One Model, Many Worlds: Cross-Lingual Fine-Tuning Can Improve Low-Resource Capabilities of Language Models

Tyler Slomianyj[1], **Rudraansh Korlakunta**[1], **Victor He**[1], **Daniel Gao**[1], **Sunishchal Dev**[1], **Kevin Zhu**[1], **Aryan Shrivastava**[2]

[1]Algoverse AI Research
[2]University of Chicago

Abstract

Multilingual language models (LLMs) have demonstrated strong cross-lingual reasoning and comprehension capabilities. However, substantial performance disparities persist between high- and low-resource languages due to unbalanced availability of training data and linguistic diversity. This paper examines fine-tuning efficacy to determine the relative importance of language, domain, and resource-level, exploring how we can reduce these disparities in performance. Using `gpt-4.1-nano-2025-04-14`, we conducted experiments on three domains: STEM, Medical, and Humanities from the Global-MMLU dataset, focusing primarily on cross-lingual transfer. We find substantial accuracy improvements when transferring from high to low resource settings ($\approx +16\%$), but large performance degradation when transferring in the opposite direction ($\approx -13\%$). Additionally, we find that only cross-lingual ($+2.61\%$) transfers demonstrate a net improvement while cross-domain (-2.44%) transfers degrade performance. These findings present preliminary evidence that training data from linguistically diverse languages can enhance model generalization and narrow the performance gap in multilingual language models, even when low-resource language data is scarce or absent altogether.

Introduction

Multilingual transformer-based language models have achieved remarkable progress in recent years, delivering strong results across a wide range of languages (Singh et al. 2024). These token-based systems share parameters across languages, enabling substantial cross-lingual transfer, particularly in high-resource settings. However, they continue to struggle with low-resource languages, where limited training data makes it difficult to capture diverse morphological and lexical patterns (Thangaraj et al. 2024). This performance gap remains a major obstacle to equitable language technology.

In this study, we analyze fine-tuning efficacy to determine the relative importance of language, domain, and resource-level, exploring how we can improve performance on low-resource languages across multiple domains. Specifically, we seek to understand how effectively knowledge transfers across languages and domains when models are fine-tuned on modest amounts of training data from different resource levels. We do this through controlled fine-tuning experiments on `gpt-4.1-nano-2025-04-14` using three languages—Igbo (low-resource), Hebrew (mid-resource), and Turkish (high-resource)—and three domains—STEM, Medical, and Humanities—from the Global-MMLU dataset (Singh et al. 2024). We evaluate both within-language/domain performance and multiple transfer scenarios. Specifically, we test whether model knowledge obtained in one language or domain via fine-tuning can be applied effectively to another, revealing insights into factors that enhance or limit cross-lingual and cross-domain transfer in multilingual models.

Our findings show clear patterns across domains and language boundaries. Specifically, we find that fine-tuning on high-resource languages leads to substantial accuracy improvements ($\approx +16\%$ gain) when evaluating on low-resource, whereas upward transfer from low-resource to high-resource settings generally results in large performance degradation ($\approx -13\%$). Additionally, when the effect of resource level is isolated, we find that both cross-domain (-2.44%) and cross-lingual ($+2.61\%$) transfers demonstrate a net improvement. Generally, models fine-tuned on diverse, well-resourced data enhances model robustness. This work advances the understanding of how fine-tuning strategies can be designed to enhance the performance of multilingual models to underrepresented languages, helping narrow gaps in performance in multilingual technologies. Models fine-tuned only on high-resource data still improve on unseen low-resource languages, indicating that robust cross-lingual transfer can occur without any target-language fine-tuning. Additionally, cross-lingual transfer demonstrated consistent improvement, whereas cross-domain adaptation alone did not reliably produce positive gains, underscoring that linguistic diversity contributes strongly to generalization more than domain matching.

Related Works

Multilingual and Cross-Lingual Evaluation

Low-resource settings present a persistent challenge for large language models as limited data and unbalanced pretraining corpora result in poor generalization and biased model behavior (Li et al. 2025; Hangya, Saadi, and Fraser

2022; Conneau et al. 2020). Large multilingual language models such as mBert, XLM-R, and BLOOM have demonstrated that joint multilingual pretraining enables substantial cross-lingual transfer (Devlin et al. 2019; Conneau et al. 2020; Workshop et al. 2022). However, performance asymmetries between high and low-resource remain a central challenge (Hu et al. 2020). Recent benchmarks such as XTREME, XGLUE, and Global-MMLU have extended this evaluation to typologically diverse languages and specialized knowledge domains (Hu et al. 2020; Liang et al. 2020; Singh et al. 2024). Even more recently, the introduction of ATLAS expanded evaluations through more optimized scaling of cross-lingual transfer across models (Longpre et al. 2025). Despite these advances, few studies have examined how resource level, specialized domains, and linguistic similarity altogether affect cross-lingual and cross-domain transfer dynamics. To fill this gap, our study analyzes these factors together, providing a more complete picture of what enables multilingual models to be successfully fine-tuned in low-resource, specialized contexts.

Methodology

Data

Across all experiments, we utilize Global-MMLU (Singh et al. 2024), a multilingual and multi-domain adaptation of the original MMLU benchmark (Hendrycks et al. 2020). The dataset maintains a multiple-choice question format while covering various languages and specialized domains. We chose three specific languages from this dataset, based on varying resource availability levels. For low-resource, we used Igbo; for mid-resource, Hebrew; for high resource, Turkish. We also selected three domains: Humanities (literature, history, philosophy, cultural studies), STEM (mathematics, physics, chemistry, engineering), and Medical (clinical medicine, anatomy, pharmacology). These domains were chosen because they differ substantially in vocabulary, linguistic context, and structure of reasoning. Additionally, these domains had a larger proportion of questions in the dataset, allowing for more robust training and evaluation due to increased data availability. This allows us to evaluate how models generalize across both structural and interpretative linguistic areas. This creates nine language-domain combinations of which `gpt-4.1-nano-2025-04-14` is fine-tuned: Humanities-ig, Humanities-he, Humanities-tr, STEM-ig, STEM-he, STEM-tr, Medical-ig, Medical-he, and Medical-tr. Dataset sizes vary by language-domain pair, ranging from approximately 1,400 to 4,000 examples after removing duplicate questions identified within the original dataset. Duplicate questions and questions that weren't present all subsets of test languages were then removed, ensuring each language's set contained the same question instances. Each sample consists of a question prompt in the target language; four answer choices labeled A, B, C, and D in the target language; one correct answer designation; and domain and language metadata. 70/30 train-test splits were created in the specific domain-language subsets for evaluation. The multiple-choice question structure remains unchanged from the original Global-MMLU format to ensure consistency and reproducibility. Figure 3 illustrates the domain transfer matrix, revealing substantial variation in cross-domain fine-tuning effectiveness.

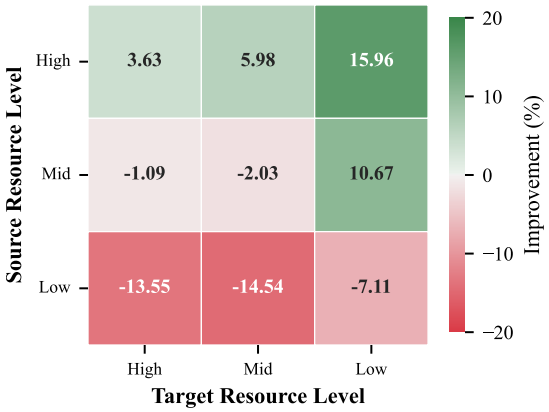

Figure 1: Cross-domain transfer performance matrix showing mean accuracy improvement over baseline. Fine-tuning on the medical domain demonstrates exceptional transfer capabilities, particularly to the STEM domain (+9.61%), suggesting that medical training develops general skills beyond domain-specific knowledge. Medical-to-Humanities transfer also performs well (+3.41%), while Humanities fine-tuning shows the weakest transfer performance. Within-domain improvements (diagonal) range from -2.63% (Humanities) to +5.09% (STEM). Values represent mean improvement over baseline across all languages tested.

Fine-Tuning Experiment

We fine-tuned models on specific language-domain pairs and then evaluated their ability to transfer knowledge across both languages and domains. Nine separate models are fine-tuned, each model trained on one specific language-domain combination using OpenAI's standard fine-tuning workflow. We use a learning rate multiplier of 0.4, 1 training epoch, and batch size 8. We conducted cross-lingual evaluations which tested models on target languages different from those used in fine-tuning (e.g. Humanities-ig $\rightarrow$ Humanities-he). Similarly, we conducted cross-domain evaluations where models were tested on subject areas differing from those used in fine-tuning (e.g. Humanities-he $\rightarrow$ Medical-he). These evaluations assessed generalization and transfer learning capabilities. Furthermore, to measure the robustness of generalization capabilities beyond the training distribution, we evaluated models in cross-both, where both language and domain differ from the data used to finetune the models. In addition, models were tested in typologically similar and typologically different languages to measure generalization between language groups. This specifically allows us to analyze how transferable domain and linguistic specific data is across different resource levels based on language.

Evaluation

All model-generated answers were graded using the base `gpt-4.1-nano-2025-04-14` as the evaluator. The

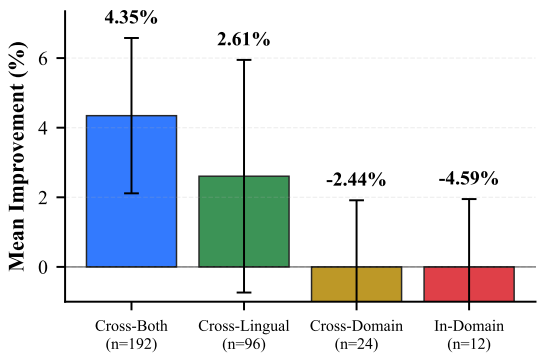

Figure 2: Performance improvement by transfer type. Cross-both (simultaneous cross-domain and cross-lingual transfer) achieves +4.35% mean improvement. Cross-lingual transfer shows +2.61% improvement, while cross-domain and in-domain transfers show negative results (-2.44% and -4.59% respectively), indicating that domain matching without linguistic diversity may harm performance. Error bars represent 95% confidence intervals (n=324 experiments).

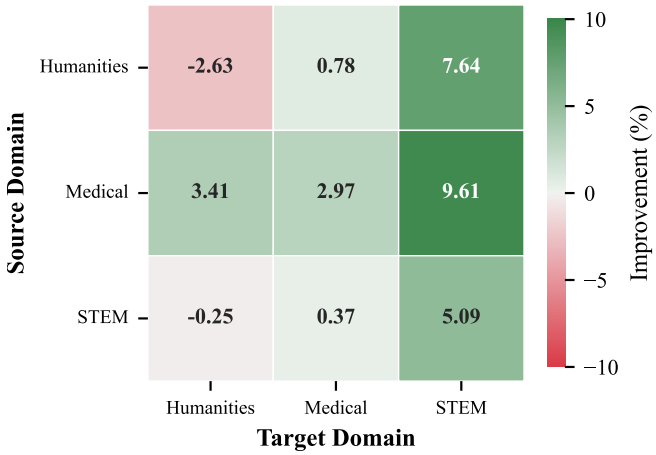

Figure 3: Cross-domain transfer performance matrix showing mean accuracy improvement over baseline. Medical fine-tuning demonstrates the strongest cross-domain transfer, particularly to STEM (+9.61%), while Humanities-to-STEM also shows promising transfer (+7.64%). Within-domain performance varies substantially: STEM achieves +5.09%, Medical +2.97%, while Humanities shows degradation (-2.63%). Values represent mean improvement over baseline across all languages tested.

grader was prompted to extract the models answer (A, B, C, or D) and cross-checked against the ground-truth answers from the dataset. All evaluations were evaluated after fine-tuning with no additional inference or fine-tuning updates. Prompt formats and hyperparameters were implemented the same throughout all experiments to ensure fair and consistent comparisons.

Results

Cross-Lingual and Cross-Domain Transfer Performance

We evaluated the impact of fine-tuning in four settings, with results displayed in Figure 2: cross-lingual, cross-domain, cross-both, and in-domain. Each evaluation instance corresponds to a model-test pair, where a model fine-tuned on one language-domain combination is tested on another. Across the 324 total evaluation instances, cross-both transfer yields the highest mean improvement of $+4.35\%$. Cross-lingual fine-tuning yields a mean improvement of $+2.61\%$, while cross-domain (-2.44%) and in-domain (-4.59%) transfers show negative means. However, overall performance of the macro-categories—cross-lingual ($+2.61\%$) and cross-both ($+4.35\%$)—is positive, driven by primarily by transfers involving high-resource languages. These results show that transfer cannot be reliably predicted from language or domain similarity alone.

When we further analyze the resource-level difference between the fine-tuning language and evaluation language, a clear distinction emerges. Nearly all positive transfer results come from downward transfer (high→low or medium→low), while upward transfer (low→high) consistently causes negative percentage yields. This indicates that, within the scope of our experiments, the target source relationship is the dominant factor in shaping transfer results. Consequently, overall transfer metrics on their own do not explain transfer effectiveness. Future work involving these metrics should control for resource direction when interpreting cross-domain and cross-lingual fine-tuning performance.

Medical fine-tuning provides the strongest cross-domain transfer, achieving +9.61% improvement when transferred to STEM evaluations and +3.41% when transferred to Humanities. This exceptional transferability suggests that fine-tuning on the medical domain develops robust capabilities beyond domain-specific knowledge. In contrast, STEM fine-tuning shows ineffective cross-domain transfer (+0.37% to Medical, -0.25% to Humanities), while Humanities fine-tuning demonstrates the most varied transfer performance across all domains. Notably, Humanities fine-tuning degrades performance within its own domain (-2.63%). These findings suggest structured domains like Medical and STEM yield more transferable representations. Within-domain performance (diagonal of Figure 3) also varies substantially, with STEM showing the strongest within-domain improvement (+5.09%), Medical showing moderate gains (+2.97%), and Humanities showing degradation (-2.63%).

Effects of Resource Level

Furthermore, we examine the effects of resource levels between source and target languages and how they influence fine-tuning effectiveness. The results reveal a strong downward transfer from high-resource to low-resource languages with mean gains of +15.96%. Similarly, downward transfer from medium to low-resource languages produced +10.67% mean improvements. In contrast, upward transfer led to performance degradation with average losses between -14.54% and -1.09%. Lateral transfer yields moderate improvement between the ranges +1-5%, indicating a positive but limited

Table 1: Full Fine-tuning Experiment Accuracy Matrix

Model	Hum am	STEM am	Med am	STEM yo	STEM ig	STEM ky	STEM tr	STEM he	Med ar	STEM de	STEM ar	Hum ig	Hum yo	STEM en	Hum ky	Med tr	Med ky	Hum ar	Med ig	Med he	Med de	Med yo	Hum tr	Med en	Hum he	Hum de	Hum en
Base	0.57	0.60	0.69	0.75	0.76	0.80	0.81	0.81	0.82	0.83	0.83	0.84	0.85	0.85	0.85	0.86	0.86	0.86	0.87	0.87	0.88	0.89	0.89	0.90	0.91	0.92	0.93
Hum ig	0.66	0.64	0.63	0.67	0.72	0.67	0.69	0.74	0.73	0.65	0.71	0.68	0.66	0.75	0.59	0.68	0.68	0.63	0.76	0.73	0.69	0.64	0.59	0.79	0.67	0.58	0.71
STEM ig	0.66	0.58	0.66	0.51	0.51	0.50	0.56	0.52	0.58	0.59	0.53	0.51	0.51	0.63	0.51	0.62	0.52	0.53	0.58	0.54	0.64	0.54	0.57	0.75	0.55	0.58	0.61
Med ig	0.99	0.97	0.99	0.92	0.77	0.89	0.87	0.88	0.89	0.89	0.90	0.87	0.97	0.89	0.93	0.87	0.90	0.94	0.77	0.87	0.88	0.87	0.94	0.91	0.95	0.95	0.92
Hum he	0.88	0.91	0.90	0.82	0.80	0.84	0.82	0.84	0.85	0.79	0.86	0.77	0.81	0.87	0.79	0.81	0.84	0.80	0.81	0.85	0.82	0.81	0.71	0.88	0.72	0.67	0.86
STEM he	0.97	0.97	0.99	0.94	0.92	0.92	0.90	0.92	0.90	0.89	0.90	0.94	0.91	0.93	0.89	0.91	0.93	0.91	0.93	0.93	0.91	0.94	0.90	0.90	0.87	0.92	0.94
Med he	0.94	0.93	0.94	0.89	0.88	0.83	0.85	0.83	0.81	0.84	0.84	0.88	0.89	0.89	0.82	0.80	0.84	0.80	0.88	0.84	0.83	0.86	0.82	0.87	0.79	0.85	0.86
Hum tr	0.98	0.98	0.99	0.95	0.94	0.94	0.91	0.94	0.95	0.91	0.93	0.90	0.86	0.93	0.94	0.90	0.91	0.93	0.94	0.92	0.92	0.94	0.82	0.92	0.93	0.85	0.93
STEM tr	0.98	0.95	0.97	0.93	0.93	0.92	0.92	0.94	0.95	0.90	0.94	0.96	0.97	0.93	0.98	0.91	0.95	0.96	0.94	0.96	0.92	0.95	0.93	0.95	0.96	0.95	0.96
Med tr	0.96	0.95	0.96	0.85	0.82	0.87	0.79	0.87	0.88	0.84	0.87	0.84	0.77	0.85	0.85	0.79	0.85	0.90	0.89	0.88	0.86	0.82	0.71	0.87	0.89	0.86	0.78
Hum en	1.00	1.00	1.00	0.98	0.98	0.97	0.95	0.97	0.93	0.95	0.95	0.96	0.99	0.95	0.97	0.90	0.96	0.93	0.95	0.96	0.92	0.97	0.93	0.92	0.94	0.94	0.95
STEM en	1.00	0.99	0.98	0.96	0.94	0.95	0.91	0.94	0.93	0.89	0.93	0.95	0.97	0.91	0.97	0.89	0.94	0.95	0.92	0.93	0.91	0.95	0.83	0.91	0.92	0.93	0.92
Med en	0.96	0.96	0.98	0.94	0.91	0.91	0.88	0.89	0.88	0.88	0.88	0.88	0.93	0.89	0.89	0.87	0.90	0.85	0.91	0.89	0.89	0.92	0.85	0.90	0.88	0.85	0.93

Resource Levels: Low: am, yo, ig, ky, ar; Mid: he; High: en, tr, de

adaptation capacity (Figure 1).

On the Effect of Language Family and Typological Similarity

The effectiveness of cross-lingual transfer is not solely dependent on the language's resource level but is also significantly influenced by the typological proximity of the source and target languages. Concurrent work has similarly shown that typologically similar can substantially strengthen ability for cross lingual generalization in multilingual models (Longpre et al. 2025). Our selected languages span three distinct families: Igbo (Niger-Congo, highly morphological), Hebrew (Afro-Asiatic, non-concatenative morphology/abjad script), and Turkish (Turkic, agglutinative). We observe that fine-tuning success is mediated by these typological differences. While downward transfer from high-resource Turkish to low-resource Igbo is highly effective ($+15.82\%$ mean gain, per Figure 1), this robust transfer likely stems from the high quality of the Turkish training data and generalizable knowledge rather than linguistic similarity. This suggests that high-resource fine-tuning can overcome significant typological distance by better refining general capabilities within the model.

Conversely, languages with complex, distinct morphological systems, such as the highly agglutinative Turkish or the highly tonal Igbo, can present challenges. We hypothesize that low-resource fine-tuning on Igbo, due to its sparse and typologically distant data, leads to overfitting on specific token patterns, hindering generalization to other languages, which aligns with the observed negative upward transfer (-5.95%). These results suggest that language-family dynamics and linguistic structure remain as key bottlenecks for upward or lateral knowledge transfer. We encourage future work to rigorously evaluate how typological similarity interacts with resource levels to shape transfer effectiveness. We view concurrent work such as Longpre et al. (2025) as a positive step in this direction.

Conclusion

Limitations

While our findings demonstrated clear patterns in how resource levels and linguistic diversity effects transfer performance, our focus on only three language families limits our ability to make broad generalization claims about cross-lingual behavior across the full multilingual spectrum. Additionally, due to compute constraints, we only trained on one base model (`gpt-4.1-nano-2025-04-14`), making it unclear whether our findings generalize to other multilingual models. The evaluations are limited to a multiple-choice task setting, which may not generalize to free-form tasks such as QA or translation. The focus on only three languages on which the model was finetuned also limits our ability to make broad claims about how typology affects transfer. Testing more languages would strengthen the generality of our findings. Moreover, evaluating transfer on additional languages within the same language families as our selected languages would provide a clearer picture regarding the effectiveness of typological similarity on knowledge transfer.

Future Work

Future research should expand on this evaluation to broader sets of languages, particularly those with more typological, morphological, and script based differences. Doing so would clarify the extent in which syntactic structure, word order variation, and morphological complexity affect transfer across languages. Furthermore, scaling experiments that include larger model and fine-tuning data sizes should be explored to investigate whether the resource-direction confounding pattern persists in larger models. Future work should consider evaluation on different question-answering beyond solely multiple-choice benchmarks to better understand the practical implications of these limitations. We encourage future research on investigating methods to improve the inherent ability of multilingual LLMs on low-resource languages.

Summary

This work demonstrates that high-resource fine-tuning data can significantly improve performance on low-resource languages across multiple domains. We show that even downward transfer provides strong improvements and that cross-domain transfer provides reliable positive results as well. In conclusion, these findings exhibit the potential of leveraging high-resource and diverse data to narrow performance gaps in LLMs, paving the way for further development of more reliable, accessible, and equitable linguistic technologies.

References

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8440–8451.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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)*, 4171–4186.

Hangya, V.; Saadi, H. S.; and Fraser, A. 2022. Improving Low-Resource Languages in Pre-Trained Multilingual Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11993–12006. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. Xtreme: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, 4411–4421. PMLR.

Li, Z.; Shi, Y.; Liu, Z.; Yang, F.; Payani, A.; Liu, N.; and Du, M. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28186–28194.

Liang, Y.; Duan, N.; Gong, Y.; Wu, N.; Guo, F.; Qi, W.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; et al. 2020. XGLUE: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Longpre, S.; Kudugunta, S.; Muennighoff, N.; Hsu, I.; Caswell, I.; Pentland, A.; Arik, S.; Lee, C.-Y.; Ebrahimi, S.; et al. 2025. ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality. *arXiv preprint arXiv:2510.22037*.

Singh, S.; Romanou, A.; Fourrier, C.; Adelani, D. I.; Ngui, J. G.; Vila-Suero, D.; Limkonchotiwat, P.; Marchisio, K.; Leong, W. Q.; Susanto, Y.; et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Thangaraj, H.; Chenat, A.; Walia, J. S.; and Marivate, V. 2024. Cross-lingual transfer of multilingual models on low resource African Languages. *arXiv preprint arXiv:2409.10965*.

Workshop, B.; Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.