

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

Alexander Arno Weber<sup>1,2</sup> Klaudia Thellmann<sup>3</sup> Jan Ebert<sup>4</sup> Nicolas Flores-Herr<sup>1</sup>

Jens Lehmann<sup>3</sup> Michael Fromm<sup>1,2</sup> Mehdi Ali<sup>1,2</sup>

<sup>1</sup>Lamarr Institute <sup>2</sup>Fraunhofer IAIS <sup>3</sup>TU Dresden <sup>4</sup>FZ Jülich

## 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 instruction-tuned on different language compositions on *parallel 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 and a large, multilingual LLMs by instruction-tuning them 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 9.9%. 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 (Ding et al., 2023). 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 (Mishra et al., 2022). 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 multilingual benchmarks targeting instruction-tuned models represents a major gap, as previous instruction-tuned multilingual models are only evaluated on zero- or few-shot, 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 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 par-

allel benchmark *MT-Bench-X* and systematically investigate how the language and size of instruction datasets impact the instruction-tuning of pre-trained, mid-sized, and large 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 of whether multilingual models pre-trained with a focus on Indo-European languages 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*<sup>1</sup>, a high-quality, complex, parallel corpus comprising 1030 instructions for each English, German, French, Italian, and Spanish (Section 3).
- Creation of *MT-Bench-X*<sup>2</sup>, a parallel, multilingual, human-curated evaluation dataset for evaluating instruction-tuned LLMs (Section 4).
- Study about the effect of i.) the dataset size, ii.) the dataset nature, i.e., human-curated or synthetic data, iii.) the dataset language composition, and iv.) the importance of the predominant language during pre-training on multilingual instruction for a medium-sized and large language model (Section 5).
- Correlation analysis of the agreement levels between human experts and GPT-4 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 letting 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. 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 3 of the preliminary study in the Appendix 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;

<sup>1</sup><https://hf.co/datasets/lamarr-org/Lima-X>

<sup>2</sup><https://github.com/Modalities/MT-Bench-X>

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).

**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<sup>3</sup>.

Singh et al. (2024); Üstün et al. (2024), 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.

The work of Li et al. focuses on low-resource language instruction-tuning on the English-centric Llama-2 (Touvron et al., 2023) model. In contrast, in this work, we focus on the composition of language mixture datasets of high-resource European languages on multilingual models by design and underline the importance of thoroughly evaluated multilingual human-curated evaluation datasets.

### 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 mid-sized (7B) and large (8x7B) pre-trained models that have been trained with a focus on Indo-European languages.
- We investigate whether an instruction tuning dataset needs to be represented in each language in parallel, has to be split across languages, or should be monolingual.
- We inspect the role of multilingual dataset quality by comparing models trained on human-curated and translated datasets with models trained on same-sized synthetically generated datasets.

- We evaluate our models on manually human-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 or datasets contain shorter, less complex responses (Muennighoff et al., 2023), as a preliminary study shown in the Appendix A reveals. 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)<sup>4</sup>.

### 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% (64K samples per language) of the data as the training dataset and use the remaining 5% as the validation dataset (3K samples per language). We additionally down-sample the dataset to the same size as *LIMA*.

### 3.2 Lima-X

Each sample in *LIMA* is carefully 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 B.

As we focus on Indo-European languages in our study, we chose to utilize DeepL as a translator performing well in these languages (Yulianto

<sup>3</sup><https://paperswithcode.com/dataset/mt-bench>

<sup>4</sup>The license details of our created datasets are described in Appendix E

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 that were written entirely in a language other than English, and iv.) cultural aspects of English that are not transferable to the target language. 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, jokes and poems are not directly translatable, which we see as a downside of the translation approach.

### 3.3 Language Mixtures

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. In the same manner, we create DEFITES, which does not contain the English part of the datasets and thus only quadruples the dataset size. Additionally, we create a variant, maintaining the same semantics of the questions as in the monolingual original but distributed equally across the five languages within the dataset, resulting in the concrete datasets ENDEFRITES-sampled and *DEFITES-sampled*. For example, originating from the English dataset example ID *100*, ENDEFRITES-sampled will contain either one of *100\_X* by choice, whereas *X* stands for the fine-tuning language *EN*, *DE*, *FR*, *IT*, *ES* and so that no ID is duplicated across languages and all languages are represented equally often. To analyze the role of the dataset nature, i.e., human-curated vs. synthetic datasets, we down-sample the *Bactrian-X* dataset variants to contain as many examples as *Lima-X*, utilizing the same seed across languages, resulting in the dataset variants *Bactrian-X-small*.

## 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 translated MT-Bench (Zheng et al., 2023) (Section 4.1) into the five target languages, followed by human quality assurance for each language to create the multilingual MT-Bench variant MT-Bench-X (Section 4.2). We employed MT-Bench-X to conduct a machine evaluation and a human evaluation (Section 4.3) to assess the reliability and limitations of MT-Bench-X.

### 4.1 MT-Bench

For evaluating instruction-tuned models within the conversational format, human evaluation is considered the gold standard. However, with *MT-Bench* (Zheng et al., 2023) a cost-effective 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 eight 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 its 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. 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 judgment prompts as developed by Zheng et al. (2023) but translated into the new benchmark language. 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 a translation engine to translate the questions, 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*, which spans 400 two-turn examples in total, i.e., 80 two-turn benchmark examples per language. Publishing details are in the Appendix E. We investigate the performance of DeepL in the Appendix C. While we consider DeepL as an appropriate choice as a translation tool, there are still problematic cases that we let manually edit for correctness and amend wording for both questions and references across all languages by humans with at least a graduate and who are 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 were 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 inspecting *MT-Bench* are the requirement of i.) translation capabilities of the LLM-as-a-judge models, e.g. from Chinese to the translated language, and ii.) up-to-date knowledge e.g. mentioning of GPT-4. In addition to the user requests and references, we also translate the prompts within *MT-Bench* to not mix languages systematically during evaluation with *MT-Bench-X*. Through the manual correction of the translated *MT-Bench-X* dataset and the evaluation of *MT-Bench-X* regarding its reliability and limitations, described in the following, we offer a high-quality instruction-tuning evaluation benchmark resource to the community.

### 4.3 Evaluation Approach of *MT-Bench-X*

To assess *MT-Bench-X*'s suitability, we conduct a human evaluation and validate the correlation between human and LLM-as-a-judge ratings. 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 a user interface inspired by Zheng et al. (2023), as can be seen in Figure 8 in the Appendix F.3, to let human expert judges rate answers of models in a pair-wise response comparison. For this, we translated the prompts provided by Zheng et al. (2023) into German, see Appendix F.2. 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. The authors of *MT-Bench* (Zheng et al., 2023) minimized the risk for several biases that could occur when using a LLM as a judge. In that spirit, we randomly select the display side for each model newly for each turn to omit positional bias.

## 5 Multilingual Performance Analysis

To answer the question of whether polyglot models require multilingual or monolingual instructions, we conducted several experiments. We describe the experimental setup in Section 5.1, followed by the evaluation by *GPT-4-as-a-judge* of single score ratings and pair-wise evaluation (Section 5.2). Moreover, we analyze the role of dataset quality by comparing same-sized synthetic with curated datasets in Section 5.3 and elaborate the *Superficial Alignment Hypothesis* for the multilingual setup in Section 5.4. Additionally, we conducted a qualitative analysis (Appendix G).

### 5.1 Experimental Setup

We employ the datasets described in Section 3 to conduct our experiments. This includes instruction-tuning on each monolingual dataset and all four multilingual dataset mixtures for each dataset source, i.e., *Lima-X*, *Bactrian-X* and *Bactrian-X-small*. As base models, we use a mid-sized multilingual 7B model that closely follows the Llama 2 (Touvron et al., 2023) architecture, which was trained on a large number of non-English samples and the large, multilingual Mixture of Experts model *Mixtral-8x7B-v0.1*. Further details are specified in the Appendix D. We follow a similar hyperparameter 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. Due to the high computational demand of full-weight fine-tuning of *Mixtral-8x7B*, we omitted fine-tuning *Mixtral-8x7B* with the large *Bactrian-X* datasets. The number of epochs and the learning rate are determined by early stopping based on the validation loss, which resulted in a learning rate of  $1e-6$  for the fully parallel datasets for *Mixtral-8x7B*.

## 5.2 The Impact of Language

In the following, we investigate the performance of models trained on different language mixtures by both monolingual evaluation and cross-lingual evaluation.

**Monolingual Performance** For automatic evaluation with GPT-4, we display the summarized results within Figure 4, alongside fine-grained results in the Appendix F.1. The Figure shows the average absolute results obtained on *MT-Bench-X* for each instruction-tuned model.

The results for *24EU-7B* in Figure 4 highlight that the single language instruction-tunings Bactrian-DE, Bactrian-IT and Bactrian-ES are within their respective evaluation language inferior compared to Bactrian-ENDEFRITES. In contrast, Bactrian-EN achieves strong performance, as Figure 4 shows. We hypothesize 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*. The fully parallel *DEFRITES*-based and *ENDEFRITES*-based language mixture fine-tunings show no consistent superiority. *Bactrian-X-small* shows a coarsely similar, albeit overall lower performance compared to *Bactrian-X*.

*Mixtral-7x8B*-based models showcase an overall higher absolute score within each language, compared to *24EU-7B*-based models, supposedly due to its larger size. For the *Mixtral-7x8B*-based results, a consistent effect of monolingual performance increase or at least comparative performance for instruction-tuning with cross-lingual datasets compared to monolingual datasets can be observed for *Bactrian-X-small* as well as *Lima-X*-based models. However, as with the *24EU-7B* model results, either *DEFRITES* or *ENDEFRITES* language mixture based models show this effect inconsistently. There is no clear picture when evaluating model performance in a single language only.

**Cross-lingual Performance** We compare the average performance across languages between parallel dataset mixtures and monolingual datasets in Figure 1. As can be inferred from Figure 1a, for *24EU-7B*-based models the language mix *ENDEFRITES*-sampled, multilingual instruction-tuning improvements for *Lima-X* and the same sized *Bactrian-X-small* are notable, but the opposite for *Bactrian-X*. However, the same cannot be observed for *DEFRITES*-sampled, as here, all mod-

els trained with parallel data mixtures are inferior compared to their monolingual counterparts. The inconsistency within these results might come from the number of samples per language, which is five times as small within *ENDEFRITES*-sampled and *DEFRITES*-sampled compared to the full monolingual dataset. Here, *Lima-X* and *Bactrian-X-small* only contain 1030 samples in total, i.e. 206 samples per language within *ENDEFRITES*-sampled.

For *Mixtral-8x7B* models in Figure 1b, we note consistent improvement of parallel instruction-tuning, i.e., for the sampled as well as full-sized datasets, with only a few, but consistent and less pronounced exceptions for the same language and dataset combination. In the same manner, the *24EU-7B* models trained based on the full-sized, parallel language mix *ENDEFRITES* and *DEFRITES* show performance improvements across languages for *Bactrian-X* and *Lima-X*, when compared to their monolingual counterparts.

Furthermore, the performance improvement when including the predominant pre-training language during parallel instruction-tuning (*ENDEFRITES* vs *DEFRITES*) highlights the importance of its inclusion during instruction-tuning. However, the down-sampled variants show decreased performances for both parallel language mixtures. We assume this is due to the synthetic nature of *Bactrian-X*, as it may show different levels of hallucination depending on the language skill of the generative model. Therefore, the sampled setting could have a lower signal-to-noise ratio than the *ENDEFRITES* setting. In contrast, *LIMA*-sampled, a carefully curated dataset, does not express a performance decrease, providing evidence for the above-mentioned hypothesis. We assume the increased world knowledge associated with a larger model size allows *Mixtral-8x7B* to compensate for the assumed low signal-to-noise ratio in *Bactrian-X-small*.

We observe a maximum performance gain of 9.9% (0.99 points on the *MT-Bench* scale ranging from 1 to 10) by Bactrian-*ENDEFRITES* compared to Bactrian-*IT*.

Despite the emergence of multilingual capabilities during pre-training, we found fine-tuning on parallel data improves multilingual instruction-following performance stronger across languages compared to non-parallel multilingual data as well as monolingual data for mid-sized and large LLMs.

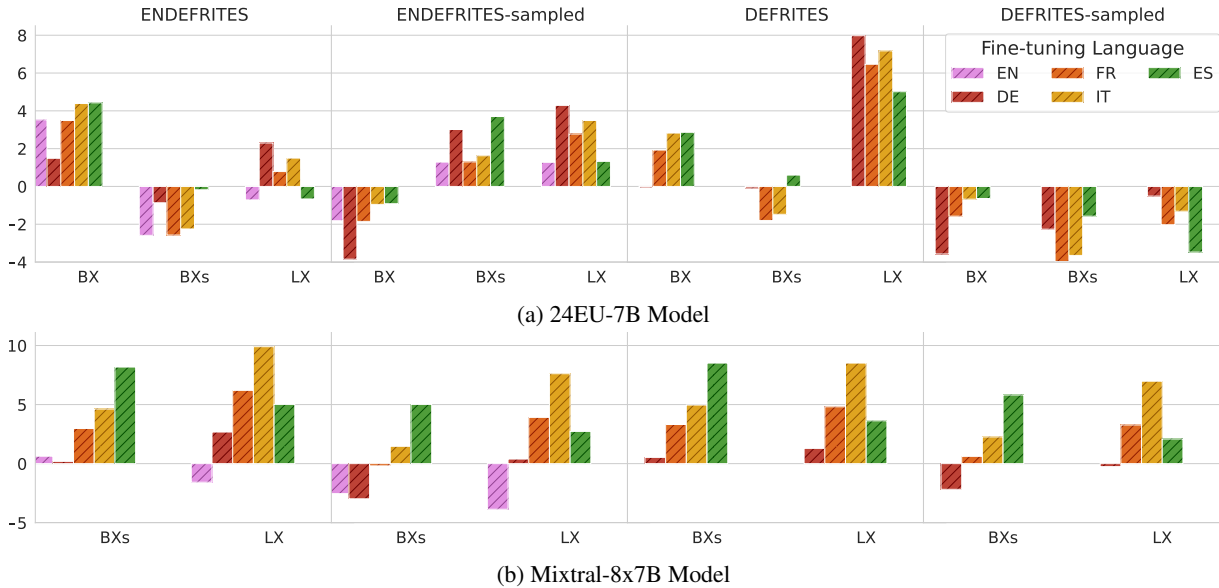


Figure 1: Percentage improvement for turn one averaged across *MT-Bench-X* languages of models fine-tuned on parallel mixed language instruction-tuning datasets over single language fine-tunings.

### 5.3 Synthetic vs Curated Training Datasets

To factor out the impact of training dataset size, we down-sampled the *Bactrian-X* datasets to the same size as the human-curated *LIMA-X* datasets, resulting in the dataset *Bactrian-X-small*. As shown in Figure 4 in the Appendix, the *Bactrian-X*-based synthetic datasets outperform *LIMA-X* for both dataset magnitudes. When comparing *LIMA-X* with the cross-lingual performance of models tuned on *Bactrian-X-small* or *Bactrian-X* in Figure 1, the same overall performance trend i.e. improved cross-lingual performance for parallel *Bactrian-X*-based instruction-tunings, is visible. These results underline the beneficial effect of employing synthetic, semantically parallel datasets for multilingual instruction-tuning.

### 5.4 Superficial Alignment Hypothesis

We additionally investigated the superficial alignment hypothesis, which states that only a few examples per task are required to teach a model to follow instructions. Figure 4 in the Appendix highlights that *24EU-7B* models instruction-tuned on *Bactrian-X* as well as *Bactrian-X-small* generally outperform models instruction-tuned on *Lima-X* datasets. Our results show that the *Superficial Alignment Hypothesis* (Kirstain et al., 2022; Zhou et al., 2023) does not generally hold for mid-sized LLMs. However, with *Mixtral-8x7B*, we see high performances for synthetic as well as human-curated data, indicating that the effectiveness of the

*Superficial Alignment Hypothesis* increases with larger model size or respectively with greater pre-trained model capabilities.

## 6 Human Evaluation

Human evaluation is the gold standard for evaluating the output of generative models, as responses can be complex and highly diverse. Following the human evaluation as described in Section 4.3 on answers of the *24EU-7B* model tuned on *Bactrian-X*, we: i.) inspect the role of positional bias in *MT-Bench-X* (Section 6.1), ii.) compare the voting distributions between humans and GPT-4 in Section 6.2, and iii.) analyze the correlation of agreement levels between humans and GPT-4 in Section 6.3.

### 6.1 Positional Bias

We inspect potential limitations of utilizing *GPT-4-as-a-judge* exemplary for German text in Table 1. We observe a high level of positional bias for the categories *Stem*, *Humanities* and *Writing*. For the following correlation analysis, we mitigate the effect of positional bias, in which case no result is provided for a given benchmark example, by substituting missing values through the results of the following run, where possible. Albeit judgment generation in *MT-Bench-X* is conducted by greedy search and the evaluation runs were executed immediately after another, we mitigated the positional bias by up to 16.25%.

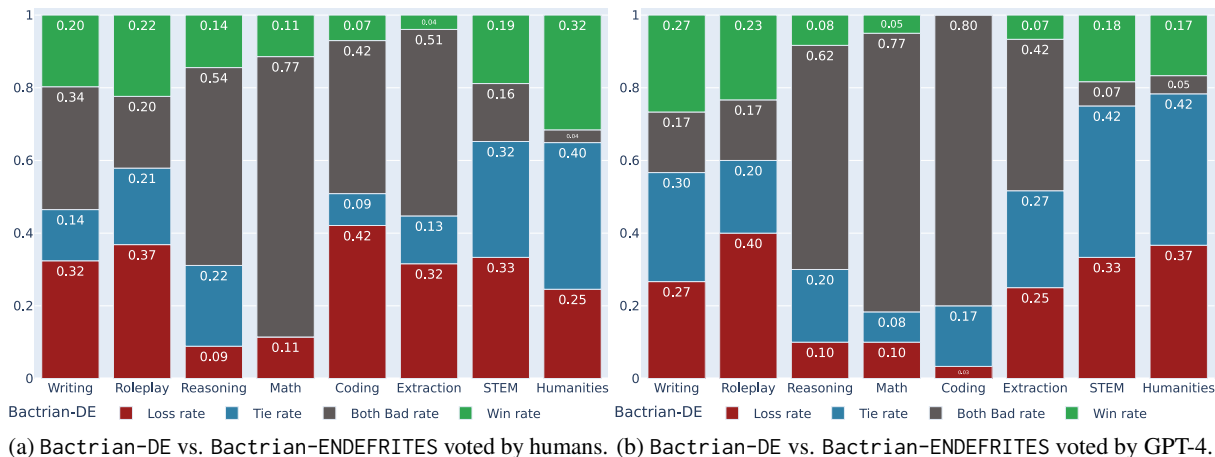


Figure 2: Pair-wise MT-Bench-DE quality assessment by humans and GPT-4, including voting option "both bad".

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

Table 1: Average percentage of positional bias (PB).

## 6.2 Voting Distribution

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 2. As can be seen in Figure 2b, Bactrian-DE loses more frequently against Bactrian-ENDEFRITES, except for the *Reasoning* category. 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 in capabilities learned during pre-training. Within Figure 2a it is evident that human evaluators tend to vote less often for "Tie" and "Both Bad". For the categories *Math*, *Reasoning* and *Coding*, the underperformance shown by automatic per-category single evaluation scores in Figure 5 in the Appendix is confirmed by the human evaluation in Figure 2a.

## 6.3 Correlation Analysis

The correlation between agreement levels of humans and GPT-4 is shown in Table 2. As can be seen, there is an agreement of 39% between humans and GPT-4 for turn one and 56% agreement between humans. To incorporate the agreement by chance, we calculate the Fleiss' kappa

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

<sup>a</sup> *RCEMR*: Roleplay, Coding, Extraction, Math and Reasoning

Table 2: Percentage of agreement between Human-GPT-4 and Human-Human with subscripted upper limit of Human-Majority-Human agreement.

scores (Fleiss and Cohen, 1973) among humans, resulting in the coefficient of 0.37 (p-value:  $3 \times 10^{-8}$ ) on 66 samples for the first turn and 0.36 (p-value:  $6 \times 10^{-8}$ ) 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 between humans and 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. Of the 242 model responses, hu-



man judges prefer to 62.8% the model responses of Bactrian-ENDEFITES, 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. Across a selection of major Indo-European languages, we thoroughly analyzed instruction-tuned models using parallel instruction-tuning benchmarks. Hereby, the creation of multilingual benchmarks and datasets are crucial to investigate multilingual instruction-tuning capabilities. Our findings highlight the benefits of instruction-tuning on parallel datasets, showcasing improvements of up to 9.9% in cross-lingual instruction-following capabilities compared to monolingual corpora. Additionally, our results challenge the generalizability of the *Superficial Alignment Hypothesis*, showing that extensive instruction-tuning datasets are necessary for mid-sized multilingual models, but not necessarily for larger models. We extended the findings of Zheng et al. (2023) for German and analyzed the similarities and differences between human evaluation and evaluation with GPT-4-as-a-judge. We identify disparities between human evaluations and those generated by GPT-4 in multilingual chat scenarios.

By illuminating these challenges, we emphasize 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 the following limitations.

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.

Secondly, due to involved costs when evaluating with *GPT-4-as-a-judge*, 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.

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 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.

## Acknowledgments

We would like to thank Dr. Joachim Köhler, Ines Wendler, Joe Jamison, and Valentina Ciar dini (Fraunhofer IIS) for their invaluable support for insightful discussions and participation for the quality assessment of created resources. We would like to extend our gratitude to the Fraunhofer IAIS team for their valuable contributions to this project, particularly their involvement in human evaluation. This work was funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence, LAMARR22B as well as by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D) and by the European Union’s Horizon 2020 research and innovation program under grant agreement No

101135671 (TrustLLM). The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. ([www.gauss-centre.eu](http://www.gauss-centre.eu)) for funding this project by providing computing time on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC) as well as the Center for Information Services and High Performance Computing [Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)] at TU Dresden for providing its facilities for automatic evaluation computations.

## 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.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Enhancing chat language models by scaling high-quality instructional conversations.](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051. Association for Computational Linguistics.
- 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.
- Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.
- 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. [Mixtral of experts.](#) *CoRR*, abs/2401.04088.
- 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, Arnab 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.
- 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. 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), 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.
- OpenGPT-X. 2024. [Progress report: Towards european llms](#).
- 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, Elie 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 D’souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Matciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzeminski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Chien Vu, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. [Aya dataset: An open-access collection for multilingual instruction tuning](#). *CoRR*, abs/2402.06619.
- 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, Shrutli 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 finetuned chat models](#). *CoRR*, abs/2307.09288.
- Matthias Uhlig, Sigurd Schacht, and Sudarshan Kamath Barkur. 2024. [Phoenix: Open-source language adaption for direct preference optimization](#).
- 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. 2024. [Aya model: An instruction finetuned open-access multilingual language model](#). *CoRR*, abs/2402.07827.
- 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.

## A Preliminary Dataset Study

For selecting appropriate multilingual datasets to instruction-tune the multilingual models, we conducted a preliminary study to investigate the language distribution and length statistics. The results, depicted in Table 3 show a high skew towards English-focused datasets.

## B LIMA Validation Dataset Creation

We utilize the open-source project *stackexchange-dataset*<sup>5</sup> 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 that 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, Min-Hashing, and LSH over the *LIMA* training dataset split. In total, we reduced the 62021 examples to only 84, which we then carefully inspected and manually curated by rewriting or deleting samples. This leads to the final 52 samples, which is roughly the size of the validation dataset reported by Zhou et al. (2023).

## C Translation Quality

Most similar to our benchmark translation efforts is the dataset *MT-Bench-TrueGerman*. The authors of *MT-Bench-TrueGerman*<sup>6</sup>, 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

<sup>5</sup><https://github.com/EleutherAI/stackexchange-dataset>

<sup>6</sup><https://huggingface.co/VAG0solutions/SauerkrautLM-7b-v1>

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*.

## D Multilingual Pre-Trained Models

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 a prerequisite for minimal instruction-following capability. To the best of our knowledge, only three existing, openly available model families are multilingual European ones. This includes BLOOM (Scao et al., 2022) and Nemotron<sup>7</sup> and *Mixtral-8x7B*. 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. While for *Mixtral-8x7B* no details about the tokenizer training are known, it shows high performance across languages for downstream evaluation tasks.

This study is based on an intermediate checkpoint of a multilingual pre-trained model that was trained on 1T tokens at the point when the experimentation for this study was conducted and is presented in (OpenGPT-X, 2024). The tokenizer was specifically trained to support each of the 24 EU languages equally. 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<sup>8</sup> Tokenizer with a large vocabulary of 250680 tokens. The pre-training datasets utilized 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. Regarding architectural

<sup>7</sup><https://developer.nvidia.com/nemotron-3-8b>

<sup>8</sup><https://github.com/google/sentencepiece>

choices, the architecture of Llama 2 (Touvron et al., 2023) was followed. Regarding the multilingual *Mixtral-8x7B-v0.1*, which supports all languages inspected within this study, we refer to the details provided by Jiang et al. (Jiang et al., 2024).

## E 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. This is consistent with the intended use of the source dataset *MT-Bench* (Zheng et al., 2023).

## F Evaluation

### F.1 Single Score Evaluation

With the bar plots in Figure 4, we compare differences of monolingual and multilingual instruction-tuned models across all languages within *MT-Bench-X* for all pre-trained models instruction-tuned on all dataset variants. For the main discussion of these results, we refer the reader to the monolingual performance analysis in Section 5.2.

Furthermore, with the radar plots in Figure 5, we compare differences of monolingual and multilingual instruction-tuned models across all languages and instruction categories within *MT-Bench-X* for *Bactrian-X* and *Lima-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 *24EU-7B* model is 2.2, which renders its instruction-following capabilities insufficient across languages. While scores increase marginally, the *24EU-7B* base model instruction-tuned on *Lima-X* shows 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 *24EU-7B* models, a pattern is evident, showing the 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 6. The same pattern as for turn one within Figure 1 in Section 5.2 is evident. Additionally, we report the absolute cross-lingual *MT-Bench-X* scores by averaging across the evaluation results for all five languages for turn one in Figure 7.

**Detailed Tabular View** In the Tables 9, 10 and 11, the detailed category-wise scores of each model variant are displayed. Despite being an automatic evaluation and less costly, still high costs are involved in 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.

### F.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 3, an imbalance between the voting options occurs, 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.

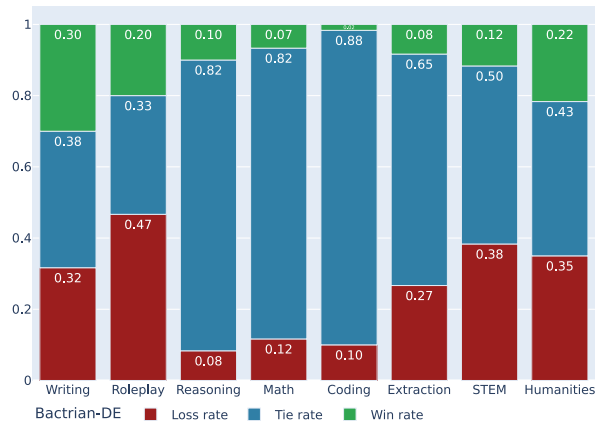


Figure 3: Bactrian-DE vs. Bactrian-ENDEFRITES. Pair-wise *MT-Bench-DE* quality assessment by *GPT-4*.

### F.3 Human Evaluation

We conduct a human evaluation on pair-wise model assessment on the German *MT-Bench-DE*. Therefore, we provide the annotation platform depicted in Figure 8, in which academic, native-speaking Germans pair-wise assess model responses by selecting 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 is 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).*

## G Qualitative Analysis

In the following sections, we conduct a qualitative analysis regarding cross-lingual vs. monolingual model performance (Cf. Appendix G.1) and the performance of GPT-4 as a judge (Cf. Appendix G.2).

### G.1 Cross-lingual vs. Monolingual

Within the examples shown in Table 6 and Table 7, minor differences in the format of the answers of Bactrian-ENDEFRITES and Bactrian-DE can be observed. While the multilingual fine-tuned model shows a format and placeholders as one would expect, the model Bactrian-DE shows incorrect formatting. We also noted that responses of Bactrian-ENDEFRITES are often more elaborate compared to monolingual Bactrian-X models.

### G.2 GPT-4-as-a-judge

While for both examples, reasonably high ratings are given by *GPT-4-as-a-judge* during single mode evaluation and reasonable justifications for not scoring model outputs higher were given by GPT-4 (not shown), we can observe inconsistencies and incorrect ratings, as marked red in Table 8 - even when a reference solution is given - for a lemon-picked example. While this exemplifies that human evaluation should still be considered the gold standard, it also shows that a coarse impression of quantitative model performance by *GPT-4-as-a-judge* is a cost- and time-effective evaluation option despite not being as accurate for some examples.

## H Infrastructure & Computational Costs

We trained each of our 27 7B parameter models and 18 8x7B parameter models on NVIDIA H100 GPUs, and the training and answer generation of the models took between 337 (LIMA-X) and 540 (Bactrian-X) and 337 (Bactrian-X-small) GPU hours. Therefore, the total training and answer generation costs amounted to  $\approx 1214$  GPU hours.

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

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 <i>re-membre where is her</i> purse, so I <i>thinks</i> 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)



English Original	Translated German
<b>Instruction</b>	
<p>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.</p>	<p>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.</p>
<b>Prompt Template</b>	
[User Question] {question}	[Frage des Nutzers] {question}
[The Start of Assistant A’s Answer] {answer_a}	[Beginn der Antwort von Assistent A] {answer_a}
[The End of Assistant A’s Answer]	[Das Ende der Antwort von Assistent A]
[The Start of Assistant B’s Answer] {answer_b}	[Beginn der Antwort von Assistent B] {answer_b}
[The End of Assistant B’s Answer]	[Ende der Antwort von Assistent B]

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

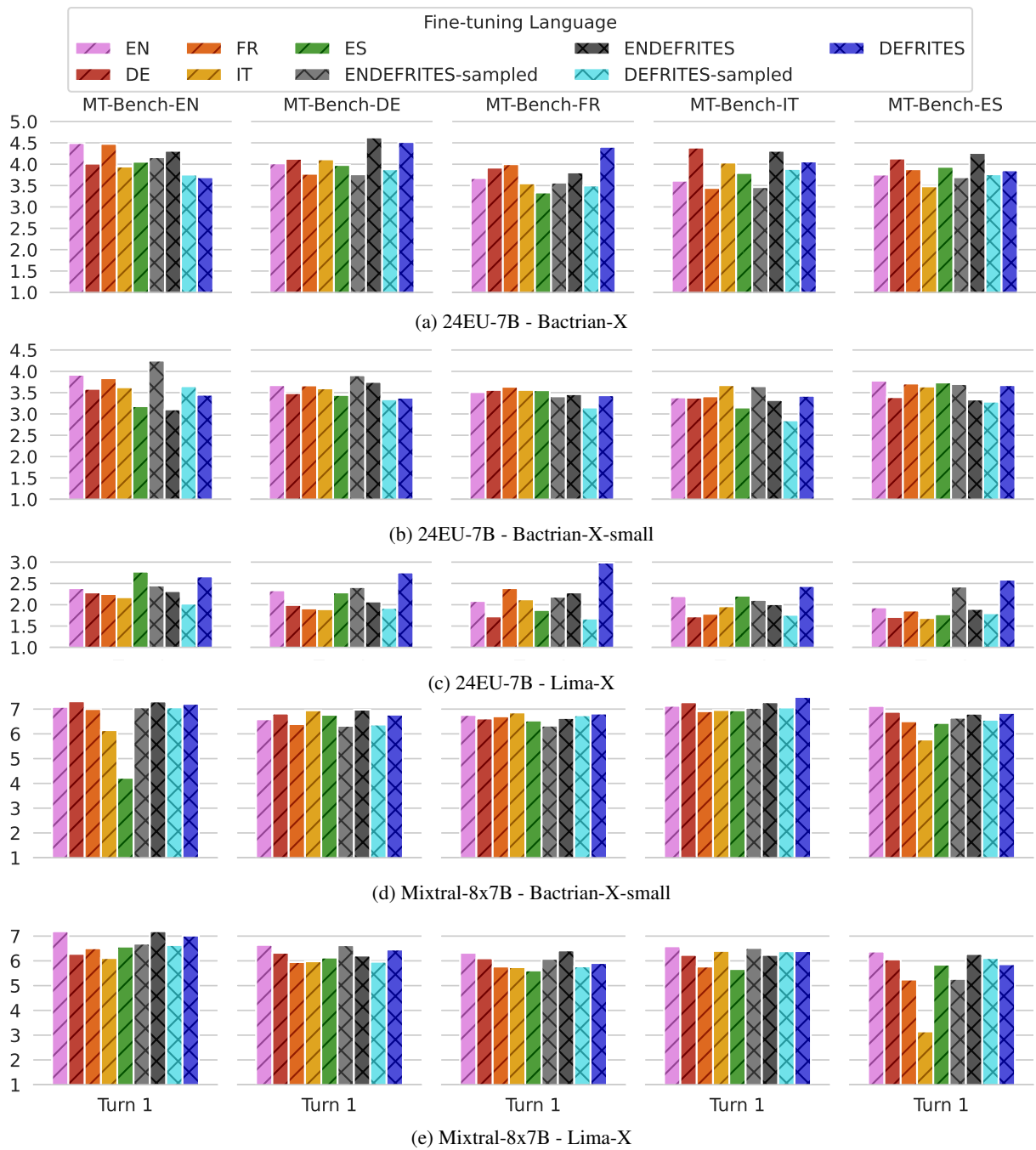
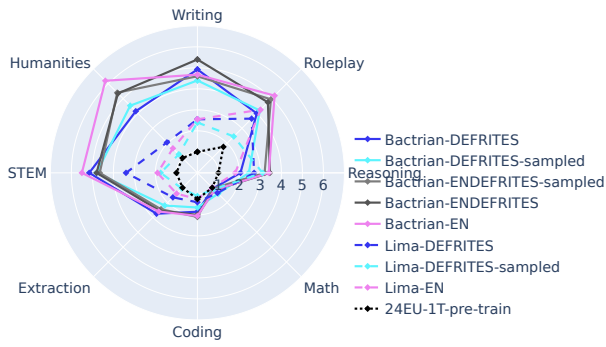
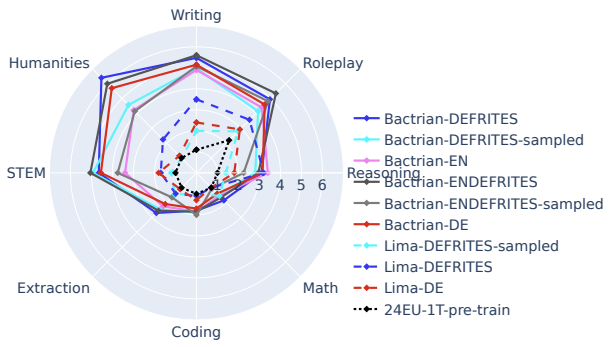


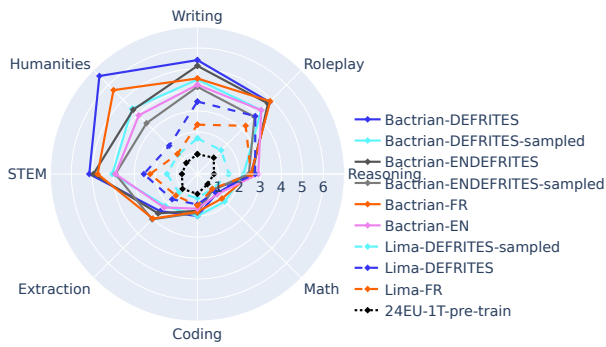
Figure 4: GPT-4-as-a-judge single evaluation average scores for each language mix dataset variant on MT-Bench-X.



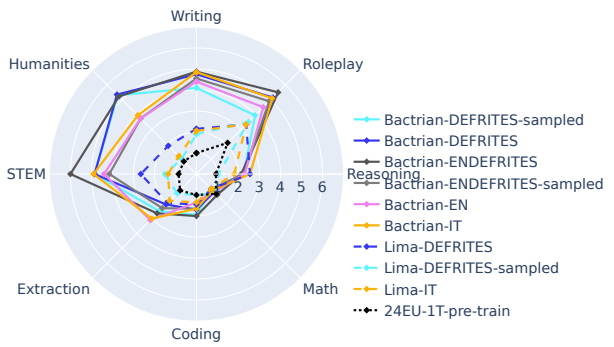
(a) Category-wise MT-Bench-EN single-score by GPT-4.



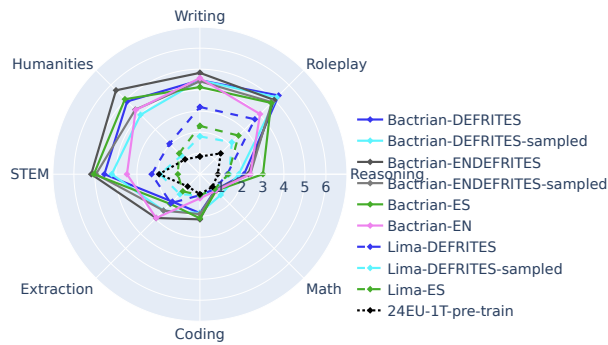
(b) Category-wise MT-Bench-DE single-score by GPT-4.



(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.

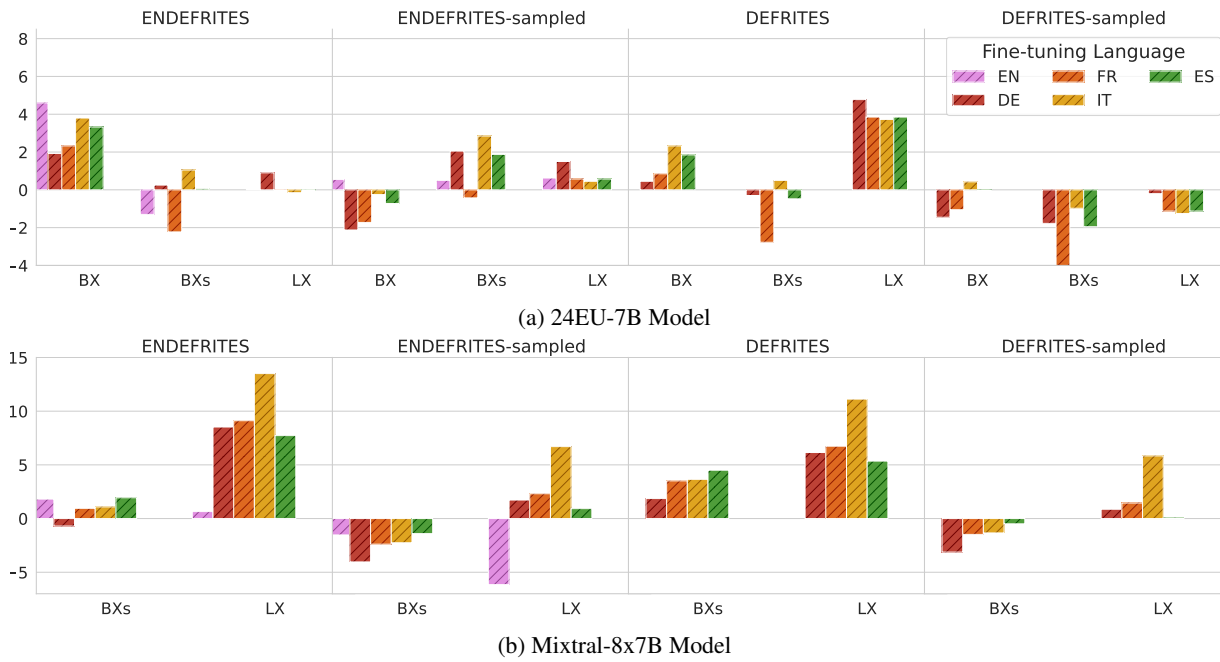


Figure 6: Percentage improvement for turn two averaged across *MT-Bench-X* languages of models fine-tuned on parallel mixed language instruction-tuning datasets over single language fine-tunings.

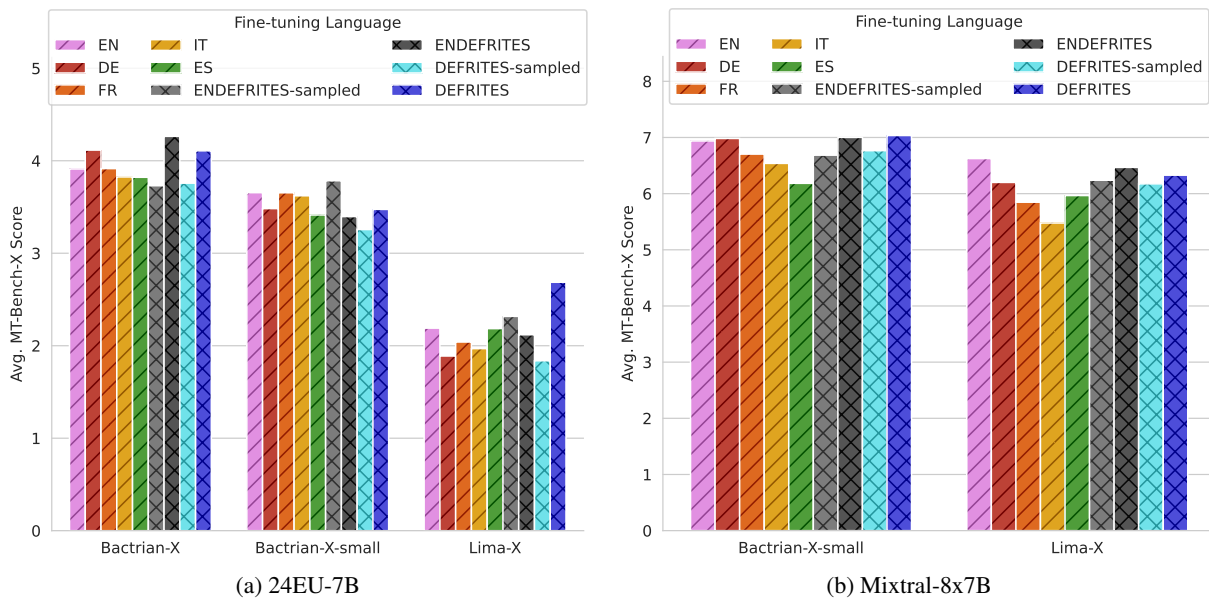


Figure 7: Absolute cross-lingual MT-Bench-X scores across all five languages for turn one.

## MT-Bench-X Bewerter

Bitte bewerten Sie als unparteiischer Richter die Qualität der Antworten von zwei KI-Assistenten auf die unten angezeigte Benutzerfrage. Sie sollten den Assistenten auswählen, der die Anweisungen des Benutzers befolgt und die Frage des Benutzers besser beantwortet. Bei Ihrer Bewertung sollten Sie Faktoren wie Hilfsbereitschaft, Relevanz, Genauigkeit, Tiefe, Kreativität und Detailgenauigkeit der Antworten berücksichtigen. Vermeiden Sie jegliche Voreingenommenheit und stellen Sie sicher, dass die Reihenfolge, in der die Antworten präsentiert wurden, Ihre Entscheidung nicht beeinflusst. Lassen Sie sich bei Ihrer Bewertung nicht von der Länge der Antworten beeinflussen. Seien Sie so objektiv wie möglich. Die Benutzerfragen und eventuelle Referenzantworten sind maschinell vom Englischen ins Deutsche übersetzt wurden. Bitte sehen Sie über eventuelle Übersetzungsungenauigkeiten hinweg. Geben Sie abschließend Ihr Urteil durch Anklicken einer Schaltfläche an.

[> Rate only turn 1](#)

**User**

Welches Wort gehört nicht zu den anderen? Reifen, Lenkrad, Auto, Motor

<b>Assistant A - Turn 1</b>	<b>Assistant B - Turn 1</b>
Auto	Auto

**Reference Solution - Turn 1**

Das Auto gehört nicht dazu, denn alle anderen sind Bestandteile eines Autos.

[> Rate only turn 1](#)

A is better     B is better     Tie     Both bad

Figure 8: 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 (Rating 9)**

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 (Rating 10)**

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 answers of the 24EU 7B model fine-tuned on Bactrian-ENDEFRITES and Bactrian-DE for *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 (Rating 7)** 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 (Rating 8)** 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 (Rating 5)** 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 (Rating 6)** 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 answers of the 24EU 7B model fine-tuned on Bactrian-ENDEFRITES and Bactrian-DE for *MT-Bench-EN*.

---

**Request (reasoning) - Turn 1** David has three sisters. Each of them has one brother. How many brothers does David have?

**Reference Solution** David has no brother. He is the one brother of his three sisters.

---

**Bactrian-DE** David has two brothers.

**Judgment** The assistant’s answer is incorrect. [...] Rating: [[2]]

**Bactrian-DEFRITES** David has six brothers.

**Judgment** The assistant’s answer is incorrect. [...] Rating: [[1]]

**Bactrian-DEFRITES-sampled** David has two brothers.

**Judgment** The assistant’s answer is incorrect. [...] Rating: [[2]]

**Bactrian-ENDEFRITES** David has one brother.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

**Bactrian-EN** David has one brother.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

**Bactrian-ENDEFRITES-sampled** David has one brother.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

---

**Request (reasoning) - Turn 2** If we change the previous question and assume that each sister of David has two brothers, how many brothers would David have?

**Reference Solution** David has one brother.

---

**Bactrian-DE** David has one brother.

**Judgment** The assistant’s response to the second question is incorrect. [...] Rating: [[1]]

**Bactrian-DEFRITES** David has fourteen brothers.

**Judgment** The assistant’s response is incorrect. [...] Rating: [[1]]

**Bactrian-DEFRITES-sampled** David would have four brothers.

**Judgment** The assistant’s response to the second question is incorrect. [...] Rating: [[2]]

**Bactrian-ENDEFRITES** David would have two brothers.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

**Bactrian-EN** David has two brothers.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

**Bactrian-ENDEFRITES-sampled** David has two brothers.

**Judgment** The assistant’s answer is correct. [...] Rating: [[10]]

---

Table 8: Qualitative comparison of Bactrian-ENDEFRITES, Bactrian-ENDEFRITES-sampled, Bactrian-DEFRITES, Bactrian-DEFRITES-sampled, Bactrian-EN and Bactrian-DE on *MT-Bench-EN*, showing judgment inconsistencies of GPT-4-as-a-judge.



MT-Bench-EN	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
24EU-1T-pre-train	1.0	1.75	1.0	1.0	1.25	1.0	1.0	1.0	1.12
Bactr.-DE	5.1	5.1	2.0	1.35	1.95	2.8	4.25	5.35	3.49
Bactr.-DEFRITES	4.92	4.0	2.05	1.05	1.85	2.75	5.15	4.15	3.24
Bactr.-DEFRITES-sampled	4.4	4.2	2.45	1.4	1.65	2.2	4.8	4.53	3.2
Bactr.-EN	4.68	5.2	3.4	1.15	2.05	2.6	5.5	6.2	3.85
Bactr.-ENDEFRITES	5.4	4.75	3.45	1.15	2.0	2.55	4.8	5.35	3.68
Bactr.-ENDEFRITES-sampled	4.6	4.95	3.2	1.05	2.1	2.45	4.65	5.4	3.55
Bactr.-ES	5.5	5.1	2.2	1.35	2.0	2.6	4.05	5.48	3.53
Bactr.-FR	5.68	4.95	3.2	1.4	2.1	2.5	4.28	5.2	3.66
Bactr.-IT	4.82	4.45	2.75	1.35	1.89	2.3	4.35	4.7	3.33
Lima-DE	2.8	3.15	1.5	1.15	1.4	1.3	1.85	1.6	1.84
Lima-DEFRITES	2.55	3.65	2.7	1.35	1.4	1.65	3.4	2.05	2.34
Lima-DEFRITES-sampled	2.4	2.45	3.1	1.1	1.1	1.05	1.75	1.25	1.78
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-ENDEFRITES-sampled	2.6	3.45	2.05	1.25	1.2	1.15	2.95	2.0	2.08
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
Bactr.-X-Avg.	5.01	4.74	2.74	1.25	1.95	2.53	4.65	5.15	3.5
Lima-X-Avg.	2.59	3.34	2.26	1.13	1.3	1.36	2.11	1.74	1.98
MT-Bench-DE	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
24EU-1T-pre-train	1.1	2.2	1.0	1.0	1.0	1.0	1.0	1.0	1.16
Bactr.-DE	5.15	4.6	3.15	1.4	1.7	2.1	4.55	5.7	3.54
Bactr.-DEFRITES	5.48	4.95	3.1	1.85	1.8	2.7	4.65	6.4	3.87
Bactr.-DEFRITES-sampled	4.97	4.15	2.8	1.5	1.9	2.45	4.92	4.58	3.41
Bactr.-EN	4.9	4.4	3.4	1.55	1.85	2.25	3.4	4.22	3.25
Bactr.-ENDEFRITES	5.6	5.35	2.95	1.65	1.85	2.55	5.05	6.0	3.88
Bactr.-ENDEFRITES-sampled	5.08	4.82	2.25	1.05	2.0	1.65	3.75	4.15	3.09
Bactr.-ES	4.75	4.6	3.5	1.5	2.05	2.0	3.6	4.08	3.26
Bactr.-FR	5.3	4.3	3.05	1.95	2.25	1.8	3.35	3.3	3.16
Bactr.-IT	4.8	4.55	2.6	1.5	2.1	2.6	3.75	3.65	3.19
Lima-DE	2.4	2.92	1.8	1.0	1.3	1.1	1.8	1.15	1.68
Lima-DEFRITES	3.5	3.58	3.2	1.0	1.1	1.4	1.7	2.25	2.22
Lima-DEFRITES-sampled	2.0	2.8	1.4	1.55	1.0	1.45	1.2	1.0	1.55
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-ENDEFRITES-sampled	2.9	2.65	2.1	1.4	1.1	1.5	2.15	1.6	1.92
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
Bactr.-X-Avg.	5.11	4.64	2.98	1.55	1.94	2.23	4.11	4.68	3.41
Lima-X-Avg.	2.76	2.92	1.76	1.22	1.17	1.27	1.65	1.38	1.77

Table 9: Single evaluation results by GPT-4 for MT-Bench-EN and MT-Bench-DE for 24EU-7B-based models.

MT-Bench-FR	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
24EU-1T-pre-train	0.95	1.1	0.79	0.68	0.95	1.0	0.74	0.75	0.87
Bactr.-DE	5.42	4.8	2.35	1.0	1.7	2.45	3.9	5.05	3.33
Bactr.-DEFRITES	5.42	4.85	2.55	1.3	2.0	2.5	5.15	6.6	3.8
Bactr.-DEFRITES-sampled	4.5	4.3	2.2	1.85	2.0	2.15	4.05	4.4	3.18
Bactr.-EN	4.25	4.3	2.9	1.35	1.65	2.25	3.9	3.95	3.07
Bactr.-ENDEFRITES	5.15	4.75	2.6	1.05	1.75	2.65	4.95	4.32	3.4
Bactr.-ENDEFRITES-sampled	4.15	3.85	2.45	1.05	1.85	3.05	3.9	3.42	2.97
Bactr.-ES	4.0	4.35	2.35	1.1	2.0	2.55	3.05	3.7	2.89
Bactr.-FR	4.55	4.9	2.55	1.65	1.8	3.0	4.75	5.65	3.61
Bactr.-IT	4.8	4.8	2.65	1.4	1.75	3.7	3.85	3.1	3.26
Lima-DE	1.65	1.55	1.7	1.0	1.1	1.1	1.75	1.6	1.43
Lima-DEFRITES	3.45	3.9	2.75	1.2	1.45	1.7	2.55	1.9	2.36
Lima-DEFRITES-sampled	1.7	1.6	1.5	1.0	1.15	1.25	1.45	1.3	1.37
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-ENDEFRITES-sampled	2.25	2.75	1.8	1.05	1.4	1.4	1.85	1.45	1.74
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
Bactr.-X-Avg.	4.69	4.54	2.51	1.31	1.83	2.7	4.17	4.47	3.28
Lima-X-Avg.	2.31	2.82	2.03	1.11	1.25	1.48	1.78	1.43	1.78
MT-Bench-IT	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
24EU-1T-pre-train	1.0	2.1	0.93	1.33	1.0	1.1	0.85	0.85	1.15
Bactr.-DE	4.95	4.6	2.65	1.7	1.8	3.15	4.5	5.5	3.61
Bactr.-DEFRITES	4.75	5.15	2.15	1.2	1.7	2.05	4.85	5.35	3.4
Bactr.-DEFRITES-sampled	4.1	3.95	2.25	1.05	1.95	2.5	4.85	5.3	3.24
Bactr.-EN	4.4	4.5	2.3	1.2	1.35	3.1	4.4	3.75	3.12
Bactr.-ENDEFRITES	4.88	5.5	2.15	1.4	2.0	2.65	6.0	5.22	3.72
Bactr.-ENDEFRITES-sampled	4.55	4.9	2.25	1.0	1.65	2.3	4.15	3.75	3.07
Bactr.-ES	4.68	5.05	2.45	1.1	1.7	2.45	3.95	5.0	3.3
Bactr.-FR	4.58	4.9	2.15	1.05	1.3	2.95	3.75	4.0	3.08
Bactr.-IT	4.85	5.1	2.55	1.1	1.65	3.0	4.9	3.95	3.39
Lima-DE	1.85	2.2	1.1	1.35	1.15	1.7	1.1	1.35	1.48
Lima-DEFRITES	2.15	3.35	2.55	1.0	1.45	1.95	2.65	1.9	2.12
Lima-DEFRITES-sampled	1.9	3.5	1.05	1.3	1.0	1.3	1.5	1.0	1.57
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-ENDEFRITES-sampled	1.85	2.9	2.4	1.0	1.2	1.6	1.35	1.5	1.72
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
Bactr.-X-Avg.	4.64	4.85	2.32	1.2	1.68	2.68	4.59	4.65	3.33
Lima-X-Avg.	2.01	2.91	1.69	1.17	1.29	1.61	1.57	1.45	1.71

Table 10: Single evaluation results by GPT-4 for MT-Bench-FR and MT-Bench-IT for 24EU-7B-based models.

MT-Bench-ES	Wr.	Role.	Reas.	Math	Code	Extr.	STEM	Hum.	Avg.
24EU-1T-pre-train	0.84	1.4	0.85	0.85	0.95	0.82	1.94	1.0	1.08
Bactr.-DE	4.5	5.18	2.6	1.05	1.6	2.95	4.5	5.6	3.5
Bactr.-DEFRITES	4.47	5.3	2.1	1.05	1.85	1.85	4.55	4.9	3.26
Bactr.-DEFRITES-sampled	4.47	5.18	1.85	1.4	1.85	2.45	4.2	4.0	3.18
Bactr.-EN	4.58	4.05	2.4	1.1	1.15	2.95	3.48	4.32	3.0
Bactr.-ENDEFRITES	4.82	5.0	2.3	1.05	2.15	2.95	5.18	5.65	3.64
Bactr.-ENDEFRITES-sampled	4.42	4.85	2.4	1.05	1.9	2.45	4.95	4.35	3.3
Bactr.-ES	4.15	4.8	3.0	1.1	2.1	2.0	5.05	5.05	3.41
Bactr.-FR	4.97	4.68	2.45	1.05	1.95	2.05	4.4	5.28	3.35
Bactr.-IT	4.72	4.15	2.4	1.15	1.65	2.1	4.5	4.25	3.12
Lima-DE	2.35	2.3	1.25	1.25	1.1	1.05	1.35	1.25	1.49
Lima-DEFRITES	3.2	3.7	1.3	1.0	1.0	1.95	2.3	2.05	2.06
Lima-DEFRITES-sampled	1.8	2.15	1.4	1.0	1.0	1.35	1.9	1.25	1.48
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-ENDEFRITES-sampled	2.65	2.85	1.65	1.05	1.0	1.25	2.95	1.75	1.89
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
Bactr.-X-Avg.	4.57	4.8	2.39	1.11	1.8	2.42	4.53	4.82	3.3
Lima-X-Avg.	2.23	2.65	1.32	1.08	1.07	1.38	1.64	1.48	1.61

Table 11: Single evaluation results by GPT-4 for MT-Bench-ES for 24EU-7B-based models.