Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed?

Tannon Kew[1*] **Florian Schottmann**[2,3] **Rico Sennrich**[1,4]
[1]University of Zurich, [2]Textshuttle, [3]ETH Zurich, [4]University of Edinburgh
`{kew,sennrich}@cl.uzh.ch, schottmann@textshuttle.com`

Abstract

The vast majority of today's large language models (LLMs) are English-centric, having been pretrained predominantly on English text. Yet, in order to meet user expectations, models need to be able to respond appropriately in multiple languages once deployed in downstream applications. This requires strong cross-lingual transfer abilities. In this work, we investigate the minimal amount of multilinguality required during finetuning to elicit cross-lingual generalisation in English-centric LLMs. In experiments across four LLMs, we find that multilingual instruction tuning with as few as two to three languages is both necessary and sufficient to elicit effective cross-lingual generalisation, with the limiting factor being the degree to which a target language is seen during pretraining. Evaluations on five different tasks further reveal that multilingual instruction tuning is most beneficial for generative tasks that assume input/output language agreement, such as in chat settings, while being of less importance for highly structured classification-style tasks.[1]

1 Introduction

Conversational instruction tuning is a popular method for aligning large pretrained language models (LLMs) with user expectations such that they can effectively respond to a user's input query and follow natural language instructions (Ouyang et al., 2022; Wang et al., 2023; Chiang et al., 2023; Zhou et al., 2023). An implicit expectation of conversational chatbots is that the language of a model's response should match that of the user's input query. For instance, unless otherwise specified, a German-language input should result in a German-language output. However, since the vast majority of training data is in English, many instruction-tuned LLMs struggle to respond consistently in other languages (Ouyang et al., 2022; Touvron et al., 2023; Chen et al., 2024; Ye et al., 2023; Zhang et al., 2024b).

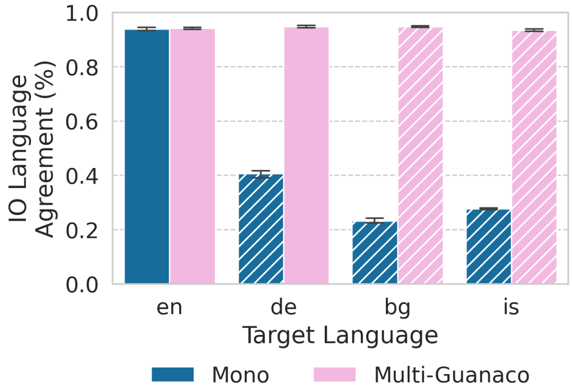


Figure 1: Input/output (IO) language agreement for English (en), German (de), Bulgarian (bg), and Icelandic (is) given English-only instruction tuning (Mono) and multilingual instruction tuning (Multi-Guanaco). Striped bars indicate that the target language is not seen during finetuning (i.e. the 0-shot setting). Error bars show a confidence interval of 95%.

Despite limited exposure, however, English-centric LLMs such as Llama 2 seemingly achieve near-perfect input/output (IO) language agreement when tuned with relatively few multilingual conversational instructions. Figure 1 depicts IO language agreement, as measured by OpenLID (Burchell et al., 2023), and compares multilingual tuning on the language-diverse Guanaco dataset (Dettmers et al., 2023; Köpf et al., 2023) with monolingual tuning on an English-only subset of instructions. As can be seen, multilingual tuning elicits strong IO language agreement – within the bounds of OpenLID's error rates – for non-English languages seen to varying degrees during pretraining and finetuning without degrading performance on English.

This observation raises two major questions, which we aim to address in this paper:

---

[*] Work partially conducted during a research internship at Textshuttle.

[1] Our code and data is available at `https://github.com/ZurichNLP/multilingual-instruction-tuning`.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 13097–13124
November 12-16, 2024 ©2024 Association for Computational Linguistics

Q1: How much multilinguality is required during finetuning to elicit 0-shot cross-lingual generalisation in English-centric LLMs?
Q2: Which languages and tasks benefit most from multilingual instruction tuning of English-centric LLMs?

To investigate these questions, we instruction tune four distinct English-centric LLMs with varying degrees of multilinguality and evaluate performance on a diverse set of up to 19 languages across five different tasks. Specifically, we consider high-, medium-, and low-resource target languages with regard to a model's pretraining data distribution and generative tasks, such as single-turn dialogue, sentence simplification, and extractive question answering, as well as more structured tasks aimed at assessing commonsense reasoning and language understanding.

Our results indicate that multilingual instruction tuning is crucial for eliciting cross-lingual transfer on generative tasks that assume IO language agreement while being less important in structured tasks that are commonly used to benchmark LLM performance. Furthermore, we empirically show that only a small number of finetuning languages is required to promote cross-lingual transfer. This highlights that tuning data for all potential target languages is not necessary to derive capable polyglot chat models from English-centric LLMs.

2 Related Work

2.1 Instruction Tuning LLMs

Unlike task-specific finetuning, instruction tuning aims to promote cross-task generalisation, allowing for a 'generalist' model that is capable of completing any text-based task on the basis of natural language instructions provided at inference time (Mishra et al., 2022; Wei et al., 2022; Wang et al., 2022; Sanh et al., 2022; Longpre et al., 2023). Meanwhile, framing instructions in a conversational manner and over multiple dialogue turns has been shown to be effective at deriving performant chat models (Taori et al., 2023; Conover et al., 2023; Chiang et al., 2023; Dettmers et al., 2023; Ding et al., 2023). Furthermore, LLM instruction tuning remains effective given relatively limited labelled data (Ouyang et al., 2022; Touvron et al., 2023; Zhou et al., 2023), parameter efficient training strategies (Hu et al., 2021; Zhang et al., 2024a) and model quantisation (Dettmers et al., 2023; Li et al., 2023b).

2.2 Cross-lingual Transfer in English-centric LLMs

The vast majority of today's publicly available LLMs are English-centric. For instance, GPT-3's training data consisted of approximately 93% English documents with the remaining 7% pertaining to other languages (Brown et al., 2020).[2] This trend is further reflected in popular open-weight LLMs (see Table 1). One potential reason for this could be the "the curse of multilinguality" (Conneau et al., 2020) which arises from having to share a finite model capacity across more languages (Lin et al., 2022; Le Scao et al., 2022). Consequently, cross-lingual transfer abilities of performant English-centric models is highly valuable.

Despite this, proprietary models such as GPT-3 and its derivatives have shown impressive performance in multilingual settings across a range of NLU and NLG benchmarks (Lai et al., 2023a; Holmström et al., 2023; Armengol-Estapé et al., 2022; Hendy et al., 2023; Lu et al., 2023; Jiao et al., 2023; Bang et al., 2023; Laskar et al., 2023). Meanwhile, there is a growing body of research dedicated to studying similar capabilities of open-weight LLMs (Bawden and Yvon, 2023) and the benefits of multilingual tuning (Ye et al., 2023; Muennighoff et al., 2023). For instance, large-scale multilingual instruction tuning has been shown to improve performance on open-ended chat tasks in multiple target languages for both English-centric and multilingual LLMs (Chen et al., 2024; Li et al., 2023a; Weber et al., 2024). In concurrent work to ours, Shaham et al. (2024) report that this holds given minimal multilingual instruction tuning with the multilingual PaLM 2 model (Anil et al., 2023).

Our work closely relates to that of Chen et al. (2024) and Shaham et al. (2024) insofar as we explore how multilingual instruction tuning can improve cross-lingual transfer in LLMs. However, in contrast to these works, we focus on the *minimal amount of multilinguality* needed to elicit cross-lingual transfer in *open-weight English-centric LLMs*, and we evaluate on a range of tasks and target languages that cover the full distribution of a model's pretraining data.

3 Experimental Setup

To explore the multilingual capabilities of English-centric LLMs, we instruction tune a series of mod-

---

[2] `https://github.com/openai/gpt-3/blob/master/dataset_statistics`

els on a fixed-size set of examples, varying the number of languages available. Following this, we evaluate the resulting models in multiple target languages on five distinct tasks that are representative of how LLMs may be used in downstream applications.

3.1 English-centric LLMs

A prevailing trend in the development of recent LLMs is a clear focus on scaling up the size of the pretraining corpus (Hoffmann et al., 2022; Sardana and Frankle, 2023). For instance, open-weight LLMs such as Falcon (Almazrouei et al., 2023) and Llama 2 (Touvron et al., 2023) were pretrained on 1.5 and 2 trillion tokens, respectively. Yet, while this far surpasses the 300 billion tokens used to train GPT-3 (Brown et al., 2020), the distribution of language coverage remains similar with more than 90% pertaining to English (see Table 1). For our main experiments, we focus on Llama 2 7b. We then consider Llama 2 70b (§5.2) to study the effect of model scaling on multilinguality. Additionally, we also test whether our findings generalise to other LLMs using Falcon 7b and Llama 3 8b (MetaAI, 2024), which employ different pretraining approaches (see Appendices F and G).

3.2 Instruction-tuning Data

For instruction tuning, we take inspiration from Dettmers et al. (2023) and finetune on high-quality conversations from the OpenAssistant dataset (Köpf et al., 2023). These conversations comprise multiple dialogue turns between crowdworkers who were asked to either interact with or assume the role of a helpful AI assistant. In contrast to Dettmers et al. (2023), who use all 9,846 top-rated conversations to train their 'Guanaco' models, we subsample training instances from the Guanaco dataset in order to control the amount of multilinguality. Specifically, we sample 3,200 unique English instances as an initial monolingual dataset, which we refer to as 'Mono'. To construct datasets for multilingual finetuning we sample 200 unique training examples from each of the five most frequent non-English languages in Guanaco (Spanish, Russian, German, Chinese, and French). Given these subsets, we incrementally substitute English examples in Mono for non-English ones, one language at a time, following the order of how frequently each language appears in Guanaco. The resulting multilingual datasets are denoted as Multi-i, where i indicates the number of distinct languages included. Table 3 provides a detailed summary of the makeup of these datasets. For comparison, we also train models on the full Guanaco dataset, which includes more than 30 distinct languages.

3.3 Instruction-tuning and Inference Settings

For instruction tuning, we train LoRA adapters ($R = 64$, $\alpha = 16$) (Hu et al., 2021) using Hugging Face's TRL library[3]. All models are trained for 2k update steps on sequences of 1024 tokens using an effective batch size of 64 and a learning rate of $1\text{e}{-5}$.[4]

For efficient inference, we employ vLLM (Kwon et al., 2023). We use nucleus sampling (Holtzman et al., 2019) ($p = 0.9$) with a temperature of 0.8 for open-ended generation tasks and a temperature of 0.001 for the more constrained QA-style tasks to encourage more deterministic behaviour. For all tasks, we report the mean performance across three inference runs using different random seeds.

3.4 Evaluation Tasks and Languages

As evaluation tasks, we consider single-turn dialogue, sentence simplification, extractive question answering, commonsense reasoning and natural language inference. Since all of these tasks differ in terms of the availability and representation of ground-truth labels, we describe the specific evaluation strategies in the following section.

We select target languages based on the makeup of Llama 2's pretraining data (see Table 1). In doing so, we aim to study the effect of multilingual instruction tuning in both supervised and 0-shot cross-lingual settings (Wu and Dredze, 2019), and report results on high-, medium-, and low-resource languages with regard to the pretrained LLM, covering distinct language families and scripts. German (de), French (fr), Swedish (sv), Chinese (zh),[5] Spanish (es), Russian (ru), and Italian (it) represent **high-resource** languages with an estimated 3.4 to 2.2 billion tokens. Portuguese (pt), Vietnamese (vi), Korean (ko), Finnish (fi), Hungarian (hu), Norwegian (no),[6] Bulgarian (bg), and Slovenian (sl) represent **medium-resource** languages, with an estimated 1.8 to 0.2 billion tokens. And finally,

---

[3] `https://github.com/huggingface/trl`

[4] Details on hardware are provided in Appendix C.

[5] Note, Touvron et al. (2023) do not distinguish between Simplified and Traditional Chinese in their reporting. For our purposes, we explicitly use Simplified Chinese.

[6] Again, lacking specification between Norwegian Bokmål and Nynorsk in Touvron et al. (2023), we explicitly consider the former.

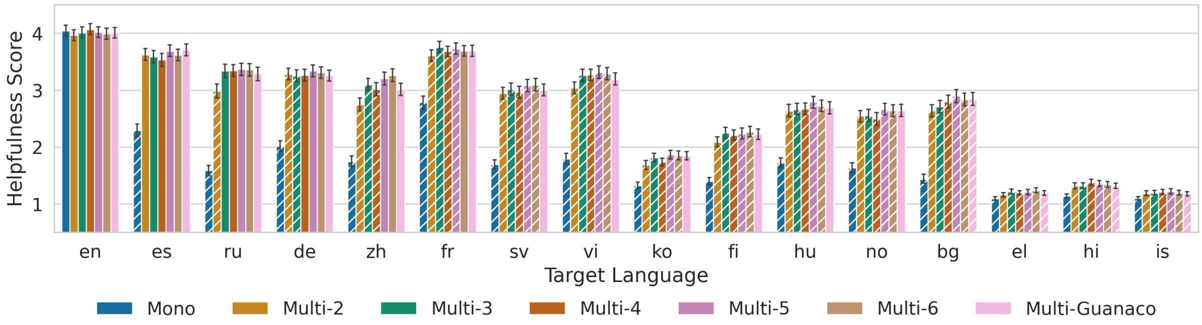


Figure 2: Average helpfulness of single-turn dialogue responses from Llama 2 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

we select Greek (el), Hindi (hi), and Icelandic (is) as **low-resource** languages, whose frequency in Llama 2's pretraining data is not known but are likely appear in small amounts due to contamination (Blevins and Zettlemoyer, 2022). Unfortunately, existing benchmark evaluation datasets do not cover all of these languages and thus we limit the target languages in those tasks to the available subset of our target languages.

4 Experiments

4.1 Single-turn Dialogue

General-purpose chatbots are a popular application of instruction-tuned LLMs. To evaluate performance in this type of setting, we make use of the AlpacaEval prompt dataset (Dubois et al., 2023), which includes a diverse set of prompts for open-ended questions, creative writing, brainstorming, and other tasks. We randomly sample 300 prompts and translate these into each target language. As a translation engine, we follow Lai et al. (2023b) and use GPT-3.5-Turbo.[7] In contrast to using dedicated translation systems, employing GPT-3.5-Turbo for this purpose has the advantage of being able to explicitly specify instructions that allow for preserving code blocks, tables, and terminology, which we include as part of our translation prompt (see Figure B.1). Furthermore, since GPT-3.5-Turbo is trained on instruction- and conversational-style data, we expect it to perform well at translating in this domain.

To automatically evaluate open-ended responses, we leverage LLM-as-a-judge (Zheng et al., 2023). Following (Zhou et al., 2023), given an input prompt and model's response, we ask GPT-3.5-Turbo[8] to grade the *helpfulness* of the response on a 6-point Likert scale (see Figure 20 for the prompt used). For each evaluation instance, we provide the prompt and model-generated response *directly* in the target language, which we found to be on par with evaluating via first translating responses into English (cf. Hada et al., 2024, see Appendix E for more details). As noted by Chen et al. (2024), GPT-3.5-Turbo sometimes ignores the fact that the output language differs from the input language. Therefore, we force a score of 1 (indicating least helpful) if the language of the response does not match the intended target language according to OpenLID (Burchell et al., 2023).

Results Figure 2 shows the helpfulness scores assigned by the LLM judge on 16 target languages. For English, performance remains uniform across all multilingual instruction tuning settings. For high- and medium-resource non-English target languages, performance increases significantly when moving from monolingual to bilingual instruction tuning (Multi-2). Meanwhile, performance plateaus when training with as few as three languages (Multi-3), with no substantial differences observed between supervised and 0-shot settings. For low-resource languages, performance remains low despite multilingual finetuning. Manual inspection reveals these outputs, while initially convincing and sufficient for language identification (see Figure 1), are mostly nonsensical. These results indicate that instruction tuning with as few as two to three languages is necessary and sufficient to elicit 0-shot cross-lingual generalisation among high- and medium-resource languages on this task, with language exposure during pretraining being the main limitation to this generalisation ability.

---

[7] To translate prompts from AlpacaEval, we use `gpt-3.5-turbo-0613`.

[8] For evaluation, we use `gpt-3.5-turbo-1106` due to its longer context window.

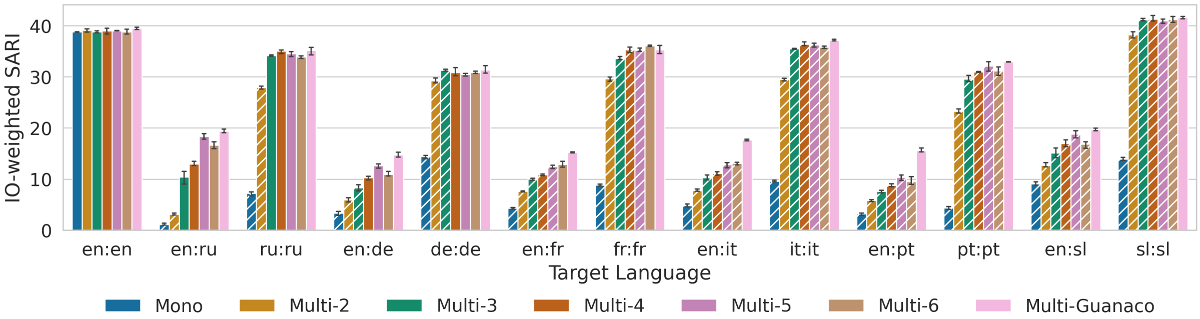


Figure 3: SARI weighted by IO language agreement for Llama 2 7b given incremental multilingual instruction tuning. Results are shown for both cross-lingual prompting (en:xx) and monolingual prompting (xx:xx). Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

4.2 Sentence Simplification

Sentence simplification aims to make complex sentences easier to read and understand. Due to a lack of high-quality supervised training data, automatic sentence simplification remains a challenging task (Štajner, 2021), and thus stands to benefit from the few- and 0-shot generalisation capabilities of instruction-following LLMs. To assess performance on this task, we use the MultiSim benchmark (Ryan et al., 2023), which includes sentence-aligned parallel datasets in multiple languages. Since the individual datasets in MultiSim are taken from distinct sources, the amount of available data varies across languages. We sample 1,371 complex-simple sentence pairs for each language, except for Slovenian, where we use all 939 instances available (see Appendix D for details).

To prompt models for sentence simplification, we repurpose the detailed instructions given to crowdworkers for the creation of the ASSET corpus (Alva-Manchego et al., 2020), which has been shown to be effective in few-shot settings (Kew et al., 2023). In its original form, this prompt explicitly states that the output should be suitable for “learners of English”. We remove this language specification so that the model is not explicitly instructed on the target language and instead is expected to infer it from the complex input sentence provided. To translate the prompt into each of our considered target languages, we used a free-tier machine translation service.[9] We also consider both a cross-lingual prompt setting (en:xx) and a monolingual prompt setting (xx:xx). The former presents the task instruction in English and provides the complex source sentence in the relevant target language, while the latter is entirely in the target language. An example of the prompt template is shown in Table 4.

For evaluation, we report SARI (Xu et al., 2016) as an indicator of simplicity and adequacy given the reference simplification. To account for IO language agreement, we weight a model’s corpus-level SARI scores by the percentage of IO language agreement it achieved. This penalises a model if it consistently produces outputs in a language that does not match the source sentence. For example, if the model generates English outputs given German input sentences at a rate of 50%, the final score is reduced by half.

Results Figure 3 shows the performance according to IO-weighted SARI on each of the target languages considered given incremental multilingual instruction tuning. Similar to the results on the single-turn dialogue task, performance remains uniform for English as the ratio of multilingual instructions increases. For all non-English target languages, we observe a large discrepancy between the cross-lingual prompting strategy and the monolingual prompting strategy. While the former tends to improve as the number of languages increases, it fails to match that of the monolingual prompting strategy under all conditions. In contrast, the results for the monolingual prompting strategy closely resemble those of the single-turn dialogue task across all available target languages: performance gains are most pronounced when moving from monolingual to bilingual instruction tuning, and they generally plateau with as few as three instruction tuning languages.

4.3 Extractive Question Answering

In contrast to open-ended questions commonly used to query LLMs in chat settings, extractive

---

[9] `https://www.deepl.com/translator`

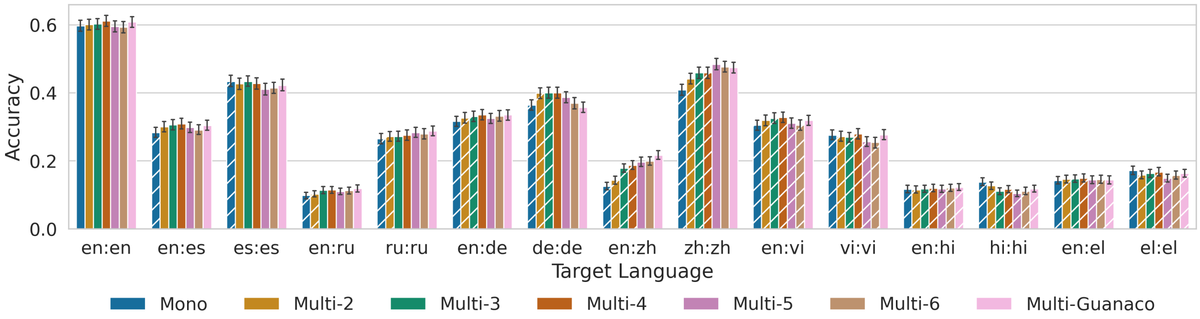


Figure 4: XQuAD results for Llama 2 7b given incremental multilingual instruction tuning. Results are shown for both cross-lingual prompting (en:xx) and monolingual prompting (xx:xx). Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

question answering requires the model to identify relevant answer spans within longer context passages provided as part of the prompt. Such a task closely resembles a retrieval augmented generation (RAG) setting, which is a popular method for extending an LLM's knowledge with additional data not available during training (Lewis et al., 2020; Izacard and Grave, 2021). To evaluate our models on this task in multiple target languages we use XQuAD (Artetxe et al., 2020).[10]

As a starting point, we borrow the English prompt from Lai et al. (2023a) and manually translate it into each of the target languages considered. Additionally, we include a standardised response prefix as part of the prompt, effectively force-decoding the response "Based on the passage, the answer to the question is". This allows us to better isolate the relevant answer string in the generative model's output. A response is considered correct if the ground truth answer can be matched with the beginning of the model's response after minimal post-processing.[11] Again, we consider both cross-lingual (en:xx) and monolingual (xx:xx) prompting strategies. An example of our prompting strategy for this task and model outputs is shown in Table 5.

Results Figure 4 shows the performance on each of our target languages in XQuAD given incremental multilingual instruction tuning. Again, we observe consistent performance for English under multilingual instruction tuning settings. Meanwhile, performance on high-resource languages such as German and Chinese exhibits marginal gains when moving from monolingual to bilingual instruction tuning and plateaus with relatively few languages. For low-resource languages (e.g., Hindi and Greek), performance remains consistently low across all settings, indicating that the model's ability to generalise to these languages is limited by the lack of exposure during pretraining. Comparing the cross-lingual and monolingual prompting strategies, we observe that the latter generally leads to better performance. This agrees with the results from the sentence simplification task (§4.2) and highlights the benefit of prompting under an IO language agreement assumption. While performance gains on this task are generally less pronounced than on the previous tasks considered, we note that 0-shot extractive QA is inherently challenging for LLMs tuned on conversational instructions as they tend to generate verbose responses rather than the single word or entity that correctly matches the ground truth.

4.4 Commonsense Reasoning

'Commonsense knowledge' is the term frequently used to refer to the set of general facts that reflect how concepts can relate to one another in the real world (Li et al., 2022; Talmor et al., 2019). Effectively understanding natural language requires some representation of these concepts and relationships, making commonsense reasoning a key skill for LLMs. To evaluate how well English-centric models can apply commonsense reasoning across languages, we use the X-CSQA dataset (Lin et al.,

---

[10] We report results measured on the validation split of XQuAD since labels for the test split are not publicly available. This provides 1,190 QA pairs that were professionally translated into different languages.

[11] We find that some post-processing of model outputs is necessary for certain languages. Specifically, when queried with German and Russian prompts, models consistently repeated the question before providing the extracted answer. To handle such cases, we strip away the repeated question and truncate the response to a maximum of 50 characters or the first line break, whichever comes first.

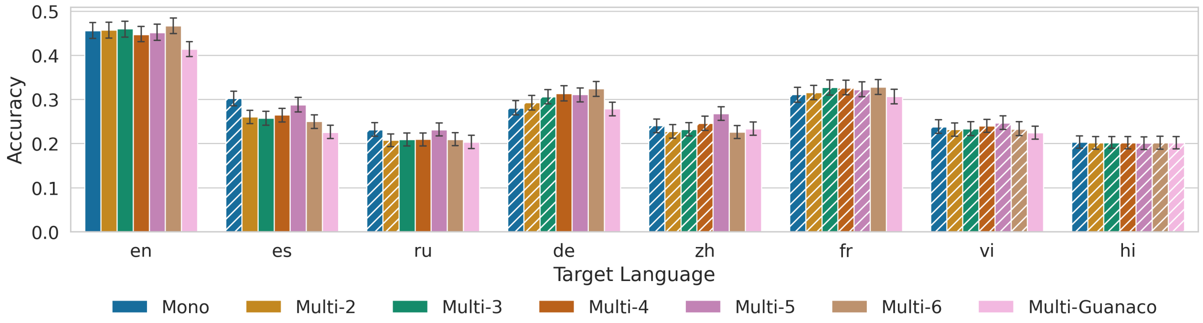


Figure 5: X-CSQA results for Llama 2 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

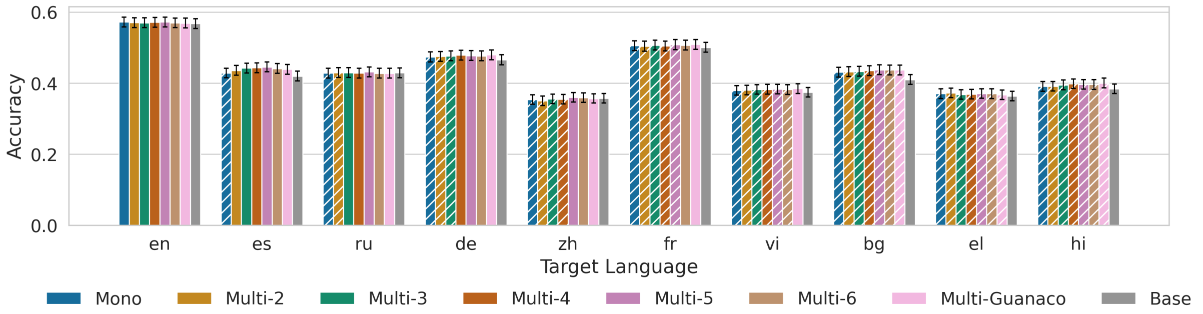


Figure 6: XNLI results for Llama 2 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

2021).[12] This dataset contains questions paired with multiple choice answers which aim at assessing general, language-agnostic world knowledge involving different types of commonsense reasoning. Given a question and a set of five possible answers from A-E, we prompt the model to output the letter corresponding to the most suitable answer. Again, we borrow the English prompt template for this task from Lai et al. (2023a) as a starting point and translate it into each target language. For this task we consider only the monolingual prompting strategy (xx:xx), which proved most effective in the previous tasks. An example of the prompt used is given in Table 6.

Results Figure 5 shows the accuracy on X-CSQA given incremental multilingual instruction tuning. Again, for English, we observe that multilingual instruction tuning does not significantly degrade performance. However, in contrast to the previous tasks considered, we do not observe any consistent gains for non-English target languages on the basis of multilingual instruction tuning, suggesting that there are limitations on what tasks this benefits.

4.5 XNLI

Natural language inference (NLI) is an important skill for LMs, as it ensures textual coherence, which is particularly critical as input and output sequences grow in length. Given two sentences, this task aims to recognise a relationship between them either as *entailment*, *contradiction*, or *neutral*. To assess an LLM's ability to solve this task given multilingual instruction tuning, we use XNLI (Conneau et al., 2018) and evaluate performance on the official test split.

For this task, we use the implementation in the LM Evaluation Harness (Gao et al., 2023).[13] Instead of querying the model to generate the desired label, we use rank classification (Sanh et al., 2022; Brown et al., 2020), which allows for direct comparison to the base model. In this setting, multiple queries are constructed for each test instance given a predefined template and subsequently scored by the model. The most probable query sequence under the model is then chosen as the final answer. Prompt templates for this task are language-specific making it equivalent to monolingual prompting

---

[12] We report results for X-CSQA measured on the validation set of 1,000 questions.

[13] `https://github.com/EleutherAI/lm-evaluation-harness`

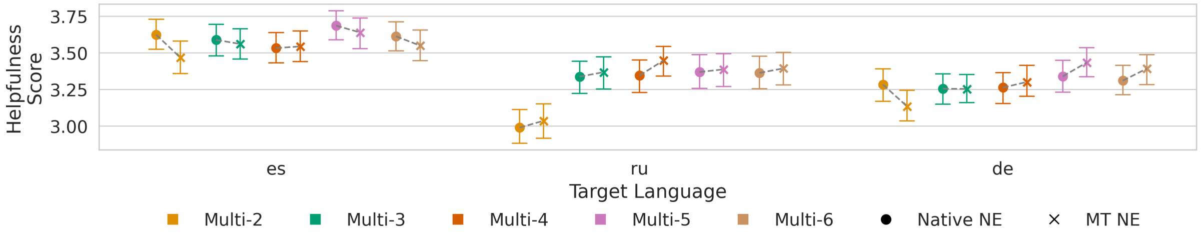


Figure 7: Comparison of Llama 2 7b given multilingual instruction tuning using native non-English (Native NE) examples vs. translated non-English (MT NE) examples. Error bars show a 95% confidence interval.

(xx:xx). An example of this task is provided in Table 7.

Results Figure 6 shows the accuracy measured on our target languages from XNLI. Strikingly, performance for all target languages remains uniform regardless of the amount of multilinguality used during finetuning. Furthermore, the comparable performance across all training settings suggests that, in general, instruction tuning fails to provide measurable gains in this challenging task.

5 Further Analysis

Our experimental results on five distinct tasks demonstrate that multilingual instruction tuning benefits open-ended single-turn dialogue and sentence simplification in non-English languages most strongly. In this section, we focus on single-turn dialogue to investigate the impact of instruction diversity vs. language diversity and model scaling.

5.1 Instruction Diversity vs. Language Diversity

Diversity is a key factor for sample efficient instruction tuning (Zhou et al., 2023). Since our experiments in §4 make use of native non-English training instances from the OpenAssistant dataset, a potential confounding factor could be that adding more languages during finetuning also introduces more diverse training instructions. To investigate this, we retrain Multi-i models using translated instruction tuning data from Mono in place of native non-English examples, following the same incremental recipe as described in §3.2. This ensures that the data distribution remains constant as multilinguality increases. As a translation engine, we again use GPT-3.5-Turbo[14] and the prompt template provided in Figure 10.

Figure 7 compares the resulting performance on the single-turn dialogue task for a subset of our target languages (results for the remaining target languages are provided in Figure 19). Notably, we observe no significant differences between tuning with distinctly native non-English examples compared to those derived via automatic translation, indicating that gains attributed to increased multilinguality are not conflated with an increase in the diversity of instructions. Following this, in Appendix H, we dive deeper to investigate the role of language diversity given a fixed budget of non-English training examples. From these ablation experiments, we find that the impact of increasing the number of languages is stronger than that of increasing the number of examples for a given language.

5.2 Scaling up Model Size

In order to investigate the effect of model scaling, we repeat our single-turn dialogue experiments using Llama 2 70b as the underlying base model. Figure 8 shows the resulting helpfulness scores assigned by the LLM judge, with a direct comparison to Llama 2 7b. Most notably, performance on high- and medium-resource non-English target languages is dramatically improved, often matching that of English. However, despite these gains, multilingual instruction tuning with two languages remains essential to elicit cross-lingual generalisation on this task, further underpinning our main findings. In addition, we observe that the larger model's performance on most non-English target languages tends to plateau with just two finetuning languages, unlike the smaller model that typically required three. Finally, while the performance on low-resource languages remains poor, it exhibits a substantial relative improvement compared to the 7-billion-parameter model. These results indicate that model scaling is extremely beneficial for exploiting the multilingual capabilities in English-centric models

---

[14] Since conversational training instances can be quite long, sometimes exceeding the default 4k token context window, we use `gpt-3.5-tubo-1106` for this translation task due to its longer context window (16k).

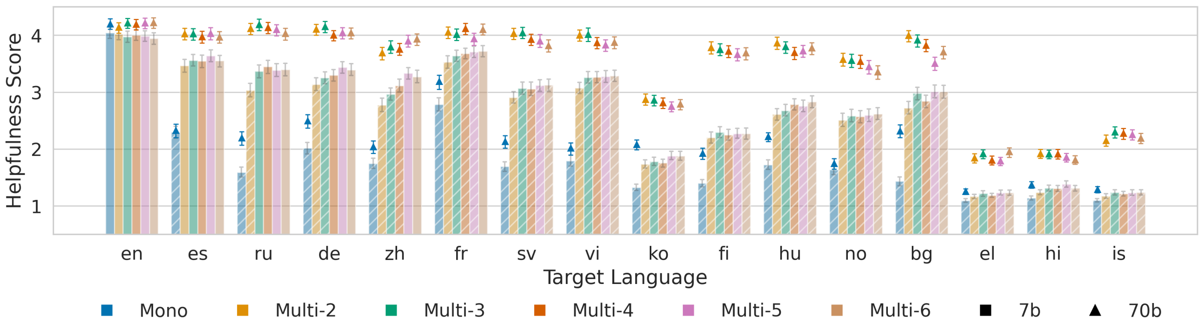


Figure 8: Average helpfulness of single-turn dialogue responses from Llama 2 70b (denoted by triangular points) and Llama 2 7b (semi-transparent bars) given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

and aligns with findings from Shaham et al. (2023).

6 Discussion

Our findings show that multilingual instruction tuning aids cross-lingual transfer in English-centric LLMs, though its effectiveness on downstream performance varies across tasks and the degree to which a specific target language is seen during pretraining. For high- and medium-resource languages, we observe substantial performance gains on single-turn dialogue and sentence simplification with monolingual prompting, as well as marginal gains on extractive QA with monolingual prompting. Meanwhile, for low-resource languages and highly structured tasks that impose a strict constraint on the output space regardless of the input language (e.g., multiple choice QA), multilingual instruction tuning has little impact. This distinction highlights that multilingual instruction tuning is most beneficial for more open-ended generative tasks that assume IO language agreement, and crucially, that this benefit is limited to high- and medium-resource languages with respect to model's pretraining language distribution.

Most surprisingly, we find that **multilingual instruction tuning with just two to three languages is necessary and sufficient to promote this generalisation** in all four of the English-centric LLMs considered. Furthermore, we consistently see that performance tends to plateau rather quickly and adding more instruction tuning languages – including the target language itself – typically provides no significant gains. This underscores that the performance gains are largely the result of cross-lingual transfer and suggests that there may be diminishing returns associated with scaling up multilingual instruction tuning beyond just a handful of languages.

We posit that our findings align with the *superficial alignment hypothesis* put forward by Zhou et al. (2023), which states that a model acquires its knowledge and abilities during pretraining, while instruction tuning simply guides the model towards a desirable 'subdistribution of formats' to use when prompted. We refine this hypothesis by adding that just a small amount of multilingual instruction data also encourages the model to learn a simple mapping between input and output language. This guides the model towards **language-specific subdistributions**, leading to better performance on tasks where IO language agreement is required.

7 Conclusion

We investigated the minimal amount of multilingual instruction tuning required to elicit multilingual capabilities of four distinct English-centric LLMs. Our results show that finetuning with as few as two to three languages is necessary and sufficient to promote effective cross-lingual transfer, allowing models to better exploit the relatively small amounts of non-English data seen during pretraining. Experiments on five distinct tasks revealed that this can lead to significant performance improvements for high- and medium-resource languages on open-ended generative tasks that assume input/output language agreement.

While the effectiveness of cross-lingual transfer is indeed good news for LLM developers, future work could explore methods to reduce the performance gap between English and non-English languages, particularly for smaller models, as well as investigating tasks for which language-specific instruction tuning may be of greater importance, such as tasks involving cultural awareness.

Limitations

Our experimental results demonstrate that a relatively small amount of multilingual instruction tuning data can elicit highly valuable cross-lingual transfer, leading to improved performance in open-ended generation tasks that assume IO language agreement. However, we have not explored tasks involving more complicated mappings between languages, such as tasks involving extensive cross-lingual processing or translation.

In this work we rely on automatic language identification to construct multilingual training data and in evaluating model outputs. To this end, we employed the OpenLID model from Burchell et al. (2023). Despite low error rates achieved by this model, language identification is not perfect and can lead to some texts being misidentified. To mitigate the risk of unintentional language contamination (Blevins and Zettlemoyer, 2022) in our finetuning datasets we include training examples whose language is identified with a confidence threshold $\geqslant 0.8$.

In §5.1 we investigate the impact of multilingual diversity versus training example diversity. While our findings reveal that there is no significant difference between these two settings, we note that even when finetuning with the original native non-English examples, task diversity may be inherently limited by design of the data collection. For instance, regardless of the language used, crowdworkers were asked to follow the same set of guidelines[15] when creating the data.

Finally, to evaluate the quality of model-generated responses in the single-turn dialogue task, we rely on LLM-based evaluation methods. While a number of works have shown that LLM-based evaluation provides a decent proxy for assessing the quality of dialogue responses, achieving strong correlations with human judgements (Liu et al., 2023; Kocmi and Federmann, 2023; Zheng et al., 2023), we stop short of empirically establishing this agreement on our own model outputs. To ensure the validity of model outputs we manually assessed the adequacy of randomly sampled responses in three target languages (en, es, and de). This assessment revealed a similar trend to our LLM-based evaluation, namely that Spanish-language outputs were generally slightly worse than their English-language counterparts but slightly better than those in German.

Ethical Considerations and Risks

This work aims to evaluate multilingual capabilities of English-centric LLMs. In doing so, we acknowledge potential ethical considerations and risks associated with our research. Firstly, LLMs have been shown to inadvertently perpetuate biases present in their training data, which can lead to unexpected and unfair outcomes when these models are used in real-world applications. Therefore, measures must be taken to minimise this risk (e.g., using additional alignment strategies, rigorous testing in multilingual settings) before deploying public-facing models. Secondly, when building on top of English-centric LLMs, there is a risk of cultural homogenization, where nuances and diversity of different languages could potentially be lost due to their under-representation in the training data.

Acknowledgements

We sincerely thank our friends and colleagues, Fabian Aiolfi, Thea Bächler, Anastassia Shaitarova, Finnur Ágúst Ingimundarson, Cui Ding, Andrianos Michail, Janine Aeberhard, Arnisa Fazla, Farhad Nooralahzadeh, Omnia Ibrahim, Xuan Lam, Manasi Muglikar, Anne Göhring, and Alessia Battisti for helping us with language-specific questions and validating translations used in our experiments, as well as Lena Sophia Bolliger and Patrick Haller for providing valuable feedback on earlier versions of this paper. Rico Sennrich acknowledges funding by the Swiss National Science Foundation (project MUTAMUR; no. 213976).

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, California. Association for Computational Linguistics.

---

[15] `https://projects.laion.ai/Open-Assistant/docs/guides/guidelines`

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

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

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, Austin, Texas. Association for Computational Linguistics.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations.

Anna Dmitrieva, Antonina Laposhina, and Maria Lebedeva. 2021. A quantitative study of simplification strategies in adapted texts for L2 learners of Russian. In *Proceedings of the International Conference "Dialogue"*, pages 191–203.

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

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Sabina Gorenc and Marko Robnik-Šikonja. 2022. Slovene text simplification dataset SloTS. Slovenian language resource repository CLARIN.SI.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Oskar Holmström, Jenny Kunz, and Marco Kuhlmann. 2023. Bridging the resource gap: Exploring the efficacy of English and multilingual LLMs for Swedish. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 92–110, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, 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, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. 2023b. Loftq: Lora-fine-tuning-aware quantization for large language models.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*, pages 2511–2522, Singapore. 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.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

MetaAI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Martina Miliani, Serena Auriemma, Fernando Alva-Manchego, and Alessandro Lenci. 2022. Neural readability pairwise ranking for sentences in Italian administrative language. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 849–866, Online only. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for German language. arXiv preprint 1904.07733.

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

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue*, pages 607–617.

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

Nikhil Sardana and Jonathan Frankle. 2023. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws.

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

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2304–2317, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Sara Tonelli, Alessio Palmero Aprosio, and Francesca Saltori. 2016. SIMPITIKI: A simplification corpus for Italian. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it) and the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Napoli, Italy. CEUR-WS.

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

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions?

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

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

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and

Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. 2024a. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024b. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Pretraining Data for English-centric LLMs

With the notable exception of BLOOM (Le Scao et al., 2023) and the recent OLMo models (Groeneveld et al., 2024), very limited information is offered about the distribution of languages represented in datasets used to train LLMs and the subsequent performance on non-English languages. Table 1 provides an overview of the document-level language distributions of the LLMs used in this paper. For Llama 2, the information is taken from the original paper (Touvron et al., 2023), in which the authors analyse the training data using a fastText language classifier on corpus documents with a threshold of 0.5. For GPT-3, we use the official dataset statistics made available on GitHub which provide document-level language identification information.[16] To gather statistics for Falcon, we inspected a sample of the RefinedWeb dataset (Penedo et al., 2023) that was constructed to train this family of models.[17] Using the OpenLID fastText model from Burchell et al. (2023), we identify the most frequent document-level languages based on approximately 320 million examples from this corpus. Following (Touvron et al., 2023), languages are identified based on a confidence threshold of 0.5 and predictions below this threshold are aggregated under ‘unknown’. Note, that while this language distribution has a higher concentration of English data than both Llama 2 and GPT-3, Almazrouei et al. (2023) combine the RefinedWeb with additional curated corpora to train the Falcon models, and thus the true language distributions for this model may differ from these estimates.

B Translating Training and Evaluation Data

As part of our experiments and evaluation, we generate translations of existing datasets. All original datasets are licensed under the Apache License 2.0. Since translations are generated with GPT-3.5-Turbo, OpenAI’s usage policy applies to the resulting versions, which we make available for future research.

B.1 AlpacaEval Prompts

For the evaluation of single-turn dialogue in non-English languages, we translate AlpacaEval prompts from English into each of our considered target languages using GPT-3.5-Turbo and the template in Figure 9. Manual inspection of a sampled subset of these translated prompts in various languages revealed that the translations were typically decent, although often included literal translations for metaphorical expressions rather than how a native speaker might express themselves. For instance, English ‘bullet points’ was translated literally into Hindi, rather than an arguably more appropriate phrasing such as ‘important points’. Additionally, in languages that distinguish between formal, informal, or gendered pronouns (e.g., German, French, etc.), the formal and male forms are dominant. While these characteristics may be not truly representative of how native speakers actu-

---

[16] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv

[17] https://huggingface.co/datasets/tiiuae/falcon-refinedweb

Llama 2			Falcon			GPT-3		
#	Language	Percent	#	Language	Percent	#	Language	Percent
1	**English**	89.70%	1	**English**	95.751%	1	**English**	93.69%
2	unknown	8.38%	2	unknown	3.902%	2	**German**	1.20%
3	**German**	0.17%	3	Standard Malay	0.045%	3	**French**	1.02%
4	**French**	0.16%	4	Swahili	0.038%	4	**Portuguese**	0.64%
5	**Swedish**	0.15%	5	**French**	0.019%	5	**Italian**	0.58%
6	**Chinese**	0.13%	6	Sardinian	0.014%	6	**Spanish**	0.51%
7	**Spanish**	0.13%	7	**Portuguese**	0.014%	7	Dutch	0.37%
8	**Russian**	0.13%	8	Dutch	0.01%	8	Polish	0.25%
9	Dutch	0.12%	9	Chinese (Traditional)	0.013%	9	Japanese	0.25%
10	**Italian**	0.11%	10	**Spanish**	0.010%	10	Danish	0.16%
11	Japanese	0.10%	11	**Italian**	0.010%	11	**Norwegian**	0.15%
12	Polish	0.09%	12	**German**	0.009%	12	Romanian	0.13%
13	**Portuguese**	0.09%	13	Danish	0.008%	13	**Finnish**	0.13%
14	**Vietnamese**	0.08%	14	Indonesian	0.007%	14	**Chinese (Simplified)**	0.12%
15	Ukranian	0.07%	15	Somali	0.006%	15	**Russian**	0.11%
16	**Korean**	0.06%	16	**Swedish**	0.006%	16	Czech	0.10%
17	Catalan	0.04%	17	**Russian**	0.006%	17	**Swedish**	0.06%
18	Serbian	0.04%	18	Venetian	0.005%	18	**Hungarian**	0.06%
19	Bahasa Indonesian	0.03%	19	**Vietnamese**	0.005%	19	Chinese (Traditional)	0.04%
20	Czech	0.03%	20	Northern Uzbek	0.005%	20	Bahasa Indonesian	0.04%
21	**Finnish**	0.03%	21	Limburgish	0.005%	21	Croation	0.04%
22	**Hungarian**	0.03%	22	Tagalog	0.005%	22	Turkish	0.04%
23	**Norwegian**	0.03%	23	**Chinese (Simplified)**	0.005%	23	Catalan	0.03%
24	Romanian	0.03%		⋮		24	**Vietnamese**	0.03%
25	**Bulgarian**	0.02%	27	**Norwegian Bokmal**	0.003%	25	**Slovenian**	0.03%
26	Danish	0.02%		⋮		26	Estonian	0.02%
27	**Slovenian**	0.01%	33	**Finnish**	0.003%	27	Slovak	0.02%
28	Croatian	0.01%		⋮		28	**Korean**	0.02%
	⋮		35	**Greek**	0.002%		⋮	
n/a	**Greek**	n/a		⋮		31	**Greek**	0.017%
	⋮		46	**Hungarian**	0.002%		⋮	
n/a	**Hindi**	n/a		⋮		45	**Hindi**	0.004%
	⋮		51	**Slovenian**	0.001%		⋮	
n/a	**Icelandic**	n/a		⋮		47	**Icelandic**	0.004%
	⋮		68	**Bulgarian**	0.001%		⋮	
				⋮		48	**Bulgarian**	0.003%
			78	**Icelandic**	0.001%		⋮	
				⋮				
			86	**Korean**	0.0003%			
				⋮				
			102	**Hindi**	0.0001%			

Table 1: Distribution of document languages in pretraining datasets for popular English-centric LLMs. For Llama 2 and GPT-3, statistics are taken from the original works. For Falcon, statistics are estimated based on a sample of the RefinedWeb corpus (roughly 320 million documents). Note that for Llama 2 and Falcon, 'unknown' indicates texts identified below a predefined confidence threshold of 0.5. Touvron et al. (2023) state that this data partially pertains to programming code. Target languages considered in our experiments are highlighted in bold.

ally interact with LLMs, we consider these to be potentially valid queries for general-purpose chat models.

B.2 Guanaco Training Examples

To investigate the effect of language diversity compared to instruction diversity (§5.1), we translate a subset of Guanaco's English training examples into non-English target languages and use these to create MT-based Multi-i instruction-tuning datasets. By default, speaker roles in Guanaco are denoted with '### Human:' and '### Assistant'. To ensure that these are never translated and the dialogue structure is maintained, we substitute them with special tokens '<S1>' and '<S2>' and explicitly tell the model to leave these tokens intact (see Figure 10). Before training, we map the special placeholder tokens back to their original form.

```
You are a helpful assistant.

Translate the following text into
{{target_language}}.
Keep the structure of the original text and
preserve things like code and names.
Please ensure that your response contains
only the translated text.
The translation must convey the same meaning
as the original and be natural for native
speakers with correct grammar and proper word
choices.
Your translation must also use exact
terminology to provide accurate information
even for the experts in the related fields.

Original:

{{source_text}}

Translation into {{target_language}}:
```

Figure 9: Prompt template used to translate AlpacaEval with GPT-3.5-Turbo. Values in curly braces represent placeholders. The value specified for 'target language' is the English name of the language (e.g., 'German') we are translating into, except for Greek and Chinese, where we explicitly specify standard modern Greek and Mandarin Chinese, respectively.

```
You are a helpful assistant.

Translate the following conversation
between a human and an AI assistant
into {{target_language}}.
Keep the structure of the original text
and preserve things like code, names
and role labels (e.g. <S1>, <S2>).
Please ensure that your response
contains only the translated text.
The translation must convey the same
meaning as the original and be natural
for native speakers with correct
grammar and proper word choices.
Your translation must also use exact
terminology to provide accurate information
even for the experts in the related fields.

Original:

{{training_instance}}

Translation into {{target_language}}:
```

Figure 10: Prompt template used to translate Guanaco training instances from English into various target languages to investigate the effect of language diversity vs. instruction diversity.

C Hardware Requirements for Instruction Tuning

As mentioned in §3, we use LoRA (Hu et al., 2021) for parameter-efficient finetuning. As hyperparameters, we set $R = 64$, $\alpha = 16$. This results in the number of trainable parameters being roughly 2.3% of the original model size for Llama 2 7B, 1.2% for Llama 2 70B and 1.8% for Falcon 7B. For 7-billion parameter models, we use two NVIDIA GeForce RTX 3090 with 24GB of memory. The time required for each training run is approximately 8 hours. For the larger 70-billion parameter model, discussed in §5.2, we use the same hyperparameters and train on four NVIDIA A100 GPUs with 80GB of memory each. Here, a single training run takes approximately 20 hours.

D MultiSim Datasets

MultiSim is composed of 34 distinct text simplification data sets aligned at either the document- or sentence-level (Ryan et al., 2023), covering 13 different languages. Due to restrictive licensing on some datasets, only a subset are openly accessible. An overview of these datasets is provided in Table 2. As can be seen, domain coverage and the amount of available data vary considerably across languages. To balance out the number of evaluation instances between the different languages, we randomly sample up to 1,371 complex-simple sentence pairs from all available corpora for each language, except for Slovenian, where we use all 939 available instances. While we prioritise sampling from the test and validation splits shared by Ryan et al. (2023), we also draw samples from the designated train splits for Portuguese, German, and Slovenian in order to have a sufficiently large evaluation set for our experiments.

E Direct vs. Translated Evaluation for Non-English Dialogue Responses

While using a powerful LLM to evaluate the outputs of other models has been shown to achieve reasonable agreement with human judgements in English (Zheng et al., 2023; Chiang and Lee, 2023), it is unclear whether this agreement transfers to all languages under investigation. Recent work by Hada et al. (2024) has shown that agreement between human and LLM judges tends to be lower for non-English languages, especially in the case of low-resource and non-Latin scripted languages, where the LLM judge tends to be overly optimistic in its assessment. However, for certain assessment criteria, such as linguistic acceptability and general content quality, they also confirm that inter-

Language	Source corpora	# instances avail. (test / val / train)	# eval. instances
en	ASSET (Alva-Manchego et al., 2020)	359 / 100 / 19,000	1,371
	WikiAuto (Jiang et al., 2020)	5,002 / 4,988 / 576,126	
ru	RSSE (Sakhovskiy et al., 2021)	1,083 / 97 / 3,182	1,371
	RuAdapt Encyclopedia (Dmitrieva et al., 2021)	839 / 840 / 7,782	
	RuAdapt Fairytale (Dmitrieva et al., 2021)	31 / 31 / 248	
	RuWikiLarge (Sakhovskiy et al., 2021)	312 / 678 / 246,978	
de	GEOLino (Mallinson et al., 2020)	81 / 82 / 958	1,371
	TextComplexityDE (Naderi et al., 2019)	26 / 28 / 208	
fr	CLEAR (Grabar and Cardon, 2018)	100 / 294 / 4,196	1,371
	WikiLarge FR (Cardon and Grabar, 2020)	345 / 878 / 296,402	
pt	PorSimples (Aluísio and Gasperin, 2010)	420 / 420 / 6,290	1,371
it	AdminIT (Miliani et al., 2022)	49 / 48 / 588	1,371
	SIMPITIKI Wiki (Tonelli et al., 2016)	160 / 146 / 1,436	
	PaCCSS-IT (Brunato et al., 2016)	1,061 / 1,061 / 60,485	
	Teacher (Brunato et al., 2015)	17 / 17 / 136	
	Terence (Brunato et al., 2015)	101 / 102 / 809	
sl	SloTS (Gorenc and Robnik-Šikonja, 2022)	96 / 94 / 749	939

Table 2: Overview of datasets included in the MultiSim benchmark (Ryan et al., 2023). # instances avail. denotes the total number of complex-simple sentence pairs available for each corpus in the MultiSim benchmark. # eval. instances denotes the number of items sampled from all test, validation, and train splits (where necessary) for language used to evaluate performance on sentence simplification.

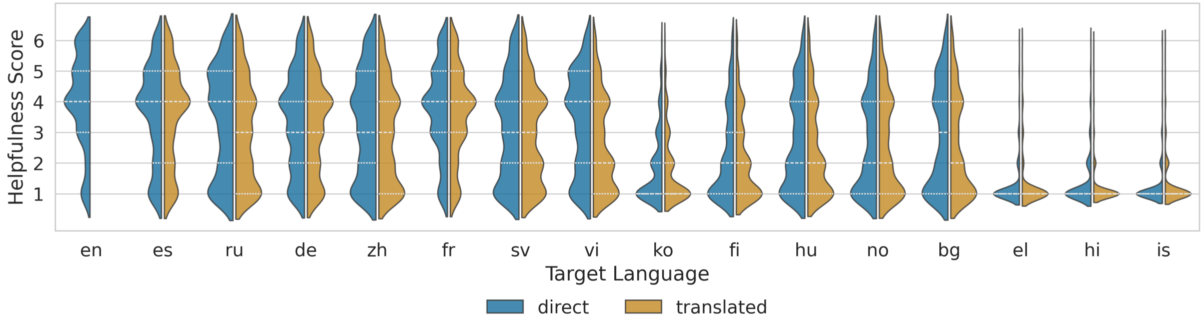


Figure 11: Distribution of helpfulness scores assigned by our LLM judge, GPT-3.5-Turbo, using direct evaluation in the specified target language and after translating target-language responses to English.

annotator agreement between LLM-based evaluators and humans is in line with that of multiple human annotators.

To investigate this potential bias, we compared scores assigned by the LLM judge on model-generated responses *directly* in each non-English target language and on their *English translations*. For each non-English prompt-response pair, we translate it into English using GPT-3.5-Turbo (`gpt-3.5-turbo-1106`) and the prompt shown in Figure 9. We pair the resulting translated responses with their corresponding English AlpacaEval prompts and repeat our LLM judge evaluation.

Figure 11 shows that the distribution of assigned scores in the direct and translated evaluation settings is very similar for most languages. For languages that use non-Latin scripts (e.g., Chinese, Russian, Korean, etc.), we observe that the LLM judge tends to assign slightly higher scores more frequently when evaluating directly on the non-English prompt/response pairs. This finding agrees with those from Hada et al. (2024) and indicates that LLM-based evaluations in non-English languages can be overly optimistic and should be considered with caution. Nevertheless, we observe that the discrepancy between direct and translated evaluation is relatively minor for the languages considered and leads to negligible differences in the overall average helpfulness score. Based on these results, we opt to use the direct evaluation strategy for our experiments, which has the added benefit of avoiding the introduction of potential translation errors and keeping the cost of evaluation to a minimum.

F Results with Falcon 7b

In order to assess whether our findings generalise to other LLMs, we repeat our experiments using Falcon 7b (Almazrouei et al., 2023).

F.1 Single-turn Dialogue

Figure 12 shows the helpfulness scores assigned by the LLM judge for Falcon 7b given incremental multilingual instruction tuning across all target languages. Similarly to our results with Llama 2 7b (cf. Figure 2), cross-lingual transfer can be elicited after finetuning with relatively few languages, and no additional gains observed when including more than three languages. Interestingly, Falcon 7b appears to show strong performance on French, even without multilingual finetuning, indicating that, despite being an English-centric model, it has strong capabilities in French out of the box. For Spanish, German, and Chinese, performance is comparable to that of Llama 2 7b. However, for all other languages, responses are often ranked least helpful, indicating that Falcon 7b's multilingual capabilities are limited strictly to major European languages and Chinese.

F.2 Sentence Simplification

Figure 13 depicts Falcon 7b's performance on the sentence simplification task given multilingual instruction tuning. In general, the results emulate those seen with Llama 2 7b. Performance on English remains uniform as multilinguality increases, while for non-English target languages, the largest jump in performance happens when moving from monolingual to bilingual instruction tuning, and plateauing again with just three languages. The notable exceptions here are French and Italian, for which Falcon 7b appears to provide greater support even in the monolingual instruction tuning setting, echoing the results from the single-turn dialogue task.

F.3 Extractive Question Answering

Figure 14 shows the results of Falcon 7b on XQuAD. While performance is generally lower than that of Llama 2 7b on this task (cf. Figure 4), we observe a similar, albeit weaker, effect of multilingual finetuning within the supported languages (es, de, zh).

F.4 Commonsense Reasoning

Figure 15 shows the results of Falcon 7b on X-CSQA. Strikingly, in contrast to the results achieved with Llama 2 7b (cf. Figure 5), Falcon 7b fails to score above random performance across all target languages. Regarding the effect of multilingual instruction tuning, we again see that it fails to deliver any performance improvements on this highly structured task. Degradation on Russian and Vietnamese, is likely due to the model consistently failing to follow instructions and reproducing part of the question as the response.

F.5 XNLI

Figure 16 shows the results of Falcon 7b on XNLI. Similar to the results attained with Llama 2 7b (cf. Figure 6), we observe no significant differences in performance given different degrees of multilingual instruction tuning.

F.6 Discussion

These additional experiments using Falcon 7b provide further support for our main findings: instruction tuning with as few as two languages significantly improves cross-lingual generalisation, enhancing performance in open-ended single-turn dialogue for some non-English languages. However, strikingly, multilingual capabilities of Falcon are considerably narrower than that of Llama 2. This is also suggested by the statistics provided in Table 1, which show a higher concentration of English and much lower proportions of non-English languages compared with both Llama 2 and GPT-3. We suspect that this may be a result of stringent filtering of web-scraped pretraining data performed in producing the RefinedWeb corpus used to train the Falcon models (Penedo et al., 2023), which not only reduces the risk of potentially accidental contamination (Blevins and Zettlemoyer, 2022) but also language coverage. That said, the RefinedWeb corpus comprises only part of Falcon's actual training corpus. (Almazrouei et al., 2023) note that the final corpus also contains additional data drawn from curated sources including a European-focused (multilingual) Common-Crawl dataset, which could explain the strong performance on some major European languages. In light of this, we suspect that additional Chinese data is also included in Falcon's training corpus, allowing for relatively strong performance in Chinese on the single-turn dialogue task.

G Results with Llama 3 8b

In §5.2 we observed that scaling the base model from 7b to 70b parameters reduces the gap in performance between English and non-English target languages with regard to the single-turn dialogue task. In this section, we evaluate Llama 3 8b (MetaAI, 2024), which builds on Llama 2 by scaling up the pretraining data from 2 trillion tokens to 15 trillion tokens. In Addition to a much larger pretraining corpus, Llama 3 is trained with a vocabulary of 128k tokens (vs. 32k for Llama 2), bringing it much closer to traditional multilingual models which typically contain roughly 250k tokens (Le Scao et al., 2023; Xue et al., 2021; Conneau et al., 2020). While the exact language distribution of the model's pretraining data is not publicly known, it is reasonable to expect that these enhancements could lead to improvements in the model's non-English capabilities.

Figure 17 shows the performance of Llama 3 8b, which closely resembles that of Llama 2 70b across most non-English languages, even outperforming it in languages like Vietnamese, Greek, and Hindi. Most notably, we observe that our main findings still hold: multilingual instruction tuning with as few as two languages is both necessary and sufficient to elicit cross-lingual generalisation. Again, performance typically plateaus thereafter. Given the improved performance on languages like Greek and Hindi that are low-resource under Llama 2's pretraining data distribution, we suspect that their representation in the pretraining data for Llama 3 is substantially higher. Still, performance on Icelandic suggests that this language may remain in the low-resource category under this model.

H Ablation Experiments

In Section 3.2, we constructed Multi-i instruction tuning datasets of the same size by replacing a fixed number of English examples with examples from a new language. As a result, the proportion of non-English examples used for finetuning increases along with the number of languages. Table 3 shows the exact makeup of each these datasets and how they relate to one another. To rule out potential confounding factors between the proportion of non-English examples and the diversity of languages, we conduct ablation experiments in which we keep one variable fixed while incrementing the other.

Dataset	Languages	% NE	Total
Mono	en	0.00%	3,200
Multi-2	en, es	6.25%	3,200
Multi-3	en, es, ru	12.50%	3,200
Multi-4	en, es, ru, de	18.75%	3,200
Multi-5	en, es, ru, de, zh	25.00%	3,200
Multi-6	en, es, ru, de, zh, fr	31.25%	3,200

Table 3: Makeup of the incremental multilingual instruction-tuning datasets.

Figure 18a shows that when using a fixed budget of 400 non-English examples (which equates to 12.5% of the finetuning data) and incrementing only the number of languages within this budget, performance increases consistently as each new language is added, closely reflecting the results of our main experiments. In contrast, Figure 18b shows that when training with a fixed number languages (e.g., Multi-2) and incrementing the number of non-English examples (in this case, Spanish), performance tends to remain lower and generally shows less improvement. These results underscore our main findings discussed in §4.1 and §6.

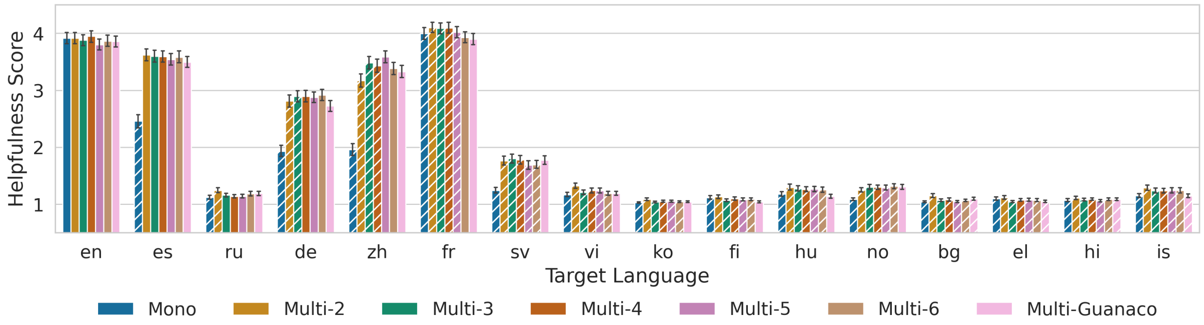


Figure 12: Average helpfulness of single-turn dialogue responses from Falcon 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

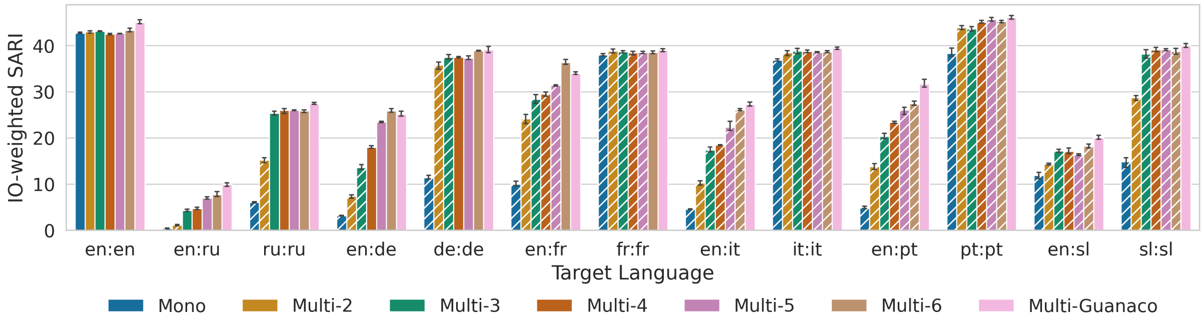


Figure 13: SARI weighted by IO language agreement for sentence simplification with Falcon 7b given incremental multilingual instruction tuning. Results are shown for both cross-lingual prompting (en:xx) and monolingual prompting (xx:xx) Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

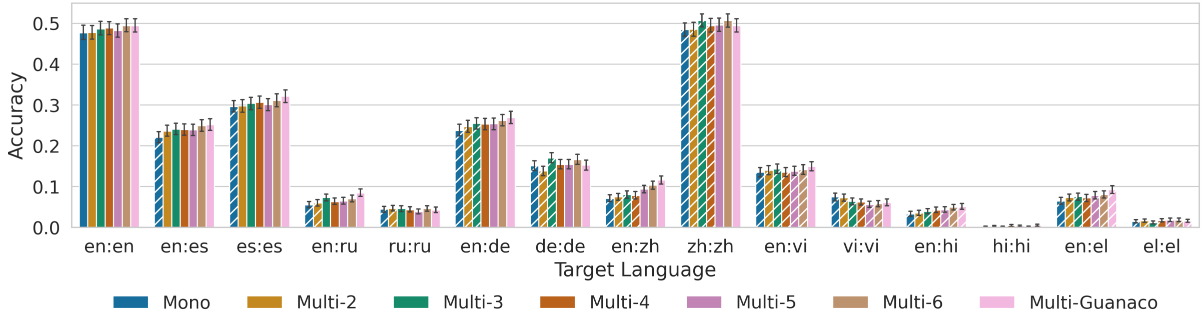


Figure 14: XQuAD results for Falcon 7b given incremental multilingual instruction tuning. Results are shown for both cross-lingual prompting (en:xx) and monolingual prompting (xx:xx). Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

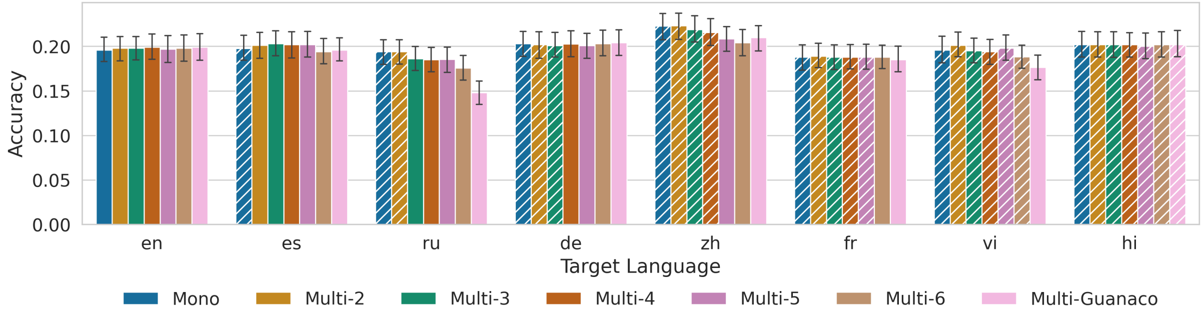


Figure 15: X-CSQA results for Falcon 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

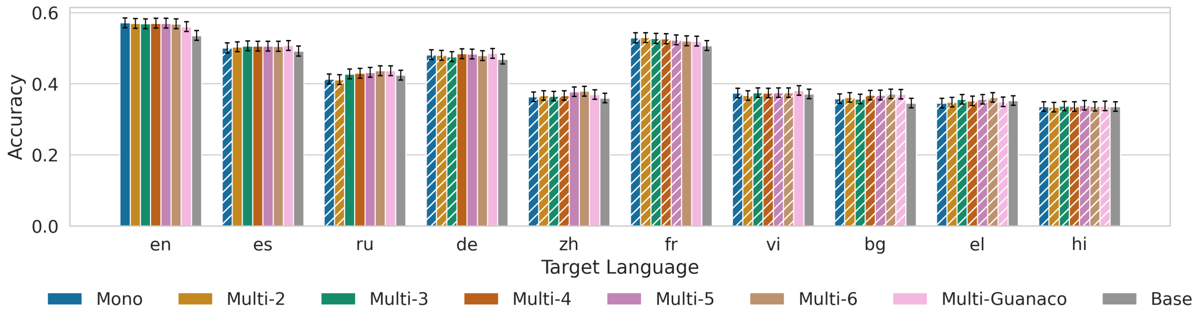


Figure 16: XNLI results for Falcon 7b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.

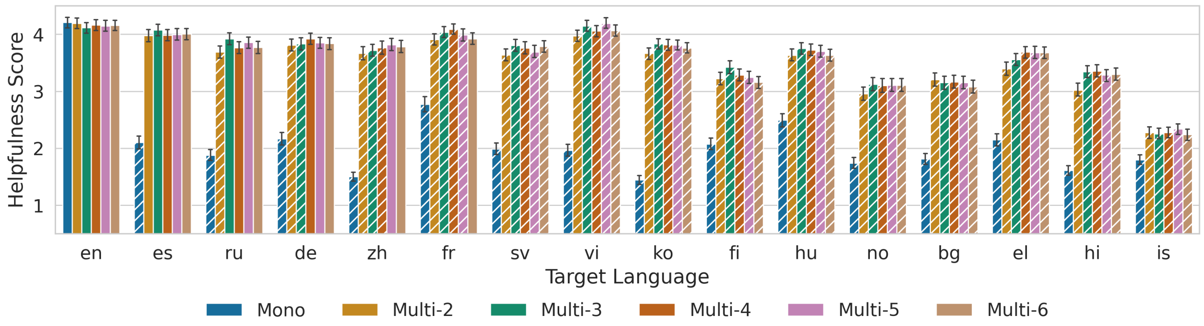
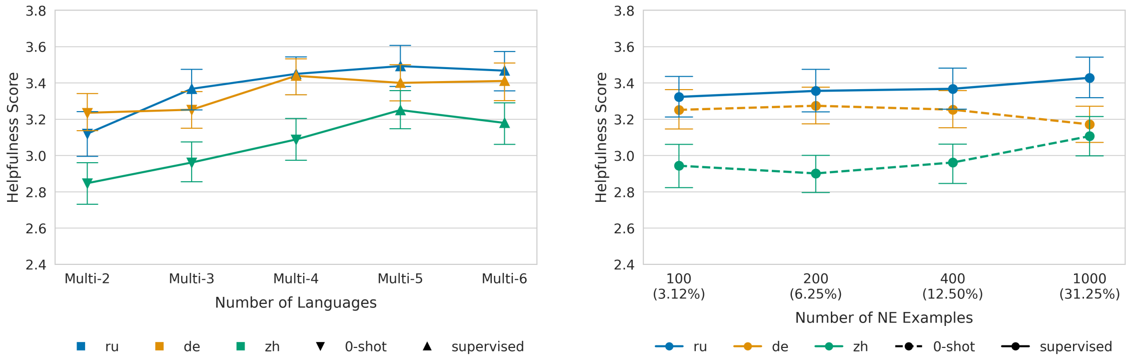


Figure 17: Single-turn dialogue performance of Llama 3 8b given incremental multilingual instruction tuning. Striped bars indicate a 0-shot setting and error bars show a confidence interval of 95%.



(a) Impact of incrementing the number of languages while keeping the total number of non-English examples fixed at 400 (12.5% of the finetuning data). 0-shot and supervised settings are distinguished using ▾ and ▴, respectively.

(b) Impact of incrementing the total number of non-English examples while keeping the number of languages fixed at three (English, Spanish, and Russian). Here, Russian is seen during finetuning, while German and Chinese are in the 0-shot setting.

Figure 18: Average helpfulness of single-turn dialogue responses from Llama 2 7b given multilingual instruction tuning on a fixed budget (limiting either the number of non-English instances 18a or the number of languages 18b). Error bars show a confidence interval of 95%.

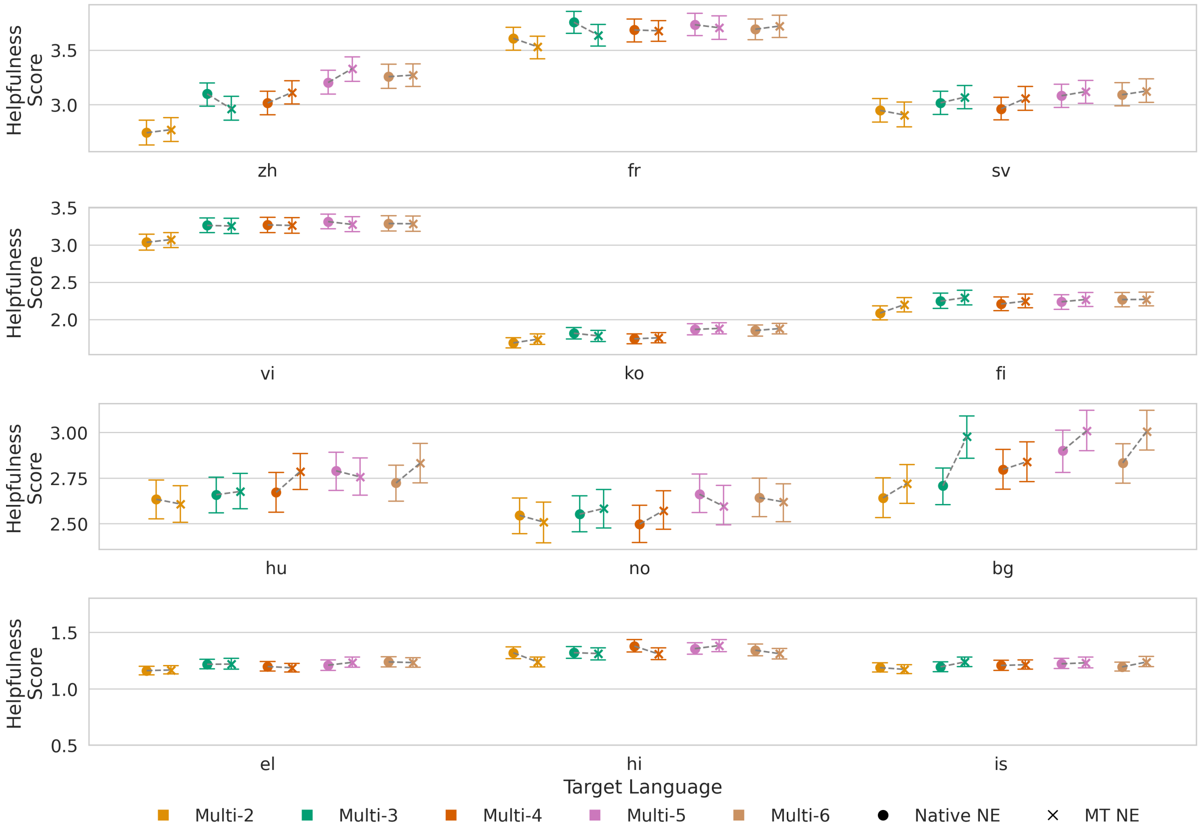


Figure 19: Comparison of Llama 2 7b with multilingual instruction tuning using native (Native NE) and translated non-English (MT NE) examples for additional target languages not shown in Figure 7. Error bars show a confidence interval of 95%.

Prompt:	`### Human: Please rewrite the following complex sentence in order to make it easier to understand. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.`
	`Complex sentence:` Sie gelten als eiskalte Fress-Maschinen.
	`### Assistant:`

Reference:	Sie gelten als Fress-Maschinen.
System:	They are seen as cold-blooded eaters.

Table 4: An example of the 0-shot sentence simplification task. Monospaced font represents the prompt template used, adapted from instructions provided to crowdworkers for the creation of the ASSET dataset (Alva-Manchego et al., 2020). This example shows the cross-lingual prompt strategy (en:xx), where the task instruction is provided in English but the source sentence is in German. Ideally, the system language should match the source sentence, which it fails to do in this case.

```
You are an expert language evaluator.

You are evaluating a response that has been submitted for a particular task, using a
specific set of standards. Below is the data:

[BEGIN DATA]
***
[Task]: {{task}}
***
[Submission]: {{submission}}
***
[Criterion]: helpfulness:
“1”: “Not helpful - The generated text is completely irrelevant, unclear, or
incomplete. It does not provide any useful information to the user.”
“2”: “Somewhat helpful - The generated text has some relevance to the user’s question,
but it may be unclear or incomplete. It provides only partial information, or the
information provided may not be useful for the user’s needs.”
“3”: “Moderately helpful - The generated text is relevant to the user’s question, and
it provides a clear and complete answer. However, it may lack detail or explanation
that would be helpful for the user.”
“4”: “Helpful - The generated text is quite relevant to the user’s question, and it
provides a clear, complete, and detailed answer. It offers additional information or
explanations that are useful for the user. However, some of the points of the response
are somewhat repetitive or could be combined for greater clarity and concision.”
“5”: “Very helpful - The generated text is highly relevant to the user’s question, and
it provides a clear, complete, and detailed answer. It offers additional information,
explanations, or analogies that are not only useful but also insightful and valuable to
the user. However, the structured of the response is not well-organized and there is no
clear progression or logical sequence of different points in the response.”
“6”: “Highly helpful - The generated text provides a clear, complete, and detailed
answer. It offers additional information or explanations that are not only useful
but also insightful and valuable to the user. The response is also in a logical and
easy-to-follow manner by explicitly using headings, bullet points, or numbered lists
to break up the information and make it easier to read.”
***
[END DATA]

Does the submission meet the criterion? First, write out in a step by step manner your
reasoning about the criterion to be sure that your conclusion is correct. Avoid simply
stating the correct answers at the outset. Provide your response as a RFC8259 compliant
JSON following this schema:

{“reasoning”: str, “score”: int}
```

Figure 20: Prompt template used to query the LLM judge in order to assess the helpfulness of a single-turn dialogue response. This prompt closely follows the one used by Zhou et al. (2023) but extends it by specifying the return value as a valid JSON-like object.

Prompt:	`### Human: Answer the question from the given passage. Your answer should be directly extracted from the passage, and it should be a single entity, name, or number, not a sentence.`
	`Passage:`
	Die Verteidigung der Panthers gab nur 308 Punkte ab und belegte den sechsten Platz in der Liga, während sie die NFL mit 24 Interceptions in dieser Kategorie anführte und sich mit vier Pro Bowl-Selektionen rühmen konnte. Pro Bowl Defensive Tackle Kawann Short führte das Team mit 11 Sacks an, erzwang zudem drei Fumbles und erzielte zwei Fumble Recoverys. Mario Addison, ebenfalls Lineman, addierte $6\frac{1}{2}$ Sacks hinzu. Die Panthers-Line präsentierte auch den erfahrenen Defensive End Jared Allen, einen 5-fachen Pro-Bowler, der mit 136 Sacks der aktive Anführer in der NFL-Kategorie Karriere-Sacks war, sowie den Defensive End Kony Ealy, der 5 Sacks in nur 9 Starts erzielte. Nach ihnen wurden zwei der drei Linebacker der Panthers ausgewählt, um im Pro Bowl zu spielen: Thomas Davis und Luke Kuechly. Davis erzielte $5\frac{1}{2}$ Sacks, vier erzwungene Fumbles und vier Interceptions, während Kuechly das Team bei den Tackles anführte (118), zwei Fumbles erzwang und vier Pässe abfing. Carolinas Secondarys bestanden aus dem Pro Bowl-Safety Kurt Coleman, der das Team mit einem Karrierehoch von sieben Interceptions anführte und gleichzeitig 88 Tackles erzielen konnte, und Pro Bowl-Cornerback Josh Norman, der sich während der Saison zur Shutdown Corner entwickelte und vier Interceptions erzielte, von denen zwei zu Touchdowns für sein Team wurden.
	`Question:`
	Mit wie vielen Interceptions erzielte Josh Norman im Jahr 2015 Touchdowns?
	`Note: Your answer should be directly extracted from the passage and be a single entity, name, or number. Provide the answer in quotations.`
	`### Assistant: Based on the passage, the answer to the question is “`

Reference:	zwei
System:	two". In the passage, it is stated that Josh Norman “erzielte vier Interceptions, von denen zwei zu Touchdowns für sein Team wurden.” This means that Norman intercepted four passes during the season, and two of those interceptions were returned for touchdowns.

Table 5: Example input prompt, reference, and model response for the 0-shot extractive QA task using XQuAD. Monospaced font represents the prompt template used for each test instance. Here, we show the cross-lingual prompting strategy (en:xx), where the task instruction and response prefix is provided English, while the context passage and question are in the target language (in this case, German). A correct answer in this setting requires the model to continue the response by switching into the target language, which it fails to do in this particular example.

Prompt:	`### Human: Dir wird eine Frage vorgelegt, die mehrere Antwortmöglichkeiten hat. Wähle die am besten geeignete Option aus “A”, “B”, “C”, “D” oder “E” basierend auf deinem gesunden Menschenverstand. Bitte gebe deine Antwort in Form eines einzelnen Buchstabens in Anführungszeichen an.`
	
	`Frage:`
	
	Sarah stellte ihr Essen in einen Mülleimer neben ihrem Schreibtisch. Wo mag sie gewesen sein?
	
	`Antwortmöglichkeiten:`
	
	`A:` Bushaltestelle
	`B:` Ecke
	`C:` Motel
	`D:` Park
	`E:` Büro
	
	`### Assistant: “`

Reference:	E
System:	E”

Table 6: Example of the commonsense reasoning task using X-CSQA. Monospaced font represents the prompt template used for each test instance. Here, the prompt strategy is monolingual (xx:xx) with the task instruction provided in the same language as the test instance (in this case, German). A correct answer requires the model to output the letter corresponding to the ground-truth answer.

<table>
<tr><th colspan="3">XNLI prompt template: <code>{Sentence 1}, right? {label}, {Sentence 2}</code></th></tr>
<tr><th>Label</th><th>Prompt query</th><th>NLL</th></tr>
<tr><td>Entailment</td><td>In measuring effectiveness, perfection is unattainable., right? Yes, You can never be perfect.</td><td>-113.94</td></tr>
<tr><td>Neutral</td><td>In measuring effectiveness, perfection is unattainable., right? Also, You can never be perfect.</td><td>-116.79</td></tr>
<tr><td>Contradiction</td><td>In measuring effectiveness, perfection is unattainable., right? No, You can never be perfect.</td><td>-114.01</td></tr>
</table>

Table 7: Example prompt queries for XNLI given the predefined template used in the LM Evaluation Harness (Gao et al., 2023) and the three possible labels. Underlined words are the language-specific connectors corresponding to the three possible labels. The final answer is the sequence with the highest negative log-likelihood (NLL) according to the model. In this case, Llama 2 7b Mono correctly identifies the relationship between the two sentences as entailment.