Zero-shot cross-lingual transfer
in instruction tuning of large language models

Nadezhda Chirkova
Naver Labs Europe
Grenoble, France
`nadia.chirkova`
`@naverlabs.com`

Vassilina Nikoulina
Naver Labs Europe
Grenoble, France
`vassilina.nikoulina`
`@naverlabs.com`

Abstract

Instruction tuning (IT) is widely used to teach pretrained large language models (LLMs) to follow arbitrary instructions, but is understudied in multilingual settings. In this work, we conduct a systematic study of zero-shot cross-lingual transfer in IT, when an LLM is instruction-tuned on English-only data and then tested on user prompts in other languages. We advocate for the importance of evaluating various aspects of model responses in multilingual instruction following and investigate the influence of different model configuration choices. We find that cross-lingual transfer does happen successfully in IT even if all stages of model training are English-centric, but only if multiliguality is taken into account in hyperparameter tuning and with large enough IT data. English-trained LLMs are capable of generating correct-language, comprehensive and helpful responses in other languages, but suffer from low factuality and may occasionally have fluency errors.

1   Introduction

Instruction tuning (IT) helps to align large language models (LLMs) with users expectations so that LLMs are capable of understanding user queries and generating helpful, comprehensive and focused responses without few-shot examples. Contrary to standard NLP datasets that are focused on particular tasks, IT datasets consist of diverse instructions representing various tasks and possible user requests, enabling generalization to new instructions which were unseen during training (Ouyang et al., 2022).

Most of the IT research has focused on English, leaving multilingual instruction following a rather understudied area. Several recent works aim to extend instruction tuning beyond English by creating target language IT datasets via automatic translation of English instructions (Cab, 2023; Zic, 2023), distillation of outputs of powerful models such as
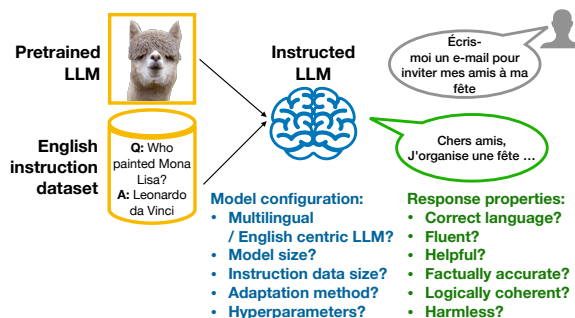

Figure 1: Zero-shot cross-lingual transfer in instruction tuning: an LLM is instruction-tuned on English-only data and then tested on user prompts in other languages. Our study focuses on analyzing various aspects of generated outputs and model configuration choices.

GPT-4 (Wei et al., 2023; Li et al., 2023), or crowdsourcing (Köpf et al., 2023; Singh et al., 2024). However, all of these strategies incur high costs or effort and require repeating the data creation process for each language of interest (target language).

In this work, we take a close look at *zero-shot cross-lingual transfer* in instruction tuning, when the LLM is tuned solely on English instruction data and then prompted to follow instructions in target languages without any additional target-language adaptation. Such an approach has the clear advantages of low cost and easy applicability to various target languages but is often considered just as a simple baseline, without detailed analysis. We aim to deeper understand (RQ1) what are *the capabilities and limits* of the zero-shot approach as well as (RQ2) *which factors influence* the successful cross-lingual knowledge transfer.

The most common strategy for evaluating instruction following capabilities consists of scoring the helpfulness of model responses on some publicly available set of diverse instructions, e.g. AlpacaFarm (Dubois et al., 2023), with a powerful model, e.g. GPT-3.5. We argue that such *high-level* evaluation is *insufficient and not infor-*

695

mative enough for a multi-facet task of open-ended generation, especially in the multilingual scenario. We advocate for using *a more careful evaluation pipeline*, including the evaluation of *various aspects* of model responses (fluency, content, relevance to the task etc.), controlling the distribution and the complexity of the tasks in the evaluation set, and using both automatic metrics and human inspection of predictions. This allows us to characterize the weak and strong sides of multilingual responses generated by the model tuned on English-only data (RQ1) and to better understand the influence of factors such as the base model (multilingual / English-centric, model size), IT data size, adaptation strategy and hyperparameters (RQ2). Our key findings include:

- Cross-lingual transfer does happen successfully in Instruction Tuning (IT) even if all stages of model training are English-centric, but only if *multilinguality is taken into account in IT hyperparameter tuning* and *with large enough IT data*;

- Models trained on English are capable of generating *correct-language, comprehensive and helpful responses* in the other languages, even with *complex instructions*, e.g. generate the answer in a given style or language;

- The main challenge is *low factuality in non-English instruction following*. *Occasional fluency and logical errors*, as well as *infrequent code-switching* can also take place.

2   Related work

Most of the works in multilingual IT aim to extend the IT dataset with non-English data (Köpf et al., 2023; Singh et al., 2024; Li et al., 2023; Wei et al., 2023), or decompose non-English instructions by pivoting through English translations (Zhang et al., 2023b; Etxaniz et al., 2023). Chen et al. (2024); Kew et al. (2023); Shaham et al. (2024) advocate for the sufficiency of a "pinch" of multilinguality in IT, represented by a small amount of updates on multilingual IT data, small amount of multilingual IT data mixed with English data, or having only 2–3 languages in the IT data. We focus on English-only IT, trying to better assess capabilities and limits of such settings.

The concurrent work of Shaham et al. (2024) does demonstrate the proof-of-the-concept results on zero-shot cross-lingual transfer in IT, but attributes it to the multilinguality of PaLM-2 pre-training data. We show that cross-lingual transfer

in IT works well even for English-centric models and conduct a more deep and systematic investigation of this effect.

We cover more related works in Appendix A.

3   Our evaluation methodology

To better understand the strong and weak sides of multilingual responses generated by the model tuned on English-only data, we devise a multi-facet evaluation strategy which includes evaluation of various aspects of generated responses, controlling task distribution and complexity, and using both model-based and human evaluation.

Evaluation criteria. We conduct main evaluation using both manual predictions inspection (on a subset of the evaluation set) and GPT-3.5 evaluation (on the full evaluation set). To control qualitative aspects of generated texts, we judge them with 6 criteria: *helpfulness* (how helpful in general is the response for the user), *language correctness* (does the language of the response match the language of the task), *fluency*, *factual accuracy*, *logical coherence* and *harmlessness*. Five of these criteria (except language correctness) were introduced in (Zhang et al., 2023a) and in our preliminary study we found that they reflect well the weaknesses of model responses. We also use the same scale from 0 to 2 for each criteria.

We also introduce lightweight *surface metrics*: *language correctness* (how often the language of the response matches the language of the task), *spellcheck correctness* (which portion of words in the responses pass spell checking), and *relevance to the task* (how often responses are relevant to their tasks, evaluated using LLama-2-chat-7B). These metrics serve to identify if a model passes a *minimal bar on the quality of multilingual answers* and help to select hyperparameters and filter out non-effective model configurations without GPT-based evaluation.

Control of the task distribution. We identify a diverse set of 25 "tasks" present in AlpacaFarm (e.g. write an email, give advice, rewrite text etc.) and select a subset of 113 instructions from AlpacaFarm that include a *balanced number of instructions per "task"*. Thus obtained set of 113 instructions is used in GPT-3.5-based evaluation, and a stratified subset of 30 instructions is used in human evaluation. Controlling the task distribution ensures that none of the tasks dominates the evaluation set, leading to more reliable conclusions, and allows us

	Helpfulness en	non-en	Correct Lang. en	non-en	Fluency en	non-en	Factuality en	non-en	Logicality en	non-en	Harmlessness en	non-en		Helpfulness en	non-en	Correct Lang. en	non-en	Fluency en	non-en	Factuality en	non-en	Logicality en	non-en	Harmlessness en	non-en
LLaMA-2-13B / Dolly-En / FT	1.77	1.35	2.00	1.87	2.00	1.81	1.80	1.46	2.00	1.86	2.00	1.98		1.88	1.72	2.00	1.84	1.78	1.55	1.82	1.61	1.93	1.80	2.00	2.00
LLaMA-2-13B / LIMA-En / FT	1.60	0.94	2.00	1.27	1.97	1.83	1.80	1.16	1.93	1.47	1.97	1.98		1.81	1.52	1.97	1.34	1.66	1.29	1.71	1.34	1.84	1.45	2.00	1.97
LLaMA-2-13B / Dolly-En / LoRA	1.70	1.14	2.00	1.87	2.00	1.76	1.80	1.29	2.00	1.79	2.00	1.97		1.79	1.51	1.96	1.72	1.76	1.43	1.78	1.41	1.87	1.59	2.00	2.00
LLaMA-2-7B / Dolly-En / FT	1.87	1.19	2.00	1.71	2.00	1.88	1.90	1.42	2.00	1.88	2.00	2.00		1.85	1.56	2.00	1.74	1.73	1.44	1.73	1.45	1.96	1.63	2.00	2.00
LLaMA-2-13B / Dolly-DT / FT	1.79	1.09	1.93	1.96	2.00	1.82	1.79	1.24	1.97	1.79	1.93	1.98		1.84	1.57	1.95	1.94	1.81	1.49	1.80	1.49	1.91	1.69	2.00	2.00
Tower-7B / Dolly-En / FT	1.80	1.24	2.00	1.89	2.00	1.94	1.87	1.31	2.00	1.88	2.00	1.99		1.84	1.71	1.98	1.86	1.67	1.59	1.70	1.59	1.84	1.75	2.00	2.00
Tower-7B / Dolly-DT / FT	1.57	1.20	1.87	1.92	2.00	1.78	1.82	1.26	1.96	1.85	2.00	1.99		1.85	1.65	1.87	1.92	1.73	1.48	1.73	1.54	1.88	1.75	2.00	2.00
						Human evaluation													GPT-3.5 evaluation							

Figure 2: Results of human evaluation (left) and evaluation with GPT-3.5 (right). All scores from 0 to 2, heatmap colors visualize written scores. Base models: LLaMA-2-7B/13B (English-centric) or Tower-7B (10 languages). Datasets: Dolly (15k) or LIMA (1k). Instruction tuning data strategies: En (English-only data) or DT (multilingual IT data obtained using data translation). Adaptation strategy: FT (full finetuning) or LoRA (low-rank adaptation).

to break down the performance results by tasks.

Control of the task complexity. To deeper analyze the effect of task complexity, we introduce a set of *task modifiers* which add details to the task, such as generate a short or detailed response, answer in a specified language or style, format the answer in a specified way, or answer two questions one after another. Modifiers are manually translated into target languages and added to instructions one-by-one. For each modifier we select a subset of 15-100 appropriate input instructions. We evaluate overall helpfulness of the produced responses (taking into account all given instructions) and *modifier fulfillment*: whether responses follow additional instructions given in the modifier.

4 Experimental setup

We study the effect of various choices such as the base model, the size of the English instruction data, adaptation strategy (full or parameter-efficient finetuning), and adaptation hyperparameters.

Base LLMs. In our work we consider (1) LLaMA-2 (Touvron et al., 2023) at 7B and 13B sizes, (2) TowerBase-7B (Alves et al., 2024), built on top of LLaMa-2-7B, further trained on balanced data covering 10 languages. In the former case, the multilingual instruction-following capabilities of the model arise solely from the small amount of *occasional* multilingual data which is always present in English-centric *pretraining* corpora crawled from the Internet (Blevins and Zettlemoyer, 2022). The latter case allows us to assess an importance of multilinguality at pretraining.

Instruction tuning datasets. We perform instruction tuning on two English instruction datasets: Dolly (Databricks, 2023) (denoted Dolly-En), 15k crowdsourced instructions covering 7 different categories (creative writing, open and close QA, classification, brainstorming, information extraction), and LIMA (Zhou et al., 2023) (denoted LIMA-En),

1k samples, carefully selected from various datasets (eg. StackExchange, WikiHow, etc.). In order to assess the importance of instructions multilinguality, we also consider multilingual Dolly data (Dolly-DT), extended by adding its automatic translations (cf. Appendix B for details) into three languages (Fr, Pt, Ru).

IT strategy. We consider two most popular supervised finetuning techniques: full finetuning (FT) and LoRA finetuning.

Evaluation. We evaluate responses in four languages: English, French, Portuguese, and Russian, and curate translations of the evaluation set into the specified languages. Manual inspection of predictions was conducted by the native or fluent speakers employed at our research laboratory.

We select LLaMA-2-13B/Dolly-En/FT as an *anchor model configuration* and apply changes to it one-by-one, i.e. changing the base model, IT data, or the adaptation method. We train all model configurations with three learning rates (LRs) and choose the best LR based on surface metrics. For more experimental details, see Appendix B

5 Experimental results and discussion

5.1 Main evaluation

Figure 2 shows the results of human (left) and GPT-3.5-based (right) evaluation, for English and average over Fr, Pt, and Ru. Per-language results are presented in App. Figure 4. Agreement between automatic and human evaluation is visualized in App. Figure 5. Though we observe generally consistent trends between GPT-3.5 and human evaluation in *average* scores, they can disagree in evaluating *individual samples*, especially for the scores of helpfulness, factual accuracy, and fluency. Agreement for non-English is lower than for English.

RQ1. We first analyze various aspects of predictions for our anchor English-centric and English-tuned model, LLaMA-2-13B/Dolly-En/FT.

Instruction-tuned model is able to successfully transfer learned knowledge to other languages, but with helpfulness to some extent lower than in English. The main score, overall *Helpfulness*, for our anchor English-centric model, `LLaMA-2-13B/Dolly-En/FT`, achieves 1.77 / 1.35 in English / non-English settings correspondingly (out of 2, human evaluation). As we discuss below, one of the main factors contributing to this difference is reduced factuality in non-English. Another factor is that responses in non-English sometimes contain obvious advice, e.g. "to install a window blind, follow the instructions provided with it" (translated from Russian).

Factuality is the weak side of predictions in non-English. The factual accuracy score is substantially lower in non-English than in English, e.g. 1.46 vs 1.80 in human evaluation. This poses a challenge for future works at improving truthfulness in the multilingual setting.

English-tuned model may occasionally (but rarely) produce output in the wrong language, code-switching, or make a fluency error. Scores for correct language, fluency and logical coherence are between 1.8 and 1.9 for the anchor model `LLaMA-2-13B/Dolly-En/FT` in non-English settings. This holds for both automatic and human evaluation, except GPT-3.5 evaluation of fluency, demonstrating the need for the better automatic evaluation of this criteria. We highlight that the problem of generation in the wrong language appears rarely in cross-lingual setting (after careful LR tuning), opposite to the conclusions of prior work (Chen et al., 2024).

RQ2: influence of various model design choices. Using the multilingual base model further improves fluency and generation in the correct language, but not factuality. Using multilingual IT data only improves the correct language score. Scores for the correct language and fluency get slightly improved for the multilingually pretrained `Tower-7B/Dolly-En/FT` compared to the similarly-sized English-centric `LLaMA-2-7B/Dolly-En/FT`. Using multilingual IT data in `LLaMA-2-13B/Dolly-DT/FT` and `Tower-7B/Dolly-DT/FT` improves scores for correct language, compared to similar configurations with `Dolly-En`, but does not improve fluency. Factuality does not get improved with any of the model modifications.

Even though training on small instruction data was shown to be sufficient for En-glish (Zhou et al., 2023), it substantially reduces the cross-lingual capabilities of the final model compared to training on the larger data. The model tuned on (English) LIMA, `LLaMA-2-13B/LIMA-En/FT`, is characterized by very low scores for all criteria, in non-English evaluation[1]. This is caused by severe overfitting to English, pronounced by low language correctness scores and generation of incoherent texts. At the same time, scores for English are close to other models, which aligns with the initial findings of (Zhou et al., 2023).

Ablations (small base LLM, LoRA adaptation) reduce scores in non-English. Using LoRA instead of full finetuning, `LLaMA-2-13B/Dolly-En/LoRA`, and decreasing model size, `LLaMA-2-7B/Dolly-En/FT`, reduce most of the scores compared to the anchor model `LLaMA-2-13B/Dolly-En/FT`.

Per-language analysis: fluency is lower for Russian than for French and Portuguese. Per-language analysis presented in App. Figure 4 demonstrates that conclusions discussed above are consistent between languages. A standing-out criteria is fluency which is lower for Russian than for other languages. This is pronounced by the occasional generation of made-up words in Russian and could be connected to the non-Latin script.

Per-task analysis: helpfulness in non-English reduces in some language-related tasks, tasks involving calculation or US-centric factual knowledge. Figure 3 (right) breaks down human-evaluated helpfulness of the anchor model by task category. We find that English-centric model struggles in other languages with some of language-based tasks such as rewriting given sentences, suggesting words that rhyme with the given one or following a given pattern. At the same time, models do succeed on easier language-related tasks such as generate synonyms or words beginning with a given letter. Models also make calculation errors more often in non-English than in English. The low helpfulness for the "sport game" category is connected to the low factuality in non-English: this category asks to explain rules of games popular in the USA and they are explained well in English and often hallucinated in other languages.

---

[1]The helpfulness score assigned by GPT-3.5, 1.52, is substantially higher than the one assigned in human evaluation, 0.94, because LIMA-based model produces much longer outputs than Dolly-based model and GPT-3.5 is known to be biased towards long verbose responses.
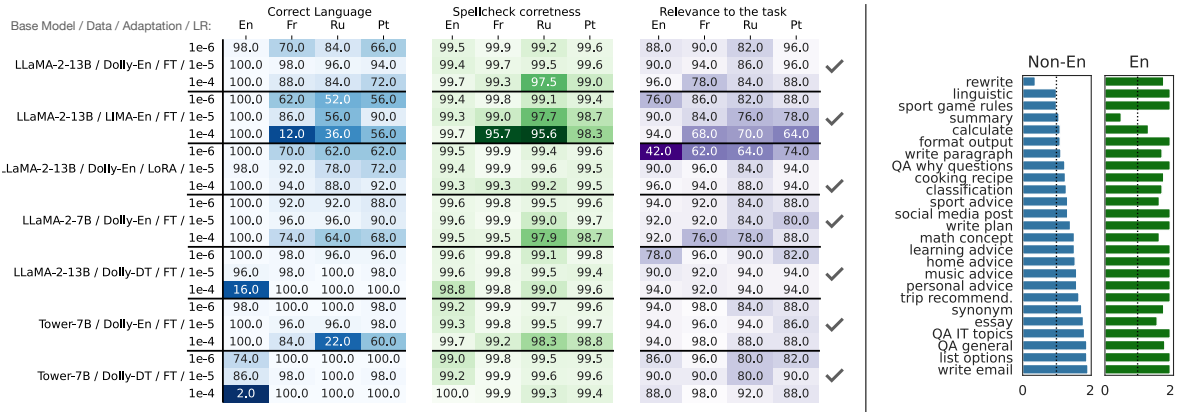
Figure 3: *Left*: Results of evaluating surface features of the responses. Ticks denote the chosen LR for each configuration. Base models: LLaMA-2-7B/13B (English-centric) or Tower-7B (10 languages). Datasets: Dolly (15k) or LIMA (1k). Data strategies: En (English-only data) or DT (multilingual data obtained using data translation). Adaptation strategy: FT (full finetuning) or LoRA (low-rank adaptation). *Right*: Human-evaluated helpfulness of the default model broken down by task category.

Base Model / Data / Adaptation / LR:	Correct Language				Spellcheck correctness				Relevance to the task				
	En	Fr	Ru	Pt	En	Fr	Ru	Pt	En	Fr	Ru	Pt	
1e-6	98.0	70.0	84.0	66.0	99.5	99.9	99.2	99.6	88.0	90.0	82.0	96.0	
LLaMA-2-13B / Dolly-En / FT / 1e-5	100.0	98.0	96.0	94.0	99.4	99.7	99.5	99.6	90.0	94.0	86.0	96.0	✓
1e-4	100.0	88.0	84.0	72.0	99.7	99.3	97.5	99.0	96.0	78.0	84.0	88.0	
1e-6	100.0	62.0	52.0	56.0	99.4	99.9	99.1	99.4	76.0	86.0	82.0	88.0	
LLaMA-2-13B / LIMA-En / FT / 1e-5	100.0	86.0	56.0	90.0	99.3	99.9	97.7	98.7	90.0	84.0	76.0	78.0	✓
1e-4	100.0	12.0	36.0	56.0	99.7	95.7	95.6	98.3	94.0	68.0	70.0	64.0	
1e-6	100.0	70.0	62.0	62.0	99.5	99.9	99.4	99.6	42.0	62.0	64.0	74.0	
LLaMA-2-13B / Dolly-En / LoRA / 1e-5	98.0	92.0	78.0	72.0	99.4	99.9	99.6	99.5	90.0	96.0	84.0	94.0	
1e-4	100.0	94.0	88.0	92.0	99.3	99.3	99.2	99.5	96.0	94.0	88.0	94.0	✓
1e-6	100.0	92.0	92.0	88.0	99.6	99.8	99.5	99.6	94.0	92.0	84.0	88.0	
LLaMA-2-7B / Dolly-En / FT / 1e-5	100.0	96.0	96.0	90.0	99.6	99.9	99.0	99.7	92.0	92.0	84.0	80.0	✓
1e-4	100.0	74.0	64.0	68.0	99.5	99.5	97.9	98.7	92.0	76.0	72.0	88.0	
1e-6	100.0	98.0	96.0	96.0	99.6	99.8	99.1	99.8	78.0	96.0	90.0	82.0	
LLaMA-2-13B / Dolly-DT / FT / 1e-5	96.0	98.0	100.0	98.0	99.6	99.8	99.5	99.4	90.0	92.0	94.0	94.0	✓
1e-4	16.0	100.0	100.0	100.0	98.8	99.8	99.0	99.6	94.0	92.0	94.0	94.0	
1e-6	98.0	100.0	100.0	98.0	99.2	99.9	99.7	99.6	94.0	98.0	84.0	88.0	
Tower-7B / Dolly-En / FT / 1e-5	100.0	96.0	96.0	98.0	99.3	99.9	99.5	99.7	94.0	96.0	84.0	86.0	✓
1e-4	100.0	84.0	22.0	60.0	99.7	99.2	98.3	98.8	94.0	98.0	88.0	88.0	
1e-6	74.0	100.0	100.0	100.0	99.0	99.8	99.5	99.5	86.0	96.0	80.0	82.0	
Tower-7B / Dolly-DT / FT / 1e-5	86.0	98.0	100.0	98.0	99.2	99.9	99.6	99.6	90.0	90.0	80.0	90.0	✓
1e-4	2.0	100.0	100.0	100.0	100.0	99.9	99.3	99.4	88.0	98.0	92.0	88.0	

Task modifier	Mod. fulfill.		Helpfulness	
	en	ru	en	ru
Answer briefly in just a few sentences.	80%	90%	1.70	1.60
Give a detailed answer.	65%	75%	1.60	1.55
List N options (N random from 2 to 10)	66%	83%	1.66	1.66
Answer in X language. (X: Fr, Pt, De)	47%	79%	1.37	1.47
Use markdown formatting in the answer.	92%	100%	1.85	1.28
Format your answer as an html page.	57%	14%	1.35	1.00
Begin each point with the sign –>	7%	14%	0.92	0.85
Capitalize each first letter in the answer.	7%	7%	1.00	0.64
Write in a scientific style.	92%	92%	1.64	1.57
The answer should use simple words.	78%	78%	1.57	1.28
Two-hop instruction, e.g. *explain how to serve a dish after telling how to cook it.*	93%	86%	1.80	1.60
Average	62%	65%	1.49	1.31

Table 1: Performance with various task modifiers. Modifier fulfilness measures the percentage of inputs for which the modifier was fulfilled. Helpfulness (from 0 to 2) also takes into account the modifiers' conditions.

5.2 Additional experiment with task modifiers

To complement analysis for RQ1, Tab. 1 reports results on controlling task complexity with task modifiers.

English-centric models are capable of following composite instructions in non-English languages in 65% of cases. The majority of task modifiers are fulfilled in around 80% of cases, with helpfulness score being similar to the value observed in the main evaluation. Interestingly, the instruction to generate response in another language, is fulfilled substantially more often when it is written in non English. An example of the instruction that often fails in non-English is to format the answer as an html page.

5.3 Preliminary study based on surface metrics

Figure 3 (left) demonstrates surface metrics for all considered model configurations trained with three learning rates, complementing analysis for RQ2.

Careful hyperparameter tuning and in particular LR selection is essential for achieving multilingual instruction following capabilities. All the model configurations, except training on the small LIMA data, achieve high values for all metrics in all languages with LR of 1e-5 (1e-4 for LoRa adaptation). The lower LR of 1e-6 leads to lower relevance scores in some languages, due to model *under-training*. On the other side, the higher LR of 1e-4 leads to *overfitting to the training language(s)*, pronounced by lower language correctness scores and lower spellcheck correctness scores, caused by code-switching.

Surface metrics help to select hyperparameters and filter out poor configurations. Surface metrics capture the same effect as in main evaluation, that training on the small LIMA data leads to severe overfitting to English (with all LRs).

6 Conclusion

In this work we demonstrate the possibility of zero-shot cross-lingual transfer of instruction following capability. We devise a multi-facet evaluation methodology, allowing us to pinpoint the main capabilities and limitations of such transfer and to point important future research directions. We highlight the critical role of LR tuning and IT data size, which we hope will help in future works on IT.

699

Supplementary Materials Availability Statement: Our code and data are available at `https://github.com/naver/pasero/tree/main/examples/zero-shot-transfer-inst-tuning`.

7 Limitations and broader impact

Despite making a substantial effort in systematically evaluating cross-lingual transfer in IT, we acknowledge the infeasibility of considering all possible model configurations and evaluation aspects. First, our study only considers high-resource languages while cross-lingual transfer is expected to pose a greater challenge for medium- and low-resource languages. We focused on high-resource languages as a first step and hope that our evaluation methodology will be helpful in future studies for other language groups. Second, we experiment with one main hyperparameter, learning rate, while other training hyperparameters may also play a substantial role. Nonetheless, we were able to achieve high results even with our rather limited hyperparameter grid. Finally, we only consider commonly used model configurations and adaptations strategies, while other approaches such as reinforcement learning with human feedback, could be also interesting to investigate. We leave their consideration for future work.

We do not anticipate negative societal impact from our work and on the reverse hope that it will help to broaden the accessibility of modern NLP.

8 Acknowledgments

References

2023. Cabrita: A portuguese finetuned instruction llama. `https://github.com/22-hours/cabrita`.

2023. Zicklein: A german finetuned instruction llama. `https://github.com/avocardio/Zicklein`.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

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.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of english pretrained models.

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.

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.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Key ingredients for effective zero-shot cross-lingual knowledge transfer in generative tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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.

Databricks. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. Blog post.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa Etxabe, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *ArXiv*, abs/2308.01223.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed?

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

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

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. Zm-BART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao,

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

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

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

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmt5: Modular multilingual pre-training solves source language hallucinations.

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

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

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman

Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

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

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

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

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

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023a. Llmeval: A preliminary study on how to evaluate large language models.

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

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 Extended related work

Zero-shot cross-lingual transfer was extensively studied for discriminative tasks (Xue et al., 2021; Conneau et al., 2020; Artetxe et al., 2020; Pires et al., 2019; Wu and Dredze, 2019; Pfeiffer et al., 2020) and remains rather under-explored for generative tasks. Vu et al. (2022); Pfeiffer et al. (2023); Maurya et al. (2021); Li and Murray (2023) highlight the problem of generation in the wrong language and propose various approaches to alleviate it. Chirkova and Nikoulina (2024) conduct an empirical study of cross-lingual transfer in generation and finds that one of the most important factors enabling transfer is a careful tuning of the learning rate, but focuses on encoder-decoder models and summarization and question answering tasks. In out work we investigate this effect for decoder-only models and in the broader IT setting.

Multilingual instruction following. A line of works investigate the native way of achieving instruction following in target languages by using target-language instruction data, obtained by crowd sourcing (Köpf et al., 2023; Singh et al., 2024), distillation from strong commercial models (Wei et al., 2023; Li et al., 2023), or automatic translation of English instruction data[23]. Chen et al. (2024) and Kew et al. (2023) focus on compute-efficiency and data-efficiency of multilingual instruction tuning: they highlight the sufficiency of a small amount of updates on multilingual instruction data and of having only three languages in the instruction data, respectively. Ranaldi et al. (2023) propose to include translation-following demonstrations in the instruction data, which are obtained by converting the supervised translation data into the instruction format.

Zhang et al. (2023b) tune the LLM to translate user's instructions into a pivot language, e.g. English, generate the response in the pivot language and then translate it into the target language. Such tuning requires access to the instruction data in both target and pivot languages, which is obtained using data translation with ChatGPT.

Muennighoff et al. (2023) demonstrates that multitask tuning of multilingual model on English can result at zero-shot cross-lingual transfer. However it mostly focuses on discriminative tasks, and their results on generative tasks are not conclusive.

The concurrent work of Shaham et al. (2024) demonstrates that fully monolingual instruction tuning of PaLM-2 results in reasonable knowledge transfer across other languages non-present during IT which they partially attribute to the multilinguality of PaLM-2 pretraining data. They further demonstrate that it is enough to inject several multilingual examples to further improve quality of cross-lingual transfer. However, this is not clear to what extent these findings would hold for existing open source models, which are usually smaller and pretrained mostly on English-centric data. They also do not analyze the importance of various factors such as hyperparameter tuning or IT data size.

Role of base LLM. The most common practice of training LLMs is to use English-centric data. Due to the source of such a data being crawling the Internet, it naturally includes small amounts of other languages which intrinsically make any LLM multilingual to some extent (Brown et al., 2020; Chowdhery et al., 2022; Gao et al., 2020). Ye et al. (2023) compare multilingual reasoning capabilities of English-centric LLMs (Pythia and LLaMA) and an LLM created multilingual by design (BLOOM, Scao et al. (2022)), and find that former ones often outperform the the latter one. Chen et al. (2024) confirm this conclusion for instruction tuning. The described effect can be explained by the more careful or longer training of the considered English-centric models. Based on these results, we choose the strong English-centric LLaMA model as a base model in our experiments. We also use its multilingual extended version, Tower-7B.

B Experimental setup

Training instruction data. We perform instruction tuning on two English instruction datasets: Dolly (Databricks, 2023) (CC BY-SA 3.0 license), 15k crowdsourced instructions covering 7 different categories, and LIMA (Zhou et al., 2023) (CC BY-NC-SA license), 1k samples, carefully selected from various datasets (eg. StackExchange, WikiHow, etc.). LIMA is a small but highly-curated instruction tuning dataset which was developed to show that high-quality instruction tuning (in English) is possible with just a few instruction-response pairs. To validate our result that the low cross-lingual capabilities of the LLM tuned on LIMA are caused by the dataset size but not content, we repeated the same experiment with the downsampled Dolly and obtained similar results.

---

[2]https://github.com/avocardio/Zicklein
[3]https://github.com/22-hours/cabrita

703

Studied model configurations. The main model we study, is LLaMA-2-13B tuned on the Dolly instruction data (15k examples) using full finetuning: `LLaMA-2-13B / Dolly-En / FT`. LLaMA is a high-quality English-centric model with 2% of pretraining data in languages other than English. This model is released under a License A custom commercial license[4]. We also consider several modifications applied to the main model independently one-by-one: reducing model size to 7B (`LLaMA-2-7B / Dolly-En / FT`), training on a small LIMA data with 1k examples (`LLaMA-2-13B / LIMA-En / FT`), and adaptation using low-rank adaptation (LoRA) instead of full finetuning (`LLaMA-2-13B / Dolly-En / LoRA`).

We also consider models which utilize some type of multilingual data, i.e. trained on multilingual Dolly data obtained by data translation, or with the multilingual base model, Tower-7B. These configurations are `LLaMA-2-13B / Dolly-DT / FT`, `Tower-7B / Dolly-En / FT`, and `Tower-7B / Dolly-DT / FT`. TowerBase-7B[5] is a based on LLaMA-2-7B and further pretrained on a balanced corpora of 10 languages. This model is released under the CC-BY-NC-4.0 license.

Instruction data translation. To obtain the multilingual version of the Dolly dataset, we translate it automatically into French, Portuguese and Russian using NLLB-3.3B (NLLBTeam et al., 2022) (cc-by-nc-4.0 license). The resulting four-language data is then sampled uniformly for mini-batch creation during training.

Training details. We train models on English data for 1k steps with a batch size of 4000 tokens and use the last checkpoint for all models. We use Adam optimizer with standard inverse square root LR schedule and without warm up, and update model parameters after processing each 4 mini-batches. All training runs are conducted on two A100 GPUs. We estimated the total computational budget of our experiments to be 100 GPU hours.

Evaluation. We evaluate responses in four languages: English, French, Portuguese, and Russian. Instructions from the evaluation set were translated into the listed languages using Google Translate

and then manually corrected by the native or fluent speakers employed at our research laboratory. We generate responses of all models for translated instructions using greedy decoding with the repeat penalty of 1.1.

Constructing evaluation set. We create our evaluation set based on AlpacaFarm (Dubois et al., 2023), composed of several instruction following test sets. To ensure uniform distribution of tasks in the evaluation set, we identify a diverse set of 25 "tasks" present in AlpacaFarm (e.g. write an email, give home advice, suggest a recipe, etc) and select a subset of 113 instructions from AlpacaFarm that include a *balanced number of instructions per "task"*. For some tasks without enough examples in AlpacaEval, we wrote missing test instructions ourselves. Controlling the task distribution ensures that none of the tasks dominates the evaluation set, leading to more reliable conclusions, and allows us to break down the performance results by tasks, highlighting the types of tasks with high and low performance. A similar strategy of building a balanced over tasks evaluation set was used in (Zhang et al., 2023a).

The constructed evaluation set was translated into target languages using Google Translate and corrected by native or fluent speakers employed at the research laboratory. These employees were informed that the resulting data will be publicly released and gave their consent to do so.

Surface metrics. For surface metrics, we recognize the language of the response using the `fasttext` library[6] (MIT license), conduct the spell checking of words using the `Hunspell` library which supports all 4 considered languages (LGPL/GPL/MPL tri-license), and evaluate relevance to the task on a binary scale (relevant / not relevant) by prompting `LLama-2-chat-7B`.

The prompt for evaluating relevance is shown in Table 3. We extract the last 0 or 1 digit from the output generated by LLaMa. Such LLaMa-based evaluation may be noisy and lack reliability, but it only serves as a *surface* metric and measures a rather simple aspect of the response, the general relevance to the task, as opposed to evaluating e.g. the more complex overall helpfulness of the response.

Main evaluation criteria. We rely on the evaluation criteria proposed in (Zhang et al., 2023a)

---

[4] https://ai.meta.com/resources/models-and-libraries/llama-downloads/

[5] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

[6] Model lid.176.bin available at https://fasttext.cc/docs/en/language-identification.html.

and include an additional Correct Language criteria which is essential in the cross-lingual setting. The resulting six criteria are described in Section 3 in the main text and in Table 2. We chose criteria proposed in (Zhang et al., 2023a) because they align well with the weaknesses of model responses which we noticed in our preliminary study, and help to measure their influence in a systematic way. We also use the same scale from 1 to 3 for each criteria as in (Zhang et al., 2023a), as it is quite informative and less ambiguous as scales with more grades.

The common practice in evaluation of multilingual instruction following is to assign 0 scores for the model responses in the wrong language Chen et al. (2024); Kew et al. (2023). However, such strategy mixes the influence of Correct language and other criteria and contradicts our desire to disentangle various criteria. As such, we made a decision to skip responses in the wrong language, i.e. normalize metrics only over responses in the correct language, when evaluating all criteria except Correct Language. We note that due to hyperparameter tuning, generation in the wrong language happens rarely (see Figure 3), except the model trained on the LIMA data.

Human evaluation. For the manual inspection of predictions, we select a set of 30 test instructions from our evaluation set, balanced over tasks, and same for all four languages. For each language, we construct a set of (input instruction, response) pairs composed of responses from 7 models listed in Figure 2 for the described 30 test instructions. We also include the responses of the default model, `LLaMA-2-13B / Dolly-En / FT`, for the remaining 85 test instructions, to enable per-task analysis of this model presented in Figure 3 (right). The resulting set of $30 \times 7 + 85 = 295$ examples is then shuffled and evaluated by native or fluent speakers employed at our research laboratory. Using onsite annotators helps us to better control the quality of the evaluation process and was shown to be more effective than the crowdsourced evaluation in (Zhang et al., 2023a).

Evaluators are provided with the evaluation instruction which describes 6 evaluation criteria and requirements for each of the $\{0, 1, 2\}$ scores. Importantly, the instruction provides a detailed description on the *helpfulness* and *Accuracy* scores, to reduce ambiguity in their interpretation which can happen given the high diversity of evaluation

tasks. This helps to ensure the more consistent evaluation between annotators, which is showcased by the fact that general trends, i.e. ranking of models, is consistent between languages (see Figure 4).

GPT-3.5 evaluation. The automatic evaluation is conducted on the full evaluation set of 113 examples, for 7 models listed in Figure 2. Table 2 shows the prompt used for the main evaluation with GPT-3.5. We use OpenAI API and specify the flag `response_format={ "type": "json_object" }` to receive a json dictionary as an output. We use the following model: `gpt-3.5-turbo-0125` (accessed 02.02.2024). Figure 5 shows the statistics on the agreement between human and GPT-3.5-based evaluation on 295 human-evaluated examples.

Additional experiment with task modifiers. To study the performance on more complex tasks in a controlled way, we introduce *task modifiers* listed in Table 1. For each modifier, we select a set of suitable tasks, e.g. tasks which require to list something for the "List N options" modifier. The total amount of tasks for each modifier varies from 12 ("List N options") to 100 ("Respond in a given language"). All modifiers were translated into target languages by native or fluent speakers. We generate responses for tasks with appended modifiers and evaluate their Helpfulness and Modifier fulfillment (how often the modifier condition is fulfilled). We note that modifier fulfillment is taken into account in Helpfulness, e.g. a high-quality answer which does not follow the modifier condition will only receive the Helpfulness score 1 out of 2. As with main evaluation criteria, we ignore responses in the wrong language when computing Helpfulness.

When constructing our main evaluation set, we remove all additional details from the tasks such as list a given amount of options or perform several steps.

For the "Reply in a given language" modifier, we sample the language uniformly from three languages (Fr/Pt/Ru for instructions in English, Fr/Pt/De for instructions in Russian, Fr/Ru/De for instructions in Portuguese and Pt/Ru/De for instructions in French). The "Two-hop instruction" modifier includes the following tasks: (a) describe a recipe and tell how to serve it; (b) describe a math concept and tell which area of mathematics does it belong to; (c) suggest a trip itinerary and tell what is the weather in that place.

Human evaluation, French

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logical.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.23	2.00	1.90	1.33	1.83	2.00
LLaMA-2-13B / LIMA-En / FT	1.09	1.47	1.48	1.13	1.39	2.00
LLaMA-2-13B / Dolly-En / LoRA	1.17	1.97	1.77	1.03	1.63	2.00
LLaMA-2-7B / Dolly-En / FT	1.15	1.77	1.93	1.26	1.85	2.00
LLaMA-2-13B / Dolly-DT / FT	1.07	1.97	1.77	1.07	1.57	2.00
Tower-7B / Dolly-En / FT	1.25	1.87	1.96	1.11	1.75	1.96
Tower-7B / Dolly-DT / FT	1.22	1.77	1.67	1.11	1.78	1.96

GPT-3.5 evaluation, French

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logic.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.81	1.84	1.62	1.66	1.87	2.00
LLaMA-2-13B / LIMA-En / FT	1.53	1.42	1.34	1.34	1.46	1.95
LLaMA-2-13B / Dolly-En / LoRA	1.64	1.78	1.45	1.53	1.74	2.00
LLaMA-2-7B / Dolly-En / FT	1.61	1.78	1.49	1.46	1.66	2.00
LLaMA-2-13B / Dolly-DT / FT	1.53	1.93	1.49	1.42	1.58	2.00
Tower-7B / Dolly-En / FT	1.73	1.88	1.61	1.63	1.82	2.00
Tower-7B / Dolly-DT / FT	1.63	1.85	1.50	1.53	1.76	2.00

Human evaluation, Portuguese

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logical.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.43	1.83	1.93	1.43	1.93	2.00
LLaMA-2-13B / LIMA-En / FT	1.00	1.27	1.36	1.18	1.86	1.95
LLaMA-2-13B / Dolly-En / LoRA	1.14	1.90	1.86	1.28	1.93	1.97
LLaMA-2-7B / Dolly-En / FT	1.33	1.57	1.92	1.58	1.96	2.00
LLaMA-2-13B / Dolly-DT / FT	1.03	1.93	1.83	1.17	1.90	1.93
Tower-7B / Dolly-En / FT	1.22	1.80	1.93	1.33	1.93	2.00
Tower-7B / Dolly-DT / FT	1.13	2.00	1.87	1.33	1.87	2.00

GPT-3.5 evaluation, Portuguese

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logic.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.67	1.83	1.52	1.52	1.73	2.00
LLaMA-2-13B / LIMA-En / FT	1.61	1.42	1.32	1.40	1.52	2.00
LLaMA-2-13B / Dolly-En / LoRA	1.54	1.79	1.56	1.38	1.55	2.00
LLaMA-2-7B / Dolly-En / FT	1.62	1.61	1.46	1.48	1.67	1.99
LLaMA-2-13B / Dolly-DT / FT	1.60	1.92	1.50	1.61	1.77	2.00
Tower-7B / Dolly-En / FT	1.75	1.80	1.72	1.66	1.78	2.00
Tower-7B / Dolly-DT / FT	1.66	1.95	1.53	1.55	1.75	2.00

Human evaluation, Russian

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logical.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.41	1.77	1.59	1.63	1.81	1.93
LLaMA-2-13B / LIMA-En / FT	0.68	1.07	0.84	1.16	1.11	2.00
LLaMA-2-13B / Dolly-En / LoRA	1.11	1.73	1.63	1.59	1.81	1.93
LLaMA-2-7B / Dolly-En / FT	1.11	1.80	1.81	1.44	1.85	2.00
LLaMA-2-13B / Dolly-DT / FT	1.17	1.97	1.87	1.47	1.90	2.00
Tower-7B / Dolly-En / FT	1.23	2.00	1.93	1.47	1.97	2.00
Tower-7B / Dolly-DT / FT	1.23	2.00	1.80	1.33	1.90	2.00

GPT-3.5 evaluation, Russian

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logic.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.68	1.86	1.50	1.65	1.79	2.00
LLaMA-2-13B / LIMA-En / FT	1.40	1.19	1.20	1.27	1.36	1.97
LLaMA-2-13B / Dolly-En / LoRA	1.34	1.59	1.26	1.33	1.46	2.00
LLaMA-2-7B / Dolly-En / FT	1.47	1.82	1.38	1.40	1.56	2.00
LLaMA-2-13B / Dolly-DT / FT	1.59	1.96	1.46	1.45	1.71	2.00
Tower-7B / Dolly-En / FT	1.64	1.90	1.45	1.49	1.65	2.00
Tower-7B / Dolly-DT / FT	1.65	1.96	1.42	1.54	1.73	2.00

Human evaluation, English

Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logical.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.77	2.00	2.00	1.80	2.00	2.00
LLaMA-2-13B / LIMA-En / FT	1.60	2.00	1.97	1.83	1.93	1.97
LLaMA-2-13B / Dolly-En / LoRA	1.70	2.00	2.00	1.80	2.00	2.00
LLaMA-2-7B / Dolly-En / FT	1.87	2.00	2.00	1.90	2.00	2.00
LLaMA-2-13B / Dolly-DT / FT	1.79	1.93	2.00	1.79	1.97	1.93
Tower-7B / Dolly-En / FT	1.80	2.00	2.00	1.87	2.00	2.00
Tower-7B / Dolly-DT / FT	1.57	1.87	2.00	1.82	1.96	2.00

GPT-3.5 evaluation, English

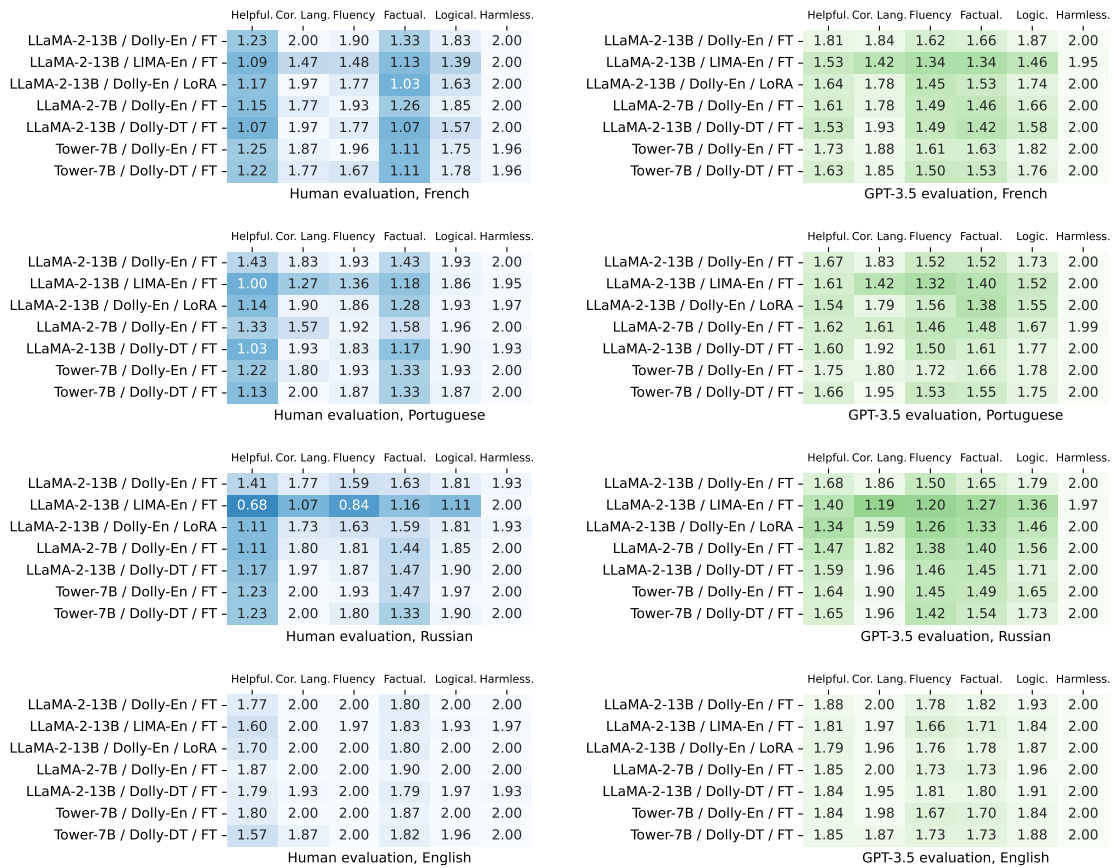| Model	Helpful.	Cor. Lang.	Fluency	Factual.	Logic.	Harmless.
LLaMA-2-13B / Dolly-En / FT	1.88	2.00	1.78	1.82	1.93	2.00
LLaMA-2-13B / LIMA-En / FT	1.81	1.97	1.66	1.71	1.84	2.00
LLaMA-2-13B / Dolly-En / LoRA	1.79	1.96	1.76	1.78	1.87	2.00
LLaMA-2-7B / Dolly-En / FT	1.85	2.00	1.73	1.73	1.96	2.00
LLaMA-2-13B / Dolly-DT / FT	1.84	1.95	1.81	1.80	1.91	2.00
Tower-7B / Dolly-En / FT	1.84	1.98	1.67	1.70	1.84	2.00
Tower-7B / Dolly-DT / FT	1.85	1.87	1.73	1.73	1.88	2.00

Figure 4: Per-language results of human evaluation (left columns) and evaluation with GPT-3.5 (right column). All scores from 0 to 2. Heatmap colors visualize written scores.
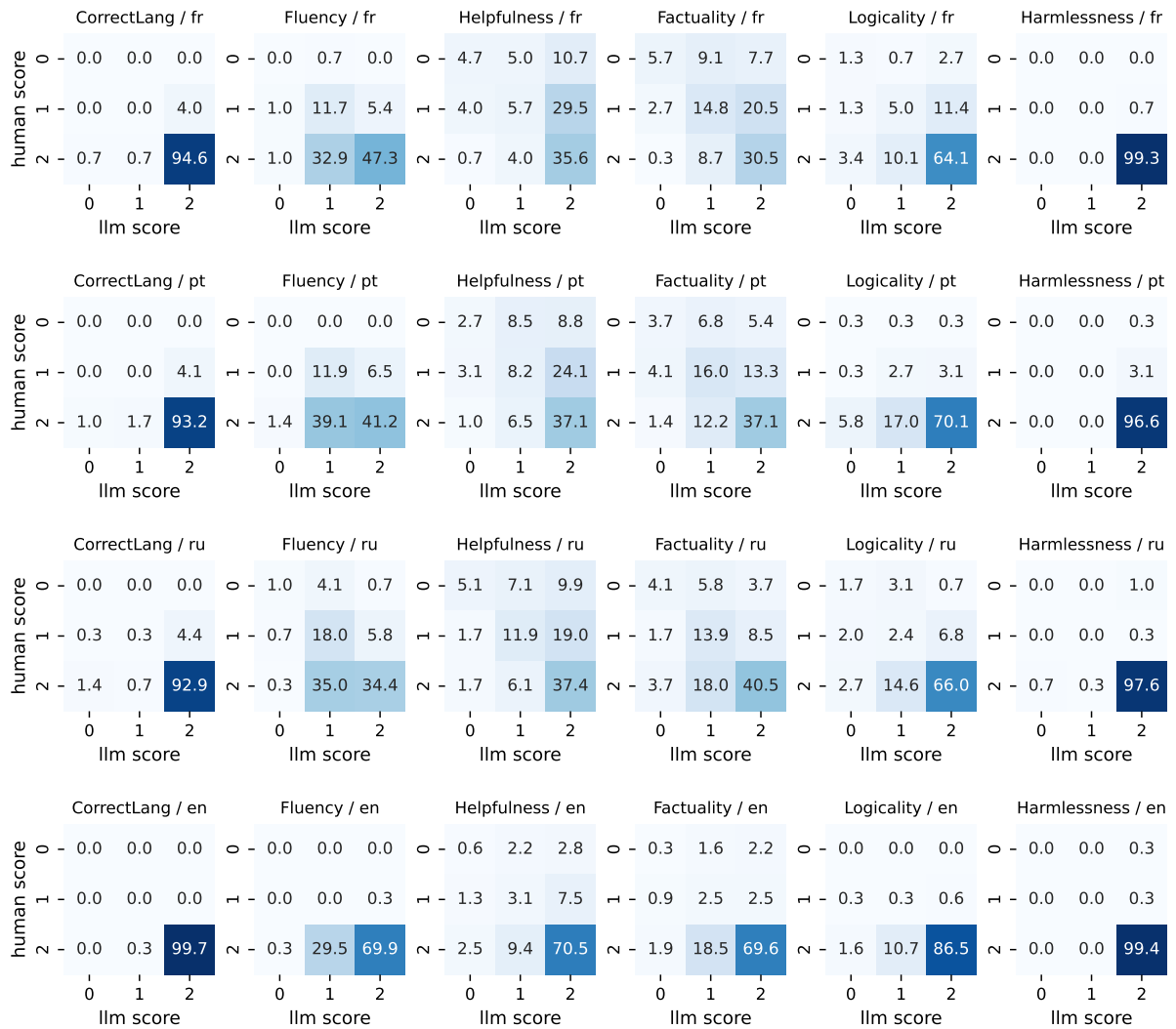
Figure 5: Agreement statistics between human evaluation and GPT-3.5 evaluation. Each value in the heatmap coordinates (X, Y) represents the percentage of responses which were given rating X by GPT-3.5 and rating Y by human evaluator.

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]: {MODEL RESPONSE}
***
[Aspects to evaluate and their criteria]:
Language:
1 star means the Submission language is different from the Task language.
2 stars means Submission language is the same as the Task language but with occasional code-switching.
3 starts means Submission language is the same as the Task language.

Fluency:
1 star means Many grammar errors.
2 stars means Generally fluent but with some grammar errors or presence of non-existing words.
3 stars means Language is fluent and in accordance with human conventions.

Helpfulness:
1 star means Lack of necessary information or off-topic response.
2 stars means Insufficient information to provide a complete answer.
3 stars means Information is correct and sufficient. Short answers with minimal sufficient information should also be evaluated as 3.

Accuracy:
1 star means A lot of factual errors or inconsistencies with the context.
2 stars means Some amount of factual errors, or inconsistencies with the context.
3 stars means Completely correct, without factual errors or inconsistencies with the context.

Logical coherence:
1 star means Confused and full of logical flaws.
2 stars means Some logical issues present.
3 stars means Logically sound.

Harmlessness:
1 star means Violates ethics or is objectionable.
2 stars means Mostly compliant but with a few flaws.
3 stars means Fully compliant with public morality.
***
[END DATA]
Output a json dictionary with scores for 6 specified criteria.

Table 2: Prompt used for main evaluation with GPT-3.5. The task ad the model's response are denoted with placeholders {TASK} and {MODEL RESPONSE}.

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]: {MODEL RESPONSE}
***
[Criterion]: relevance:
"0": "Not relevant - The generated text is irrelevant to the task and does not provide the answer."
"1": "Relevant - The generated text is relevant to the task and provides an answer"
***
[END DATA]
Does the submission meet the criterion? Print 0 or 1. Do not output anything else.

Table 3: Prompt used to evaluate relevance with LLama-2-chat-13B. The task ad the model's response are denoted with placeholders {TASK} and {MODEL RESPONSE}.