Investigating Multilingual Instruction-Tuning: Do Polyglot Models Demand for Multilingual Instructions?

Anonymous ACL submission

Abstract

The adaption of multilingual pre-trained Large Language Models (LLMs) into eloquent and helpful assistants is essential to facilitate their use across different language regions. In that spirit, we are the first to conduct an extensive study of the performance of multilingual models on *parallel*, *multi-turn instruction-tuning* benchmarks across a selection of the most-spoken Indo-European languages. We systematically examine the effects of language and instruction dataset size on a mid-sized, multilingual LLM by instruction-tuning it on parallel instruction-tuning datasets. Our results demonstrate that instruction-tuning on parallel instead of monolingual corpora benefits cross-lingual instruction following capabilities by up to 4.6%. Furthermore, we show that the *Superficial Alignment Hypothesis* does not hold in general, as the investigated multilingual 7B parameter model presents a counter-example requiring large-scale instruction-tuning datasets. Finally, we conduct a human annotation study to understand the alignment between human-based and GPT-4-based evaluation within multilingual chat scenarios.

1 Introduction

LLMs have a significant impact on the daily work of many, as they are practical to use and assist in solving natural text problems ranging from creative writing to math problems. One of the primary reasons for their fast adoption as assistants is their facilitated usage by simply instructing the model to conduct a specific task. The training of such an assistant involves multiple stages of model training. First, an extensive, compute-intensive pre-training over large document corpora is conducted where the model is typically trained to predict the next token in a sequence. While pre-trained LLMs already exhibit significant capabilities, there is a discrepancy between their pre-training objective and the users' need for an instruction-following model. Therefore, in the second step, fine-tuning is performed where the model is trained to follow the users' instructions within a chat environment. The second step is crucial for the model to solve complex, multi-turn user requests.

With the availability of strong open-source English-centric models (Touvron et al., 2023), many English-centric instruction-tuning datasets emerged (Wang et al., 2023; Longpre et al., 2023; Zhou et al., 2023). While there are adoptions of monolingual English models for other languages (Uhlig et al., 2024), the choice and training of tokenizers is critical for the fair representation of languages (Petrov et al., 2023; Ali et al., 2023) within pre-trained models, as otherwise underrepresented languages are not only more resource intensive regarding training and inference, but also the performance for those languages drop. A fundamental problem is the availability of appropriate open-source, multilingual datasets and benchmarks for training and assessing instruction-tuned LLMs. Here, especially the lack of multi-turn multilingual benchmarks targeting instruction-tuned models represents a major gap, as previous instruction-tuned multilingual models are only evaluated on zero- or few-shot, single-turn, academic benchmarks targeting pre-trained LLMs (Muennighoff et al., 2023; Holmström and Doostmohammadi, 2023). However, it is essential to evaluate the multilingual instruction-following capabilities of the model on instruction benchmarks to realistically assess the helpfulness of a model as a chat assistant.

To the best of our knowledge, there is no study investigating multilingual instruction-tuning focusing on multilingual multi-turn user request performance where the base model has been pre-trained with a large amount of multilingual data where English is represented only by around 42%. We tackle this research gap by translating *MT-Bench* into the parallel benchmark *MT-Bench-X* and systematically investigate how the language and size

of instruction datasets impact the instruction-tuning of pre-trained, mid-sized multilingual LLMs for the Germanic and Italo-Western language family, including English, German, French, Italian, and Spanish, on this novel benchmark dataset.

To answer the research question, whether multilingual models pre-trained with a substantial amount of data for each language require instruction-tuning in all target languages to show competitive instruction-following capabilities across target languages, we make the following contributions:

- Creation of Lima-X, a high-quality, complex, parallel corpus comprising 1030 instructions for each English, German, French, Italian, and Spanish (Section 3).
- Creation of MT-Bench-X, a parallel, multilingual, human-curated evaluation dataset for evaluating instruction-tuned LLMs (Section 4).
- Multilingual instruction-tuning study with a focus on multilingual multi-turn user request performance (Section 5).
- Correlation analysis of the agreement levels between humans and machine on MT-Bench-X (Section 6).

2 Related Work

This section provides an overview of instruction-tuning datasets and aspects important for their utilization.

2.1 Instruction-tuning

Template-based Datasets Several English-focused instruction-tuning datasets have been introduced to broaden the scope of tasks and response formats by incorporating diverse sets of instructions (Iyer et al., 2022; Longpre et al., 2023). Primarily, many of these datasets revolve around Natural Language Processing (NLP) benchmarks that are refined through the application of either single or multiple prompt templates for responses and requests (Longpre et al., 2023; Muennighoff et al., 2023).

Synthetic Datasets An alternative approach involves extending only requests of NLP benchmarks by templates, but let sophisticated instruction-tuned models predict responses (Zhang et al., 2023). Examples here are *OASST* (Köpf et al., 2023) and *LIMA* (Zhou et al., 2023). The latter introduces the *Superficial Alignment Hypothesis* (Kirstain et al., 2022; Zhou et al., 2023). It states that only a few examples per task or instruction format are required to teach a LLM the response style. At the same time, most of the capabilities and knowledge are acquired during pre-training. While gaining great performance advancements with instructional data ranked by user preferences (Uhlig et al., 2024; Lai et al., 2023), this study focuses on the role of language mixtures during instruction-tuning.

Multilingual Datasets Muennighoff et al. (2023) translate classical NLP tasks , which were refined by prompt templates to create the *xP3mt* dataset. With experiments involving the dataset the authors indicate, that fine-tuning solely in English is adequate for a multilingual pre-trained LLM to adapt and perform well across various tasks in other pre-trained languages. However, these results were evaluated solely on downstream evaluation tasks for pre-trained LLMs and not on evaluation schemes developed for evaluating instruction-tuned models. On the other hand, Holmström and Doostmohammadi (2023) translate and evaluate instruction-tuning datasets for Swedish and their results indicate translated instructions significantly improve zero-shot performance of models and strong foundation in the target language benefits model performance, which contradicts the findings of Muennighoff et al. (2023). This discrepancy might be introduced by the lack of response diversity (Li et al., 2023) of template-based datasets and their dissimilarity to human created and generated responses as Table 1 suggests.

Bactrian-X (Li et al., 2023) is based on machine translated user requests, whereas corresponding responses were generated by OpenAI's `gpt-3.5-turbo`.

2.2 Evaluation

Pre-training Benchmarks Most often multilingual benchmarks, such as XCOPA (Ponti et al., 2020) and XWinograd (Tikhonov and Ryabinin, 2021) are utilized (Muennighoff et al., 2023; Li et al., 2023; Lai et al., 2023) for evaluating instruction-tuned models across languages. While these benchmarks measure specific aspects of *pre-trained* LLMs by accuracy regarding a gold truth often only spanning only a few words, they fail to capture the complex diversity instruction responses may offer (Zheng et al., 2023).

Dataset	Source	Samples per Language					∅ Words	
		EN	DE	FR	IT	ES	User	Sys.
xP3mt (Muennighoff et al., 2023)	template	32.7M	0	5M	0	5.1M	147	47
HH-RLHF (Bai et al., 2022)	gen.	112k	13	10	8	9	118	67
OASST (Köpf et al., 2023)	gen.	20.2k	1.5k	1.1k	359	13.4k	20	114
Bactrian-X (Li et al., 2023)	gen.	67k	67k	67k	67k	67k	27	90
LIMA (Zhou et al., 2023)	human	1k	4	0	0	2	39	419

Table 1: Language distribution for different instruction datasets for English (EN), German (DE), French (FR), Italian (IT) and Spanish (ES).

Instruction-tuning Benchmarks With *MT-Bench*, Zheng et al. (2023) developed an automatic benchmark, that encompasses the diversity of instruction responses, by utilizing *LLMs-as-a-judge*. Despite the availability of recent alternatives (Liu et al., 2023; Ferron et al., 2023; Ye et al., 2023), *MT-Bench* was broadly adopted by the community[1].

Despite these dataset releases, we utilize the same translation and quality assurance pipeline for all target languages, to allow for the same quality across translated benchmarks.

The concurrent work of the Aya Project (Singh et al.; Üstün et al.), introduces the *Aya Evaluation Suite*, which covers multilingual open-ended generation evaluation across 101 languages. While their prompts are suited for the conversational setup, a key difference to *MT-Bench-X* is that it only covers single turns.

2.3 Demarcation

While works exist addressing multilingual fine-tuning, our work differs from others in central aspects:

- We conduct our instruction tuning based on a pre-trained model that has been trained with a substantial amount of data for each language, has been trained with a large number of overall tokens (1T tokens), and relies on a fair tokenizer (Petrov et al., 2023; Ali et al., 2023). This ensures that we obtain reliable results in our multilingual setting.

- We investigate whether the structural formal of an instruction tuning dataset needs to be represented parallelly in each language, has to be split across languages, or should be monolingual.

- We evaluate our models on manually curated multilingual benchmarks that specifically assess the model's instruction-following capabilities.

3 Multilingual Instruction-Tuning Data

To investigate the defined research questions, we require high-quality parallel instruction-tuning datasets of different sizes. While there exist multilingual instruction datasets, the distribution of languages is highly skewed towards English, as Table 1 reveals[2] or contain shorter, less complex responses (Muennighoff et al., 2023). An exception here is Bactrian-X (Li et al., 2023), a large-scale, parallel instruction-tuning dataset.

Therefore, we select Bactrian-X (Section 3.1) as a large-scale multilingual instruction-tuning dataset and translate a smaller English-only instruction-tuning dataset *LIMA* (Zhou et al., 2023) into our target languages (Section 3.2).. For both datasets, we created different multilingual compositions (Section 3.3)[3].

3.1 Bactrian-X

The large-scale instruction-tuning dataset *Bactrian-X* (Li et al., 2023) was created by translating the original English instructions and generating responses by ChatGPT (`gpt-3.5-turbo`). We selected English, German, Italian, French and Spanish as target languages. We keep 95% (63667 samples per language) of the data as the training dataset and use the remaining 5% as the validation dataset (3350 samples per language).

[1] https://paperswithcode.com/dataset/mt-bench

[2] The language was classified by FastText `lid.176.bin` https://fasttext.cc/docs/en/language-identification.html

[3] The license details of our created datasets are described in Appendix D

3.2 Lima-X

Each sample in *LIMA* is highly curated, which is one benefit of its manageable size of 1030 samples.

Despite the creation of a validation set with high standards of curation by Zhou et al. (2023), the authors did not publish their validation dataset. Simply sampling the validation dataset from a training data split might remove samples providing important learning signals that are potentially not redundant within the remaining few samples. We thus adapt the curation steps and create a novel validation dataset, that is described in the Appendix A.

As we focus on Indo-European languages in our study, we chose to utilize DeepL as a translator performing well in these languages (Yulianto and Supriatnaningsih, 2021; Jiao et al.; Hidalgo-Ternero, 2020). We translate *LIMA* and the novel validation dataset into German, French, Italian, and Spanish. Before translating, we manually reviewed all 1030 training instances and marked the ones that could lead to problematic translations. The reasons here could be i.) mixed language usage in text ii.) code snippets, where code comments should be translated into other languages, but control statements not iii.) samples which were written entirely in a different language other than English and iv.) cultural aspects of English that are not transferable to the target language, e.g., *Here is a joke about numbers: Why was the number 10 scared? Because 7 8 9*. We mark 66 such cases in total and investigate whether DeepL can handle those for German. The *LIMA* dataset has 12 entries of non-English language, e.g., Portuguese, Spanish, or German, which are not problematic for DeepL. However, variable names in code snippets were partially translated. Furthermore, riddles, jokes, and poems are not directly translatable, which we see as a downside of the translation approach. We mark the erroneous cases.

3.3 Language Mixtures

Additionally, we compose multilingual variants of the translated monolingual datasets in our five target languages that make up *Lima-X* and our language selection of *Bactrian-X*. This involves creating `ENDEFRITES`, which encompasses all five languages, effectively increasing the dataset size five-fold. Additionally, we create a variant called `sampled`, maintaining the same semantics of the questions as in the monolingual original but distributed equally across the five languages within the dataset.

4 Multilingual Benchmarks

To evaluate the multilingual instruction-following capabilities of the models, a comprehensive multilingual benchmark for our target languages is indispensable. Thus, we created, based on the existing benchmark MT-Bench (Section 4.1), our multilingual variant MT-Bench-X (Section 4.2) that is translated into the five target languages. We employed MT-Bench-X to conduct a machine evaluation and a human evaluation (Section 4.3).

4.1 MT-Bench

For evaluating instruction-tuned models, human evaluation is considered gold standard. However, with *MT-Bench* (Zheng et al., 2023) an automatic evaluation alternative, that correlates with human judgment was developed for English. For automation *MT-Bench* utilizes LLMs-as-a-judge. The benchmark consists of 80 high-quality, two-turn user requests across 8 categories, whereas complex categories come along with reference answers. A LLM-as-a-judge is then prompted to assess model responses either in a pair-wise mode i.e. comparing two model responses to determine the better answer or a tie, or in a single scoring mode, where a score between 1 to 10 is to issue. In both modes, the model is also asked to explain it's decision. The pair-wise mode allows to check for positional bias by prompting the judge the same task twice but with reversed model response positions. For both modes, the judgment is generated by greedy search. The benchmark allows to reduce the cost of evaluation, as the authors showed the correlation to human evaluation agreement levels for English. The benchmark covers a diverse set of use-cases including *Writing*, *Math*, *Coding*, *Reasoning* and *Extraction* among others. To answer the research question above, we prompt GPT-4 with the translated prompts, whereas the English original of Zheng et al. (2023) was utilized. Thus, the focus of evaluation with *MT-Bench* is to assess "the quality of the response provided by an AI assistant", especially in terms of "helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response", as quoted from the prompt to user and machine.

4.2 MT-Bench-X

Similarly to the translation of *Mulima-X* we chose DeepL as translation engine to translate the ques-

tions, reference answers and judge prompts of *MT-Bench* from originally English to German, Spanish, Italian and French. Along with the original English *MT-Bench*, this leads to a novel multilingual benchmark called *MT-Bench-X*, whereas publishing details are in the Appendix D.

We investigate the performance of DeepL as translation engine in the Appendix B. While we consider DeepL as appropriate choice as a translation tool, there are still problematic cases, that we let manually edit for correctness and correct wording for both question and references across all languages with at least a graduate and fluent in the corresponding language. For German 31 cases were edited, many of them minor, for French 36, for Spanish 37 and for Italian 72. While French and German was correctly translated into the polite form, the Italian personal pronouns within user requests were translated into plural, which made many corrections necessary. Furthermore, for programming related tasks, in some cases variable names and control sequences are translated. Other aspects noticeable when going through the questions and relevant for evaluation, are the requirement of i) translation capabilities of the LLM to evaluate e.g. from Chinese to the translated language, and ii) up-to-date knowledge e.g. mentioning of GPT-4. Additional to the user requests and references, we also translate the prompts within *MT-Bench* to not mix languages during evaluation with *MT-Bench-X*. Through the manual correction of the translated *MT-Bench-X* dataset, we offer a high-quality instruction-tuning evaluation benchmark resource to the community.

4.3 Evaluation Approach

We utilize the currently best model available, GPT-4, which was shown to correlate best to human evaluation for English (Zheng et al., 2023). Furthermore, it was reported that GPT-4 is proficient in the languages we target in our study (Jiao et al.; OpenAI, 2023). We provide an user interface inspired by Zheng et al. (2023), as can be seen in Figure 7 in the Appendix E.3, to let human expert judges rate answers of models in a pair-wise response comparison. Given a random question, we first set the first turn of each model response against each other and let the user choose between the options i) Assistant A is better, ii) Assistant B is better, iii) Tie, iv) both answers are not helpful or v) to skip this turn. To reduce evaluation time, we let the second turn directly follow in the same manner. During the design of the *MT-Bench*, Zheng et al. (2023) minimised the risk for several biases which could occur when using a LLM as a judge. To omit positional bias, we randomly select the display side for each model newly for each turn.

5 Multilingual Performance Analysis

We first describe the experimental setup in Section 5.1, followed by the evaluation of single score ratings by *GPT-4-as-a-judge* (Section 5.2) and pair-wise evaluation by *GPT-4-as-a-judge*. We conclude this section with a qualitative analysis in Section 5.3.

5.1 Experimental Setup

For answering the question whether a mix of languages is needed for multilingual fine-tuning or if monolingual tuning suffices, we conduct several fine-tunings with the datasets described in Section 3. This includes, instruction-tuning on each monolingual dataset i.e. English, German, French, Italian and Spanish and both variants of the multilingual datasets i.e., `sampled` and `ENDEFRITES`. To also clarify the effect of instruction dataset size, we conduct fine-tunings for both *Lima-X* as well as *Bactrian-X*.

We use a multilingual 7B model, that closely follows the Llama 2 (Touvron et al., 2023) architecture, further details are specified in the Appendix C. The model is fine-tuned by using either the *Lima-X*, or the *Bactrian-X* dataset. We follow a similar hyper-parameter setting to Zhou et al. (2023), i.e., we utilize a learning rate of 1e-5 with linear decrease, use a batch size of 64, utilize the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 0.1. The number of epochs is determined based on the validation loss resulting in seven epochs for the *Lima-X* dataset and three epochs for *Bactrian-X*.

5.2 The Impact of Language

Single Language Benchmark With *GPT-4-as-a-judge* (Zheng et al., 2023), we display summarised results within Figure 1. It shows the average absolute *MT-Bench-X* scores for each model variant and evaluation language across categories. Further fine-grained results per category of all our model variants across all languages within *MT-Bench-X* in the Appendix E.1.

We note the single language fine-tunings `Bactrian-DE`, `Bactrian-IT` and `Bactrian-ES`

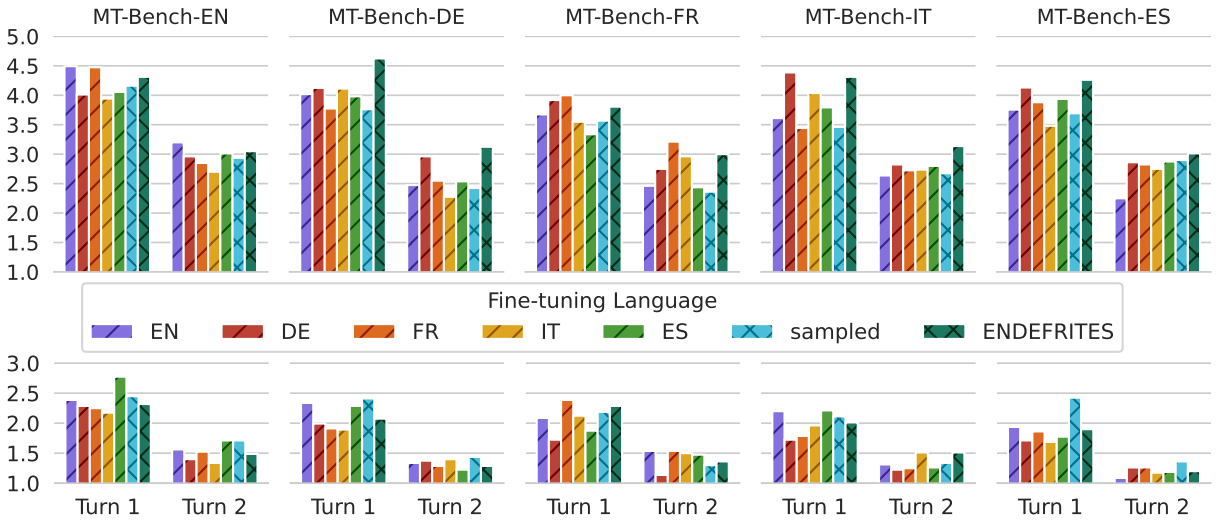


Figure 1: GPT-4-as-a-judge single evaluation average scores for each language mix dataset variant on MT-Bench-X. Top: *Bactrian-X*. Bottom: *Lima-X*.

are within their respective evaluation language inferior compared to `Bactrian-ENDEFRITES` for both turns. For `Bactrian-EN` this is not the case. We assume this is due to the large fraction of English data within the pre-training corpus. Also `Bactrian-FR` shows to outperform `Bactrian-ENDEFRITES` within *MT-Bench-FR*. In most cases fine-tuning for one single language is not the optimal configuration, even when it is aimed for optimizing performance regarding that specific language.

Superficial Alignment Hypothesis As evident from Figure 1, fine-tuned models based on the larger instruction-tuning dataset *Bactrian-X* generally outperform models based on the smaller *Lima-X* dataset. All models trained on *Lima-X* also show weak performance on the absolute measure. With this result, we show that the *Superficial Alignment Hypothesis* (Kirstain et al., 2022; Zhou et al., 2023), which states only a few examples per task or instruction-format are needed to teach a LLM the style of the response, while the capabilities and knowledge are to very large parts already gained during pre-training, does in general not hold for mid-sized LLMs.

Cross-lingual Benchmark As can be inferred from Figure 2, for the language mix strategy `sampled`, multi-lingual instruction-tuning improvements for *Lima-X* are notable, but the opposite for *Bactrian-X*. The inconsistency within these results might come from the number of samples per language, which is five times as small within *sampled*

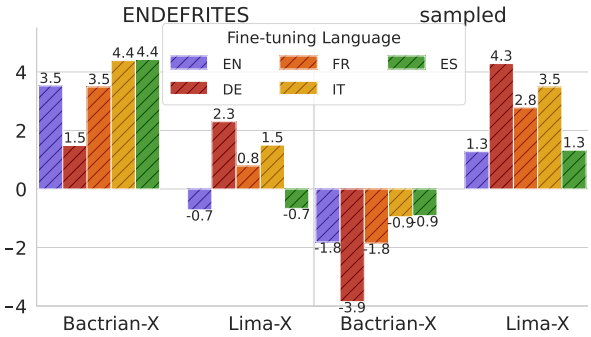


Figure 2: Percentage improvement for turn one averaged across *MT-Bench-X* languages of models fine-tuned on parallel mixed language instruction-tuning datasets (`ENDEFRITES`, `sampled`) over single language fine-tunings.

compared to the full monolingual dataset. Here, *Lima-X* only contains 1030 samples per language i.e. 206 samples per language within `sampled`.

However, the models based on the full-sized, parallel language mix `ENDEFRITES` show performance improvement across datasets, when compared to their monolingual counterparts. Hereby the maximum performance gain of `Bactrian-ENDEFRITES` is 4.1% compared to `Bactrian-EN` and also `Bactrian-IT`. This corresponds to 0.41 points on the MT-Bench scale ranging from 1 to 10. We therefore conclude that training with full-sized parallel multilingual datasets increases the cross-lingual instruction-tuning performance, while equal-sized mix-language datasets are inconsistent in their performance gain, presumably due to the decreased amount of total samples per language.

Multi-turn Evaluation Furthermore, across all dataset variants scores of the second turn shows to be lower compared to scores of first turns in Figure 1. This is expected, as *Bactrian-X* contains no multi-turn examples and *Lima-X* only 30 multi-turn examples.

5.3 Qualitative Analysis

For monolingual and multilingual models we visualize radar plots within the corresponding *MT-Bench-X* language in Figure 5 in the Appendix E.1. As visible in Figure 5b, already the small instruction-tuning dataset LIMA, translated into German (*Lima-DE*), improves the scores on *MT-Bench-DE*, compared to the multilingual pre-trained model. Also apparent is the consistent under-performance across models and datasets regarding the categories Reasoning, Math, Coding and Extraction. We assume this shows the lack of learning these capabilites during pre-training and either improved datasets, more pre-training tokens or very large-scale high-quality instruction-tuning datasets might be necessary to improve.

For a further qualitative analysis, we inspect the *MT-Bench-EN* answers of the models trained on `Bactrian-ENDEFRITES` and `Bactrian-DE` in Table 6 and Table 7 in the Appendix F. While the multilingual fine-tuned model shows a format and placeholders as one would expect, the `Bactrian-DE` shows incorrect formatting. We also noted that responses of `Bactrian-ENDEFRITES` are often more elaborate compared to monolingual `Bactrian-X` models.

6 Human Evaluation

Human evaluation is the gold standard for evaluating output of generative models, as responses can be highly divers and tasks may require a high degree of creativity to be solved. Furthermore, when conducting human evaluation it is important to mitigate for subjective ratings by including a large set of expert annotators.

To conduct the correlation analysis for *MT-Bench-DE*, we translate the prompts provided by Zheng et al. (2023) into German, see Appendix E.2.

To evaluate model pairs with *GPT-4-as-a-judge*, we first inspect potential limitations of utilizing *GPT-4-as-a-judge* for German text. We observe a high level of positional bias for the categories *Stem*, *Humanities* and *Writing* across, as shown in Table 2. For the following correlation analysis we mitigate the effect of positional bias by the substitution of missing values through results of a following run, where possible. Albeit judgment generation in *MT-Bench-X* is conducted by greedy search and the evaluation runs were executed immediately one after another, we can mitigate the positional bias by up to 16.25%.

Category	Positional Bias [%]
Stem	30.00
Humanities	27.50
Writing	20.00
Extraction	16.25
Roleplay	15.00
Coding	12.50
Reasoning	7.50
Math	6.25

Table 2: Percentage of positional bias occurrences averaged across three evaluation runs.

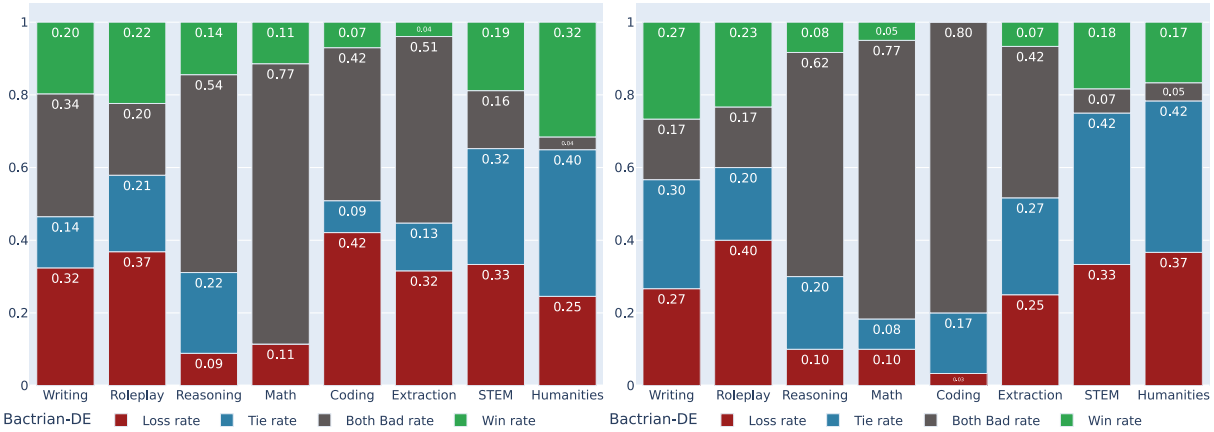
We averaged the results of three evaluation runs for the models `Bactrian-DE` and `Bactrian-ENDEFRITES` on *MT-Bench-DE* and display the results in Figure 3b.

As can be seen from the win rate, loss rate, tie rate and "both bad" rate in Figure 3b, `Bactrian-DE` looses more frequently against `Bactrian-ENDEFRITES`, except for the *Reasoning* category.

For the categories *Math*, *Reasoning* and *Coding* underperformance was already shown with the single evaluation scores. As can be seen in Figure 3b, especially for the categories *Math*, *Reasoning* and *Coding* the model performance is insufficient and thus a performance comparison is infeasible. We hypothesize this shows a gap of capabilities learned during pre-training.

To compare the results of the pair-wise automatic evaluation with *GPT-4* to human preferences, we conduct a human evaluation, as described in the Appendix E.3. Within Figure 3a it is evident that human evaluators tend to vote less often for "Tie" and "Both Bad".

This also results in a less high agreement between human and machine, as is visible from Table 3. Only 39% agreement between humans and GPT-4 for turn 1 and 56% agreement between humans for evaluation is observable. To incorporate the agreement by chance, we calculate the Fleiss' kappa scores among humans, resulting in the coefficient of 0.37 (p-value: 3×10^{-8}) on 66 samples



(a) `Bactrian-DE` vs. `Bactrian-ENDEFRITES`. Pair-wise MT-Bench-DE quality assessment by humans with both bad.

(b) `Bactrian-DE` vs. `Bactrian-ENDEFRITES`. Pair-wise MT-Bench-DE quality assessment by GPT-4 with both bad.

Figure 3: In-depth MT-Bench-DE quality assessment by GPT-4.

Turn	1		2	
Category	GPT-4	Hum.	GPT-4	Hum.
All	39_{59}	56	40_{62}	52
RCEMR[a]	50_{60}	57	46_{71}	58
Coding	47_{80}	61	11_{36}	33
Extraction	51_{48}	48	29_{81}	62
Humanities	0_{88}	77	44_{46}	66
Math	62_{92}	85	70_{88}	77
Reasoning	43_{54}	60	70_{83}	66
Roleplay	46_{45}	33	33_{70}	41
Stem	20_{60}	54	26_{42}	26
Writing	16_{49}	46	18_{51}	29

[a] *RCEMR* describes the agreement when only considering Roleplay, Coding, Extraction, Math and Reasoning

Table 3: Percentage of voting agreement between Human-GPT-4 and Human-Human. The subscripted numbers denote the upper limit of agreement defined through the agreement between human votes and the majority-vote of humans.

for the first turn and 0.36 (p-value: 6×10^{-8}) on 62 samples for the second turn. This indicates fair agreement between annotators.

We assume the low scores are effects of the difficulty of the evaluation task itself, as even human judges agree only to 56% to each other. Due to the disagreement among human judges only the upper bound of 59% agreement can be reached for human-GPT-4 agreement.

Interestingly, *Humanities*, *Writing* and *Stem* contribute significantly to the disagreement level of Human-GPT-4. We attribute this to the positional bias, which was especially observable within categories that involve creativity and thus are more subjective to assess (cf.Table 2). Of the 242 model responses where either one or the other model was selected as preference, human judges prefer to 62.8% the model responses of `Bactrian-ENDEFRITES`, which further supports our findings in Sec. 5.2.

7 Conclusion & Future Work

We see this work as a fundamental step towards supportive multilingual assistants. We comprehensively examined fine-tuned models on parallel, multi-turn instruction-tuning benchmarks across a selection of major Indo-European languages. Our findings highlight the benefits of instruction-tuning on parallel datasets, showcasing improvements of up to 4.6% in cross-lingual instruction-following capabilities compared to monolingual corpora. Additionally, our research challenges the *Superficial Alignment Hypothesis*, showing that extensive instruction-tuning datasets are necessary for mid-sized multilingual models. We identify disparities between human evaluations and those generated by GPT-4 in multilingual chat scenarios.

We illuminate these challenges, emphasizing the need for future research to address them. Additionally, we recognize the need to explore the impact of multilingual multi-turn dataset variants, which we leave as an avenue for future exploration. By addressing these challenges head-on, we can improve the performance of generative assistants in real-world communication contexts, advancing the field of natural language processing for practical applications.

8 Limitations

While our study offers valuable insights into instruction-tuning for multilingual LLMs, it is essential to acknowledge several limitations that may impact the generalizability and completeness of our findings.

Firstly, our research does not aim to push the boundaries of state-of-the-art performance. Instead, we focus on exploring the effectiveness of different instruction-tuning settings in guiding pre-trained multilingual LLMs to follow instructions within multi-turn conversation datasets.

Secondly, due to resource constraints, we conducted single-score evaluations for each model variant across various languages in the *MT-Bench-X* dataset only once. While this approach provided initial insights, it limited our ability to calculate comprehensive statistical measures like mean and standard deviation.

Additionally, while alignment and preference learning are crucial aspects of LLM development, our study concentrates solely on the preceding step of multilingual instruction-tuning.

Moreover, our research scope is confined to languages within the Germanic and Italo-Western language families due to resource constraints. Consequently, the generalizability of our findings to languages from more distant language families remains to be determined. Despite these limitations, our study lays the groundwork for exploring whether multilingual instruction-tuning benefits languages beyond those examined in this research, opening avenues for further investigation and refinement of multilingual LLM fine-tuning methodologies.

9 Ethical And Broader Impact

Instruction-following LLMs offers an efficient way of solving natural language problems by simply instructing the model to perform the tasks. With our work we highlight the importance of investigating the multilingual aspect throughout the creation process of helpful LLMs, as this becomes an important feature for democratizing this technology. While this allows users to become proficient in various areas, pre-trained and instruction-tuned models are not restricted out-of-the-box to a certain set of content and do not follow a specific set of values. Thus an important next step is to investigate the generalizability of the alignment to human curated values embedded within moderated datasets across multiple languages.

References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2023. Tokenizer choice for LLM training: Negligible or crucial? *CoRR*, abs/2310.08754.

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 B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2078–2100. Association for Computational Linguistics.

Carlos Manuel Hidalgo-Ternero. 2020. Google translate vs. deepl. *MonTI. Monografías de Traducción e Interpretación*, pages 154–177.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-english languages: A case study on swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics, NoDaLiDa 2023, Tórshavn, Faroe Islands, May 22-24, 2023*, pages 634–642. University of Tartu Library.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.

Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. 2022. A few more examples may be

worth billions of parameters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1017–1029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *CoRR*, abs/2304.07327.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu 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, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 318–327. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *CoRR*, abs/2305.15425.

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, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning.

Alexey Tikhonov and Max Ryabinin. 2021. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3534–3546, Online. Association for Computational Linguistics.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Matthias Uhlig, Sigurd Schacht, and Sudarshan Kamath Barkur. 2024. Phoenix: Open-source language adaption for direct preference optimization.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. FLASK: fine-grained language model evaluation based on alignment skill sets. *CoRR*, abs/2307.10928.

Ahmad Yulianto and Rina Supriatnaningsih. 2021. Google translate vs. deepl: a quantitative evaluation of close-language pair translation (french to english). *AJELP: Asian Journal of English Language and Pedagogy*, 9(2):109–127.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

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. *CoRR*, abs/2306.05685.

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. *CoRR*, abs/2305.11206.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya Model: An Instruction Fine-tuned Open-Access Multilingual Language Model.

A LIMA Validation Dataset Creation

We utilise the open-source project *stackexchange-dataset*[4] to scrape 18 different stackexchange topics (bioinformatics, biology, buddhism, chemistry, coffee, cooking, crafts, datascience, diy, earthscience, fitness, health, math, mythology, physics, poker, softwareengineering, travel). Despite already filtering for quality by stackexchanges scoring method , we end up with 62021 Question Answering (QA) pairs. Additionally, we filter answers by phrases such as "my", "as mentioned", "stack exchange", "referenced", "figure", "image", among others, to exclude examples not written in the style of a helpful assistant or referencing images, which cannot be represented in our unimodal models. We also filter by the length of QA pairs i.e. only allowing pairs which count more words than 1200, but do not exceed 4096 words. Additionally, we filter by consistent language across question and answer and perform near deduplication with Shingling, MinHashing, and LSH over the *LIMA* training dataset split. In total we reduce the 62021 examples to only 84, which we then carefully inspected and manually curate by rewriting or deleting samples. This leads to final 52 samples, which is roughly the size of the validation dataset reported by Zhou et al. (2023).

B Translation Quality

Most similar to our benchmark translation efforts is the dataset *MT-Bench-TrueGerman*. The authors of *MT-Bench-TrueGerman* [5], report translation problems when using `GPT-4` as a translation engine. To assess the translation quality of *MT-Bench-X*, we compare their findings with our translations by DeepL. While `GPT-4` can translate across various languages, it falls short compared to specialized translation engines such as DeepL. We showcase this in Table 4, by comparing the failure cases reported by *MT-Bench-TrueGerman* authors.

DeepL offers a more realistic translation than GPT-4 for the anglicism problem and we find the translation of simile accurate. With the exception of the translation errors due to intentionally grammatically incorrect sources we cannot support the findings of *MT-Bench-TrueGerman*.

[4] https://github.com/EleutherAI/stackexchange-dataset

[5] https://huggingface.co/VAGOsolutions/SauerkrautLM-7b-v1

C Pre-trained Model

To investigate multilingual instruction-tuning performance, we require the pre-trained model to have been i) trained on multilingual data including our target languages, ii) trained with a fair tokenizer, i.e., an equal amount of tokens per language was used for the tokenizer training, and iii) trained on at least 1T tokens to ensure a minimal set of learned aptitudes as perquisite for minimal instruction-following capability. To the best of our knowledge, only two existing, openly available model families are multilingual European ones. This includes BLOOM (Scao et al., 2022) and Nemotron [6]. However, BLOOM was not pre-trained on German data and only on 366B tokens for 46 languages, and for Nemotron, no details about the tokenizer training, nor details about the dataset language composition are available.

Thus, we adopt a multilingual LLM with 7B parameters pre-trained on 1T tokens. The pre-training datasets exhibit an English-dominated share of all 24 European languages (1T token dataset: 43.88% EN, 8.65% FR, 7.63% ES, 8.48% DE, 4.64% IT and 4.51% Code) and are filtered and deduplicated. The tokenizer was trained on a dataset where each of the 24 languages contributed equally to support each of these languages fairly.

D Licenses

The created *LIMA-X* datasets are licensed by CC BY-NC-SA (Lima-X) or stricter as required by Zhou et al. (2023) and the used *Bactrian-X* dataset was released under CC BY-NC 4.0 are thus both non-commercial, but usable as intended by the source datasets.

We license the created resource *MT-Bench-X* under Apache License 2.0. It is consistent with the intended use of the source dataset *MT-Bench* (Zheng et al., 2023).

E Evaluation

E.1 Single Score Evaluation

With the radar plots in Figure 5, we compare differences of monolingual and multilingual instruction-tuned models across all languages within *MT-Bench-X*. *GPT-4-as-a-judge* was utilized to rate the model responses for each individually and turn-based on a range from 1 to 10. The highest observed score for the pre-trained model is

[6] https://developer.nvidia.com/nemotron-3-8b

Problem	Original (Zheng et al., 2023)	GPT-4 MT-Bench-TrueGerman	DeepL MT-Bench-DE
Anglicism	Now you are a machine learning engineer ...	Jetzt bist du ein Maschinenlern-Ingenieur. ...	Jetzt sind Sie ein Ingenieur für maschinelles Lernen. ...
Grammatically incorrect source	...correct any grammatical errors: She didn't *remembre where is her* purse, so I *thinks* its	...grammatikalische Fehler zu korrigieren: Sie erinnerte sich nicht daran, wo ihre Geldbörse ist, also denke ich, ...	...grammatikalische Fehler zu korrigieren: Sie erinnerte sich nicht, wo ihre Handtasche war, also dachte ich, ...
Reference mismatch	Please assume the role of an English translator, ...Regardless of the language I use, ...respond ...in English.	Bitte nehmen Sie die Rolle eines englischen Übersetzers an ...auf Englisch antworten.	Bitte schlüpfen Sie in die Rolle eines Englisch-Übersetzers ...auf Englisch antworten.
None	Can you rephrase your previous answer and incorporate a metaphor or simile in each sentence?	Kannst du deine vorherige Antwort umformulieren und in jedem Satz eine Metapher oder ein Gleichnis einbauen?	Können Sie Ihre vorherige Antwort umformulieren und in jeden Satz eine Metapher oder ein Gleichnis einbauen?

Table 4: EN-DE translation comparison by GPT-4 (MT-Bench-True-german) and DeepL (MT-Bench-DE, Ours)

2.1, which renders its instruction following capabilities as insufficient across languages. While scores increase marginally, the models instruction-tuned on `Lima-X` show to perform only on the lower end of the scale. However, we note spikes in the category *Writing* and *Roleplay* across languages. For `Bactrian-X`-based models a pattern is evident showing best performance for `Bactrian-ENDEFRITES` or models trained directly in the language of the benchmark. However, here no consistent superior fine-tuning is evident across categories and languages.

For completeness, we display the evaluation results of the percentage average performance improvement of models fine-tuned on parallel mixed language instruction-tuning datasets across *MT-Bench-X* languages for turn two in Figure 4. The same pattern as for turn one within Figure 2 in Section 5.2 is evident.

Detailed Tabular View In the following Tables 8, 9 and 10, the detailed category-wise scores of each model variant are displayed. Despite being an automatic evaluation and less costly, still high costs are involved utilizing the OpenAI API for evaluation with GPT-4-as-judge. Because of these high costs involved, we did not average the results across multiple evaluation runs.

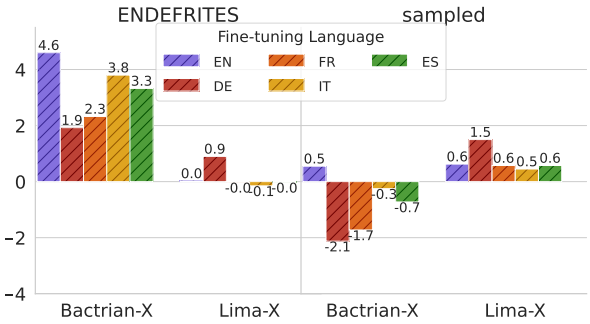


Figure 4: Percentage improvement for turn two averaged across *MT-Bench-X* languages of models fine-tuned on parallel mixed language instruction-tuning datasets (`ENDEFRITES`, `sampled`) over single language fine-tunings.

E.2 Pair-wise Evaluation

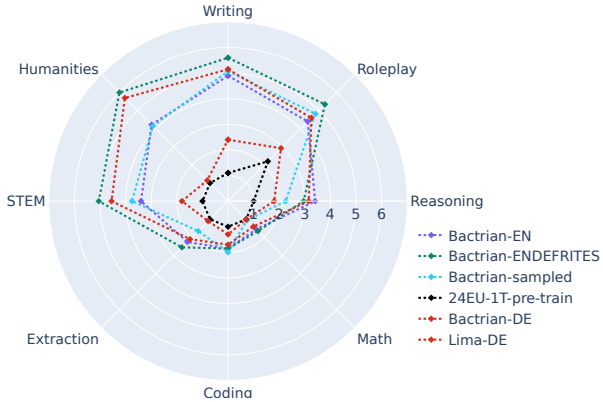
For the pair-wise mode of evaluating according to *MT-Bench* (Zheng et al., 2023) in German, we translate the English prompts for *GPT-4-as-a-judge* as displayed in Table 5.

The initial experiment according to the methods of Zheng et al. (2023), involved only the three voting options *Model A is better*, *Model B is better* and *Tie*. However, as evident from Figure 6, a im-

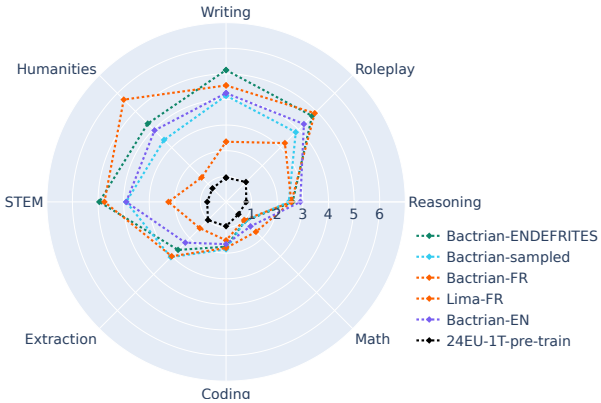
English Original	Translated German
Instruction	
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.	Bitte beurteilen Sie als unparteiischer Richter die Qualität der Antworten von zwei KI-Assistenten auf die unten dargestellte Benutzerfrage. Sie sollten den Assistenten auswählen, der die Anweisungen des Nutzers befolgt und die Frage des Nutzers besser beantwortet. Bei Ihrer Bewertung sollten Sie Faktoren wie Hilfsbereitschaft, Relevanz, Genauigkeit, Tiefe, Kreativität und Detailgenauigkeit der Antworten berücksichtigen. Beginnen Sie Ihre Bewertung mit einem Vergleich der beiden Antworten und geben Sie eine kurze Erklärung ab. Vermeiden Sie jegliche Voreingenommenheit und stellen Sie sicher, dass die Reihenfolge, in der die Antworten präsentiert wurden, keinen Einfluss auf Ihre Entscheidung hat. Lassen Sie sich bei Ihrer Bewertung nicht von der Länge der Antworten beeinflussen. Bevorzugen Sie nicht bestimmte Namen der Assistenten. Seien Sie so objektiv wie möglich. Geben Sie nach Ihrer Erklärung Ihr endgültiges Urteil ab, indem Sie sich strikt an dieses Format halten: "[[A]]", wenn Assistent A besser ist, "[[B]]", wenn Assistent B besser ist, "[[C]]" bei Gleichstand und "[[D]]" wenn beide Assistenten gleich schlecht sind.
Prompt Template	
[User Question] {question} [The Start of Assistant A's Answer] {answer_a} [The End of Assistant A's Answer] [The Start of Assistant B's Answer] {answer_b} [The End of Assistant B's Answer]	[Frage des Nutzers] {question} [Beginn der Antwort von Assistent A] {answer_a} [Das Ende der Antwort von Assistent A] [Beginn der Antwort von Assistent B] {answer_b} [Ende der Antwort von Assistent B]

Table 5: English original prompt for pair-wise evaluation (Cf. (Zheng et al., 2023)) and it's translation and extension by the fourth option [[D]]: *[. . .] and "[[D]]" if both assistants are equally bad.*

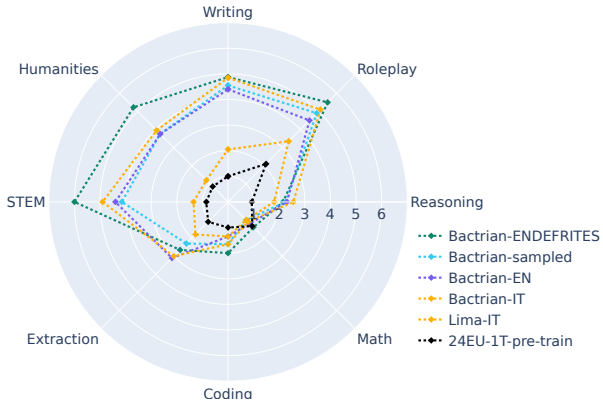
(a) Category-wise MT-Bench-EN single-score by GPT-4.



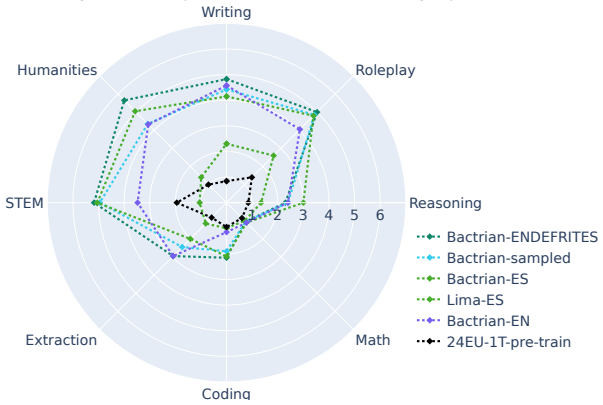
(b) Category-wise MT-Bench-DE single-score by GPT-4.



(c) Category-wise MT-Bench-FR single-score by GPT-4.



(d) Category-wise MT-Bench-IT single-score by GPT-4.



(e) Category-wise MT-Bench-ES single-score by GPT-4.

Figure 5: In-depth MT-Bench-X quality assessment by GPT-4.

balance between the voting options occur and it is unclear whether the large fraction of *Tie* votes result from equally bad or good answers. To clarify whether the large fraction of *Tie* votes occurs due to insufficient overall performance, we repeated the experiment with the additional option to select when both models are equally bad, as presented in Section 6.

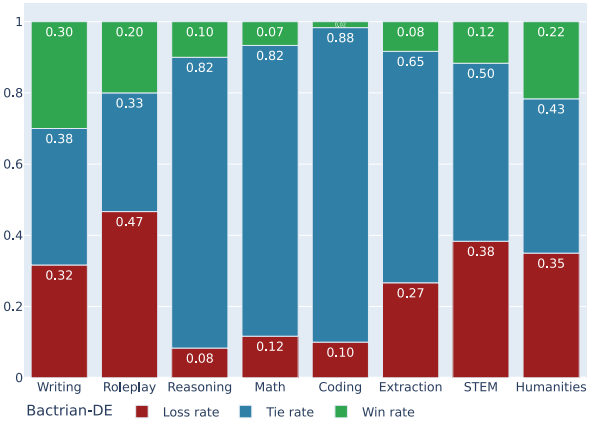


Figure 6: `Bactrian-DE` vs. `Bactrian-ENDEFRITES`. Pair-wise MT-Bench-DE quality assessment by GPT-4.

E.3 Human Evaluation

We conduct an human evaluation on pair-wise model assessment on the German *MT-Bench-DE*. Therefore, we provide the annotation platform depicted in Figure 7, in which academic, native speaking Germans pair-wise assess model responses by selection either a clear winner, denoting a tie or signaling that both models were not helpful at all in answering the multi-turn user request. The 12 participants are all considered to be experts, as all are working within the field of generative artificial intelligence. All participants were informed that the purpose of the annotations was to investigate research questions for a paper to be published. No payment was involved. Consent for the human evaluation study as well as the manual curation of *MT-Bench-X* was obtained through video calls. To protect the annotators participating in this study, we anonymized the data collected. The translation for the user directives in English are as follows:

> "*MT-Bench-X Evaluator Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. Finally, indicate your verdict by clicking one button. Please use the tie button sparsely.*" Cf. (Zheng et al., 2023).

F Qualitative Analysis

G Multilingual Pre-Trained Model

This study is based on a multilingual pre-trained model, that was trained on 1T tokens and is in the processes of being open-sourced. The tokenizer was specifically trained to support each of the 24 EU languages equally fair. The choice of Tokenizer training parameters and implementation are hereby inferred from results of Ali et al. (2023), who recommend for multilingual models an Unigram (Kudo, 2018) SentencePiece[7] Tokenizer with a large vocabulary of 250680 tokens. The pre-training datasets utilised exhibit an English-dominated share of all 24 European languages. Thereby the five languages in question for this study are contained to the following proportions within the pre-training dataset: 43.88% EN, 8.65% FR, 7.63% ES, 8.48% DE, 4.64% IT and 4.51% Code). The dataset was filtered and deduplicated. Within Table X, the datasources are reported.

Regarding architectural choices, the architecture of Llama 2 (Touvron et al., 2023) was followed.

H Infrastructure & Computational Costs

We trained each of our 14 7B parameter models on NVIDIA H100 GPUs, and the training and answer generation of the models took between 162 (LIMA-X) and 540 (Bactrian-X) GPU hours. Therefore, the total training and answer generation costs amounted to $\approx$ 702 GPU hours.

[7] https://github.com/google/sentencepiece

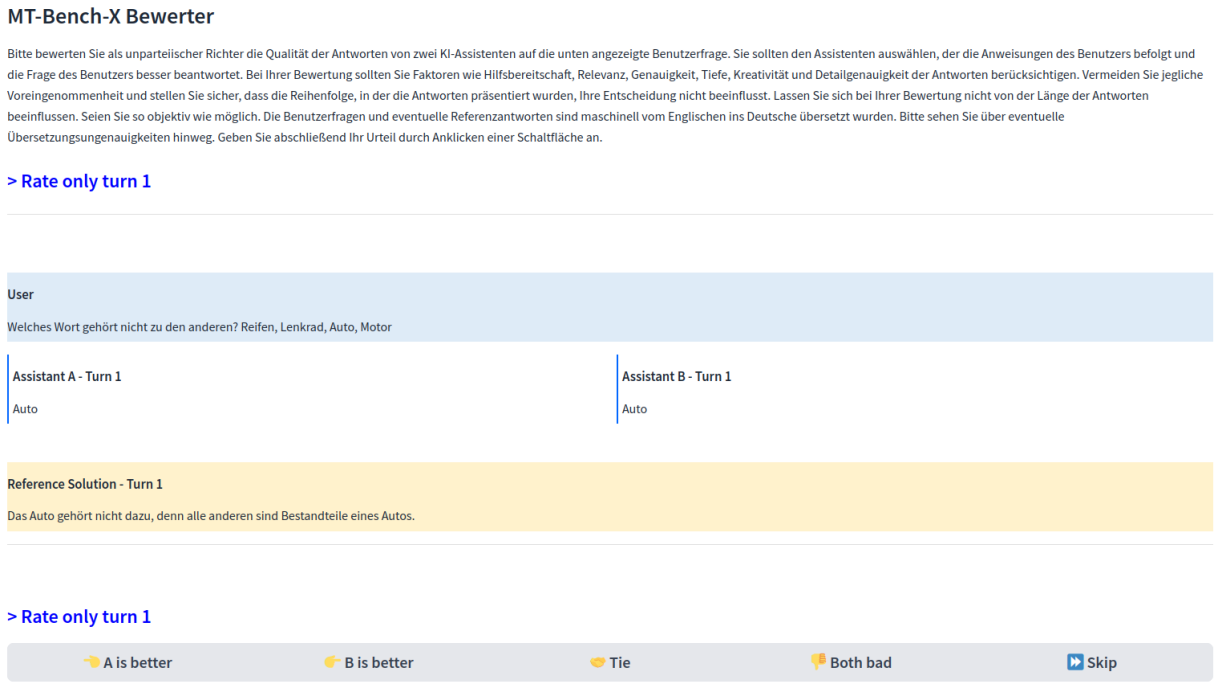


Figure 7: User interface for human evaluation. The model response sides are selected randomly for each turn. Each evaluation instance is sampled randomly and comprises of the two-turn responses of two models, which are evaluated one after the other. Based on Zheng et al. (2023).

Request (writing) Draft a professional email seeking your supervisor's feedback on the 'Quarterly Financial Report' you prepared. Ask specifically about the data analysis, presentation style, and the clarity of conclusions drawn. Keep the email short and to the point.

`Bactrian-DE`

Dear [name of supervisor], I am writing to formally request your feedback on my Quarterly Financial Report. The report was prepared in accordance with the required data analysis and presentation style, as per our company's guidelines. I would appreciate if you could provide me with your feedback on the following:

- Data analysis: Can you please comment on the accuracy of the data presented in the report?
- Presentation style: How do you feel about the visual presentation of the data? Is it easy to understand and visually appealing?
- Conclusions drawn: Please provide your insights on the conclusions drawn from the report. Were they clear and easy to understand, or do you have any concerns or suggestions? I appreciate your time and consideration in reviewing my work, and I look forward to your response. Thank you, [Your Name]

`Bactrian-ENDEFRITES`

Dear [Supervisor],

I hope this email finds you well. I am writing to request your valuable feedback on my Quarterly Financial Report. Please take a moment to review the report and let me know if there are any areas you would like me to improve on.

In regards to the data analysis, I have used a combination of quantitative and qualitative methods to analyze the data and present the results in an easy-to-understand manner. The report clearly demonstrates the financial performance of the company during the quarter, its trends, and the implications for the future. I would appreciate any feedback or suggestions you may have on how to further improve the data analysis.

In terms of the presentation style, I have included charts, graphs, and tables to provide visual aids to the report. The report is well-organized with a clear heading structure and a concise conclusion section. I would appreciate any comments on the clarity of the conclusions drawn.

Overall, I am satisfied with the quality of the report, but I am always eager to learn from your feedback. Thank you for your time.

Sincerely,

[Your Name]

Table 6: Qualitative comparison of `Bactrian-ENDEFRITES` and `Bactrian-DE` on *MT-Bench-EN*.

Request (stem) - Turn 1 The city of Vega intends to build a bridge that will span the Vegona River, covering a distance of 1.8 kilometers. The proposed location falls within a seismically active area that has experienced several high-magnitude earthquakes. Given these circumstances, what would be the best approach to constructing the bridge?

`Bactrian-DE` The best approach to constructing the bridge would be to use a combination of seismic engineering measures and construction techniques that are well-suited to the seismically active area. This would include the use of reinforced concrete and high-strength steel, as well as the use of advanced design and construction methods that are able to withstand the effects of earthquakes. The bridge would also need to be built in a way that allows for easy access for emergency response teams in the event of an earthquake.

`Bactrian-ENDEFRITES` The best approach to constructing the bridge in a seismically active area would be to use advanced engineering techniques and materials that are designed to withstand high seismic activity. This could include using reinforced concrete, high-strength steel, and advanced seismic design principles to ensure the bridge is able to withstand the forces of an earthquake. Additionally, the bridge could be built on a solid foundation, such as a concrete pier, to further increase its stability. It would also be important to carefully monitor the construction process and ensure that the bridge is built to the highest safety standards.

Request (stem) - Turn 2 What are the key disadvantages or flaws of your solution? Please perform calculations and use numbers to illustrate them.

`Bactrian-DE` The key disadvantages or flaws of the proposed solution are as follows:

- The cost of the project would be significantly higher than if the bridge were built in a different location.
- The construction process would be more complex and time-consuming, as the bridge would need to be built in a seismically active area.
- The bridge would be more vulnerable to earthquakes, as it would be built in a seismically active area.
- The construction process would be more dangerous for workers, as they would be working in an area that is prone to earthquakes.

`Bactrian-ENDEFRITES` The key disadvantages of the proposed solution are as follows:

1. High cost: The use of advanced engineering techniques and materials would increase the cost of the bridge significantly, making it unaffordable for the city of Vega.
2. Long construction time: The construction process would take a long time, as the bridge would need to be built on a solid foundation and carefully monitored during the construction process.
3. Environmental impact: The construction process would have a significant impact on the environment, as the bridge would require the clearing of trees and other vegetation, which could have negative consequences for the local ecosystem.
4. Safety concerns: The bridge would be built in a seismically active area, which could increase the risk of an earthquake damaging the bridge or causing injuries to construction workers.

Table 7: Qualitative comparison of `Bactrian-ENDEFRITES` and `Bactrian-DE` on *MT-Bench-EN*.

MT-Bench-EN	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
Bactrian-DE	5.1	5.1	2.0	1.35	1.95	2.8	4.25	5.35	3.49
Bactrian-EN	4.68	5.2	3.4	1.15	2.05	2.6	5.5	6.2	3.85
Bactrian-ENDEFRITES	5.4	4.75	3.45	1.15	2.0	2.55	4.8	5.35	3.68
Bactrian-ES	5.5	5.1	2.2	1.35	2.0	2.6	4.05	5.48	3.53
Bactrian-FR	5.68	4.95	3.2	1.4	2.1	2.5	4.28	5.2	3.66
Bactrian-IT	4.82	4.45	2.75	1.35	1.89	2.3	4.35	4.7	3.33
Bactrian-sampled	4.6	4.95	3.2	1.05	2.1	2.45	4.65	5.4	3.55
Lima-DE	2.8	3.15	1.5	1.15	1.4	1.3	1.85	1.6	1.84
Lima-EN	2.55	4.25	1.8	1.0	1.25	1.4	1.9	1.65	1.98
Lima-ENDEFRITES	2.48	3.85	1.45	1.0	1.4	1.6	1.5	1.95	1.9
Lima-ES	2.8	3.35	3.55	1.2	1.35	1.65	2.35	1.7	2.24
Lima-FR	2.5	3.65	2.1	1.0	1.25	1.35	1.5	1.75	1.89
Lima-IT	2.6	2.3	2.1	1.1	1.35	1.1	1.8	1.7	1.76
Lima-sampled	2.6	3.45	2.05	1.25	1.2	1.15	2.95	2.0	2.08
24EU-1T-pre-train	1.0	1.75	1.0	1.0	1.25	1.0	1.0	1.0	1.12
Bactrian-X-Avg.	5.11	4.93	2.89	1.26	2.01	2.54	4.55	5.38	3.58
Lima-X-Avg.	2.62	3.43	2.08	1.1	1.31	1.36	1.98	1.76	1.96
MT-Bench-DE	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
Bactrian-DE	5.15	4.6	3.15	1.4	1.7	2.1	4.55	5.7	3.54
Bactrian-EN	4.9	4.4	3.4	1.55	1.85	2.25	3.4	4.22	3.25
Bactrian-ENDEFRITES	5.6	5.35	2.95	1.65	1.85	2.55	5.05	6.0	3.88
Bactrian-ES	4.75	4.6	3.5	1.5	2.05	2.0	3.6	4.08	3.26
Bactrian-FR	5.3	4.3	3.05	1.95	2.25	1.8	3.35	3.3	3.16
Bactrian-IT	4.8	4.55	2.6	1.5	2.1	2.6	3.75	3.65	3.19
Bactrian-sampled	5.08	4.82	2.25	1.05	2.0	1.65	3.75	4.15	3.09
Lima-DE	2.4	2.92	1.8	1.0	1.3	1.1	1.8	1.15	1.68
Lima-EN	2.95	3.55	1.1	1.3	1.45	1.1	1.6	1.65	1.84
Lima-ENDEFRITES	2.8	2.7	1.55	1.35	1.2	1.25	1.55	1.05	1.68
Lima-ES	2.7	2.45	1.4	1.35	1.25	1.3	2.0	1.6	1.76
Lima-FR	2.75	2.6	1.85	1.05	1.05	1.15	1.35	1.0	1.6
Lima-IT	2.88	3.05	1.4	1.0	1.05	1.15	1.5	1.15	1.65
Lima-sampled	2.9	2.65	2.1	1.4	1.1	1.5	2.15	1.6	1.92
24EU-1T-pre-train	1.1	2.2	1.0	1.0	1.0	1.0	1.0	1.0	1.16
Bactrian-X-Avg.	5.08	4.66	2.99	1.51	1.97	2.14	3.92	4.44	3.34
Lima-X-Avg.	2.77	2.85	1.6	1.21	1.2	1.22	1.71	1.31	1.73

Table 8: Single evaluation results by GPT-4 for MT-Bench-EN and MT-Bench-DE

MT-Bench-FR	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
Bactrian-DE	5.42	4.8	2.35	1.0	1.7	2.45	3.9	5.05	3.33
Bactrian-EN	4.25	4.3	2.9	1.35	1.65	2.25	3.9	3.95	3.07
Bactrian-ENDEFRITES	5.15	4.75	2.6	1.05	1.75	2.65	4.95	4.32	3.4
Bactrian-ES	4.0	4.35	2.35	1.1	2.0	2.55	3.05	3.7	2.89
Bactrian-FR	4.55	4.9	2.55	1.65	1.8	3.0	4.75	5.65	3.61
Bactrian-IT	4.8	4.8	2.65	1.4	1.75	3.7	3.85	3.1	3.26
Bactrian-sampled	4.15	3.85	2.45	1.05	1.85	3.05	3.9	3.42	2.97
Lima-DE	1.65	1.55	1.7	1.0	1.1	1.1	1.75	1.6	1.43
Lima-EN	2.4	3.35	1.7	1.35	1.1	1.6	1.7	1.3	1.81
Lima-ENDEFRITES	2.85	2.4	1.85	1.2	1.3	2.2	1.4	1.4	1.82
Lima-ES	1.7	3.45	1.95	1.1	1.1	1.2	1.6	1.3	1.67
Lima-FR	2.35	3.25	2.55	1.0	1.5	1.45	2.25	1.35	1.96
Lima-IT	2.45	3.1	2.45	1.1	1.15	1.45	1.5	1.3	1.81
Lima-sampled	2.25	2.75	1.8	1.05	1.4	1.4	1.85	1.45	1.74
24EU-1T-pre-train	0.95	1.1	0.79	0.68	0.95	1.0	0.74	0.75	0.87
Bactrian-X-Avg.	4.62	4.54	2.55	1.23	1.79	2.81	4.04	4.17	3.22
Lima-X-Avg.	2.24	2.84	2.0	1.11	1.24	1.49	1.72	1.39	1.75
MT-Bench-IT	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
Bactrian-DE	4.95	4.6	2.65	1.7	1.8	3.15	4.5	5.5	3.61
Bactrian-EN	4.4	4.5	2.3	1.2	1.35	3.1	4.4	3.75	3.12
Bactrian-ENDEFRITES	4.88	5.5	2.15	1.4	2.0	2.65	6.0	5.22	3.72
Bactrian-ES	4.68	5.05	2.45	1.1	1.7	2.45	3.95	5.0	3.3
Bactrian-FR	4.58	4.9	2.15	1.05	1.3	2.95	3.75	4.0	3.08
Bactrian-IT	4.85	5.1	2.55	1.1	1.65	3.0	4.9	3.95	3.39
Bactrian-sampled	4.55	4.9	2.25	1.0	1.65	2.3	4.15	3.75	3.07
Lima-DE	1.85	2.2	1.1	1.35	1.15	1.7	1.1	1.35	1.48
Lima-EN	2.45	3.15	1.5	1.1	1.45	1.4	1.8	1.2	1.76
Lima-ENDEFRITES	2.45	2.7	1.7	1.15	1.35	1.6	1.5	1.65	1.76
Lima-ES	1.5	2.9	1.75	1.3	1.4	1.7	1.8	1.55	1.74
Lima-FR	1.85	2.15	1.35	1.35	1.25	1.4	1.1	1.7	1.52
Lima-IT	2.05	3.35	1.8	1.0	1.35	1.8	1.35	1.2	1.74
Lima-sampled	1.85	2.9	2.4	1.0	1.2	1.6	1.35	1.5	1.72
24EU-1T-pre-train	1.0	2.1	0.93	1.33	1.0	1.1	0.85	0.85	1.15
Bactrian-X-Avg.	4.7	4.94	2.36	1.22	1.64	2.8	4.52	4.45	3.33
Lima-X-Avg.	2.0	2.76	1.66	1.18	1.31	1.6	1.43	1.45	1.67

Table 9: Single evaluation results by GPT-4 for MT-Bench-FR and MT-Bench-IT

MT-Bench-ES	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
Bactrian-DE	4.5	5.18	2.6	1.05	1.6	2.95	4.5	5.6	3.5
Bactrian-EN	4.58	4.05	2.4	1.1	1.15	2.95	3.48	4.32	3.0
Bactrian-ENDEFRITES	4.82	5.0	2.3	1.05	2.15	2.95	5.18	5.65	3.64
Bactrian-ES	4.15	4.8	3.0	1.1	2.1	2.0	5.05	5.05	3.41
Bactrian-FR	4.97	4.68	2.45	1.05	1.95	2.05	4.4	5.28	3.35
Bactrian-IT	4.72	4.15	2.4	1.15	1.65	2.1	4.5	4.25	3.12
Bactrian-sampled	4.42	4.85	2.4	1.05	1.9	2.45	4.95	4.35	3.3
Lima-DE	2.35	2.3	1.25	1.25	1.1	1.05	1.35	1.25	1.49
Lima-EN	2.35	2.2	1.1	1.05	1.05	1.35	1.45	1.55	1.51
Lima-ENDEFRITES	1.95	2.7	1.3	1.05	1.0	1.6	1.6	1.2	1.55
Lima-ES	2.3	2.6	1.35	1.0	1.0	1.15	1.05	1.4	1.48
Lima-FR	1.85	2.9	1.3	1.3	1.5	1.4	1.15	1.1	1.56
Lima-IT	1.6	2.45	1.2	1.05	1.0	1.3	1.05	1.8	1.43
Lima-sampled	2.65	2.85	1.65	1.05	1.0	1.25	2.95	1.75	1.89
24EU-1T-pre-train	0.84	1.4	0.85	0.85	0.95	0.82	1.94	1.0	1.08
Bactrian-X-Avg.	4.6	4.67	2.51	1.08	1.79	2.49	4.58	4.93	3.33
Lima-X-Avg.	2.15	2.57	1.31	1.11	1.09	1.3	1.51	1.44	1.56

Table 10: Single evaluation results by GPT-4 for MT-Bench-ES