Multilingual Instruction Tuning With Just a Pinch of Multilinguality

Anonymous ACL submission

Abstract

As instruction-tuned large language models (LLMs) gain global adoption, their ability to follow instructions in multiple languages becomes increasingly crucial. In this work, we investigate how multilinguality during instruction tuning of a multilingual LLM affects instruction-following across languages from the pre-training corpus. We first show that many languages transfer some instruction-following capabilities to other languages from even monolingual tuning. Furthermore, we find that only 40 multilingual examples integrated in an English tuning set substantially improve multilingual instruction-following, both in seen and unseen languages during tuning. In general, we observe that models tuned on multilingual mixtures exhibit comparable or superior performance in multiple languages compared to monolingually tuned models, despite training on 10x fewer examples in those languages. Finally, we find that diversifying the instruction tuning set with even just 2-4 languages significantly improves cross-lingual generalization. Our results suggest that building massively multilingual instruction-tuned models can be done with only a very small set of multilingual instruction-responses.

1 Introduction

Instruction tuning is a fundamental aspect of building modern general-purpose large language models (LLMs), involving fine-tuning a pre-trained model on pairs of instructions and corresponding responses (Mishra et al., 2022; Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022). For these models to be globally applicable, they must operate on a wide range of languages, yet, most instruction tuning datasets are typically limited to English. While curating naturally occurring instructions and responses for every language is challenging, cross-lingual transfer has emerged as a promising approach, in which a model is fine-tuned using one language, and acquiring similar abilities in another (Pires et al., 2019; Wu and Dredze, 2019; Artetxe and Schwenk, 2019; K et al., 2020; Conneau et al., 2020a,b). The ability to follow instructions for languages seen only at pre-training can significantly expand the applicability of LLMs, allowing them to be used by more people worldwide. In this work, we show that instruction-tuning of multilingual LLMs transfers across languages better than previously known, and that even minimal language diversity in the tuning set can further unlock instruction-following generalization to languages that are unseen during instruction tuning.

We investigate the effect of multilingual data on instruction-following across languages using an LLM pre-trained on hundreds of languages (Anil et al., 2023), and high-quality, open-ended instructions and responses (Zhou et al., 2023; Köpf et al., 2023) translated into 11 languages, across different families and writing systems. Initially, we examine the transferability of monolingual instruction tuning across different languages. Naturally, tuning using each language individually enhances performance within that language. Notably, we find that this also translates into instruction-following capabilities across other languages, and that tuning with English, Italian, or Spanish yields the best average multilingual performance.

Inspired by this result, we turn to ask how much multilingual data is required to improve multilingual instruction-following, while preserving English performance. We find that replacing even just 40 English training examples with multilingual examples, significantly improves instruction-following in those languages. Surprisingly, this small amount of language-diverse examples also improves performance for languages that are only seen during pre-training and are not represented in the instruction tuning set at all.

The next question we tackle is whether increasing the number of languages in the tuning set can

enhance generalization to new languages from the pre-training corpus. We find that tuning using a few languages enables better performance for languages unseen during tuning, compared to monolingual tuning with the same number of examples.

Finally, we test two potential factors that might influence the degree of cross-lingual transfer: language similarity and the amount of language-specific pre-training data, but find no significant correlations. Overall, our results provide recipes for multilingual instruction tuning that improves cross-lingual generalization, while preserving performance on English, under a fixed budget. In particular, we find that capable multilingual instruction-following models can be tuned even with a minimal amount of multilingual data.

2 Measuring Multilingual Instruction-Following

Our objective is to discover how multilinguality during instruction tuning affects general-purpose instruction-following across languages. We break this down to multiple questions, including how well can monolingual instruction tuning transfer to other languages, how many multilingual examples can enhance multilingual instruction-following while preserving English performance, and whether increasing the number of languages can result in improved cross-lingual generalization. In this section we elaborate on the data, evaluation protocol, models we use, and the human annotation process to ensure the models quality.

Data We use datasets of high-quality open-ended instructions and responses, rather than classic task-specific datasets. Our training data contains 1,000 English instructions and responses from LIMA (Zhou et al., 2023) and 3,640 from OpenAssistant[1] (Köpf et al., 2023). These examples resemble real world scenarios of users interacting with chatbots, with queries like "Can you explain Fermat's Last Theorem?" and "How to keep a dog hydrated?", that enable efficient tuning even with a small training set (Zhou et al., 2023). For evaluation, we use 617 instructions from AlpacaFarm (Dubois et al., 2023), originated from Self-Instruct (Wang et al., 2023), Vicuna (Chiang et al., 2023), Koala (Geng et al., 2023), and hh-rlhf (Bai et al., 2022).[2]

We use the Google Translate API[3] to translate the instruction-response pairs of the training set and the instructions of the evaluation set to 11 languages, creating parallel training and evaluation sets in Arabic, Chinese, Czech, English, Estonian, Finnish, Hebrew, Hindi, Italian, Russian, Spanish, and Swahili.[4] While translated data is different from naturally sourced data per language, it allows for more control as the data size and semantics are similar for all languages. A overview of the languages, their language codes, families and scripts is described in Table 2 in Appendix A.

Evaluation We conduct a side-by-side automatic evaluation protocol (Bubeck et al., 2023; Dubois et al., 2023; Dettmers et al., 2023; Gudibande et al., 2023; Zheng et al., 2023), in which an LLM assesses two responses for the same instruction, with the goal of identifying the superior one. We follow the common practice of presenting both responses to the model twice, alternating the order of the two responses (Zheng et al., 2023; Zhang et al., 2023). The exact prompt we use is shown in Figure 9 in Appendix B. We define a "win" for a certain response if the judge selects it twice irrespective of the order, and a "tie" if the model selects a different response for each order. We use a discounted-tie (Zhou et al., 2023) scoring method, in which a model receives a score of 1 for a win, 0.5 for a tie, and 0 for a loss. We average the scores of individual instructions to get the score over the evaluation set and present it in percentages. To validate that the LLM judge decisions align with human preferences across languages, we conduct a human annotation study and find good aggregated agreement scores of 79.5% for English, 77% for Spanish, and 76.5%, and 75% for Russian and Hebrew, receptively. Further details on validating the LLM judge are provided in Appendix D.

Instruction-Following Score Per Language Throughout this work we measure instruction-following per language by comparing the performance of a model that was tuned on some training set D, to a model that was monolingually tuned on the target language $\mathcal{L}$, by using the full training

[1]We focus on single-instruction/single-response interactions so we keep only the first prompt and response from conversations in OpenAssistant similarly to Li et al. (2023).

[2]We exclude AlpacaFarm's evaluation instructions from OpenAssistant, as we tune using its training set.

[3]https://cloud.google.com/translate/docs/reference/api-overview

[4]Languages are selected from Table 21 in Anil et al. (2023), describing the top-50 languages the model (§2) was pre-trained on.

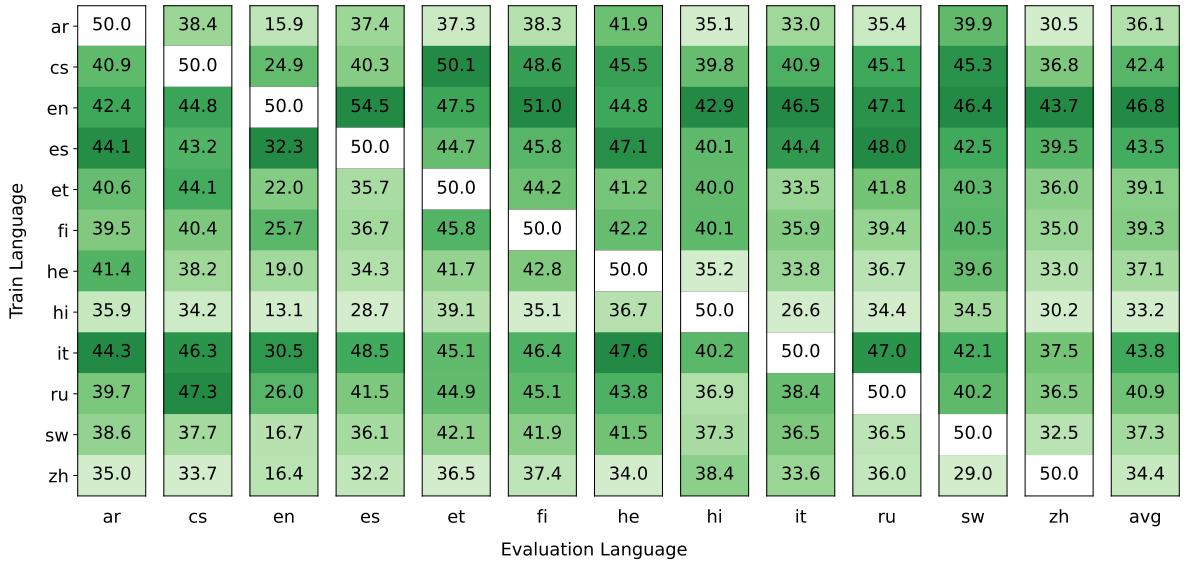

Figure 1: Per language instruction-following scores of models instruction-tuned on monolingual data. Each row represents a model tuned using a different language, and each column is an individual heatmap of the scores of all models on the same evaluation language. Scores are the discounted-ties weighted average of the side-by-side scores against the model tuned on the evaluation language. The scores along the diagonal are 50 as they are the result of comparing generations to themselves, and are excluded from the heatmap coloring.

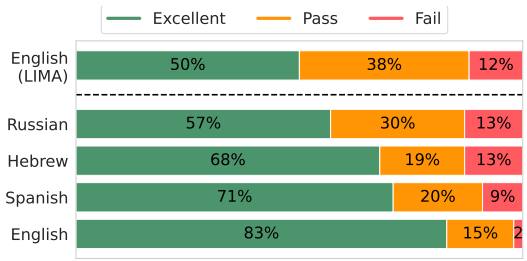

Figure 2: Human annotators rating distributions of models responses across languages. Each row describes evaluation in its corresponding language of the model tuned monolingually using that language. Numbers in the first row are reported by Zhou et al. (2023).

set in this language, $D_{\mathcal{L}}$. Formally, we define our instruction-following (IF) metric for language $\mathcal{L}$:

$$IF_{\mathcal{L}}(M_D) = S{\times}S(M_{D_{\mathcal{L}}}, M_D)$$

Where $S{\times}S(\cdot,\cdot)$ is the side-by-side protocol applied on $M_{D_{\mathcal{L}}}$ and M_D, which are the models instruction-tuned on $D_{\mathcal{L}}$ and D, respectively. A score of 0% means that M_D loses on all $\mathcal{L}$ instructions, and 50% means the performance of M_D and $M_{D_{\mathcal{L}}}$ in $\mathcal{L}$ are indistinguishable when aggregated over the evaluation set.

Model We use the PaLM 2 model family of Transformer-based (Vaswani et al., 2017) LLMs that were pre-trained on hundreds of languages (Anil et al., 2023). We use PaLM 2-S as our pre-trained model for all the instruction tuning experiments, and an instruction-tuned PaLM 2-L as the judge for the side-by-side evaluation. The training and inference hyperparameters we use are described in Appendix C.

Human Validation Our evaluation protocol relies on the quality of our monolingually tuned models. To validate their usage as high bar baselines in their respective languages, we conduct a human annotation study in 4 languages: English, Spanish, Russian and Hebrew. Namely, we sample 50 random instructions per language, and ask 2 native speakers to assign a score of excellent, pass, or fail (Zhou et al., 2023) to the responses generated by the model that was monolingually tuned using that language. Results in Figure 2 show that our tuned models indeed demonstrate strong instruction-following abilities. Notably, the scores across languages are similar or better than the reported numbers by Zhou et al. (2023) in English.[5]

[5]The differences can be attributed both to the pre-trained model and to the size of the instruction tuning dataset.

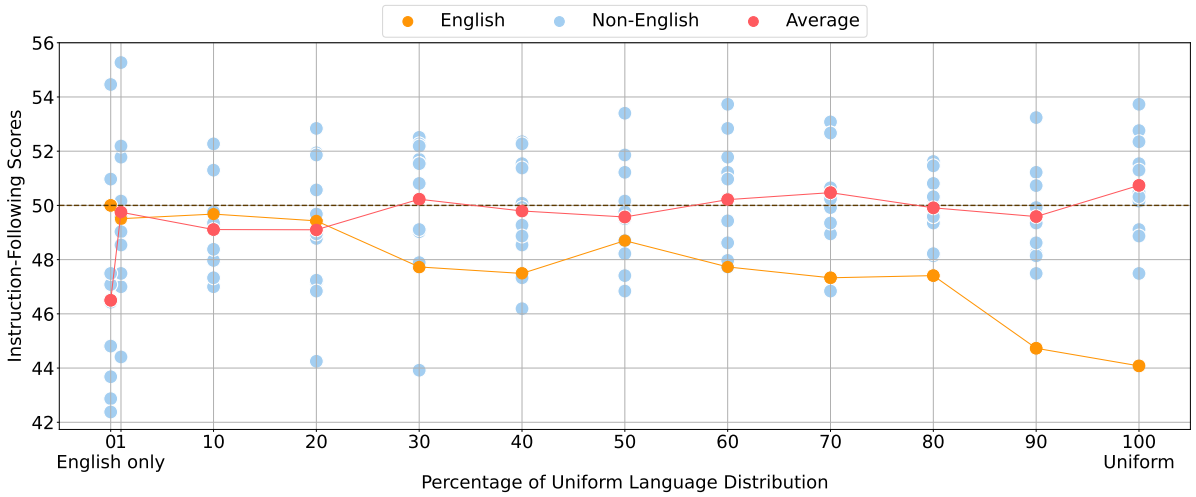

Figure 3: Instruction-following scores of models trained using when $P\%$ of the training set is distributed uniformly across 12 languages and an $(100 - P)\%$ is English only. Each X axis tick represents a tuning mixture, scores over individual non-English languages are in blue, and their averages are in red. English scores are in orange.

3 How Much Multilinguality Is Needed For Multilingual Instruction Tuning?

We now describe our controlled experiments, designed to quantify the effect of multilingual data during instruction tuning of multilingual LLMs, following the research questions defined in §2.

3.1 Monolingual Instruction Tuning Yields Multilingual Abilities

To explore zero-shot cross-lingual transfer of instruction tuning in multilingual LLMs, we tune models on a single language and evaluate them on all of the rest. We find that all of those models are able to transfer non-negligible instruction-following abilities to other languages.

Setup We instruction-tune 12 models, each one using the full train set in a different language. We generate responses using every such model to the evaluation instructions in all other languages. Finally, we calculate their per language scores as described in §2.

Results Figure 1 shows the results, where rows represent training languages and every column is an independent heatmap of the results over a single evaluation language. Most importantly, tuning using each single language yields a model with some multilingual instruction-following capabilities across languages. For context, even the model with the lowest average score, the one tuned on Hindi, achieves a score of over 30% in 9 out of 11 cases.[6] The model with the best average score is the one tuned on English, when Italian and Spanish also enable consistently high scores.

Notably, we manually inspect the generations and find that our tuned models consistently respond in the same language as their instruction, regardless of the language they were instruction-tuned on, in contrast with findings in previous work (Touvron et al., 2023a; Chen et al., 2023). We hypothesize that this comes from the multilingual nature of PaLM 2s' pre-training, compared to the more English-centric LLaMA (Touvron et al., 2023a), further details are in Appendix E. In addition to our main setup, we also compare the generations of these models to the ones of the pre-trained model that was not instruction-tuned. Results shown in Figure 10 in Appendix F further demonstrate that instruction tuning in every language separately, greatly improves instruction-following abilities across different languages.

3.2 A Few Dozen Examples Improve Multilingual Instruction-following

Naturally, multilingual tuning, as opposed to English-exclusive tuning under a fixed training examples budget, should result in better downstream performance for non-English languages, and might hurt performance on English. Therefore, we ask how many multilingual examples can improve the instruction-following abilities across languages,

[6] For example, a score of 30% can be obtained by wining 30% of the instructions and losing 70%, or by achieving a tie on 60% of the instructions and losing 40%.

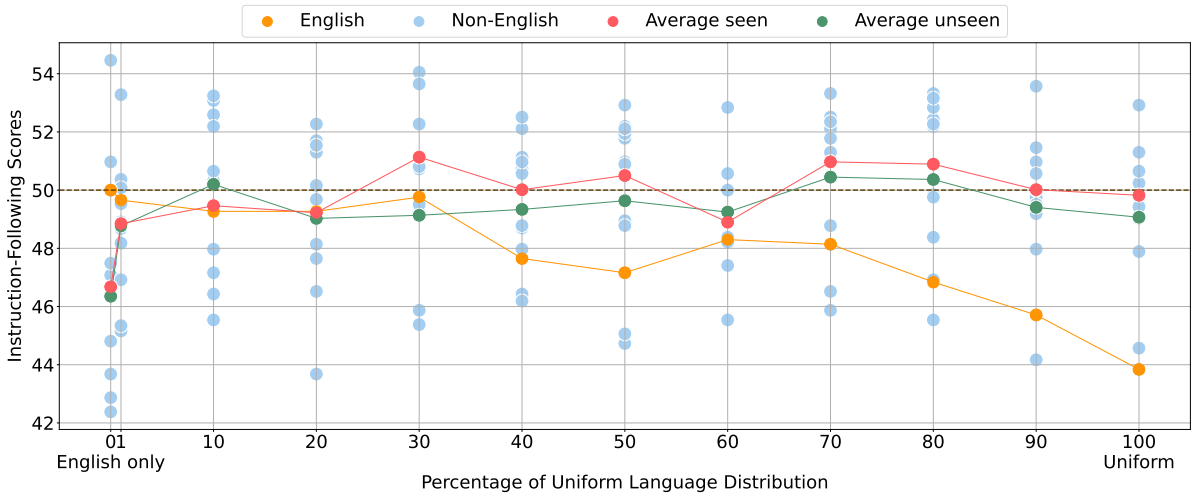

Figure 4: Instruction-following scores of models tuned when $P\%$ of the training set is distributed uniformly across 6 languages and an $(100 - P)\%$ is English only. Each X axis tick represents such a tuning set, scores over individual non-English languages are in blue and English scores are in orange. Average scores of the 5 non-English languages in the tuning set are in red, and the average scores of the 6 languages not seen during tuning are in green.

while preserving English performance. To that end, we tune models on subsets of the English examples combined with subsets of multilingual examples in different ratios. We find a significant boost in multilingual instruction-following abilities even when using just a few dozen multilingual examples.

Setup We create data mixtures with $P\%$ examples that are evenly split among all 12 languages, and the rest $(100 - P)\%$ English examples.[7] We create such a train set for every P from 10 to 100, incremented by tens, and also for $P = 1$, for which only 40 multilingual examples are included from across all 11 non-English languages, and the rest are English examples. Finally, we evaluate every tuned model on every one of the 12 languages as defined in §2.

Results Figure 3 visualizes the results. As expected, multilingual examples in the train set improve the score on their languages (Red), and diluting the number of English examples hurts the performance in English (Green). Notably, the significant multilingual improvement comes from replacing only 1% of the English examples by multilingual ones, which translates to 40 examples evenly distributed across the training languages. These results on the effect of such a small amount of language-diversity extend findings regarding task-diversity by Zhou et al. (2023), which demonstrated that a capable monolingual instruction-following model can be tuned using only 1,000 high-quality examples. A second trend is that these models often outperform their monolingually-tuned counterparts on the very language the latter were exclusively tuned on (blue markers above the 50 line). For example, the model tuned using the uniform set ($P = 100$) preforms similarly or better than the individual monolingually-tuned models in 8 of 12 languages, despite being trained on 12 times less instruction-response pairs for each language. This suggests that for some languages, multilingual tuning can enable better instruction-following abilities compared to a traditional monolingual tuning with the same number of examples.

3.3 A Few Dozen Examples Improve Cross-lingual Generalization

Combining the lessons on cross-lingual generalization from monolingual tuning and the effect of a small amount of multilingual examples from previous sections, we turn to examine how multilingual examples in the tuning set affect language generalization. Specifically, we conduct a similar experiment to the one in §3.2, this time using only half of the languages for tuning while the rest of languages are unseen. In line with the results from §3.2, we find that a very small amount of multilingual examples also improve performance on languages that were not in the tuning set.

Setup We repeat the setup from §3.2, this time with only English and 5 more languages: Arabic,

[7] Every example appears exactly once in every mixture, in a single language.

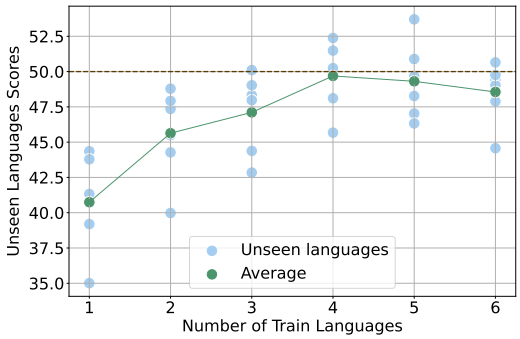

Figure 5: Instruction-following scores in Czech, Estonian, Hebrew, Hindi, Spanish, and Chinese of models instruction-tuned using various subsets of Arabic, English, Finnish, Italian, Russian, and Swahili. Blue markers are the average scores per evaluation languages across models tuned with the same number of languages. The averages of those individual languages scores are in green.

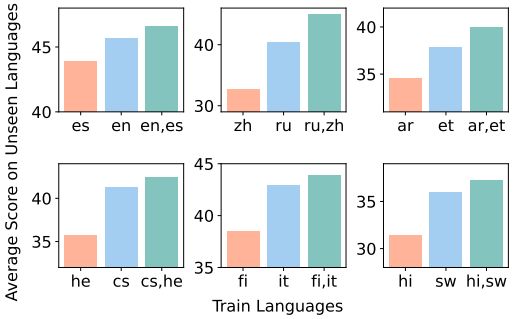

Figure 6: Average instruction-following scores of languages not seen during instruction tuning. For example, the top-left corner describes the scores of 3 models instruction-tuned on 100% Spanish, 100% English, and 50% Spanish and 50% English. The Y axis of this subfigure is the average score across all language excluding Spanish and English.

Finnish, Italian, Russian, and Swahili, and evaluate models again on all 12 languages.

Results Results in Figure 4 show similar trends to the ones in Figure 3. Specifically, the average score over non-English training languages (red) again improves very quickly, even with $P = 1$. Strikingly, this is also true for languages that the model has only seen during pre-training, and are not represented at all in the instruction tuning dataset (orange). This suggests that very few multilingual examples can not only improve performance for the languages of those examples, but also enable better cross-lingual instruction-following generalization.

3.4 Even a Small Number of Languages Improves Cross-Lingual Generalization

Given the results on the impact of a small number of multilingual *examples* from a fixed set of languages, we ask whether a small number of *languages* can also enhance cross-lingual generalization. We experiment with different numbers of languages in the tuning set and indeed observe that the transfer to languages only seen during pre-training improves from the very first additional languages.

Setup We instruction-tune models on a single language and up to 6 languages. At each step, we add a language to the tuning set, and split the same examples budget uniformly among the current set of languages. We use the 6 training languages from §3.3, and follow 3 different permutations that determine the order in which we add languages to the mix. These permutations are shown in Table 4 in Appendix G. We evaluate every model on each of the remaining 6 languages, and average scores per evaluation language across models that are tuned using the same number of languages.

Results Results on Figure 5 show that adding languages to the tuning set improves cross-lingual generalization. The average score (red) increases from tuning on monolingual data to tuning on bilingual data, and even more when using 3 and 4 languages, where the average score gets to almost 50. At that point, there is an indication for saturation, as more languages does not seem to improve transfer further. These findings demonstrate that diversifying the instruction tuning data with only a few different languages can improve cross-lingual transfer to new languages, only seen during pre-training.

Bilingual Tuning Sets To show this holds for even more combinations of languages, we randomly split all languages to pairs, and tune models using 50% of the examples in the one language and 50% in the other. We evaluate each of these models on the remaining 10 languages, and compare their score to the ones of the two models tuned using the full monolingual sets. Results on Figure 6 reveal that bilingual tuning helps generalize to new languages better than monolingual tuning.

4 Potential Factors of Transferability

Following the results from the previous sections, a natural question arises: what factors can predict the

Language	Code	Slavic Family	Script	Mutually Intelligible
Russian	ru	East	Cyrillic	-
Serbian	sr	South	Cyrillic	Croatian
Croatian	hr	South	Latin	Serbian
Slovenian	sl	South	Latin	-

Table 1: Languages used for language similarity experiment, along with their language code, subfamily, script, and the language they are mutually intelligible with.

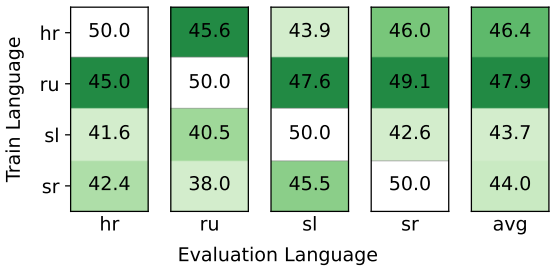

Figure 7: Instruction-following scores per language of models tuned monolingually. Each row represents a model trained using a different language, and each column is an individual heatmap of the scores of all models on the same evaluation language. The scores along the diagonal are excluded from the heatmaps coloring.

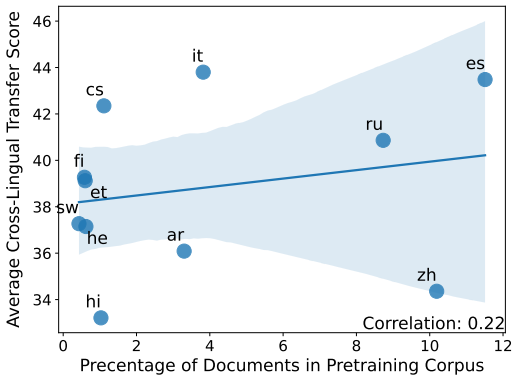

Figure 8: Weak Pearson correlation between the percentage of documents in the pre-training corpus (excluding English), and the average instruction-following score across languages for every training language. Blue area around the line is the confidence interval.

degree of cross-lingual transfer? We explore two immediate candidates. Initially, we examine the relation of various aspects of language similarity to transferability within language pairs. Next, we look into whether the proportion of language-specific data in the pre-training corpus correlates with the amount of cross-lingual transfer of instruction tuning using the given language.

4.1 Language Similarity

A intuitive hypothesis is that aspects of language similarity like the script or mutual intelligibility might affect the levels of instruction tuning cross-lingual transfer between languages. We test this using a case study of 4 Slavic languages, looking into possible effects of such aspects. However, we do not find a signal indicating these factors strongly correlate with cross-lingual transfer for this setting.

Setup We train models on monolingual versions of the data in Russian, Serbian, Croatian, and Slovenian, and evaluate their transfer to each other. These languages can be divided along several linguistic lines that are summarized in Table 1. First, Russian is East Slavic, and the rest are South Slavic. Second, Russian and Serbian both use the Cyrillic script, while Croatian and Slovenian use Latin. Moreover, Serbian and Croatian share a significant degree of mutual intelligibility.

Results Results are displayed on Figure 7. As shown, there is no a strong signal indicating that any of the aspects above is correlated with better mutual cross-lingual transfer. Russian tend to transfer instruction-following abilities best, and even though Russian and Serbian both use Cyrillic, it is Croatian that transfers capabilities to Russian better in our study. Moreover, Despite being largely mutually intelligible, Croatian and Serbian do not seem to share cross-lingual abilities more than the others. Our results align with recent findings that language similarity does not impact transferability or interference in machine translation given sufficient data and model capacity (Fernandes et al., 2023; Shaham et al., 2023).

4.2 Fraction of Data in Pre-training

A second possible predictor of the degree of cross-lingual transfer from a particular language is the extent to which the model was exposed to it during pre-training. Generally, a model's downstream performance on a specific language correlates with the fraction of data in that language in the pre-training corpus (Muennighoff et al., 2023). In contrast, Figure 8 suggests this is not necessarily the case for the cross-lingual transfer from a specific language. We find a weak Pearson correlation of 0.22 between the average cross-lingual score of each language and the number of documents in that language in pre-training corpus (Table 21 in Anil et al. (2023)).

5 Related work

Cross-lingual Transfer The success of the pre-training–fine-tuning paradigm (Devlin et al., 2019) ignited a new line of work on cross-lingual transfer. Pires et al. (2019) and Wu and Dredze (2019) showed that the multilingual variant of BERT can be fine-tuned on a specific task in one language and preform this task on another language, and Artetxe and Schwenk (2019) reported similar findings with a Recurrent Neural Network. Conneau et al. (2020a) introduced XLM-R, a multilingual pre-trained encoder with strong cross-lingual abilities. Phang et al. (2020) showed that intermediate training on an English task improves XLM-R's transfer across languages further, and Pfeiffer et al. (2020) suggested an adapter-based framework to improve cross-lingual and task generalization. Hu et al. (2020) proposed a benchmark for cross-lingual generalization consists of 40 languages across 9 NLP tasks.

K et al. (2020) found that the depth of the network matters for cross-lingual transfer, and Conneau et al. (2020b) showed that parameter sharing is more important than shared vocabulary. Choenni et al. (2023) delved into the influence of specific examples from the training data on the performance in other languages, and Malkin et al. (2022) investigated how pre-training BERT-based models using different language pairs affects cross-lingual downstream performance. Going beyond encoder-only models, Xue et al. (2021) proposed mT5, a multilingual variant of T5 (Raffel et al., 2020), and showed the significance of model scaling for cross-lingual transfer in generation tasks. Ye et al. (2023) explored trasferability in English-centric models (Touvron et al., 2023a) using four tasks.

In contrast to most cross-lingual transfer literature that is focused on task-specific fine-tuning, we explore trends of cross-lingual generalization for general-purpose instruction-following LLMs.

Multilingual Instruction Tuning Initially, works on instruction tuning (Mishra et al., 2022; Wei et al., 2022; Sanh et al., 2022) focused on cross-task generalization in English. Subsequently, a large body of work was dedicated to multilingual instruction tuning. Muennighoff et al. (2023) found that tuning models with English datasets enables zero-shot cross-lingual abilities to new languages. The authors also found that this holds for languages that the model has never intentionally seen during pre-training, and that multilingual training improves generalization to new tasks. Chen et al. (2023) investigated the effects of full parameter training vs low-rank adaptation (Hu et al., 2022) and monolingual vs multilingual instruction tuning using the Stanford Alpaca (Taori et al., 2023) data, machine translated into 5 languages. Lai et al. (2023) trained multilingual instruction-following models for 26 languages with reinforcement learning from human feedback (Ouyang et al., 2022), and Zhang et al. (2023) suggested instruction tuning LLMs by prepending the instruction and response translated into a pivot language (e.g English) to the response in the target language.

In this work, we consider transfer from monolingual instruction tuning from 12 languages, rather than exclusively on English. Furthermore, we examine multilingual instruction-following using an LLM pre-trained on hundreds of languages, which might be a key to unlocking more transfer to languages not represented during tuning. Importantly, we unveil the potential of just a small amount of language diversity in the instruction tuning set for this cross-lingual generalization.

6 Conclusion

We demonstrate that cross-lingual transfer offers a promising avenue for building multilingual instruction-following LLMs. Our findings across different languages suggest that even monolingual instruction tuning using only one language can result in improved instruction-following capabilities in other languages. Moreover, incorporating even a small set of a few dozen multilingual examples can significantly enhance instruction-following performance for both the languages the model is tuned on, and ones that were only seen during pre-training. Additionally, training on such multilingual datasets achieves comparable or even superior performance compared to monolingual tuning for some languages. We observe a similar trend when exploring the effect of total number of languages in the tuning set, as even splitting the train set to only two languages improves generalization to new languages, compared to monolingual tuning. These findings pave the way for efficient and scalable development of multilingual LLMs capable of understanding and following instructions across languages with minimal multilingual supervision.

7 Limitations

Limitations of our work include the use of translation for expanding datasets to multilingual settings, the number of languages we evaluated on, and number of models we experimented with. We now discuss each of them.

Translated data One limitation of our work is that our data is translated using the Google Translate API, and not originally sourced by native speakers. Automatic translation is inherently imperfect and may introduce noise to the tuning sets. However, translation also allows to for a controlled setup with parallel data, in which the content of all training and evaluation examples is the same for all languages.

Number of languages A second limitation is that we use 12 languages in our main experiments (§3), with 3 additional languages in the language similarity experiment (§4.1). Clearly, multilingual instruction-following models need to successfully operate in many more languages, and we leave work on scaling this number to future work.

Number of models Lastly, we experiment with PaLM 2, and results may vary with different LLMs. Nevertheless, our focus on PaLM 2 highlights the potential of multilingual pre-training for future advancements in LLMs.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023. Monolingual or multilingual instruction tuning: Which makes a better alpaca.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. 2023. Scaling laws for multilingual neural machine translation.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. 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.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

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`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023. Plug: Leveraging pivot language in cross-lingual instruction tuning.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.

A Languages

The languages we use, their language families, scripts ,and language codes are shown in Table 2.

Language	Code	Family	Script
Arabic	ar	Afro-Asiatic	Arabic
Chinese	zh	Sino-Tibetan	Chinese
Czech	cs	Indo-European	Latin
English	en	Indo-European	Latin
Estonian	et	Uralic	Latin
Finnish	fi	Uralic	Latin
Hebrew	he	Afro-Asiatic	Hebrew
Hindi	hi	Indo-European	Devanagari
Italian	it	Indo-European	Latin
Russian	ru	Indo-European	Cyrillic
Spanish	es	Indo-European	Latin
Swahili	sw	Niger-Congo	Latin

Table 2: Languages used in our main experiments.

B Side-By-Side Evaluation

Figure 9 shows the prompt given the the LLM judge for the side-by-side evaluation.

Below is an instruction and two answers. Choose your preferred answer, which can be subjective.

The instruction:
{instruction}

Answer1:
{response 1}

Answer2:
{response 2}

Which one is better, Answer1 or Answer2?
Only write a single digit as your answer, '1' for Answer1 or '2' for Answer2. Do not add any explanation.

Figure 9: Side-by-side evaluation prompt.

C Training and Inference Details

We now describe the hyperparameters we use in our experiments. We tune every model for 2,000 steps, using a fixed learning rate of 1e-5, a batch size of 128, and a dropout rate of 0.05. We limit inputs to 1,024 tokens and targets to 512 tokens. We sample a development set of 250 examples from every training set and select the checkpoint based on the development RougeL (Lin, 2004) score. During inference, we generate responses of up to 512 tokens using nucleus sampling (Holtzman et al., 2020) with $p = 0.9$ and temperature of 0.7. For the judge, we use greedy decoding to generate the ID of the better response (1 or 2).

D Judge-Human Agreement

To measure PaLM 2-L agreement with human judgments across language, we conduct a human annotation process on four languages, English, Spanish, Russian, and Hebrew. For every language we sample 50 instructions and let two native speakers select the better response out of two options, similarly to the task we assign the LLM judge (Figure 9). We always present the response by the model that was monolingually tuned using the evaluation language, alongside a response by model selected at random from the of the monolingually tuned ones described in §3.1. The agreement score on a single instruction is 1 if the LLM judge and human agree, 0.5 if exactly one of them selects a tie, and 0 if each selects a different response (Zhou et al., 2023). Table 3 shows the results. Overall, the LLM judge agreement with humans is strong for

Language	Human-Model	Human-Human
English	79.5	85.0
Spanish	77.0	80.0
Russian	76.5	79.0
Hebrew	75.0	82.0

Table 3: Judges agreement scores per language.

all four languages, yet there is some room of 2.5-7 points from inter human agreement in all languages. As expected, the models' highest agreement with humans is in English with 79.5%,. In the rest of the languages the agreement is a few points lower.

E Response Language

When a user prompts a model in a specific language, they usually expect to receive a response in that same language. However, pre-trained LLMs often respond in a different language than the language of their prompt (Touvron et al., 2023a; Chen et al., 2023). This poses a challenge also for evaluation of open-ended queries, since those are commonly evaluated with an LLM-as-a-judge (Zheng et al., 2023) protocol, and the judges often ignore whether the response language match the prompt language, even when instructed not to (Chen et al., 2023). Usually, this is handled by forcing the lowest score to such response (Chen et al., 2023), which does not account for all cases.[8] To verify our trained models respond in the same language as their prompt, we manually annotate the language of responses to evaluation instructions in all languages. For every language, we randomly sample 20 responses from the pool of models tuned monolingually in other languages, to end up with a total of 240 generations from various models. We find that 239 responses are in the same language as the prompt, as desired. This is a major difference in the behavior of our PaLM 2-based instruction-tuned models and the commonly used (Chen et al., 2023) LLaMA-based ones (Touvron et al., 2023a,b). We hypothesize this stems from the multilingual emphasis in the pre-training of PaLM 2, compared to the more English-centric LLaMA.

F Comparison to The Base Model

The scores of models of model instruction tuned monolingually compared to the pre-trained model that was not instruction tuned, as opposed to our main evaluation setup, are shown in Figure 10. As evident, instruction tuning the model on each of the languages separately unlocks instruction-following abilities across all languages.

[8] For example, a response in English to a prompt in French can still be very helpful, or when the prompt is a request for translation or code.

1	2	3	4	5	6
fi	fi,en	fi,en,ru	fi,en,ru,it	fi,en,ru,it,sw	all six
sw	sw,it	sw,it,ar	sw,it,ar,en	sw,it,ar,en,fi	all six
it	it,fi	it,fi,en	it,fi,en,ar	it,fi,en,ar,ru	all six

Table 4: Subsets of languages used to tune models for the experiment described in Section 3.4. Each cell represents a version of the training set, for which all examples are uniformly split between the languages in that cell.

G Languages Permutations

We use 3 different permutations of 6 languages to determine the order in which we add languages to the tuning set in the experiment described Section 3.4. The permutations are displayed in Table 4.

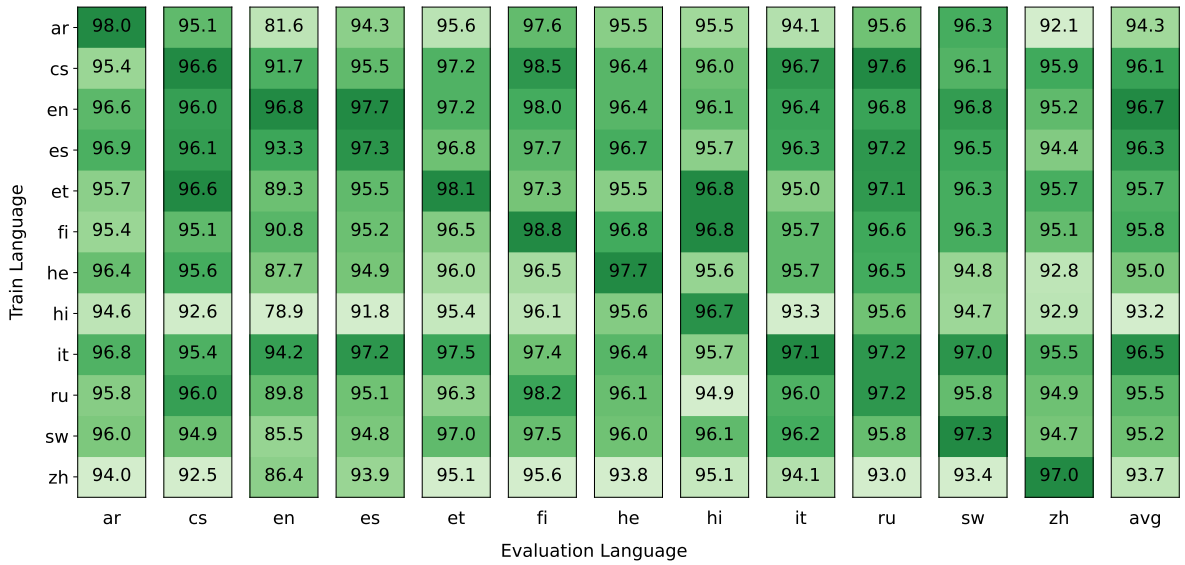

Figure 10: Per language instruction-following comparisons of models instruction-tuned on monolingual data to the pre-trained model that was not instruction tuned. Each row represents a model tuned using a different language, and each column is an individual heatmap of the scores of all models on the same evaluation language. Scores are the discounted-ties weighted average of the side-by-side scores against the pre-trained model.