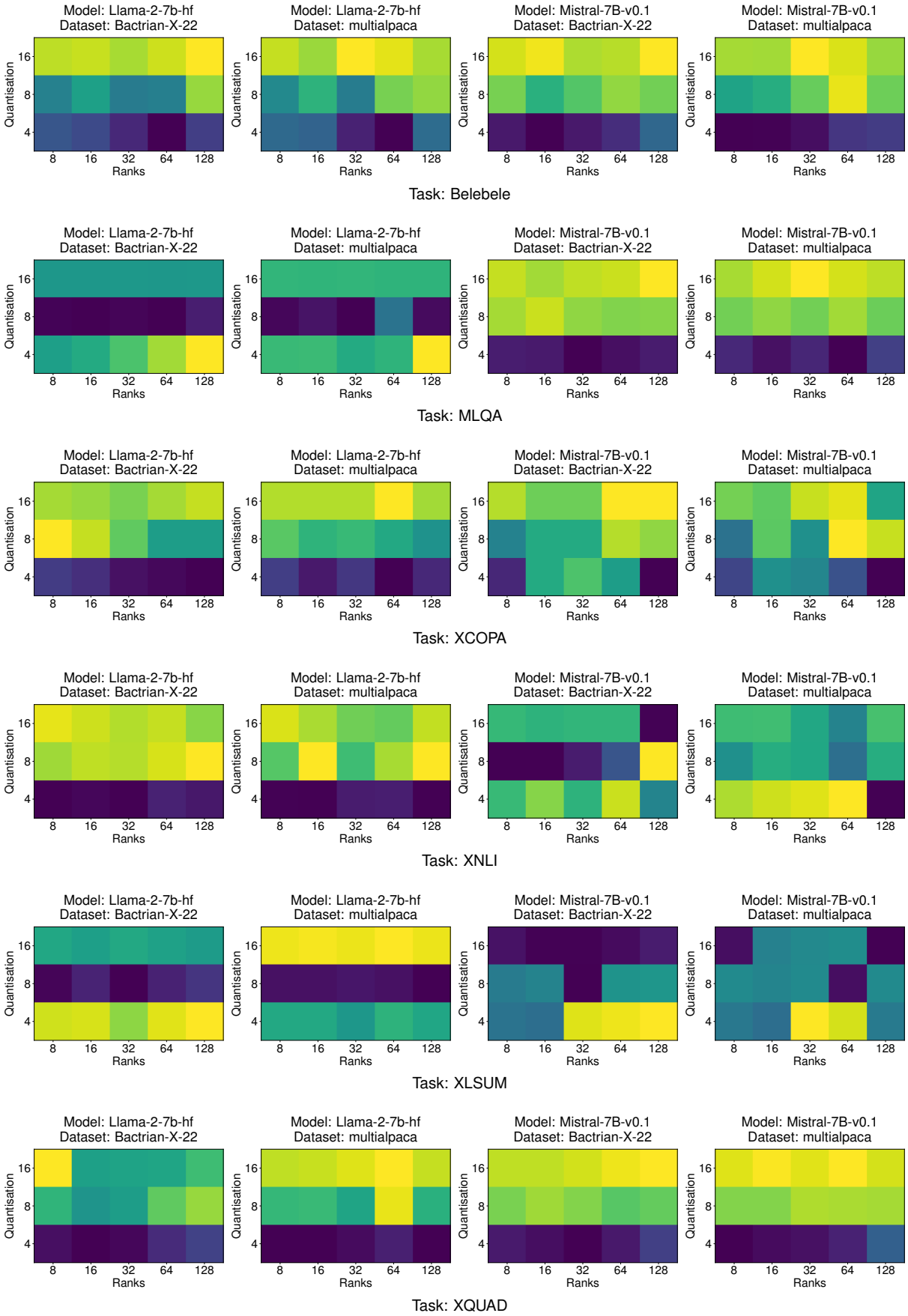

Figure 14: Task-wise performance of MISTRAL-7B and LLAMA-2-7B fine-tuned on BACTRIAN-X-22 and MULTIALPACA averaged across languages on all rank-quantisation configurations.

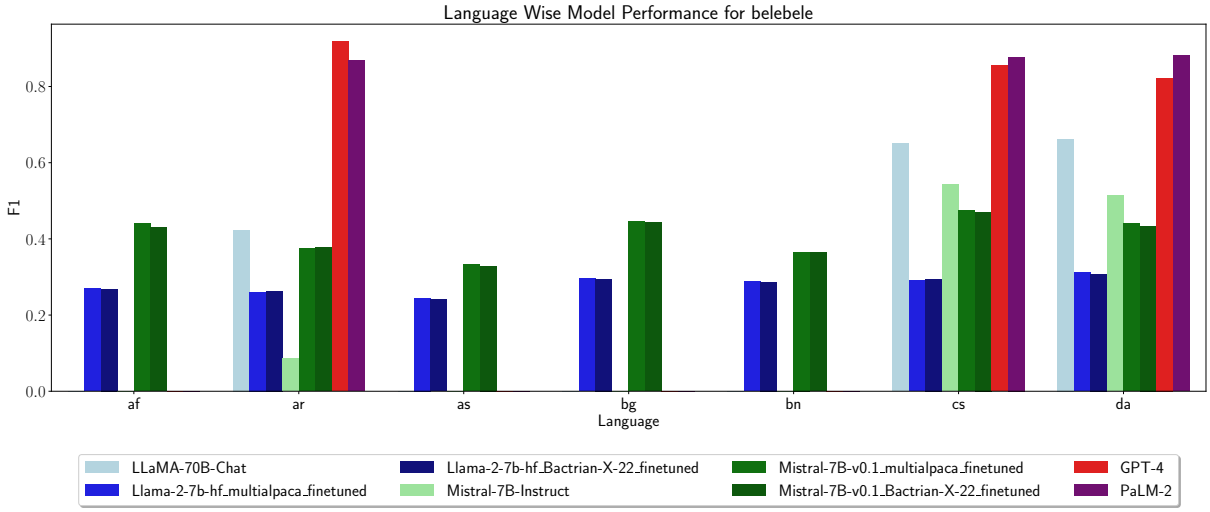

Figure 15: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Afrikaans, Arabic, Assamese, Bulgarian, Bengali, Czech and Danish for Belebele (Bandarkar et al., 2023).

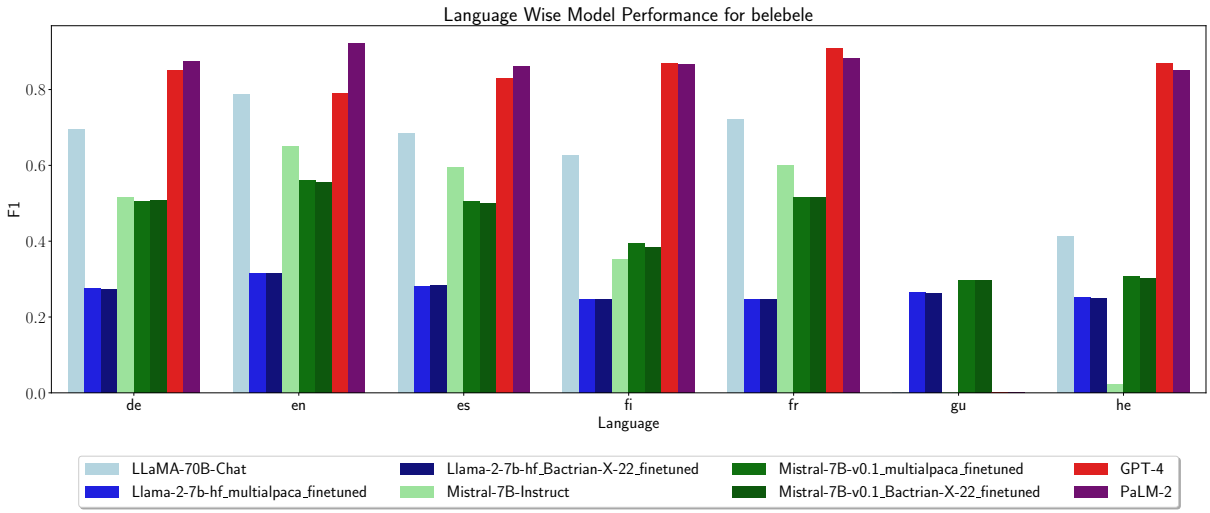

Figure 16: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on German, English, Spanish, Finnish, French, Gujarati and Hebrew for Belebele (Bandarkar et al., 2023).

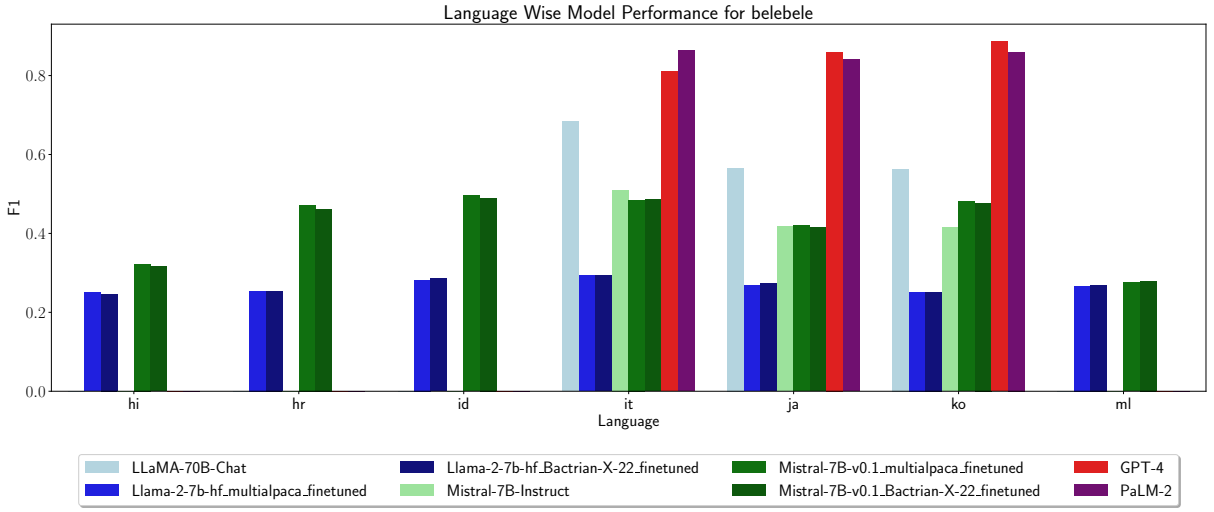

Figure 17: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Hindi, Croatian, Indonesian, Italian, Japanese, Korean and Malayalam for Belebele (Bandarkar et al., 2023).

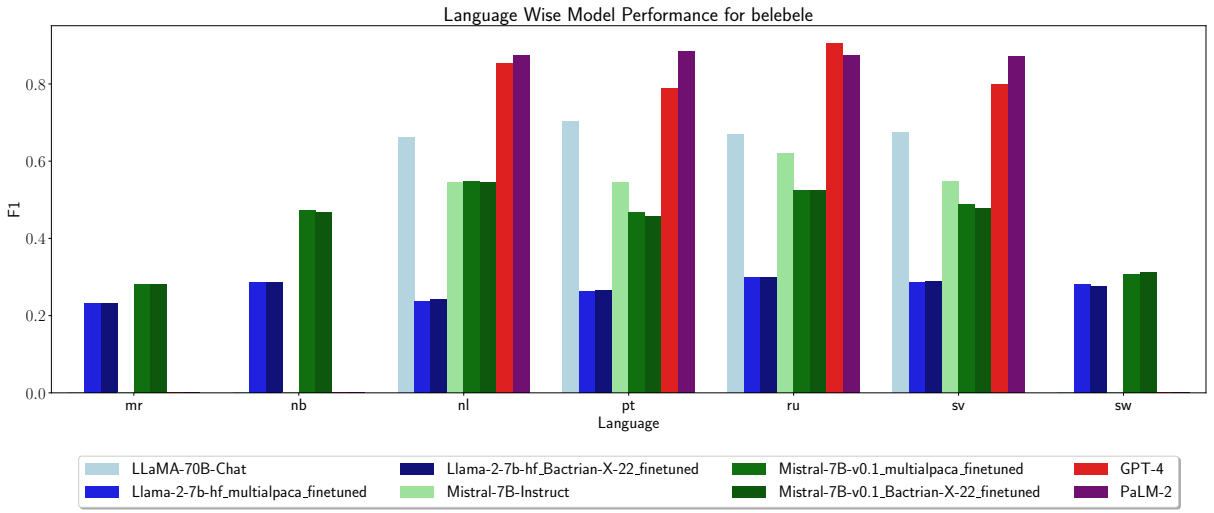

Figure 18: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Marathi, Norwegian, Dutch, Portuguese, Russian, Swedish and Swahili for Belebele (Bandarkar et al., 2023).

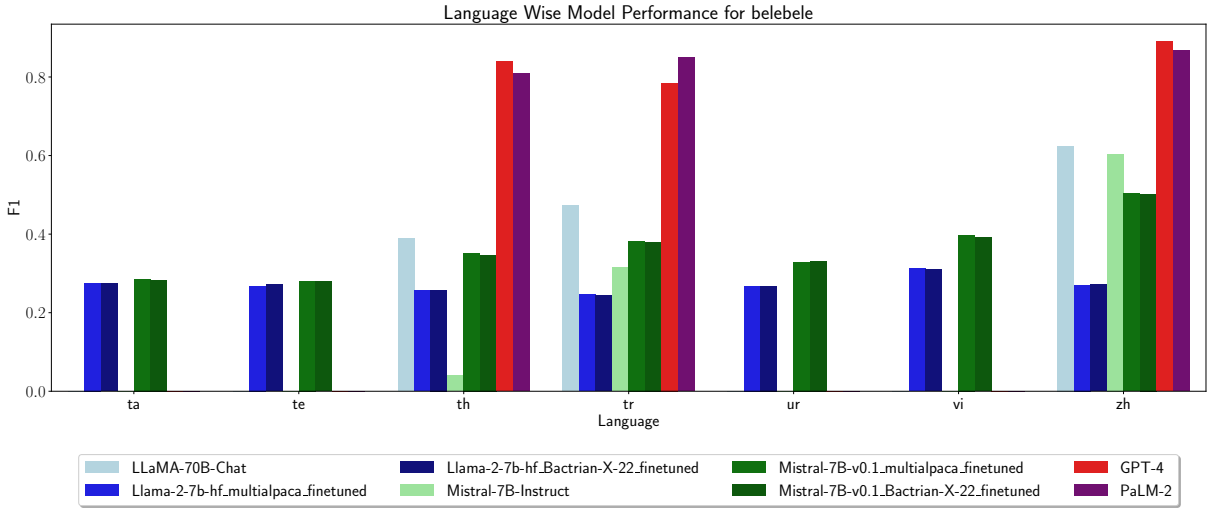

Figure 19: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Tamil, Telugu, Thai, Turkish, Urdu, Vietnamese and Chinese-Simplified for Belebele (Bandarkar et al., 2023).

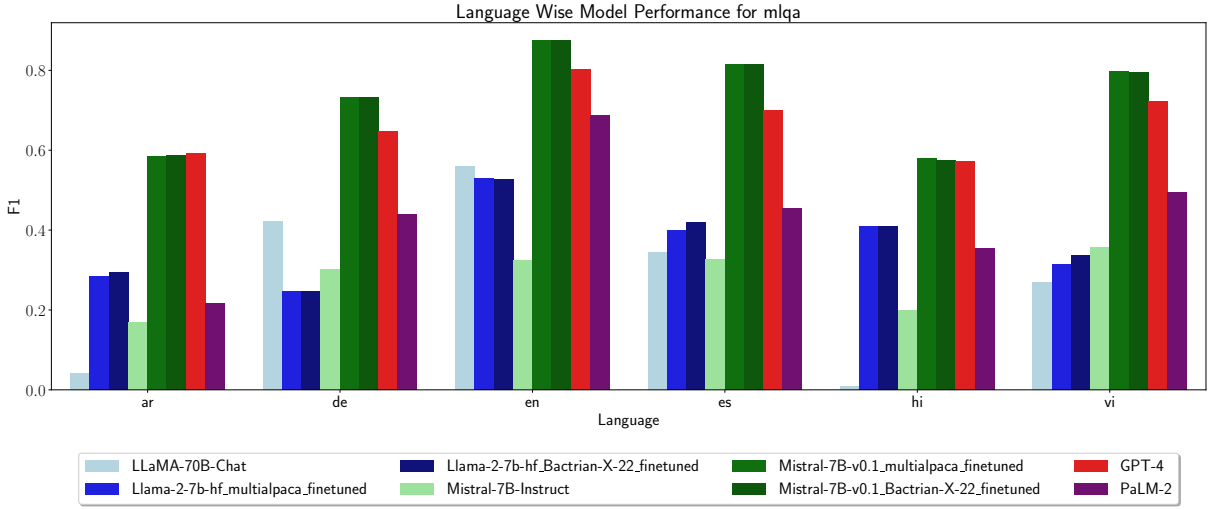

Figure 20: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Arabic, German, English, Spanish, Hindi and Vietnamese for MLQA (Lewis et al., 2020).

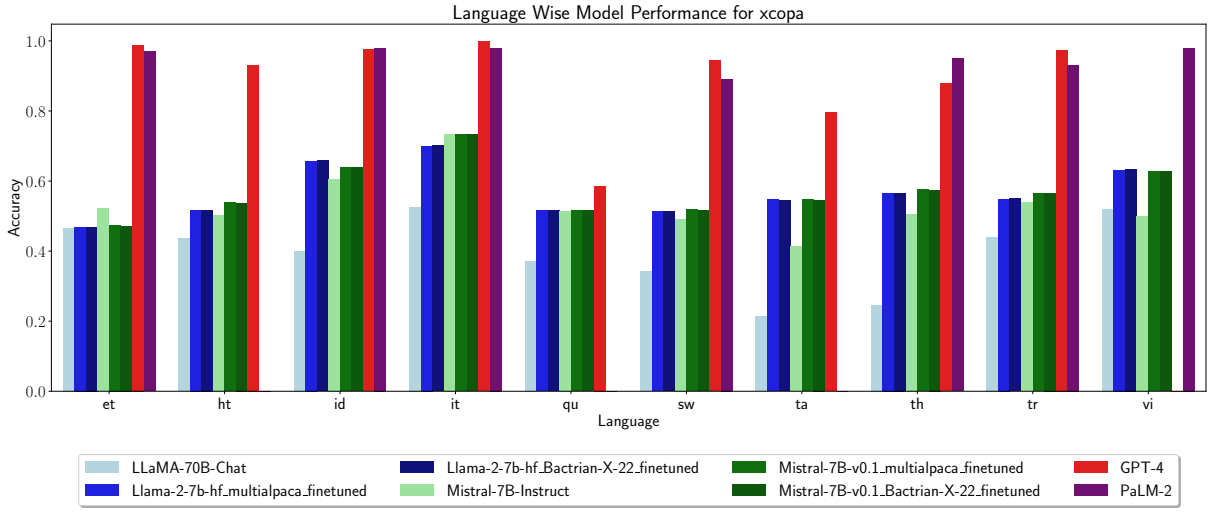

Figure 21: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Estonian, Haitian, Indonesian, Italian, Quechua, Swahili, Tamil, Thai, Turkish and Vietnamese for XCOPA (Ponti et al., 2020).

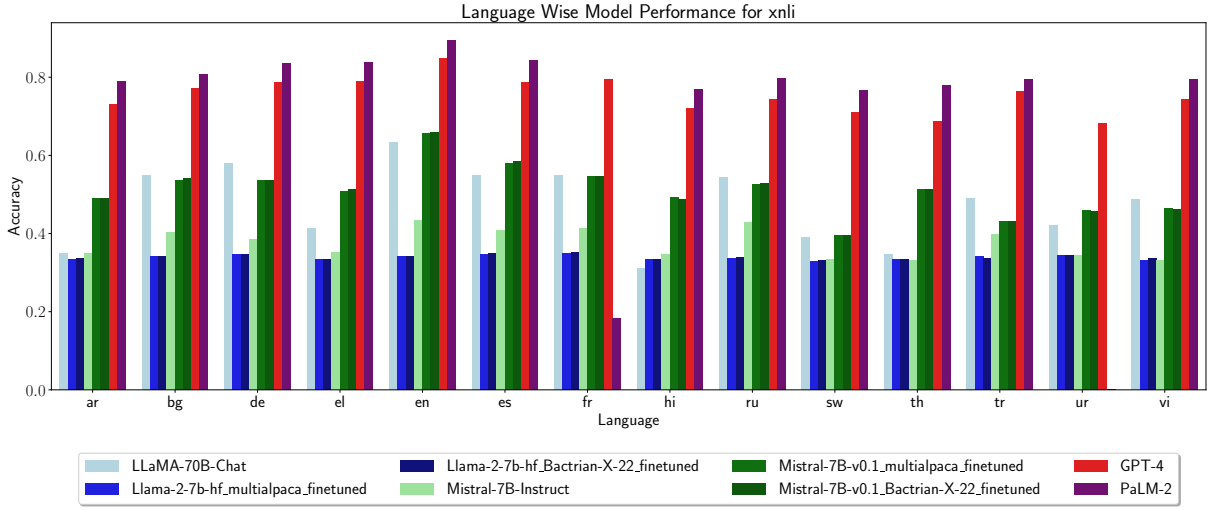

Figure 22: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Arabic, Bulgarian, German, Greek, English, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu and Vietnamese for XNLI (Conneau et al., 2018).

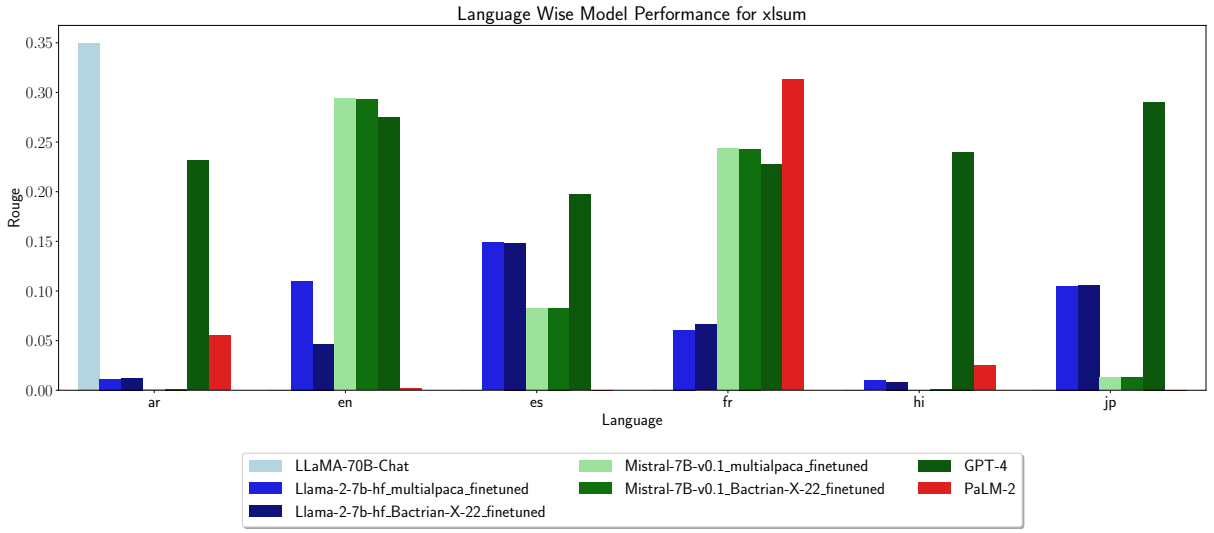

Figure 23: Detailed language-wise comparison of our fine-tuned MISTRAL-7B and LLAMA-2-7B models with other baselines (Ahuja et al., 2023b) on Arabic, English, Spanish, French, Hindi and Japanese for XLSUM(Hasan et al., 2021).

Model	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
64	8	0.27	0.27	0.22	0.33	0.29	0.30	0.28	0.30	0.33	0.30	0.25	0.26	0.24	0.24	0.24	0.24	0.26	0.29	0.26	0.25	0.28	0.23	0.30	0.23	0.26	0.29	0.28	0.29	0.29	0.27	0.26	0.26	0.27	0.30	0.27	0.27
64	16	**0.29**	0.25	**0.27**	0.31	0.29	0.32	0.29	0.27	0.32	0.27	0.24	0.25	**0.27**	0.26	**0.27**	**0.29**	**0.29**	0.28	0.28	0.26	0.27	**0.24**	0.29	0.26	0.27	0.31	0.29	0.27	0.28	0.26	0.27	0.24	**0.28**	0.31	0.27	0.28
128	8	0.27	0.24	0.23	**0.34**	**0.30**	0.30	0.28	0.28	0.32	0.29	0.24	0.25	0.26	0.23	0.25	0.25	0.27	0.29	0.27	0.26	**0.31**	0.23	**0.31**	0.23	0.27	0.30	0.31	**0.29**	**0.30**	**0.29**	0.29	0.25	0.27	0.31	0.27	0.28
128	16	**0.29**	0.24	**0.27**	0.31	0.29	0.32	0.29	0.27	0.32	0.27	0.24	0.25	**0.27**	0.26	**0.27**	**0.29**	**0.29**	0.27	0.28	0.26	0.27	0.24	0.29	0.26	0.27	0.31	0.29	0.27	0.29	0.26	0.27	0.24	**0.28**	**0.32**	0.27	0.28

Table 4: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on Belebele (Bandarkar et al., 2023).

Model	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
64	8	0.49	0.39	0.30	0.49	**0.41**	0.49	0.47	0.56	0.60	0.51	0.40	0.53	0.32	0.33	0.35	0.50	0.50	0.53	0.44	0.50	**0.31**	0.29	0.49	0.58	0.52	0.54	0.53	0.31	0.29	0.28	0.34	0.41	0.33	0.40	0.56	0.44
64	16	0.49	0.43	**0.33**	0.50	0.39	0.49	0.48	0.56	0.60	0.51	0.40	0.54	**0.34**	0.32	0.36	**0.51**	0.51	0.53	0.44	0.53	**0.31**	0.29	0.49	0.60	0.52	0.54	0.55	**0.33**	0.28	**0.30**	0.36	0.40	0.34	**0.40**	0.56	0.44
128	8	**0.50**	0.42	0.30	0.48	0.37	0.51	0.47	0.56	0.59	0.54	0.40	0.53	0.30	0.35	0.34	0.49	**0.54**	0.53	0.44	0.51	0.31	0.27	**0.50**	0.61	0.51	0.52	0.56	0.29	**0.30**	0.27	0.35	0.41	0.34	0.40	0.58	0.44
128	16	0.46	0.43	0.31	**0.51**	0.39	0.51	0.48	0.56	0.61	0.54	0.43	0.53	0.31	0.33	**0.36**	0.50	**0.54**	0.54	0.45	0.53	0.29	**0.30**	0.49	0.61	0.54	0.55	0.55	0.31	0.29	0.28	0.38	0.41	**0.37**	0.39	0.55	0.45

Table 5: Detailed performance of various ALPACA finetuned MISTRAL-7B models on Belebele (Bandarkar et al., 2023).

Model	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
8	4	0.26	0.29	0.24	0.27	0.29	0.27	0.37	0.29	0.29	0.29	0.25	0.23	**0.28**	0.26	0.24	0.23	0.28	0.29	0.26	0.22	0.26	0.23	0.28	0.24	0.26	0.27	0.26	0.29	0.27	0.26	0.23	0.25	0.26	0.29	0.30	0.27
8	8	0.26	0.24	0.25	0.31	0.29	0.27	0.30	0.26	0.31	0.27	0.26	0.26	0.26	0.23	0.26	0.26	0.28	0.28	0.27	0.27	0.25	0.24	**0.31**	0.22	0.25	0.31	0.28	0.29	0.28	**0.29**	0.26	0.24	0.26	0.33	0.26	0.27
8	16	0.28	0.24	**0.26**	0.30	0.30	0.32	0.30	0.26	0.33	0.27	0.24	0.25	**0.28**	0.26	0.27	**0.29**	0.29	0.28	0.28	0.26	0.27	0.23	0.30	0.26	0.27	0.31	0.29	0.27	0.28	0.27	0.27	0.24	0.28	0.33	0.27	0.28
16	4	0.26	0.29	0.24	0.27	0.29	0.27	0.37	0.29	0.29	0.29	0.25	0.23	**0.28**	0.27	0.24	0.22	0.28	0.29	0.26	0.22	0.26	0.22	0.28	0.24	0.26	0.28	0.26	0.29	0.27	0.26	0.23	0.25	0.26	0.29	0.30	0.27
16	8	0.28	0.27	0.25	0.29	**0.30**	0.27	0.27	0.29	0.33	0.29	0.25	0.26	0.25	0.24	0.23	0.25	0.28	0.28	0.27	0.27	0.26	**0.25**	0.29	0.23	0.27	0.29	0.29	0.28	**0.30**	**0.29**	0.26	0.25	0.26	0.32	0.27	0.27
16	16	0.28	0.24	0.26	0.30	0.30	0.31	0.30	0.27	0.32	0.27	0.23	0.26	0.27	0.26	0.27	0.27	0.28	0.28	0.28	0.27	0.27	0.23	0.30	0.25	0.27	0.31	0.30	0.28	0.28	0.27	0.28	0.25	0.28	0.32	0.26	0.28
32	4	0.26	0.29	0.23	0.28	0.29	0.28	0.35	0.29	0.29	0.28	0.24	0.23	0.26	0.26	0.24	0.24	0.28	0.31	0.26	0.22	0.26	0.22	0.27	0.24	0.26	0.29	0.27	0.28	0.24	0.26	0.23	0.24	0.24	0.29	0.29	0.26
32	8	0.28	0.24	0.24	**0.32**	0.26	0.29	0.27	0.27	0.32	0.29	0.26	0.26	0.27	0.23	0.24	0.27	0.27	0.30	0.25	0.27	0.27	0.24	0.28	0.21	0.26	0.29	0.29	0.28	0.27	0.26	0.27	0.24	0.26	0.33	0.26	0.27
32	16	0.28	0.25	0.26	0.31	0.29	0.31	0.29	0.27	0.33	0.27	0.24	0.25	0.27	0.26	**0.28**	0.27	0.29	0.30	0.28	0.27	0.27	0.23	0.30	0.24	0.27	0.32	0.30	0.29	0.28	0.26	0.28	0.25	0.28	0.33	0.26	0.28
64	4	0.24	0.29	0.22	0.29	0.28	0.28	0.35	0.28	0.30	0.26	0.24	0.23	0.26	0.25	0.24	0.23	0.28	0.32	0.27	0.22	0.26	0.23	0.26	0.23	0.26	0.28	0.26	0.26	0.25	0.26	0.23	0.24	0.24	0.29	0.27	0.26
64	8	**0.29**	0.25	**0.26**	0.32	0.28	0.30	0.29	0.28	0.32	0.29	0.26	0.26	0.24	0.24	0.22	0.24	0.30	0.30	0.26	0.26	**0.28**	0.24	0.29	0.23	0.27	0.32	0.30	0.28	0.29	0.28	0.27	0.26	0.28	0.32	0.23	0.27
64	16	0.28	0.29	0.24	0.30	0.29	0.29	0.30	0.26	0.34	0.29	0.24	0.24	0.26	0.26	0.26	0.27	**0.30**	0.30	0.28	0.28	**0.28**	0.24	0.30	0.25	0.27	0.30	0.29	0.27	0.27	0.26	0.27	0.25	**0.29**	**0.35**	0.26	0.28
128	4	0.26	0.27	0.24	0.27	0.29	0.27	0.36	0.29	0.29	0.28	0.25	0.23	**0.28**	0.27	0.24	0.23	0.28	0.30	0.26	0.22	0.26	0.22	0.27	0.24	0.26	0.29	0.29	0.29	0.27	0.27	0.23	0.25	0.26	0.29	0.30	0.27
128	8	0.28	0.24	0.25	0.30	0.29	0.29	0.30	0.29	0.33	0.30	0.26	0.25	**0.28**	0.24	0.24	0.28	0.26	0.29	0.27	0.27	**0.28**	0.23	0.27	0.23	0.26	0.30	0.31	**0.30**	0.28	0.28	0.27	0.25	0.28	0.30	0.27	0.28
128	16	0.28	0.24	**0.26**	0.30	0.30	0.31	0.28	0.27	0.32	0.27	0.23	0.25	0.26	0.26	0.27	0.26	0.28	0.29	0.28	0.27	0.27	0.23	**0.31**	0.25	0.27	0.31	0.30	0.28	0.28	0.27	0.27	0.25	0.28	0.33	0.26	0.28

Table 6: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on Belebele (Bandarkar et al., 2023).

Model	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
8	4	0.42	0.34	0.31	0.40	0.33	0.44	0.43	0.48	0.52	0.50	0.36	0.53	0.30	0.30	0.30	**0.50**	0.48	0.44	0.39	0.45	0.27	0.27	0.46	0.55	0.47	0.47	0.44	0.29	0.29	0.27	0.34	0.41	0.34	0.38	0.47	0.40
8	8	0.45	0.37	0.33	0.44	0.37	0.47	0.44	0.51	0.58	0.50	0.40	0.51	0.30	0.30	0.31	0.45	0.50	0.48	0.41	0.49	**0.30**	0.27	0.47	0.56	0.45	0.53	0.50	0.31	0.29	0.29	0.33	0.37	0.31	0.41	0.52	0.42
8	16	0.45	0.39	0.34	**0.46**	0.39	0.50	0.45	0.51	0.58	0.51	0.41	0.52	0.27	0.29	0.34	0.46	0.50	0.49	0.43	0.50	0.28	0.28	0.48	0.55	0.47	0.55	0.51	**0.34**	0.29	0.29	0.35	0.39	0.33	0.40	0.51	0.42
16	4	0.43	0.34	0.31	0.41	0.34	0.45	0.44	0.47	0.52	0.49	0.36	0.53	0.30	0.30	0.30	**0.50**	0.49	0.44	0.40	0.44	0.25	0.27	0.46	0.56	0.44	0.48	0.46	0.28	0.30	0.27	0.34	0.39	0.33	0.39	0.47	0.40
16	8	**0.47**	0.37	0.33	0.43	0.38	0.47	0.43	0.50	0.58	0.52	0.41	0.49	0.30	0.30	0.34	0.46	0.51	0.50	0.41	0.48	0.27	0.29	0.47	0.55	0.46	0.53	0.48	0.32	0.29	**0.30**	0.35	0.40	0.30	0.39	0.50	0.42
16	16	0.44	0.40	0.34	0.46	0.38	0.50	0.44	0.51	0.57	0.51	0.40	0.53	0.30	0.32	0.33	0.47	0.49	0.50	0.43	0.51	0.29	0.27	0.49	0.54	0.49	0.54	0.51	**0.34**	0.29	0.30	0.34	0.38	0.32	0.40	0.51	0.42
32	4	0.42	0.37	0.33	0.43	0.34	0.47	0.43	0.48	0.52	0.49	0.36	0.53	0.28	0.30	0.30	0.50	0.51	0.46	0.40	0.45	0.26	0.27	0.44	0.55	0.43	0.49	0.46	0.26	0.27	0.25	0.35	0.38	0.34	0.40	0.47	0.40
32	8	**0.47**	0.39	0.32	0.44	0.38	0.48	0.44	0.51	0.58	0.51	0.41	0.51	0.30	0.31	**0.35**	0.47	0.48	0.49	0.45	0.50	0.28	0.27	0.48	0.53	0.47	0.55	0.51	0.32	0.27	0.28	0.33	0.40	0.33	0.41	0.50	0.42
32	16	0.45	0.40	0.35	**0.46**	**0.39**	0.49	0.46	0.52	0.59	0.51	0.40	0.52	**0.33**	0.31	0.34	0.48	0.49	0.51	0.43	0.50	0.28	0.28	0.49	0.56	0.49	0.55	0.50	0.33	0.27	0.29	0.36	0.37	0.33	0.41	0.52	0.43
64	4	0.44	0.37	0.33	0.44	0.35	0.45	0.44	0.48	0.51	0.50	0.37	0.52	0.29	0.29	0.31	0.50	0.51	0.47	0.41	0.46	0.27	0.29	0.45	0.52	0.45	0.50	0.47	0.27	0.27	0.24	0.37	0.39	**0.35**	0.39	0.48	0.40
64	8	0.45	0.37	**0.36**	**0.46**	0.36	0.49	0.43	0.53	0.57	0.52	0.41	0.51	0.33	0.31	0.33	0.47	0.51	0.49	0.45	0.51	0.29	**0.30**	0.47	0.56	0.51	0.54	0.50	0.29	**0.30**	0.29	0.37	0.37	**0.35**	0.41	0.51	0.43
64	16	0.44	0.38	0.33	0.46	0.37	0.49	0.45	0.52	0.58	0.50	0.40	0.52	0.29	0.31	0.33	0.48	0.50	0.52	0.43	0.49	0.28	0.30	**0.50**	0.58	0.49	0.53	0.50	0.30	0.30	0.28	0.39	0.37	0.34	**0.42**	0.54	0.43
128	4	0.40	0.36	0.34	0.46	0.36	0.44	0.45	0.52	0.56	0.48	0.39	0.47	0.30	0.34	0.29	0.41	0.46	0.51	0.40	0.44	0.26	**0.30**	0.47	0.52	0.44	0.52	0.46	0.30	0.27	0.27	0.36	0.37	0.32	0.37	0.51	0.40
128	8	0.44	0.37	0.33	0.45	0.37	0.50	0.43	0.51	0.57	0.52	0.40	0.52	0.30	0.32	0.34	0.46	**0.51**	0.48	0.43	0.50	**0.30**	0.28	0.49	0.54	0.49	0.53	0.49	0.33	0.29	0.29	0.36	0.35	0.33	0.39	0.52	0.42
128	16	0.44	0.41	0.35	**0.46**	0.38	0.50	0.44	0.51	0.59	0.51	0.40	0.52	0.29	0.31	0.34	0.46	0.50	0.49	0.44	0.49	0.27	0.26	0.48	0.55	0.47	0.56	0.50	0.33	0.28	0.29	0.35	0.39	0.33	0.42	0.51	0.42

Table 7: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on Belebele (Bandarkar et al., 2023).

Model	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
8	4	0.26	0.29	0.24	0.27	0.29	0.27	0.37	0.29	0.29	0.29	0.25	0.23	**0.28**	0.27	0.24	0.22	0.28	0.29	0.26	0.22	0.26	0.22	0.28	0.24	0.26	0.28	0.27	0.29	0.27	0.26	0.23	0.25	0.26	0.29	0.31	0.27
8	8	0.26	0.25	0.25	0.30	0.29	0.29	0.26	0.27	0.31	0.27	0.25	0.27	0.24	0.25	0.23	0.26	0.27	0.29	0.27	0.26	**0.29**	0.23	0.28	0.21	0.27	0.29	0.30	**0.29**	0.28	**0.29**	0.27	0.25	0.26	0.31	0.28	0.27
8	16	0.28	0.24	**0.27**	0.31	0.29	0.32	0.29	0.26	0.32	0.27	0.24	0.25	**0.28**	0.26	0.27	**0.29**	0.29	0.28	0.28	0.26	0.27	0.23	0.30	0.26	0.27	0.31	0.29	0.27	0.28	0.26	0.26	0.24	0.28	0.32	0.27	0.28
16	4	0.26	0.28	0.24	0.27	0.29	0.27	0.36	0.29	0.29	0.29	0.25	0.23	0.27	0.27	0.25	0.23	0.28	0.29	0.26	0.22	0.26	0.23	0.27	0.24	0.26	0.28	0.27	0.28	0.26	0.27	0.23	0.25	0.26	0.29	0.30	0.27
16	8	0.28	0.25	0.24	0.29	0.26	0.29	0.28	0.27	0.33	0.29	0.23	0.26	0.26	0.23	0.22	0.25	0.29	0.29	0.26	0.26	0.28	**0.26**	0.29	0.23	0.26	0.30	0.32	0.28	**0.30**	0.28	0.28	0.23	0.27	**0.34**	0.26	0.27
16	16	0.28	0.24	**0.27**	0.31	0.29	0.32	0.29	0.27	0.32	0.27	0.24	0.25	0.27	0.26	0.27	0.29	0.29	0.27	0.28	0.26	0.27	0.23	0.30	0.26	0.27	0.31	0.29	0.27	0.29	0.26	0.27	0.24	0.28	0.32	0.27	0.28
32	4	0.26	0.28	0.24	0.28	0.28	0.27	0.35	0.29	0.29	0.29	0.25	0.23	0.27	0.26	0.24	0.23	0.28	0.31	0.26	0.22	0.26	0.23	0.27	0.24	0.26	0.29	0.28	0.28	0.25	0.27	0.23	0.24	0.25	0.29	0.29	0.27
32	8	**0.29**	0.27	0.23	0.31	0.28	0.30	0.28	0.27	0.34	0.30	0.26	0.27	0.24	0.23	0.23	0.24	0.28	0.29	0.26	0.24	0.27	0.23	0.29	0.22	0.26	0.30	0.30	0.29	0.28	0.28	0.27	0.25	0.26	0.29	0.27	0.27
32	16	0.28	0.24	0.26	0.30	**0.30**	0.32	0.29	0.26	0.32	0.27	0.24	0.25	0.27	0.26	0.27	0.29	0.29	0.28	0.28	0.26	0.27	0.23	0.30	0.26	0.27	0.31	0.29	0.27	0.28	0.26	0.27	0.24	0.28	0.33	0.26	0.28
64	4	0.24	0.29	0.23	0.28	0.29	0.28	0.35	0.28	0.30	0.28	0.25	0.23	0.27	0.26	0.24	0.23	0.28	0.30	0.27	0.23	0.26	0.22	0.26	0.24	0.26	0.29	0.27	0.26	0.26	0.27	0.23	0.23	0.24	0.30	0.27	0.26
64	8	0.26	0.27	0.22	0.30	0.29	0.29	0.27	0.26	0.32	0.29	0.27	0.24	0.24	0.21	0.23	0.26	0.28	0.30	0.28	0.27	0.26	0.24	0.30	0.23	0.27	0.30	0.30	0.29	0.27	**0.29**	0.27	0.24	0.27	0.30	0.26	0.27
64	16	**0.29**	0.24	0.26	0.30	0.30	0.31	0.29	0.27	0.32	0.27	0.23	0.24	0.26	0.26	**0.27**	0.27	0.29	0.29	0.28	0.27	0.27	0.23	**0.31**	0.26	0.27	0.31	0.30	0.28	0.28	0.26	0.27	0.26	0.28	0.32	0.26	0.28
128	4	0.24	0.27	0.22	0.28	0.28	0.28	0.34	0.28	0.29	0.28	0.25	0.24	**0.28**	0.26	0.23	0.23	0.29	0.33	0.30	0.23	0.26	0.23	0.27	0.26	0.26	0.29	0.26	0.26	0.26	0.28	0.23	0.25	0.25	0.30	0.26	0.27
128	8	0.27	0.23	0.22	0.29	**0.30**	0.29	0.29	0.26	0.33	0.30	0.26	0.27	0.24	0.24	0.25	0.26	**0.30**	0.29	0.29	0.29	0.28	0.23	0.29	0.24	0.27	0.31	0.31	0.27	0.29	0.27	0.29	0.24	0.29	0.31	0.26	0.28
128	16	0.28	0.29	0.24	**0.33**	0.29	0.31	0.29	0.27	0.33	0.29	0.24	0.24	0.26	0.26	0.26	0.27	**0.30**	0.30	0.27	0.28	0.27	0.23	0.29	0.25	0.27	0.31	0.31	0.27	0.28	0.27	0.26	0.24	**0.29**	0.34	0.25	0.28

Table 8: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on Belebele (Bandarkar et al., 2023).

Model	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
LLaMA-70B-Chat	–	0.42	–	–	–	0.65	0.66	0.69	0.79	0.68	0.63	0.72	–	0.41	–	–	–	0.68	0.57	0.56	–	–	–	0.66	0.70	0.67	0.67	–	–	–	0.39	0.47	–	–	0.62	0.61
Mistral-7B-Instruct	–	0.09	–	–	–	0.54	0.51	0.52	0.65	0.59	0.35	0.60	–	0.02	–	–	–	0.51	0.42	0.41	–	–	–	0.55	0.54	0.62	0.55	–	–	–	0.04	0.31	–	–	0.60	0.44
GPT-3.5-Turbo	–	0.69	–	–	–	0.77	0.81	0.83	0.88	0.79	0.78	0.83	–	0.64	–	–	–	0.80	0.71	0.67	–	–	–	0.80	0.83	0.78	0.82	–	–	–	0.56	0.70	–	–	0.78	0.76
GPT-4	–	**0.92**	–	–	–	0.85	0.82	0.85	0.79	0.83	**0.87**	**0.91**	–	**0.87**	–	–	–	0.81	**0.86**	**0.89**	–	–	–	0.85	0.79	**0.91**	0.80	–	–	–	**0.84**	0.78	–	–	**0.89**	0.85
PALM2	–	0.87	–	–	–	**0.88**	**0.88**	**0.88**	**0.92**	**0.86**	0.87	0.88	–	0.85	–	–	–	**0.86**	0.84	0.86	–	–	–	**0.87**	**0.88**	0.87	**0.87**	–	–	–	0.81	**0.85**	–	–	0.87	**0.87**
rank	quantisation	af	ar	as	bg	bn	cs	da	de	en	es	fi	fr	gu	he	hi	hr	id	it	ja	ko	ml	mr	nb	nl	pt	ru	sv	sw	ta	te	th	tr	ur	vi	zh	avg
8	4	0.42	0.33	0.33	0.41	0.33	0.44	0.43	0.47	0.52	0.49	0.35	0.54	**0.31**	0.29	0.29	**0.50**	0.50	0.44	0.40	0.45	0.27	0.27	0.46	0.55	0.46	0.48	0.44	0.27	0.30	0.27	0.34	0.39	0.34	0.38	0.47	0.40
8	8	**0.47**	0.38	0.31	0.44	0.37	0.49	0.44	0.52	0.57	0.50	0.41	0.49	0.27	0.31	0.30	0.45	0.47	0.49	0.41	0.51	0.29	0.30	0.47	0.54	0.49	0.55	0.49	0.34	0.29	0.29	0.33	0.39	0.33	0.39	0.52	0.42
8	16	0.45	0.39	0.34	0.46	0.37	0.49	0.43	0.51	0.57	0.51	0.40	0.51	0.29	0.29	0.33	0.47	0.49	0.49	0.43	0.49	0.29	0.30	0.47	0.56	0.48	0.56	0.50	0.32	0.30	0.29	0.33	0.37	0.32	**0.41**	0.51	0.42
16	4	0.41	0.33	0.31	0.41	0.33	0.46	0.43	0.47	0.53	0.50	0.36	0.52	0.31	0.30	0.29	**0.50**	0.49	0.45	0.39	0.45	0.27	0.27	0.46	0.56	0.45	0.48	0.44	0.29	0.28	0.26	0.33	0.38	0.34	0.38	0.47	0.40
16	8	0.44	0.38	0.31	**0.47**	0.34	0.48	0.43	0.50	0.57	0.50	0.42	0.49	0.27	0.29	0.33	0.45	0.49	0.49	0.42	0.49	0.28	0.28	0.44	0.54	0.46	0.56	0.49	0.34	0.26	0.29	0.34	0.38	0.32	0.39	0.51	0.41
16	16	0.44	0.39	0.34	0.46	0.37	0.50	0.45	0.50	0.57	0.51	0.40	0.52	0.28	0.30	0.33	0.47	0.49	0.49	0.43	0.49	0.28	0.29	0.48	0.54	0.48	0.55	0.51	**0.34**	**0.30**	0.29	0.33	0.39	0.33	0.40	0.52	0.42
32	4	0.42	0.34	0.32	0.41	0.36	0.45	0.41	0.47	0.51	0.48	0.35	0.52	0.31	0.31	0.31	0.49	0.50	0.45	0.41	0.45	0.27	0.27	0.46	0.54	0.45	0.47	0.46	0.27	0.29	0.24	0.35	0.38	0.34	0.38	0.48	0.40
32	8	0.42	0.39	0.31	0.43	**0.40**	0.47	0.43	0.52	0.58	0.51	0.41	0.52	0.31	0.30	0.33	0.45	**0.50**	0.49	0.41	0.49	**0.30**	**0.30**	0.47	0.55	0.46	0.52	0.47	0.31	0.27	0.29	0.34	0.37	0.30	0.39	0.52	0.42
32	16	0.43	0.40	0.35	0.46	0.37	0.49	0.44	0.51	0.58	0.51	0.40	0.52	0.29	0.31	**0.34**	0.46	0.49	0.50	0.42	0.49	0.27	0.26	0.48	0.55	0.47	0.54	0.50	0.33	0.28	0.28	0.34	0.37	0.32	0.41	0.51	0.42
64	4	0.41	0.36	0.33	0.43	0.36	0.46	0.39	0.49	0.51	0.48	0.34	0.52	0.31	0.30	0.30	0.47	0.49	0.46	0.41	0.45	0.27	0.29	0.46	0.53	0.43	0.50	0.45	0.28	0.27	0.26	0.36	0.39	**0.36**	0.39	0.47	0.40
64	8	0.43	0.39	0.33	0.45	**0.40**	0.47	0.41	0.52	0.57	0.51	0.38	0.51	**0.31**	0.30	0.32	0.43	0.49	0.50	0.42	0.49	0.29	0.28	**0.49**	0.55	0.46	0.55	0.49	0.32	**0.30**	**0.30**	0.34	0.38	0.33	0.39	0.54	0.42
64	16	0.44	0.40	0.33	0.46	0.37	0.48	0.44	0.53	0.58	0.49	0.40	0.51	0.31	0.31	0.33	0.45	0.49	0.52	0.43	0.49	0.28	0.26	**0.49**	0.55	0.45	0.54	0.50	0.33	0.28	0.27	0.35	0.38	0.33	0.39	0.53	0.42
128	4	0.42	0.36	0.33	**0.47**	0.36	0.46	0.43	0.55	0.51	0.49	0.37	0.53	0.29	0.31	0.29	0.43	0.47	0.50	0.41	0.44	0.28	0.27	0.46	0.54	0.43	0.50	0.46	0.31	0.27	0.27	0.36	0.38	0.33	0.40	0.49	0.40
128	8	0.43	0.40	**0.36**	0.43	0.36	0.46	0.44	0.50	0.59	0.50	0.39	0.50	0.29	0.29	0.33	0.45	0.50	0.50	0.41	0.49	0.28	0.26	0.48	0.54	0.48	0.55	0.49	0.32	0.30	0.30	0.37	0.38	0.34	0.39	0.49	0.42
128	16	0.41	0.42	0.32	0.47	0.38	0.47	0.47	0.54	0.59	0.49	0.39	0.50	0.31	0.33	0.33	0.46	0.49	0.54	0.43	0.50	0.29	**0.30**	0.47	0.55	0.43	0.52	0.49	0.31	0.27	0.29	0.38	0.38	0.34	0.40	0.51	0.42

Table 9: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on Belebele (Bandarkar et al., 2023).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
64	8	13.05
64	16	13.11
128	8	13.28
128	16	13.23

Table 10: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on AlpacaEval (Li et al., 2023b).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
8	4	11.86
8	8	12.81
8	16	13.23
16	4	11.97
16	8	13.06
16	16	13.36
32	4	11.72
32	8	13.15
32	16	13.36
64	4	11.47
64	8	13.67
64	16	13.73
128	4	11.10
128	8	13.05
128	16	0.00

Table 11: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on AlpacaEval (Li et al., 2023b).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
8	4	11.61
8	8	13.56
8	16	13.36
16	4	11.99
16	8	13.38
16	16	13.73
32	4	11.86
32	8	12.47
32	16	13.73
64	4	11.61
64	8	13.47
64	16	13.86
128	4	11.35
128	8	13.56
128	16	13.73

Table 12: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on AlpacaEval (Li et al., 2023b).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
64	8	20.47
64	16	18.89
128	8	19.55
128	16	19.33

Table 13: Detailed performance of various ALPACA finetuned MISTRAL-7B models on AlpacaEval (Li et al., 2023b).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
8	4	15.04
8	8	20.67
8	16	21.27
16	4	15.77
16	8	22.07
16	16	21.08
32	4	15.77
32	8	21.07
32	16	21.14
64	4	16.90
64	8	21.30
64	16	21.27
128	4	17.77
128	8	22.04
128	16	21.83

Table 14: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on AlpacaEval (Li et al., 2023b).

Model		win_rate
LLaMA-70B-Chat		22.36
Mistral-7B-Instruct		35.13
GPT-4		**93.78**
PALM2		79.66
rank	quantisation	win_rate
8	4	15.98
8	8	21.42
8	16	20.71
16	4	15.90
16	8	20.55
16	16	21.27
32	4	16.98
32	8	22.45
32	16	22.08
64	4	16.83
64	8	21.07
64	16	18.70
128	4	16.52
128	8	20.92
128	16	21.33

Table 15: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on AlpacaEval (Li et al., 2023b).

Models		ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat		0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat		0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat		0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct		0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003		0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo		0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4		0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6		**0.76**	**0.80**	**0.87**	**0.82**	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R		0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT		0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5		0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2		0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
64	8	0.31	0.20	0.28	0.24	0.43	0.19	0.08	0.25
64	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35
128	8	0.31	0.20	0.29	0.25	0.43	0.19	0.08	0.25
128	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35

Table 16: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on MLQA (Lewis et al., 2020).

Models		ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat		0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat		0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat		0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct		0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003		0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo		0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4		0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6		**0.76**	**0.80**	**0.87**	**0.82**	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R		0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT		0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5		0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2		0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
8	4	0.24	0.18	0.75	0.49	0.37	0.41	0.08	0.36
8	8	0.31	0.20	0.31	0.24	0.43	0.19	0.08	0.25
8	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35
16	4	0.26	0.18	0.76	0.51	0.37	0.42	0.08	0.37
16	8	0.31	0.20	0.30	0.24	0.43	0.19	0.08	0.25
16	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35
32	4	0.29	0.18	0.77	0.56	0.37	0.48	0.08	0.39
32	8	0.32	0.20	0.31	0.25	0.43	0.19	0.08	0.25
32	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35
64	4	0.31	0.18	0.79	0.65	0.37	0.54	0.08	0.42
64	8	0.32	0.19	0.31	0.25	0.42	0.19	0.08	0.25
64	16	0.27	0.36	0.49	0.43	0.43	0.33	0.15	0.35
128	4	0.33	0.18	0.80	0.72	0.38	0.61	0.08	0.44
128	8	0.37	0.20	0.34	0.26	0.42	0.20	0.08	0.27
128	16	0.28	0.36	0.49	0.43	0.43	0.33	0.15	0.35

Table 17: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on MLQA (Lewis et al., 2020).

Models		ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat		0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat		0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat		0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct		0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003		0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo		0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4		0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6		**0.76**	**0.80**	**0.87**	**0.82**	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R		0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT		0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5		0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2		0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
8	4	0.24	0.18	0.74	0.48	0.37	0.40	0.08	0.36
8	8	0.30	0.20	0.31	0.25	0.43	0.20	0.08	0.25
8	16	0.27	0.36	0.50	0.43	0.43	0.33	0.15	0.35
16	4	0.24	0.18	0.75	0.48	0.37	0.41	0.07	0.36
16	8	0.34	0.20	0.32	0.24	0.42	0.19	0.08	0.26
16	16	0.27	0.36	0.49	0.43	0.43	0.33	0.15	0.35
32	4	0.22	0.18	0.74	0.46	0.38	0.37	0.08	0.35
32	8	0.30	0.20	0.31	0.25	0.42	0.19	0.08	0.25
32	16	0.28	0.36	0.49	0.43	0.43	0.33	0.15	0.35
64	4	0.23	0.18	0.75	0.48	0.38	0.38	0.08	0.35
64	8	0.40	0.20	0.47	0.34	0.43	0.23	0.08	0.31
64	16	0.28	0.36	0.48	0.43	0.43	0.33	0.15	0.35
128	4	0.30	0.18	0.78	0.62	0.37	0.52	0.08	0.41
128	8	0.32	0.20	0.32	0.25	0.41	0.19	0.08	0.25
128	16	0.27	0.36	0.49	0.43	0.43	0.33	0.15	0.35

Table 18: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on MLQA (Lewis et al., 2020).

Models		ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat		0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat		0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat		0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct		0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003		0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo		0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4		0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6		**0.76**	**0.80**	0.87	**0.82**	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R		0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT		0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5		0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2		0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
64	8	0.61	0.73	0.89	0.81	0.60	0.80	0.41	0.69
64	16	0.61	0.74	**0.89**	0.81	0.61	0.81	0.41	0.70
128	8	0.61	0.74	**0.89**	0.81	0.60	0.80	0.41	0.69
128	16	0.61	0.74	0.88	0.81	0.60	0.81	0.41	0.70

Table 19: Detailed performance of various ALPACA finetuned MISTRAL-7B models on MLQA (Lewis et al., 2020).

Model		ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat		0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat		0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat		0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct		0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003		0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo		0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4		0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6		**0.76**	**0.80**	0.87	0.82	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R		0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT		0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5		0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2		0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
8	4	0.55	0.71	0.86	0.81	0.53	0.78	0.39	0.66
8	8	0.62	0.74	0.88	0.81	0.59	0.80	0.41	0.69
8	16	0.61	0.74	0.88	0.82	0.60	0.80	0.41	0.70
16	4	0.55	0.71	0.86	0.81	0.53	0.78	0.39	0.66
16	8	0.60	0.75	**0.89**	0.82	0.59	0.81	0.41	0.70
16	16	0.61	0.74	0.88	0.81	0.59	0.80	0.41	0.69
32	4	0.54	0.71	0.86	0.81	0.53	0.77	0.39	0.66
32	8	0.60	0.74	0.88	0.81	0.60	0.80	0.41	0.69
32	16	0.61	0.74	0.88	0.81	0.60	0.81	0.41	0.69
64	4	0.54	0.71	0.86	**0.82**	0.52	0.78	0.39	0.66
64	8	0.60	0.74	0.88	0.81	0.60	0.80	0.41	0.69
64	16	0.61	0.74	0.88	0.81	0.60	0.81	0.41	0.70
128	4	0.55	0.72	0.86	0.82	0.52	0.78	0.39	0.66
128	8	0.59	0.75	0.88	0.81	0.59	0.81	0.41	0.69
128	16	0.61	0.74	0.88	0.81	0.61	0.81	0.41	0.70

Table 20: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on MLQA (Lewis et al., 2020).

Model	ar	de	en	es	hi	vi	zh	avg
LLaMA-7B-Chat	0.05	0.45	0.70	0.52	0.00	0.42	0.07	0.32
LLaMA–3B-Chat	0.55	0.73	0.62	0.62	0.00	0.09	0.09	0.39
LLaMA-70B-Chat	0.04	0.42	0.56	0.34	0.01	0.27	0.05	0.24
Mistral-7B-Instruct	0.17	0.30	0.32	0.33	0.20	0.36	0.02	0.24
DV003	0.38	0.58	0.75	0.63	0.25	0.48	0.32	0.48
GPT-3.5-Turbo	0.48	0.51	0.73	0.54	0.51	0.59	0.57	0.56
GPT-4	0.59	0.65	0.80	0.70	0.57	0.72	0.67	0.67
TULRv6	**0.76**	**0.80**	0.87	0.82	**0.82**	**0.82**	**0.78**	**0.81**
XLM-R	0.67	0.70	0.83	0.74	0.71	0.74	0.62	0.72
mBERT	0.52	0.59	0.80	0.67	0.50	0.61	0.60	0.61
mT5	0.57	0.62	0.82	0.67	0.55	0.66	0.62	0.64
PALM2	0.22	0.44	0.69	0.45	0.36	0.49	0.06	0.39
rank	quantisation	ar	de	en	es	hi	vi	zh	avg
8	4	0.56	0.71	0.86	0.82	0.54	0.78	0.39	0.66
8	8	0.60	0.74	0.89	0.81	0.58	0.80	0.41	0.69
8	16	0.61	0.74	0.88	0.81	0.60	0.81	0.41	0.69
16	4	0.54	0.71	0.85	0.81	0.54	0.78	0.39	0.66
16	8	0.60	0.74	**0.89**	0.81	0.59	0.81	0.41	0.69
16	16	0.61	0.75	0.88	0.82	0.60	0.81	0.41	0.70
32	4	0.55	0.71	0.86	0.81	0.54	0.78	0.39	0.66
32	8	0.58	0.74	0.88	0.81	0.60	0.80	0.41	0.69
32	16	0.60	0.75	0.88	0.82	0.62	0.81	0.41	0.70
64	4	0.55	0.71	0.86	0.81	0.53	0.77	0.39	0.66
64	8	0.59	0.74	0.88	0.82	0.61	0.81	0.41	0.69
64	16	0.60	0.74	0.88	**0.82**	0.61	0.81	0.41	0.70
128	4	0.57	0.72	0.86	0.81	0.54	0.78	0.39	0.67
128	8	0.60	0.74	0.89	0.81	0.59	0.81	0.41	0.69
128	16	0.61	0.74	0.88	0.81	0.60	0.80	0.41	0.70

Table 21: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on MLQA (Lewis et al., 2020).

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat	0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat	0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct	0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003	0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo	0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4	**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6	0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ	0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM	0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5	0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2	0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**
rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
64	8	0.47	0.52	0.66	0.71	0.52	0.51	0.54	0.57	0.55	0.63	0.70	0.58
64	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.64	0.70	0.58
128	8	0.48	0.53	0.66	0.71	0.52	0.51	0.54	0.56	0.55	0.64	0.69	0.58
128	16	0.48	0.52	0.66	0.70	0.52	0.51	0.54	0.58	0.55	0.64	0.70	0.58

Table 22: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on XCOPA (Ponti et al., 2020).

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat	0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat	0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct	0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003	0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo	0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4	**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6	0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ	0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM	0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5	0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2	0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**

rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
8	4	0.46	0.51	0.66	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.58
8	8	0.47	0.53	0.66	0.71	0.52	0.52	0.54	0.57	0.56	0.63	0.69	0.58
8	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.64	0.70	0.58
16	4	0.46	0.50	0.66	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.58
16	8	0.47	0.52	0.67	0.72	0.52	0.52	0.54	0.56	0.55	0.63	0.69	0.58
16	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.64	0.70	0.58
32	4	0.46	0.50	0.66	0.69	0.52	0.51	0.55	0.56	0.55	0.63	0.69	0.57
32	8	0.47	0.53	0.65	0.71	0.51	0.52	0.54	0.58	0.55	0.64	0.69	0.58
32	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.63	0.70	0.58
64	4	0.46	0.50	0.65	0.69	0.52	0.51	0.55	0.55	0.55	0.63	0.69	0.57
64	8	0.47	0.52	0.66	0.70	0.51	0.52	0.54	0.57	0.55	0.63	0.68	0.58
64	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.64	0.70	0.58
128	4	0.46	0.50	0.65	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.57
128	8	0.47	0.52	0.66	0.70	0.52	0.52	0.54	0.57	0.55	0.63	0.70	0.58
128	16	0.47	0.52	0.66	0.70	0.51	0.51	0.54	0.57	0.55	0.64	0.70	0.58

Table 23: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on XCOPA (Ponti et al., 2020).

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat	0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat	0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct	0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003	0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo	0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4	**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6	0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ	0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM	0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5	0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2	0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**
rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
8	4	0.46	0.51	0.66	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.58
8	8	0.47	0.52	0.65	0.71	0.52	0.52	0.54	0.56	0.55	0.64	0.70	0.58
8	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.63	0.70	0.58
16	4	0.46	0.51	0.66	0.69	0.52	0.51	0.56	0.55	0.55	0.63	0.69	0.57
16	8	0.47	0.53	0.65	0.70	0.52	0.51	0.54	0.57	0.55	0.63	0.69	0.58
16	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.64	0.70	0.58
32	4	0.46	0.50	0.66	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.58
32	8	0.46	0.53	0.66	0.70	0.52	0.51	0.55	0.57	0.55	0.63	0.70	0.58
32	16	0.47	0.52	0.66	0.70	0.52	0.51	0.54	0.58	0.55	0.63	0.70	0.58
64	4	0.46	0.50	0.65	0.69	0.52	0.51	0.55	0.55	0.54	0.63	0.69	0.57
64	8	0.47	0.53	0.65	0.71	0.52	0.52	0.54	0.57	0.54	0.63	0.69	0.58
64	16	0.48	0.53	0.66	0.70	0.51	0.51	0.54	0.57	0.55	0.64	0.70	0.58
128	4	0.46	0.50	0.66	0.69	0.52	0.51	0.56	0.56	0.55	0.63	0.69	0.57
128	8	0.47	0.52	0.65	0.70	0.51	0.51	0.54	0.57	0.55	0.63	0.69	0.58
128	16	0.48	0.52	0.66	0.70	0.52	0.51	0.54	0.57	0.55	0.63	0.70	0.58

Table 24: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on XCOPA (Ponti et al., 2020).

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat	0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat	0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct	0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003	0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo	0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4	**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6	0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ	0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM	0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5	0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2	0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**
rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
64	8	0.48	0.55	0.64	0.74	0.51	0.52	0.55	0.57	0.57	0.63	0.73	0.59
64	16	0.48	0.53	0.64	0.74	0.51	0.51	0.55	0.57	0.57	0.63	0.72	0.59
128	8	0.48	0.54	0.65	0.74	0.52	0.51	0.55	0.56	0.57	0.64	0.73	0.59
128	16	0.48	0.53	0.64	0.73	0.52	0.51	0.55	0.57	0.57	0.63	0.72	0.59

Table 25: Detailed performance of various ALPACA finetuned MISTRAL-7B models on XCOPA (Ponti et al., 2020).

Model		et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat		0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat		0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat		0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct		0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003		0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo		0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4		**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6		0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ		0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM		0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5		0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2		0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**
rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
8	4	0.47	0.54	0.63	0.72	0.52	0.52	0.54	0.59	0.57	0.61	0.74	0.59
8	8	0.48	0.54	0.64	0.74	0.51	0.51	0.54	0.56	0.57	0.62	0.73	0.59
8	16	0.48	0.53	0.64	0.74	0.51	0.51	0.55	0.57	0.57	0.64	0.73	0.59
16	4	0.46	0.54	0.63	0.72	0.52	0.52	0.55	0.58	0.57	0.62	0.74	0.59
16	8	0.47	0.54	0.64	0.73	0.51	0.51	0.55	0.57	0.57	0.63	0.73	0.59
16	16	0.48	0.53	0.64	0.74	0.52	0.51	0.55	0.57	0.57	0.63	0.72	0.59
32	4	0.47	0.54	0.64	0.73	0.52	0.52	0.55	0.58	0.56	0.62	0.74	0.59
32	8	0.47	0.53	0.64	0.74	0.51	0.52	0.54	0.57	0.57	0.63	0.73	0.59
32	16	0.47	0.53	0.64	0.74	0.52	0.51	0.55	0.58	0.56	0.64	0.72	0.59
64	4	0.47	0.54	0.64	0.73	0.52	0.53	0.55	0.58	0.56	0.62	0.73	0.59
64	8	0.48	0.54	0.64	0.74	0.51	0.51	0.54	0.57	0.57	0.63	0.73	0.59
64	16	0.48	0.53	0.64	0.73	0.51	0.51	0.55	0.58	0.56	0.64	0.73	0.59
128	4	0.46	0.53	0.64	0.73	0.52	0.51	0.54	0.58	0.56	0.62	0.73	0.58
128	8	0.47	0.55	0.64	0.74	0.51	0.51	0.55	0.56	0.56	0.63	0.72	0.59
128	16	0.48	0.54	0.64	0.73	0.51	0.51	0.55	0.57	0.57	0.63	0.74	0.59

Table 26: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on XCOPA (Ponti et al., 2020).

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.51	0.51	0.59	0.71	0.50	0.51	0.50	0.52	0.53	0.58	0.59	0.55
LLaMA–3B-Chat	0.51	0.49	0.72	0.80	0.50	0.50	0.00	0.00	0.54	0.02	0.70	0.44
LLaMA-70B-Chat	0.47	0.44	0.40	0.53	0.37	0.34	0.21	0.25	0.44	0.52	0.35	0.39
Mistral-7B-Instruct	0.52	0.50	0.61	0.73	0.51	0.49	0.41	0.51	0.54	0.50	0.50	0.53
DV003	0.88	0.75	0.91	0.96	0.55	0.64	0.54	0.67	0.88	–	–	0.75
GPT-3.5-Turbo	0.91	0.72	0.90	0.95	0.55	0.82	0.59	0.78	0.91	–	–	0.79
GPT-4	**0.99**	**0.93**	0.98	**1.00**	0.59	**0.94**	0.80	0.88	**0.97**	–	–	0.90
TULRv6	0.77	0.78	0.93	0.96	**0.61**	0.69	**0.85**	0.87	0.93	–	–	0.82
BLOOMZ	0.48	0.55	0.86	0.74	0.50	0.60	0.67	0.50	0.54	–	–	0.60
XGLM	0.66	0.59	0.69	0.69	0.47	0.63	0.56	0.62	0.58	–	–	0.61
mT5	0.50	0.50	0.49	0.50	0.51	0.50	0.49	0.51	0.49	–	–	0.50
PALM2	0.97	–	**0.98**	0.98	–	0.89	–	**0.95**	0.93	**0.98**	**0.99**	**0.96**
rank	quantisation	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	avg
8	4	0.47	0.53	0.63	0.72	0.52	0.52	0.55	0.58	0.57	0.62	0.73	0.59
8	8	0.47	0.54	0.64	0.73	0.51	0.51	0.55	0.57	0.57	0.63	0.73	0.59
8	16	0.48	0.53	0.65	0.73	0.51	0.52	0.55	0.57	0.57	0.64	0.73	0.59
16	4	0.46	0.54	0.63	0.73	0.52	0.53	0.55	0.59	0.57	0.62	0.73	0.59
16	8	0.48	0.54	0.64	0.74	0.52	0.52	0.55	0.57	0.56	0.63	0.73	0.59
16	16	0.48	0.54	0.64	0.74	0.51	0.52	0.55	0.58	0.56	0.63	0.73	0.59
32	4	0.47	0.54	0.63	0.73	0.52	0.53	0.55	0.58	0.57	0.62	0.73	0.59
32	8	0.48	0.54	0.64	0.73	0.51	0.53	0.54	0.57	0.56	0.63	0.72	0.59
32	16	0.48	0.53	0.64	0.74	0.52	0.52	0.54	0.57	0.56	0.63	0.74	0.59
64	4	0.47	0.54	0.63	0.72	0.52	0.52	0.55	0.58	0.57	0.61	0.73	0.59
64	8	0.48	0.54	0.65	0.74	0.52	0.51	0.55	0.57	0.56	0.63	0.73	0.59
64	16	0.48	0.54	0.64	0.74	0.51	0.52	0.55	0.58	0.56	0.64	0.73	0.59
128	4	0.47	0.54	0.63	0.72	0.52	0.52	0.55	0.58	0.57	0.61	0.73	0.58
128	8	0.48	0.54	0.64	0.74	0.52	0.52	0.55	0.57	0.57	0.63	0.72	0.59
128	16	0.48	0.53	0.64	0.74	0.52	0.51	0.55	0.57	0.56	0.63	0.73	0.59

Table 27: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on XCOPA (Ponti et al., 2020).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76
rank quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
64 8	0.34	0.34	0.36	0.34	0.35	0.37	0.36	0.34	0.33	0.33	0.34	0.34	0.37	0.33	0.35	0.35
64 16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.33	0.34	0.35	0.35	0.34	0.36	0.35
128 8	0.34	0.34	0.36	0.34	0.34	0.36	0.35	0.34	0.34	0.33	0.34	0.34	0.36	0.34	0.36	0.35
128 16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.33	0.34	0.35	0.36	0.34	0.36	0.35

Table 28: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on XNLI (Conneau et al., 2018).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76
rank	quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
8	4	0.32	0.33	0.34	0.33	0.33	0.33	0.32	0.34	0.33	0.33	0.34	0.33	0.31	0.33	0.33	0.33
8	8	0.34	0.34	0.35	0.34	0.35	0.35	0.36	0.33	0.35	0.32	0.34	0.33	0.37	0.34	0.36	0.34
8	16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.34	0.34	0.34	0.35	0.34	0.36	0.35
16	4	0.32	0.33	0.34	0.33	0.33	0.33	0.32	0.34	0.33	0.33	0.33	0.33	0.32	0.34	0.33	0.33
16	8	0.35	0.33	0.36	0.34	0.35	0.36	0.35	0.33	0.34	0.33	0.34	0.34	0.37	0.34	0.36	0.35
16	16	0.34	0.35	0.35	0.34	0.34	0.34	0.37	0.33	0.35	0.33	0.34	0.35	0.35	0.34	0.36	0.35
32	4	0.32	0.34	0.34	0.33	0.33	0.33	0.32	0.34	0.32	0.33	0.33	0.32	0.32	0.34	0.33	0.33
32	8	0.34	0.34	0.35	0.33	0.34	0.37	0.36	0.34	0.33	0.34	0.34	0.34	0.37	0.33	0.36	0.35
32	16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.33	0.33	0.35	0.35	0.34	0.36	0.35
64	4	0.32	0.34	0.34	0.33	0.33	0.33	0.32	0.34	0.33	0.33	0.33	0.32	0.31	0.34	0.33	0.33
64	8	0.35	0.34	0.35	0.34	0.35	0.37	0.36	0.34	0.35	0.32	0.34	0.34	0.36	0.33	0.35	0.35
64	16	0.34	0.35	0.34	0.34	0.34	0.36	0.37	0.33	0.35	0.33	0.34	0.35	0.35	0.34	0.36	0.35
128	4	0.32	0.33	0.35	0.33	0.34	0.33	0.33	0.34	0.33	0.33	0.32	0.32	0.31	0.34	0.33	0.33
128	8	0.35	0.34	0.36	0.34	0.36	0.36	0.37	0.33	0.33	0.33	0.34	0.35	0.36	0.33	0.36	0.35
128	16	0.34	0.35	0.34	0.34	0.34	0.36	0.36	0.33	0.34	0.33	0.34	0.35	0.35	0.34	0.35	0.34

Table 29: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on XNLI (Conneau et al., 2018).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76
rank	quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
8	4	0.32	0.33	0.34	0.33	0.33	0.33	0.32	0.34	0.33	0.33	0.34	0.33	0.32	0.34	0.33	0.33
8	8	0.34	0.34	0.35	0.34	0.35	0.34	0.35	0.34	0.34	0.33	0.34	0.34	0.35	0.32	0.38	0.34
8	16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.34	0.34	0.35	0.36	0.34	0.36	0.35
16	4	0.32	0.33	0.34	0.33	0.33	0.34	0.32	0.34	0.33	0.33	0.34	0.33	0.32	0.34	0.33	0.33
16	8	0.34	0.35	0.36	0.34	0.34	0.38	0.37	0.34	0.33	0.32	0.34	0.35	0.37	0.33	0.36	0.35
16	16	0.34	0.35	0.34	0.34	0.34	0.35	0.37	0.33	0.34	0.34	0.34	0.35	0.35	0.34	0.36	0.35
32	4	0.32	0.34	0.34	0.33	0.34	0.33	0.32	0.34	0.32	0.33	0.33	0.33	0.31	0.34	0.33	0.33
32	8	0.35	0.34	0.34	0.34	0.34	0.35	0.35	0.33	0.34	0.33	0.34	0.35	0.36	0.32	0.35	0.34
32	16	0.34	0.34	0.34	0.34	0.34	0.35	0.37	0.33	0.34	0.33	0.34	0.35	0.35	0.33	0.35	0.34
64	4	0.32	0.34	0.34	0.33	0.34	0.33	0.33	0.34	0.33	0.33	0.33	0.33	0.31	0.34	0.33	0.33
64	8	0.35	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.34	0.32	0.34	0.35	0.37	0.33	0.34	0.35
64	16	0.34	0.34	0.35	0.34	0.34	0.36	0.36	0.33	0.35	0.33	0.34	0.35	0.35	0.33	0.35	0.34
128	4	0.32	0.33	0.34	0.33	0.33	0.33	0.32	0.34	0.33	0.33	0.33	0.33	0.32	0.34	0.33	0.33
128	8	0.34	0.34	0.37	0.34	0.35	0.36	0.36	0.34	0.34	0.32	0.34	0.35	0.36	0.33	0.37	0.35
128	16	0.34	0.35	0.35	0.34	0.34	0.35	0.37	0.33	0.35	0.33	0.33	0.35	0.35	0.33	0.36	0.35

Table 30: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on XNLI (Conneau et al., 2018).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76

rank	quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
64	8	0.47	0.55	0.56	0.50	0.69	0.59	0.55	0.50	0.54	0.37	0.52	0.45	0.47	0.47	0.59	0.52
64	16	0.47	0.57	0.57	0.53	0.68	0.62	0.57	0.51	0.54	0.38	0.53	0.45	0.46	0.47	0.60	0.53
128	8	0.46	0.53	0.54	0.51	0.68	0.58	0.54	0.49	0.53	0.36	0.52	0.43	0.46	0.46	0.57	0.51
128	16	0.48	0.56	0.55	0.51	0.66	0.61	0.56	0.50	0.53	0.37	0.53	0.43	0.46	0.45	0.59	0.52

Table 31: Detailed performance of various ALPACA finetuned MISTRAL-7B models on XNLI (Conneau et al., 2018).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76

rank	quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
8	4	0.51	0.56	0.54	0.53	0.66	0.57	0.55	0.46	0.53	0.43	0.49	0.45	0.43	0.47	0.58	0.52
8	8	0.49	0.51	0.53	0.51	0.67	0.58	0.53	0.49	0.52	0.37	0.53	0.42	0.46	0.46	0.58	0.51
8	16	0.48	0.55	0.53	0.53	0.65	0.60	0.56	0.50	0.53	0.38	0.52	0.42	0.46	0.47	0.58	0.52
16	4	0.52	0.56	0.54	0.52	0.65	0.58	0.56	0.47	0.53	0.44	0.49	0.45	0.43	0.48	0.58	0.52
16	8	0.48	0.51	0.52	0.51	0.67	0.57	0.54	0.49	0.52	0.37	0.51	0.42	0.45	0.45	0.59	0.51
16	16	0.48	0.55	0.54	0.51	0.66	0.60	0.55	0.50	0.53	0.38	0.52	0.42	0.47	0.47	0.58	0.52
32	4	0.53	0.56	0.53	0.52	0.65	0.57	0.55	0.47	0.53	0.44	0.48	0.44	0.44	0.47	0.58	0.52
32	8	0.49	0.52	0.52	0.51	0.68	0.58	0.54	0.49	0.53	0.38	0.52	0.42	0.45	0.46	0.59	0.51
32	16	0.47	0.54	0.55	0.51	0.65	0.60	0.55	0.50	0.53	0.38	0.52	0.42	0.47	0.46	0.59	0.52
64	4	0.52	0.56	0.54	0.52	0.65	0.58	0.55	0.48	0.52	0.44	0.49	0.44	0.46	0.47	0.58	0.52
64	8	0.48	0.52	0.53	0.51	0.68	0.58	0.54	0.49	0.53	0.38	0.53	0.42	0.46	0.46	0.59	0.51
64	16	0.47	0.54	0.56	0.51	0.65	0.60	0.55	0.50	0.53	0.38	0.52	0.43	0.47	0.46	0.59	0.52
128	4	0.50	0.57	0.54	0.49	0.64	0.58	0.54	0.47	0.52	0.42	0.50	0.45	0.47	0.45	0.58	0.51
128	8	0.48	0.53	0.55	0.51	0.69	0.59	0.55	0.49	0.54	0.38	0.53	0.43	0.46	0.48	0.60	0.52
128	16	0.46	0.53	0.55	0.50	0.65	0.59	0.54	0.50	0.52	0.37	0.52	0.43	0.45	0.45	0.58	0.51

Table 32: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on XNLI (Conneau et al., 2018).

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
LLaMA-7B-Chat	0.39	0.45	0.45	0.39	0.56	0.50	0.50	0.37	0.48	0.33	0.35	0.40	0.36	0.41	0.45	0.43
LLaMA–3B-Chat	0.37	0.51	0.50	0.00	0.55	0.51	0.52	0.00	0.50	0.36	0.00	0.45	0.29	0.48	0.48	0.37
LLaMA-70B-Chat	0.35	0.55	0.58	0.41	0.63	0.55	0.55	0.31	0.55	0.39	0.35	0.49	0.42	0.49	0.51	0.48
Mistral-7B-Instruct	0.35	0.40	0.38	0.35	0.43	0.41	0.41	0.35	0.43	0.34	0.33	0.40	0.34	0.33	0.40	0.38
DV003	0.52	0.62	0.66	0.60	0.80	0.71	0.66	0.48	0.62	0.50	0.51	0.58	0.50	0.56	0.58	0.59
GPT-3.5-Turbo	0.59	0.64	0.67	0.65	0.76	0.70	0.68	0.55	0.62	0.56	0.54	0.63	0.49	0.61	0.62	0.62
GPT-4	0.73	0.77	0.79	0.79	0.85	0.79	0.79	0.72	0.74	0.71	0.69	0.76	0.68	0.74	0.75	0.75
TULRv6	**0.89**	**0.91**	**0.90**	**0.90**	**0.93**	**0.91**	**0.91**	**0.86**	**0.89**	**0.85**	**0.88**	**0.88**	**0.83**	**0.89**	**0.88**	**0.89**
BLOOMZ	0.61	0.47	0.54	0.47	0.67	0.61	0.61	0.57	0.53	0.50	0.44	0.43	0.50	0.61	0.57	0.54
XLM-R	0.77	0.83	0.82	0.81	0.89	0.84	0.82	0.76	0.79	0.71	0.77	0.78	0.72	0.79	0.78	0.79
mBERT	0.64	0.68	0.70	0.65	0.81	0.73	0.73	0.59	0.68	0.50	0.54	0.61	0.57	0.69	0.68	0.65
XGLM	0.46	0.49	0.46	0.49	0.53	0.46	0.49	0.47	0.49	0.45	0.47	0.45	0.43	0.48	0.49	0.47
mT5	0.73	0.79	0.77	0.77	0.85	0.80	0.79	0.71	0.77	0.69	0.73	0.73	0.68	0.74	0.74	0.75
PALM2	0.79	0.81	0.84	0.84	0.89	0.84	0.18	0.77	0.80	0.77	0.78	0.79	–	0.79	0.80	0.76
rank	quantisation	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
8	4	0.52	0.56	0.54	0.52	0.66	0.58	0.56	0.47	0.53	0.43	0.49	0.45	0.44	0.47	0.60	0.52
8	8	0.49	0.53	0.51	0.51	0.67	0.58	0.53	0.49	0.53	0.37	0.52	0.42	0.46	0.46	0.59	0.51
8	16	0.48	0.55	0.55	0.51	0.65	0.59	0.55	0.51	0.53	0.38	0.53	0.43	0.46	0.46	0.58	0.52
16	4	0.52	0.56	0.54	0.52	0.66	0.57	0.55	0.48	0.53	0.44	0.50	0.45	0.45	0.48	0.58	0.52
16	8	0.49	0.52	0.54	0.51	0.68	0.58	0.54	0.49	0.53	0.37	0.53	0.42	0.46	0.46	0.60	0.51
16	16	0.48	0.54	0.54	0.51	0.65	0.60	0.56	0.51	0.53	0.38	0.52	0.43	0.46	0.47	0.58	0.52
32	4	0.53	0.57	0.54	0.52	0.65	0.58	0.55	0.48	0.54	0.44	0.49	0.45	0.45	0.49	0.59	0.52
32	8	0.49	0.51	0.53	0.50	0.68	0.58	0.54	0.51	0.53	0.38	0.52	0.43	0.46	0.45	0.61	0.51
32	16	0.47	0.53	0.55	0.49	0.65	0.60	0.55	0.51	0.53	0.38	0.52	0.43	0.46	0.46	0.58	0.51
64	4	0.51	0.57	0.54	0.52	0.65	0.57	0.55	0.48	0.53	0.44	0.51	0.45	0.46	0.49	0.60	0.53
64	8	0.48	0.49	0.53	0.50	0.67	0.57	0.55	0.49	0.51	0.37	0.52	0.42	0.46	0.45	0.58	0.51
64	16	0.46	0.52	0.55	0.51	0.65	0.59	0.54	0.50	0.52	0.38	0.51	0.43	0.47	0.45	0.57	0.51
128	4	0.48	0.54	0.51	0.47	0.60	0.55	0.53	0.46	0.49	0.41	0.52	0.41	0.46	0.45	0.56	0.50
128	8	0.48	0.51	0.54	0.50	0.68	0.58	0.54	0.50	0.54	0.37	0.53	0.42	0.47	0.46	0.60	0.51
128	16	0.48	0.54	0.55	0.51	0.65	0.60	0.55	0.51	0.53	0.39	0.53	0.42	0.46	0.46	0.58	0.52

Table 33: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on XNLI (Conneau et al., 2018).

Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat	0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat	0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct	0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003	0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo	0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4	0.68	0.72	0.62	0.83	0.77	0.64	**0.76**	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6	**0.86**	**0.86**	**0.86**	0.90	**0.88**	**0.86**	–	**0.87**	**0.87**	**0.84**	**0.88**	0.79	**0.86**
BLOOMZ	0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R	0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT	0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2	0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5	0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
64	8	0.44	0.76	0.58	0.86	0.66	0.50	0.68	0.61	0.33	0.58	0.80	0.56	0.61
64	16	0.44	0.75	0.58	0.85	0.79	0.50	0.72	0.66	0.34	0.58	0.78	0.55	0.63
128	8	0.44	0.76	0.58	0.86	0.66	0.49	0.68	0.62	0.33	0.59	0.79	0.54	0.61
128	16	0.44	0.75	0.58	0.85	0.79	0.50	0.72	0.66	0.34	0.58	0.78	0.55	0.63

Table 34: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on XQuAD (Artetxe et al., 2020).

Model		ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat		0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat		0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat		0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct		0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003		0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo		0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4		0.68	0.72	0.62	0.83	0.77	0.64	**0.76**	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6		**0.86**	**0.86**	**0.86**	0.90	**0.88**	**0.86**	–	**0.87**	**0.87**	**0.84**	**0.88**	0.79	**0.86**
BLOOMZ		0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R		0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT		0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2		0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5		0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
8	4	0.41	0.73	0.50	0.83	0.76	0.45	0.70	0.62	0.24	0.50	0.75	0.53	0.59
8	8	0.44	0.76	0.58	0.86	0.67	0.50	0.68	0.62	0.32	0.59	0.79	0.56	0.61
8	16	0.44	0.75	0.59	0.85	0.79	0.50	0.72	0.66	0.34	0.58	0.78	0.55	0.63
16	4	0.41	0.71	0.50	0.83	0.75	0.45	0.70	0.62	0.24	0.50	0.75	0.53	0.58
16	8	0.43	0.67	0.58	0.86	0.68	0.50	0.69	0.62	0.33	0.59	0.79	0.56	0.61
16	16	0.45	0.68	0.58	0.85	0.69	0.50	0.69	0.62	0.34	0.58	0.78	0.55	0.61
32	4	0.42	0.72	0.50	0.83	0.75	0.45	0.70	0.62	0.24	0.51	0.75	0.53	0.58
32	8	0.44	0.67	0.58	0.85	0.69	0.49	0.69	0.63	0.33	0.60	0.80	0.55	0.61
32	16	0.45	0.68	0.58	0.85	0.69	0.50	0.69	0.63	0.34	0.58	0.78	0.55	0.61
64	4	0.42	0.73	0.51	0.83	0.76	0.45	0.70	0.62	0.25	0.51	0.76	0.54	0.59
64	8	0.44	0.76	0.58	0.86	0.69	0.50	0.71	0.63	0.33	0.58	0.80	0.55	0.62
64	16	0.45	0.68	0.58	0.86	0.68	0.50	0.70	0.63	0.34	0.58	0.79	0.55	0.61
128	4	0.42	0.74	0.51	0.83	0.76	0.45	0.71	0.62	0.25	0.52	0.76	0.54	0.59
128	8	0.45	0.71	0.59	0.86	0.72	0.49	0.73	0.65	0.33	0.58	0.80	0.56	0.62
128	16	0.46	0.69	0.58	0.86	0.69	0.50	0.71	0.64	0.34	0.58	0.79	0.55	0.62

Table 35: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on XQuAD (Artetxe et al., 2020).

Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat	0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat	0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct	0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003	0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo	0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4	0.68	0.72	0.62	0.83	0.77	0.64	**0.76**	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6	**0.86**	**0.86**	**0.86**	0.90	**0.88**	**0.86**	–	**0.87**	**0.87**	**0.84**	**0.88**	0.79	**0.86**
BLOOMZ	0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R	0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT	0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2	0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5	0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
8	4	0.41	0.73	0.50	0.83	0.76	0.45	0.70	0.62	0.24	0.50	0.75	0.53	0.59
8	8	0.45	0.76	0.58	0.86	0.69	0.50	0.69	0.63	0.33	0.59	0.80	0.55	0.62
8	16	0.45	0.75	0.58	0.86	0.79	0.50	0.72	0.67	0.34	0.58	0.78	0.55	0.63
16	4	0.41	0.73	0.50	0.83	0.76	0.45	0.70	0.62	0.24	0.50	0.75	0.53	0.59
16	8	0.44	0.76	0.58	0.86	0.70	0.49	0.70	0.63	0.33	0.59	0.80	0.55	0.62
16	16	0.44	0.75	0.58	0.85	0.79	0.50	0.72	0.67	0.34	0.58	0.78	0.55	0.63
32	4	0.41	0.74	0.50	0.83	0.76	0.45	0.70	0.62	0.24	0.51	0.75	0.53	0.59
32	8	0.44	0.77	0.58	0.85	0.68	0.51	0.68	0.62	0.33	0.59	0.79	0.54	0.61
32	16	0.45	0.75	0.58	0.86	0.79	0.50	0.73	0.67	0.34	0.58	0.79	0.55	0.63
64	4	0.41	0.74	0.51	0.83	0.76	0.45	0.71	0.62	0.24	0.51	0.76	0.53	0.59
64	8	0.45	0.77	0.58	0.86	0.77	0.51	0.74	0.65	0.35	0.58	0.80	0.56	0.63
64	16	0.45	0.75	0.58	0.86	0.80	0.50	0.74	0.67	0.34	0.58	0.78	0.56	0.64
128	4	0.41	0.74	0.50	0.83	0.76	0.45	0.70	0.62	0.24	0.50	0.75	0.53	0.59
128	8	0.43	0.76	0.59	0.86	0.69	0.49	0.70	0.62	0.33	0.59	0.80	0.55	0.62
128	16	0.45	0.75	0.58	0.86	0.79	0.50	0.73	0.67	0.34	0.58	0.78	0.55	0.63

Table 36: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on XQuAD (Artetxe et al., 2020).

Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat	0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat	0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct	0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003	0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo	0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4	0.68	0.72	0.62	0.83	0.77	0.64	0.76	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6	**0.86**	**0.86**	**0.86**	0.90	0.88	**0.86**	–	**0.87**	**0.87**	**0.84**	**0.88**	0.79	**0.86**
BLOOMZ	0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R	0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT	0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2	0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5	0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
64	8	0.72	0.84	0.76	0.91	0.89	0.69	**0.86**	0.77	0.62	0.72	0.87	0.73	0.78
64	16	0.72	0.84	0.76	0.91	0.88	0.70	0.86	0.77	0.64	0.72	0.88	0.73	0.78
128	8	0.71	0.83	0.76	0.91	0.88	0.69	0.86	0.77	0.62	0.71	0.88	0.72	0.78
128	16	0.72	0.84	0.77	0.91	**0.89**	0.70	0.86	0.77	0.63	0.71	0.88	0.73	0.78

Table 37: Detailed performance of various ALPACA finetuned MISTRAL-7B models on XQuAD (Artetxe et al., 2020).

Model		ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat		0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat		0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat		0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct		0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003		0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo		0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4		0.68	0.72	0.62	0.83	0.77	0.64	0.76	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6		**0.86**	**0.86**	**0.86**	0.90	0.88	**0.86**	–	**0.87**	**0.87**	**0.84**	0.88	0.79	**0.86**
BLOOMZ		0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R		0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT		0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2		0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5		0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
8	4	0.66	0.82	0.72	0.89	0.87	0.66	0.82	0.76	0.57	0.68	0.82	0.71	0.75
8	8	0.72	0.84	0.76	0.90	0.88	0.69	0.86	0.76	0.63	0.71	0.88	0.72	0.78
8	16	0.72	0.84	0.77	0.91	**0.89**	0.70	0.86	0.77	0.62	0.71	0.88	0.73	0.78
16	4	0.66	0.82	0.71	0.89	0.87	0.66	0.82	0.75	0.58	0.68	0.82	0.70	0.75
16	8	0.72	0.84	0.76	0.91	0.88	0.70	**0.86**	0.77	0.62	0.71	0.87	0.73	0.78
16	16	0.71	0.84	0.78	0.91	**0.89**	0.71	0.85	0.77	0.62	0.71	0.88	0.73	0.78
32	4	0.65	0.82	0.71	0.89	0.87	0.66	0.82	0.75	0.57	0.67	0.81	0.70	0.74
32	8	0.71	0.84	0.76	0.91	0.88	0.69	0.86	0.76	0.62	0.71	0.86	0.73	0.78
32	16	0.73	0.84	0.78	0.91	0.89	0.70	0.85	0.77	0.62	0.71	0.88	0.73	0.78
64	4	0.65	0.82	0.71	0.89	0.87	0.66	0.83	0.76	0.58	0.68	0.82	0.71	0.75
64	8	0.71	0.84	0.76	0.91	0.88	0.69	0.85	0.76	0.62	0.71	0.87	0.71	0.78
64	16	0.73	0.84	0.78	0.91	0.88	0.71	0.86	0.76	0.63	0.72	**0.88**	0.73	0.79
128	4	0.66	0.83	0.72	0.89	0.88	0.66	0.83	0.76	0.59	0.69	0.82	0.71	0.75
128	8	0.70	0.84	0.76	0.91	0.88	0.69	0.86	0.76	0.62	0.72	0.87	0.72	0.78
128	16	0.73	0.84	0.78	0.92	0.89	0.71	**0.86**	0.77	0.63	0.72	0.88	0.73	0.79

Table 38: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on XQuAD (Artetxe et al., 2020).

Model	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
LLaMA-7B-Chat	0.02	0.12	0.02	0.19	0.14	0.01	0.10	0.46	0.01	0.05	0.07	0.07	0.10
LLaMA–3B-Chat	0.02	0.11	0.02	0.17	0.13	0.01	0.09	0.04	0.01	0.01	0.07	0.06	0.06
LLaMA-70B-Chat	0.02	0.11	0.02	0.20	0.14	0.06	0.09	0.05	0.00	0.05	0.07	0.07	0.07
Mistral-7B-Instruct	0.18	0.28	0.11	0.34	0.31	0.18	0.32	0.27	0.12	0.23	0.31	0.14	0.23
DV003	0.37	0.55	0.32	0.77	0.62	0.20	0.58	0.29	0.12	0.45	0.42	0.36	0.42
GPT-3.5-Turbo	0.60	0.71	0.49	0.79	0.70	0.54	0.70	0.58	0.42	0.62	0.69	0.50	0.61
GPT-4	0.68	0.72	0.62	0.83	0.77	0.64	0.76	0.64	0.55	0.71	0.76	0.60	0.69
TULRv6	**0.86**	**0.86**	**0.86**	0.90	0.88	**0.86**	–	**0.87**	**0.87**	**0.84**	0.88	0.79	**0.86**
BLOOMZ	0.83	0.76	0.50	**0.92**	0.87	0.83	0.71	0.66	0.21	0.51	0.87	**0.82**	0.71
XLM-R	0.69	0.80	0.80	0.86	0.82	0.77	–	0.80	0.74	0.76	0.79	0.59	0.77
mBERT	0.61	0.71	0.63	0.83	0.76	0.59	–	0.71	0.43	0.55	0.69	0.58	0.65
PALM2	0.63	0.76	0.77	0.87	0.81	0.62	–	0.70	0.62	0.68	0.73	0.49	0.70
mT5	0.64	0.74	0.60	0.85	0.75	0.60	–	0.58	0.58	0.68	0.71	0.66	0.67
rank	quantisation	ar	de	el	en	es	hi	ro	ru	th	tr	vi	zh	avg
8	4	0.66	0.82	0.72	0.89	0.87	0.66	0.82	0.74	0.57	0.67	0.82	0.71	0.75
8	8	0.71	0.83	0.76	0.91	0.88	0.70	0.86	0.76	0.62	0.70	0.88	0.72	0.78
8	16	0.73	0.84	0.77	0.91	0.89	0.70	0.86	0.76	0.62	0.71	0.88	0.73	0.78
16	4	0.66	0.82	0.72	0.89	0.87	0.66	0.83	0.75	0.57	0.67	0.82	0.71	0.75
16	8	0.71	0.84	0.76	0.91	0.88	0.69	0.86	0.76	0.62	0.70	0.87	0.73	0.78
16	16	0.73	0.84	0.77	0.91	0.88	0.70	0.86	0.77	0.63	0.71	0.88	0.73	0.78
32	4	0.66	0.82	0.72	0.89	0.87	0.67	0.82	0.75	0.58	0.67	0.83	0.71	0.75
32	8	0.72	0.84	0.76	0.91	0.89	0.70	0.86	0.76	0.62	0.72	0.88	0.72	0.78
32	16	0.72	0.84	0.77	0.91	0.88	0.71	0.85	0.76	0.63	0.72	0.88	0.72	0.78
64	4	0.66	0.82	0.71	0.89	0.88	0.68	0.83	0.74	0.58	0.67	0.83	0.71	0.75
64	8	0.72	0.84	0.76	0.91	0.88	0.70	0.86	0.76	0.62	0.71	0.87	0.73	0.78
64	16	0.71	0.84	0.77	0.92	0.89	0.71	**0.86**	0.76	0.63	0.72	**0.88**	0.73	0.78
128	4	0.69	0.82	0.73	0.89	0.88	0.68	0.83	0.75	0.58	0.68	0.84	0.71	0.76
128	8	0.72	0.84	0.76	0.91	0.88	0.70	0.86	0.76	0.63	0.70	0.87	0.73	0.78
128	16	0.72	0.84	0.77	0.91	**0.89**	0.71	0.86	0.76	0.62	0.71	0.87	0.72	0.78

Table 39: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on XQuAD (Artetxe et al., 2020).

Model	ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat	**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo	0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4	0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5	0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2	0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
64	8	0.00	0.03	0.14	0.05	0.01	0.08	0.05	0.05
64	16	0.02	0.07	0.15	0.09	0.02	0.10	0.06	0.07
128	8	0.01	0.04	0.15	0.06	0.02	0.09	0.06	0.06
128	16	0.02	0.07	0.14	0.09	0.02	0.10	0.06	0.07

Table 40: Detailed performance of various ALPACA finetuned LLAMA-2-7B models on XLSum (Hasan et al., 2021).

Model		ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat		**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo		0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4		0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5		0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2		0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
8	4	0.00	0.03	0.15	0.05	0.00	0.13	0.05	0.06
8	8	0.02	0.04	0.15	0.06	0.02	0.09	0.06	0.06
8	16	0.02	0.26	0.14	0.08	0.02	0.10	0.06	0.10
16	4	0.00	0.03	0.15	0.05	0.00	0.13	0.05	0.06
16	8	0.02	0.04	0.15	0.06	0.02	0.09	0.05	0.06
16	16	0.02	0.26	0.14	0.08	0.02	0.10	0.06	0.10
32	4	0.00	0.03	0.15	0.05	0.00	0.11	0.05	0.06
32	8	0.01	0.04	0.15	0.06	0.01	0.10	0.06	0.06
32	16	0.02	0.26	0.14	0.08	0.02	0.10	0.06	0.10
64	4	0.00	0.04	0.15	0.05	0.00	0.12	0.05	0.06
64	8	0.01	0.04	0.15	0.06	0.01	0.09	0.06	0.06
64	16	0.02	0.26	0.14	0.07	0.02	0.11	0.06	0.10
128	4	0.00	0.03	0.15	0.05	0.00	0.13	0.05	0.06
128	8	0.01	0.04	0.15	0.06	0.02	0.08	0.06	0.06
128	16	0.02	0.26	0.14	0.08	0.02	0.10	0.06	0.10

Table 41: Detailed performance of various MULTIALPACA finetuned LLAMA-2-7B models on XLSum (Hasan et al., 2021).

Model		ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat		**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo		0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4		0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5		0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2		0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
8	4	0.00	0.03	0.15	0.05	0.00	0.13	0.05	0.06
8	8	0.02	0.04	0.15	0.05	0.01	0.09	0.06	0.06
8	16	0.02	0.07	0.14	0.09	0.02	0.10	0.06	0.07
16	4	0.00	0.03	0.15	0.05	0.00	0.13	0.05	0.06
16	8	0.01	0.04	0.15	0.05	0.02	0.10	0.06	0.06
16	16	0.02	0.07	0.14	0.09	0.02	0.10	0.06	0.07
32	4	0.00	0.03	0.15	0.05	0.00	0.11	0.05	0.06
32	8	0.02	0.04	0.15	0.06	0.01	0.09	0.06	0.06
32	16	0.02	0.07	0.15	0.09	0.02	0.10	0.06	0.07
64	4	0.00	0.04	0.15	0.05	0.00	0.12	0.05	0.06
64	8	0.02	0.04	0.15	0.05	0.01	0.09	0.07	0.06
64	16	0.02	0.07	0.14	0.10	0.02	0.10	0.06	0.07
128	4	0.00	0.04	0.15	0.05	0.00	0.13	0.05	0.06
128	8	0.02	0.04	0.15	0.06	0.01	0.10	0.07	0.06
128	16	0.02	0.07	0.14	0.10	0.02	0.10	0.06	0.07

Table 42: Detailed performance of various BACTRIAN-X-22 finetuned LLAMA-2-7B models on XLSum (Hasan et al., 2021).

Model		ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat		**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo		0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4		0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5		0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2		0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
64	8	0.01	0.29	0.10	0.24	0.00	0.02	0.05	0.10
64	16	0.01	0.29	0.10	0.25	0.00	0.01	0.05	0.10
128	8	0.01	0.29	0.09	0.25	0.00	0.02	0.05	0.10
128	16	0.00	0.29	0.11	0.25	0.00	0.01	0.05	0.10

Table 43: Detailed performance of various ALPACA finetuned MISTRAL-7B models on XLSum (Hasan et al., 2021).

Model		ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat		**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo		0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4		0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5		0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2		0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
8	4	0.00	0.30	0.08	0.24	0.00	0.00	0.06	0.10
8	8	0.00	0.29	0.08	0.25	0.00	0.02	0.06	0.10
8	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
16	4	0.00	0.30	0.07	0.24	0.00	0.00	0.06	0.10
16	8	0.00	0.29	0.08	0.24	0.00	0.02	0.06	0.10
16	16	0.00	0.29	0.08	0.25	0.00	0.02	0.05	0.10
32	4	0.00	0.30	0.08	0.23	0.00	0.00	0.06	0.10
32	8	0.00	0.29	0.08	0.26	0.00	0.01	0.05	0.10
32	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
64	4	0.00	0.30	0.08	0.23	0.00	0.00	0.06	0.10
64	8	0.00	0.30	0.08	0.25	0.00	0.02	0.05	0.10
64	16	0.00	0.28	0.10	0.25	0.00	0.02	0.05	0.10
128	4	0.00	0.30	0.08	0.23	0.00	0.01	0.06	0.10
128	8	0.00	0.29	0.09	0.25	0.00	0.01	0.06	0.10
128	16	0.00	0.29	0.08	0.25	0.00	0.02	0.05	0.10

Table 44: Detailed performance of various MULTIALPACA finetuned MISTRAL-7B models on XLSum (Hasan et al., 2021).

Model		ar	en	es	fr	hi	jp	zh	avg
LLaMA-70B-Chat		**0.35**	0.00	0.00	0.00	0.00	0.00	0.20	0.08
GPT-3.5-Turbo		0.25	0.26	0.21	0.26	0.24	0.26	0.22	0.24
GPT-4		0.23	0.27	0.20	0.23	0.24	0.29	0.31	0.25
mT5		0.31	**0.35**	**0.27**	0.23	**0.34**	**0.39**	**0.40**	**0.33**
PALM2		0.06	0.00	0.00	**0.31**	0.03	0.00	–	0.07
rank	quantisation	ar	en	es	fr	hi	jp	zh	avg
8	4	0.00	0.30	0.07	0.24	0.00	0.00	0.06	0.10
8	8	0.00	0.29	0.09	0.24	0.00	0.02	0.05	0.10
8	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
16	4	0.00	0.29	0.07	0.24	0.00	0.00	0.06	0.10
16	8	0.00	0.29	0.09	0.24	0.00	0.02	0.05	0.10
16	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
32	4	0.00	0.30	0.07	0.23	0.00	0.00	0.06	0.10
32	8	0.00	0.29	0.09	0.24	0.00	0.01	0.05	0.10
32	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
64	4	0.00	0.30	0.07	0.24	0.00	0.01	0.06	0.10
64	8	0.00	0.29	0.08	0.25	0.00	0.02	0.06	0.10
64	16	0.00	0.29	0.09	0.25	0.00	0.02	0.05	0.10
128	4	0.00	0.30	0.08	0.23	0.00	0.01	0.05	0.10
128	8	0.01	0.29	0.09	0.25	0.00	0.02	0.05	0.10
128	16	0.00	0.29	0.10	0.25	0.00	0.02	0.05	0.10

Table 45: Detailed performance of various BACTRIAN-X-22 finetuned MISTRAL-7B models on XLSum (Hasan et al., 2021).