

No Optimal Language Set Exists for Multilingual Instruction Tuning: Insights from a Linguistically-Informed Study

Gürkan Soykan^{*1}, Gözde Gül Şahin^{2,3,4}

¹ Information Technology Group, Wageningen University & Research, Netherlands

² Friedrich-Alexander-Universität Erlangen-Nürnberg, Intelligent Language Systems

³ Computer Engineering Department, Koç University, Istanbul, Turkey

⁴ KUIS AI Lab, Istanbul, Turkey

<https://gglab-ku.github.io/>

Abstract

Multilingual instruction tuning (MIT) is challenged by the curse of multilinguality, data scarcity, and high computational cost. A natural hypothesis is that carefully selecting a linguistically diverse set of languages yields universally better models. We test this systematically by evaluating linguistically-informed selection strategies—based on typological, geographical, semantic, and learned features—against random baselines across three model families (mGPT, mT5-xl, BLOOM) and five multilingual benchmarks. Our key negative finding is that no universal language selection strategy emerges in our fixed-budget setting: performance is strongly task- and model-dependent, and adding more languages beyond a modest threshold triggers degradation consistent with the curse of multilinguality. We discuss implications for MIT data curation and the pitfalls of benchmark-averaged evaluation. All resources are publicly available at <https://github.com/GGLAB-KU/ling-informed-mit>.

1 Introduction

Multilingual language models exhibit uneven cross-lingual performance (Joshi et al., 2020), and instruction tuning with multilingual data has emerged as a practical remedy (Muennighoff et al., 2022). However, the curse of multilinguality (Conneau et al., 2019)—where per-language performance degrades as more languages compete for model capacity—complicates the pursuit of broad coverage, as do high computational costs and data scarcity.

A natural hypothesis follows: selecting the *right* linguistically diverse subset of languages should yield models that generalize well across languages and tasks. Prior work lends partial support—Chen et al. (2023a) and Shaham et al. (2024) show small

multilingual subsets can match or exceed monolingual tuning; Weber et al. (2024) confirm that parallel multilingual corpora improve cross-lingual instruction-following by up to 9.9% over monolingual training; and language features have aided cross-lingual transfer in related tasks (Lin et al., 2019; Üstün et al., 2020). Yet no prior work has systematically investigated whether linguistically-informed language *selection* yields a universally superior MIT strategy.

We fill this gap. Our central negative finding is that **no universal language selection strategy was observed in the fixed-budget MIT setting studied here**: the best choice depends on model family and evaluation task, which has important consequences for how the community evaluates and reports MIT results.

2 Experimental Setup

We use Bactrian-X (Li et al., 2023), a parallel multilingual instruction dataset with 3.4M instruction-response pairs across 52 languages. Following the budget-controlled setup of Chen et al. (2023a), we fix total training to 67k instances per experiment (4,786 per language for 14-language subsets). We select 14 languages per subset using k-means clustering ($k=14$) on different feature spaces: **GEO** (geographical vectors (Hammarström et al., 2023)), **TYPO** (typological vectors (Littell et al., 2017)), **SEM** (semantic typology embeddings (Chen et al., 2023b)), **LEARN** (learned language vectors (Malaviya et al., 2017)), **FAM** (one language per family), **RND** (random, baseline), and **ALL** (all 52 languages). Each is run with 3 random seeds. Table 1 lists the selected languages.

We instruction-tune three model families with LoRA (Hu et al., 2021): mGPT-1.3B (Shliazhko et al., 2024), mT5-XL-3.7B (Xue et al., 2021), and BLOOM-{1.7B, 3B, 7B} (Scao et al., 2022), eval-

^{*}E-mail: gurkan.soykan@wur.nl, Work is done during PhD studies at Koç University.

Table 1: Language subsets per strategy and seed (ISO codes).

Subset	Languages
FAM	az,en,fi,he,ja,ka,vi,ko,ml,mn,my,th,tl,xh ar,en,fi,id,ja,ka,ko,mn,sw,ta,th,tr,vi,zh af,az,et,he,ja,ka,ko,ml,mn,my,th,tl,vi,xh
GEO	af,bn,et,fr,gu,he,hr,id,ja,kk,mn,sw,ta,vi af,bn,et,fr,he,hr,id,ja,ka,kk,mn,ta,ur,vi af,bn,et,fr,he,hr,id,ja,kk,mn,sw,ta,ur,vi
LEARN	cs,hi,id,km,ko,lt,lv,my,nl,pt,sl,ta,tl,vi cs,fi,hr,km,ko,lt,lv,my,nl,pt,sl,ta,uk,vi cs,hi,id,ja,lt,lv,nl,pt,sl,sv,ta,tl,tr,vi
RND	cs,gu,hi,id,ko,lv,mk,ml,ps,pt,si,ta,vi,zh cs,en,et,fr,hi,hr,it,ja,ml,mn,mr,ne,pl,pt ar,az,de,en,es,et,he,hi,id,ml,ps,ru,sv,uk
SEM	gl,gu,ka,kk,ko,ml,pl,ps,si,sl,sv,tl,uk,vi en,fr,gl,gu,ka,kk,ko,pl,ps,sl,sv,tl,uk,vi en,gl,gu,ka,kk,ko,ml,pl,ps,sl,sv,tl,uk,vi
TYPO	az,bn,de,et,fa,hi,it,ja,mk,sw,ta,tl,vi,zh ar,az,bn,de,et,fa,it,ja,mk,sw,ta,th,ur,zh ar,az,bn,en,fi,he,hi,it,ja,mk,nl,sw,ta,th

uated zero-shot on XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2021), and XWinograd (Tikhonov and Ryabinin, 2021) using the lm-evaluation-harness (Gao et al., 2023).

3 Results

3.1 No Single Strategy Wins Across All Tasks and Models

Table 2 presents our main results for BLOOM-7B, mGPT, and mT5-xl. The best-performing strategy varies substantially across models and tasks: GEO leads for BLOOM-7B on XStoryCloze while TYPO achieves the best average for BLOOM-7B overall, FAM is best on average for mT5-xl, and GEO or TYPO lead for mGPT on NLU tasks. Crucially, **no single subset consistently achieves the highest score across all model-task combinations**.

Improvements over the RND baseline are often not statistically significant. Paired t-tests (three seeds) show significant improvements only in specific model-task combinations—e.g., GEO vs. RND for mGPT on XNLI ($p \leq 0.05$) and XStoryCloze ($p \leq 0.05$), FAM vs. RND for mT5-xl on XCOPA ($p \leq 0.05$) and XWinograd ($p \leq 0.05$)—but these do not generalize across all benchmarks or model families. Meanwhile, several linguistically-informed subsets score *significantly lower* than RND in specific configurations (red

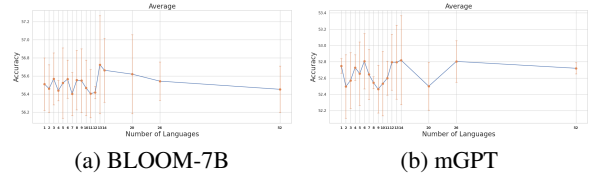


Figure 1: Average performance with confidence intervals when varying the number of GEO-selected languages. Performance peaks at 13–14 and declines beyond.

cells), underscoring that informed selection can also backfire. The effect sizes in Table 2 contextualise this: Cohen’s d_z expresses the mean gain over RND in units of its seed-to-seed variability (0.2/0.5/0.8=small/medium/large) and derives from the same paired t-tests. With $n=3$ seeds, $p \leq 0.05$ requires $|d_z| \geq 2.5$, so only very large effects can reach significance. Even these are model-specific (GEO: $d_z=3.6$ on mGPT vs. 0.5 on BLOOM-7B), and with three seeds individual d_z values are magnitude indicators rather than precise estimates.

For PAWS-X, no strategy produces significant improvement over RND in any model, and the base (non-tuned) model achieves the best PAWS-X score for both BLOOM-7B and mGPT—suggesting MIT may actively harm paraphrase identification performance.

3.2 More Languages Triggers Curse of Multilinguality

Figure 1 shows the effect of varying language count (1 to 52) using GEO on mGPT and BLOOM-7B. For BLOOM-7B, performance is stable up to 13–14 languages, peaks there, then declines consistently—replicating the curse of multilinguality (Conneau et al., 2019) under a fixed computational budget. The ALL setting (52 languages) never achieves best average performance in either model. mGPT shows wider confidence intervals but a similar trend: the 14-language default performs marginally better than larger configurations.

3.3 Cross-lingual Generalization to Unseen Languages is Model-Dependent

Figure 2 shows model performance on *unseen* languages (those in evaluation benchmarks but not in any training subset). While linguistically-informed subsets generally improve over the non-tuned base model for mGPT, mT5-xl, and BLOOM-3B, results diverge for larger BLOOM models. For BLOOM-

Table 2: Main results for BLOOM-7B, mGPT, and mT5-xl. **Bold**: highest mean per model-task. Color coding vs. RND baseline (paired t-test): **dark green** $p \leq 0.05$, **green** $0.05 < p \leq 0.1$, **light green** $0.1 < p \leq 0.15$ (significantly better); **dark red**, **red**, **light red** (significantly worse). “-” = base model (no instruction tuning). In the Avg. column, the parenthesised value is the paired effect size vs. RND (Cohen’s d_z).

Model	Subset	XNLI	XCOPA	XStoryCloze	XWinograd	PAWS-X	Avg. (d_z)
Bloom-7B	ALL	40.88 $\pm$ 0.25	58.18 $\pm$ 0.54	62.03 $\pm$ 0.13	73.71 $\pm$ 0.23	47.46 $\pm$ 0.82	56.45 $\pm$ 0.26 (-0.1)
	RND	41.63 $\pm$ 0.25	57.72 $\pm$ 0.86	61.73 $\pm$ 0.30	73.91 $\pm$ 0.31	47.36 $\pm$ 2.85	56.47 $\pm$ 0.69
	FAM	41.18 $\pm$ 0.67	57.95 $\pm$ 0.19	61.49 $\pm$ 0.72	74.00 $\pm$ 0.81	47.88 $\pm$ 2.29	56.50 $\pm$ 0.56 (0.1)
	GEO	41.24 $\pm$ 0.66	57.97 $\pm$ 0.39	62.15 $\pm$ 0.31	73.96 $\pm$ 1.69	47.99 $\pm$ 0.85	56.66 $\pm$ 0.35 (0.5)
	LEARN	41.26 $\pm$ 0.66	58.08 $\pm$ 0.22	61.96 $\pm$ 0.06	74.17 $\pm$ 0.29	46.88 $\pm$ 1.85	56.47 $\pm$ 0.21 (0.0)
	SEM	41.19 $\pm$ 0.54	58.03 $\pm$ 0.66	61.44 $\pm$ 0.69	74.16 $\pm$ 0.18	47.35 $\pm$ 0.47	56.43 $\pm$ 0.31 (-0.1)
	TYPO	41.63 $\pm$ 0.80	58.09 $\pm$ 0.54	61.94 $\pm$ 0.28	74.17 $\pm$ 1.08	47.75 $\pm$ 0.31	56.71 $\pm$ 0.24 (0.8)
mGPT	-	41.12	56.87	59.30	73.97	49.37	56.13
	ALL	41.23 $\pm$ 0.12	55.82 $\pm$ 0.14	55.98 $\pm$ 0.07	60.44 $\pm$ 0.30	50.14 $\pm$ 0.43	52.72 $\pm$ 0.07 (2.0)
	RND	40.71 $\pm$ 0.23	55.91 $\pm$ 0.62	55.97 $\pm$ 0.36	60.78 $\pm$ 0.20	48.71 $\pm$ 0.99	52.42 $\pm$ 0.36
	FAM	40.27 $\pm$ 0.08	55.82 $\pm$ 0.92	55.81 $\pm$ 0.85	60.88 $\pm$ 1.04	49.69 $\pm$ 1.72	52.49 $\pm$ 0.30 (0.7)
	GEO	41.01 $\pm$ 0.29	55.73 $\pm$ 0.56	56.35 $\pm$ 0.13	61.05 $\pm$ 0.44	49.97 $\pm$ 2.64	52.82 $\pm$ 0.55 (3.6)
	LEARN	40.64 $\pm$ 0.42	55.95 $\pm$ 0.47	56.15 $\pm$ 0.27	60.66 $\pm$ 1.03	48.92 $\pm$ 0.84	52.46 $\pm$ 0.38 (0.2)
	SEM	40.77 $\pm$ 0.21	55.61 $\pm$ 0.42	56.14 $\pm$ 0.34	60.57 $\pm$ 0.47	49.17 $\pm$ 2.17	52.45 $\pm$ 0.44 (0.4)
mT5-xl	TYPO	40.84 $\pm$ 0.31	55.93 $\pm$ 0.57	56.24 $\pm$ 0.44	60.39 $\pm$ 0.83	49.17 $\pm$ 2.06	52.51 $\pm$ 0.64 (0.7)
	-	40.90	55.04	54.43	60.55	50.30	52.24
	ALL	36.49 $\pm$ 0.69	53.64 $\pm$ 1.01	52.31 $\pm$ 0.21	50.39 $\pm$ 0.81	50.76 $\pm$ 1.15	48.72 $\pm$ 0.31 (-0.3)
	RND	36.86 $\pm$ 0.29	53.10 $\pm$ 0.48	52.46 $\pm$ 0.52	49.88 $\pm$ 0.75	51.74 $\pm$ 1.74	48.81 $\pm$ 0.55
	FAM	36.12 $\pm$ 0.44	53.72 $\pm$ 0.34	52.59 $\pm$ 0.34	51.01 $\pm$ 1.69	52.28 $\pm$ 0.76	49.14 $\pm$ 0.19 (1.4)
	GEO	36.91 $\pm$ 0.36	53.23 $\pm$ 0.23	52.61 $\pm$ 1.03	50.20 $\pm$ 0.76	52.11 $\pm$ 0.91	49.01 $\pm$ 0.50 (0.8)
	LEARN	36.69 $\pm$ 0.55	53.48 $\pm$ 0.53	52.62 $\pm$ 0.21	49.65 $\pm$ 1.49	51.69 $\pm$ 1.05	48.82 $\pm$ 0.13 (0.1)
	SEM	36.41 $\pm$ 0.57	53.35 $\pm$ 0.29	52.38 $\pm$ 0.33	50.62 $\pm$ 1.89	51.63 $\pm$ 0.66	48.88 $\pm$ 0.39 (0.2)
	TYPO	36.62 $\pm$ 0.37	53.22 $\pm$ 0.54	52.58 $\pm$ 0.34	50.09 $\pm$ 0.86	51.40 $\pm$ 0.35	48.78 $\pm$ 0.11 (-0.1)
	-	33.32	52.35	51.36	50.57	50.14	47.55

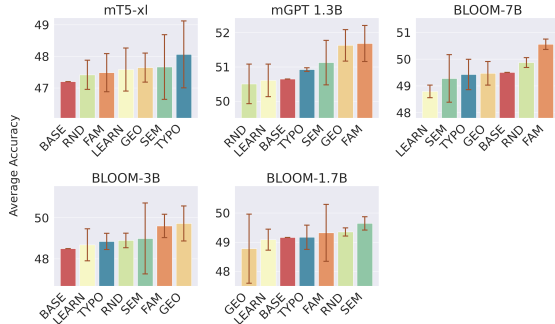


Figure 2: Average performance on unseen languages (not part of any instruction-tuning subset).

7B, only two of the five subsets improve over the base. This reveals a non-trivial interaction between pretraining corpus composition and subset selection: a strategy effective for one model’s pretraining distribution may be harmful for another’s.

3.4 Model Size Amplifies Differences—But Not Uniformly

Figure 3 shows how language subset performance evolves across BLOOM-1.7B, 3B, and 7B. While all instruction-tuned models surpass the base, the random baseline consistently ranks lowest among tuned models. GEO excels at 3B and 7B, whereas FAM starts strong at 1.7B but diminishes at scale.

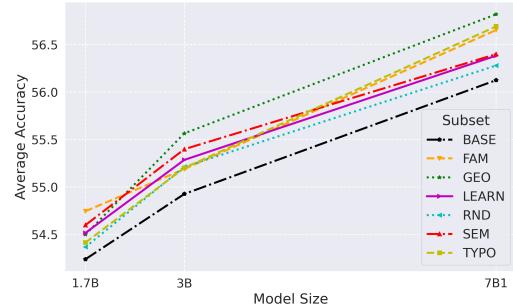


Figure 3: Average zero-shot performance across BLOOM model sizes by language subset. Strategy rankings shift with scale.

TYPO improves progressively with scale, reaching second place at 7B. Importantly, **the ranking of strategies changes across model sizes**: there is no strategy that is uniformly best as scale increases.

3.5 Monolingual Tuning Remains Competitive

In a Vietnamese case study (Figure 4), monolingual tuning (67k Vietnamese examples) ranks first for mT5 and second for mGPT on average across XNLI and XCOPA—with no statistically significant difference from the best multilingual subset (GEO) for mGPT. For BLOOM, the monolingual case consistently underperforms. This partially undermines the assumption that multilingual sub-

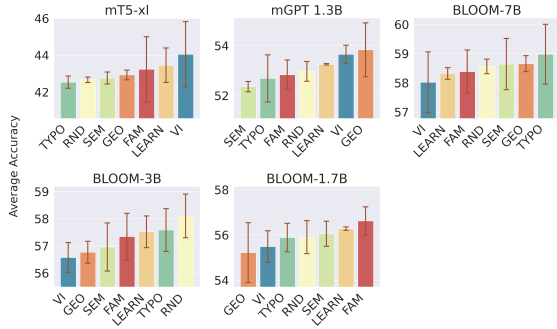


Figure 4: Monolingual (VI) vs. multilingual instruction tuning for Vietnamese on XNLI and XCOPA. Monolingual is competitive for mT5 and mGPT but not BLOOM.

sets are always preferable, and shows that the right choice is again model-dependent.

3.6 Training–Evaluation Language Overlap

Table 3 shows, for XNLI, how much overlap exists between each model’s pretraining languages (M), each training subset (S), and the evaluation languages (T). BLOOM stands out: its pretraining corpus covers only 8 of the 15 XNLI languages, and it has the smallest three-way overlaps of all models. This matches Table 2, where most informed subsets score significantly below RND on XNLI for BLOOM-7B.

mGPT instead covers 14 of the 15 XNLI languages, and the same informed subsets score significantly higher there. We find that overlap alone does not decide the outcomes, however: GEO improves XStoryCloze significantly on both BLOOM-7B and mGPT while sharing only 2 of its 11 evaluation languages. Since benchmark language sets themselves barely intersect (XNLI covers 15 languages, XWinograd 6), no single subset can align with every task.

4 Discussion and Implications

Why no optimal set exists. Three compounding factors explain the absence of a universal winner. (1) *Pretraining corpus mismatch*: languages that complement one model’s pretraining gaps differ across mGPT (61 languages), BLOOM (59, different sources), and mT5 (101). We quantify this: for BLOOM, a language’s accuracy closely tracks its share of the pretraining corpus, both before tuning (pooled Spearman $\rho=0.55\text{--}0.63$, $p<10^{-4}$) and after ($\rho=0.58\text{--}0.65$). For mT5, whose sampling flat-

Table 3: Language intersections for XNLI: counts of shared languages between the model’s pretraining corpus (M), the instruction tuning language subset (S), and the task’s evaluation languages (T). Bloom = BLOOM-7B.

Model	Subset	M-S-T	S-T	M-T	M-S
mT5-xl	ALL	13/15	13/15	15/15	50/52
mGPT	ALL	12/15	13/15	14/15	39/52
mT5-xl	RND	9/15	9/15	15/15	31/32
mT5-xl	TYPO	9/15	9/15	15/15	20/21
Bloom	ALL	8/15	13/15	8/15	18/46
mGPT	RND	8/15	9/15	14/15	24/32
mGPT	TYPO	8/15	9/15	14/15	18/21
mT5-xl	FAM	7/15	7/15	15/15	21/22
Bloom	RND	6/15	9/15	8/15	13/32
Bloom	TYPO	6/15	9/15	8/15	8/21
mGPT	FAM	6/15	7/15	14/15	20/22
Bloom	FAM	4/15	7/15	8/15	8/22
mT5-xl	GEO	4/15	4/15	15/15	15/16
mGPT	GEO	4/15	4/15	14/15	14/16
Bloom	GEO	4/15	4/15	8/15	8/16
mT5-xl	LEARN	3/15	3/15	15/15	18/20
mT5-xl	SEM	3/15	3/15	15/15	15/16
mGPT	LEARN	3/15	3/15	14/15	15/20
mGPT	SEM	3/15	3/15	14/15	11/16
Bloom	SEM	3/15	3/15	8/15	5/16
Bloom	LEARN	2/15	3/15	8/15	5/20

tens language proportions, the correlation vanishes ($\rho=-0.04$, n.s.). Yet a subset’s aggregate pretraining share does not predict tuned performance (n.s. in four of five models). Hence, pretraining explains where models perform well, but not which subset to tune on. (2) *Task–language alignment* (§3.6): high subset overlap with XNLI’s 15 languages does not imply overlap with XWinograd’s 6 or XCOPA’s 11, making universal optimality structurally impossible. (3) *Seed sensitivity*: three-seed confidence intervals are wide enough that strategy differences often fall within noise. These factors are not unique to our setting: Shimabucoro et al. (2025) reach the same conclusion at larger scale (up to 35B parameters), finding CLT dynamics vary jointly with task type, model size, and training regime.

Methodological warning. Reporting single aggregate scores across tasks and models in MIT can be deeply misleading. A strategy that appears superior in aggregate may underperform significantly on specific tasks: FAM achieves the best average for mT5-xl but scores worst on XNLI; GEO leads for mGPT overall but shows negative effects on BLOOM for the same task. Strikingly, the base non-tuned model scores best on PAWS-X for both BLOOM-7B and mGPT, suggesting that MIT can *harm* performance on certain task types.

We urge the community to always report per-task and per-model breakdowns with confidence intervals, and to avoid over-generalizing from a limited set of benchmarks or model families. Additionally, the zero-shot multiple-choice format of standard evaluation harnesses (Gao et al., 2023) has been critiqued for misaligning with real-world model use (Lyu et al., 2024), which may further explain cross-task variability.

Practical takeaway. Despite the absence of a universal optimum, linguistically-informed selection—particularly GEO and TYPO—*tends* to be a safer default than random selection, especially at larger model scales. For practitioners without resources to run full ablations, GEO-based selection via URIEL features (Littell et al., 2017) provides a compute-free heuristic that beats random on average. Alternatively, representation-space approaches such as LangGPS (Ye et al., 2026) offer a data-driven signal when model access is available. The key practical insight for dataset curators is to target 10–15 linguistically diverse languages rather than maximising language count: more languages is not better under a fixed computational budget.

5 Conclusion

We set out to find the optimal set of languages for multilingual instruction tuning. In our fixed-budget setting, we did not find one. Linguistically-informed language selection generally outperforms random selection on average, but no tested strategy is consistently best across tasks, model families, and evaluation settings. The curse of multilinguality reasserts itself beyond approximately 14 languages under a fixed computational budget; cross-lingual generalization to unseen languages is highly model-dependent; monolingual tuning can be competitive with multilingual approaches; and strategy rankings shift with model scale. Taken together, these negative findings caution against simple universal prescriptions for MIT data curation. They also highlight the need for task- and model-specific evaluation with proper statistical reporting—a standard the field has not yet widely adopted.

Limitations

Several limitations bear on the generalisability of our findings. First, we rely exclusively on LoRA (Hu et al., 2021) as the PEFT method; other fine-tuning approaches may yield different sensitivity to language subset choice. Second, Bactrian-

X (Li et al., 2023) consists largely of automatically translated English instructions, potentially introducing translation artefacts that confound language-specific effects. To gain more insights, we scored each language’s translation quality as the mean LaBSE (Feng et al., 2020) similarity between 500 English source instructions and their translations. We find that quality does not significantly correlate with per-language performance (Spearman $\rho=0.09-0.24$, n.s., vs. 0.58 for pretraining share over the same languages). Furthermore, the selection strategies barely differ in the mean quality of the languages they pick (0.862–0.877). Translation quality is therefore unlikely to explain our negative result, although LaBSE is itself an imperfect proxy.

In addition, all experiments are capped at 7B parameters due to resource constraints; behaviour at larger scales remains unknown. Fourth, only three random seeds per configuration limit statistical power, leaving confidence intervals wider than desirable—a key reason many strategy comparisons fail to reach significance. Fifth, language selection is constrained to the 52 Bactrian-X languages; a broader or different pool could yield different subsets.

Finally, the zero-shot multiple-choice evaluation format may not reflect instruction-following quality in open-ended generation settings (Lyu et al., 2024; Biderman et al., 2024), which is the more relevant real-world use case for instruction-tuned models. However, the models studied here generate outputs of too low quality for a meaningful open-ended evaluation, as we observed in preliminary LLM-as-a-judge experiments.

Acknowledgments

This work has been supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) as part of the project “Automatic Learning of Procedural Language from Natural Language Instructions for Intelligent Assistance” with the number 121C132. We also gratefully acknowledge KUIS AI Lab for providing computational support.

References

- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sid Black, Jordan Clive, Anthony DiPofi, Julien Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, and 10 others. 2024. [Lessons from the trenches on reproducible evaluation of language models](#). *ArXiv*, abs/2405.14782.
- Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023a. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.
- Yiyi Chen, Russa Biswas, and Johannes Bjerva. 2023b. [Colecx2Lang: Language embeddings from semantic typology](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 673–684, Tórshavn, Faroe Islands. University of Tartu Library.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#). In *Annual Meeting of the 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.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT Sentence Embedding. In *Annual Meeting of the Association for Computational Linguistics*.
- 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, and 5 others. 2023. [A framework for few-shot language model evaluation](#). Zenodo. Version v0.4.0, DOI: 10.5281/zenodo.10256836.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. [Glottolog database 4.8](#). Available online at <https://glottolog.org>.
- 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. *arXiv preprint arXiv:2106.09685*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. [Bactrian-X: A multilingual replicable instruction-following model with low-rank adaptation](#). *Preprint*, arXiv:2305.15011.
- 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 T. Diab, and 2 others. 2021. [Few-shot learning with multilingual language models](#). *CoRR*, abs/2112.10668.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.
- Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. [Beyond probabilities: Unveiling the misalignment in evaluating large language models](#). In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.
- Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.
- 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. 2022. [Crosslingual generalization through multitask finetuning](#). *Preprint*, arXiv:2211.01786.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020.

- [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 30 others. 2022. [BLOOM: A 176b-parameter open-access multilingual language model](#). *CoRR*, abs/2211.05100.
- Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. [Multilingual instruction tuning with just a pinch of multilinguality](#). *ArXiv*, abs/2401.01854.
- Luísa Shimabucoro, Ahmet Üstün, Marzieh Fadaee, and Sebastian Ruder. 2025. A post-trainer’s guide to multilingual training data: Uncovering cross-lingual transfer dynamics. *arXiv preprint arXiv:2504.16677*.
- Oleh Shliakhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. [mGPT: Few-shot learners go multilingual](#). *Transactions of the Association for Computational Linguistics*, 12:58–79.
- Alexey Tikhonov and Max Ryabinin. 2021. [It’s all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning](#). *Preprint*, arXiv:2106.12066.
- Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. [UDapter: Language adaptation for truly Universal Dependency parsing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. 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?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20829–20855, Miami, Florida, USA. 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.
- Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. [PAWS-X: A cross-lingual adversarial dataset for paraphrase identification](#). 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.
- Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Lei Huang, Weitao Ma, Qichen Hong, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. 2026. [LangGPS: Language separability guided data pre-selection for joint multilingual instruction tuning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.